



CERC 2021

Collaborative European Research Conference

Cork, Ireland

9 - 10 September 2021

www.cerc-conf.eu

Proceedings

Editors (in alphabetical order) :

Haithem Afli

Udo Bleimann

Dirk Burkhardt

Robert Loew

Denise Reichel

Haiying Wang

Huiru (Jane) Zheng

ISBN 978-3-96187-020-2

https://doi.org/10.48444/h_docs-pub-507



This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>.

Preface

Back in 2020, just after the first COVID-19 lockdown, we decided to host CERC 2021, conference that you are currently attending. At that stage, we were hoping to be able to meet face-to-face here in Cork, the city where everything in CERC started.

Even it was not possible to host an onsite event at the end, it is great that we are able to live stream the conference for the first time and organise virtual networking sessions in addition to the privilege of having great speakers and special track sessions.

CERC is an opportunity to welcome not just our European friends and colleagues, but also those from farther afield. Munster Technological University punches above its weight in the areas of Artificial Intelligence, Cyber security and computer science research in general, principally through National, European and international funds and collaborations. So we believe it is appropriate that CERC is being held again in our University.

We are of course grateful to everyone who submitted a paper; whether your work was selected for presentation or not, if no-one had submitted, we wouldn't have had a conference. For those of you whose work was selected for presentation online, well done!

Along the way, we have been helped greatly by the program committee and my fellow program chairs, particularly: Prof Udo Bleimann and his invaluable support throughout the conference; Prof Huiru Zheng; Prof Ingo Stengel, Dr. Haiying Wang, and Prof Denise Reichel for co-organising the conference. Dirk Burkhardt and Dr. Robert Loew put a great effort into setting up the website and conference management system and preparing the conference programme and proceedings.

We are also extremely grateful to Munster Technological University, Ulster University, Hochschule Karlsruhe and Hochschule Darmstadt for providing invaluable support to the conference.

Finally, we really hope that you all enjoy the conference, that you benefit from the excellent programme that has been assembled, and that you can virtually discover Cork during the virtual coffee breaks.

Dr Haithem Afli
Conference and Program Chair, CERC 2021
Cork, Ireland, September 2021

Conference Chairs

Dr. Haithem Afli, Munster Technological University, Ireland
Prof. Dr Udo Bleimann, Darmstadt University of Applied Sciences, Germany
Dirk Burkhardt, Darmstadt University of Applied Sciences, Germany
Dr. Robert Loew, Darmstadt University of Applied Sciences, Germany & DIPF | Leibniz Institute for Research and Information in Education, Germany
Prof. Dr. Denise Reichel, SRH Heidelberg University of Applied Sciences, Germany
Dr. Haiying Wang, Ulster University, UK
Prof. Huiru (Jane) Zheng, Ulster University, UK

International Scientific Committee

Jens-Peter Akelbein, Darmstadt University of Applied Sciences, Germany
Tatjana Archipov, University of Applied Sciences Karlsruhe, Germany
Lu Bai, Ulster University, United Kingdom
Raymond Bond, Ulster University, United Kingdom
Fiona Brown, Datatics Ltd, United Kingdom
Zoraida Callejas Carrión, Granada University, Spain
En-Chi Chang, AMG Digital Marketing & Design Company, Taiwan
Zhibin Chen, Durham University, United Kingdom
Simone Dogu, DB Systel GmbH, Germany
Markus Döhring, Darmstadt University of Applied Sciences, Germany
Lei Duan, Sichuan University, China
Klaus-Peter Fischer-Hellmann, Digamma GmbH, Germany
Marius Faßbender, Daimler AG, Germany
Matthias Feiner, University of Applied Sciences Karlsruhe, Germany
Orla Flynn, Galway-Mayo Institute of Technology, Ireland
Steven Furnell, Plymouth University, United Kingdom
Jianliang Gao, Imperial College, United Kingdom
Gunter Grieser, Darmstadt University of Applied Sciences, Germany
Vic Grout, Glyndwr University, United Kingdom
Markus Haid, Darmstadt University of Applied Sciences, Germany
Joe Harrington, Munster Technological University, Ireland
Andreas Heberle, University of Applied Sciences Karlsruhe, Germany
Andreas Heinemann, Darmstadt University of Applied Sciences, Germany
Peter Henning, University of Applied Sciences Karlsruhe, Germany
Melina Hillenbrand, University of Applied Sciences Karlsruhe, Germany
Reimar Hofmann, University of Applied Sciences Karlsruhe, Germany
Martin Knahl, University of Applied Sciences Furtwangen, Germany
Thorsten Leize, University of Applied Sciences Karlsruhe, Germany
Habin Lee, Brunel University, United Kingdom
Olive Lennon, University College Dublin, Ireland
Margot Mieskes, Darmstadt University of Applied Sciences, Germany
Kawa Nazemi, Darmstadt University of Applied Sciences, Germany
Rainer Neumann, University of Applied Sciences Karlsruhe, Germany
Thomas Pleil, Darmstadt University of Applied Sciences, Germany
Nacim Ramdani, Université of Orléans, France
Erwin Rauch, Free University of Bozen-Bolzano, Italy
Sven Rogalski, Darmstadt University of Applied Sciences, Germany
Pablo Mesejo Santiago, Granada University, Spain

Christian Schalles, DHBW Mosbach, Germany
Jan Schikofsky, LaunchMarkets Institute, Germany
Bryan Scotney, Ulster University, United Kingdom
Helen O' Shea, Munster Technological University, Ireland
Melanie Siegel, Darmstadt University of Applied Sciences, Germany
Roy Sleator, Munster Technological University, Ireland
Lina Stankovic, Strathclyde University, United Kingdom
Vladimir Stankovic, University of Strathclyde, United Kingdom
Uta Störl, Darmstadt University of Applied Sciences, Germany
Jan Stöß, University of Applied Sciences Karlsruhe, Germany
Bernhard Thull, Darmstadt University of Applied Sciences, Germany
Ulrich Trick, University of Applied Sciences Frankfurt, Germany
Hui Wang, Ulster University, United Kingdom
Christoph Wentzel, Darmstadt University of Applied Sciences, Germany
Andrea Wirth, University of Applied Sciences Karlsruhe, Germany
Matthias Wölfel, University of Applied Sciences Karlsruhe, Germany
Angela Wright, Munster Technological University, Ireland
Tuba Yilmaz Abdolsaheb, Istanbul Technical University, Turkey



Table of Content

Keynotes

Microbiome, Health and Data Science: Current Applications and Future Prospects	12
Bruno Gabriel Nascimento Andrade	
Toward Digital Enabled Connected Health and Well-Being	13
Huiru (Jane) Zheng	

Chapter 1

Visual Computing

AI-Based User Empowering Use Cases for Visual Big Data Analysis	15
Thoralf Reis, Sebastian Bruchhaus, Florian Freund, Marco X. Bornschlegl, Matthias Hemmje	
Video-Based Emotion Detection Analyzing Facial Expressions and Contactless Vital Signs for Psychosomatic Monitoring	29
Hayette Hadjar, Binh Vu, Dennis Maier, Gwendolyn Mayer, Paul Mc Kevitt, Matthias Hemmje	

Chapter 2

Data Processing and Machine Learning

Improving Machine Translation Quality Estimation Using Named-Entity Masking and Assessment Scores	38
Anthony Reidy, Sean Cummins, Kian Sweeney, George Dockrell, Pintu Lohar, Andy Way	
An Empirical Comparison Analysis on the Evolution of RNN Models Using Multiple European Languages	50
Vikram Bhutani, Farshad Ghassemi Toosi	

Data Quality Improvement and Entity Alignment Optimization for Constructing Large-Scale Knowledge Graphs	68
---	----

Keaton Sullivan, Fiona Browne, Huiru Zheng, Haiying Wang

The Choice of Reference Channel in Channel Alignment and Channel Selection	86
---	----

Ingo Stengel, Karin Pietruska, Matthias Wölfel

Chapter 3

E-Healthcare and Smart Diagnostics

A Robust Martingale Approach for Detecting Abnormalities in Human Heartbeat Rhythm	99
---	----

Jonathan Etumusei, Jorge Martinez Carracedo, Sally McClean

Artificial Neural Network for Human Activity Recognition by Use of Smart Insoles	116
---	-----

Luigi D'Arco, Haiying Wang, Graham McCalmont, XianQi Lan, Huiru Zheng

Investigation of the Use of Deep Learning and Emotion Detection for the Improvement of Text-Based Medical Conversational Agent	127
---	-----

Bing Yuan, Haithem Afli

An RT-qPCR Data Analysis Platform	143
--	-----

Thomas Krause, Elena Jolkver, Sebastian Bruchhaus, Michael Kramer, Matthias Hemmje

Diversity of the Bifidobacterial Phageome in the Infant Gut	152
--	-----

Darren Buckley, Douwe van Sinderen, Francesca Bottacini

Chapter 4

E-Learning and Competences

QBL — A Software-Technical Approach for Supporting Competence-Based Learning	154
---	-----

Matthias Then, Benjamin Wallenborn, Felix Fischman, Sebastian Lothary, Ramona Srbecky, Michael Winterhagen, Matthias Hemmje

Supporting the Mapping of Educational Game Events With Competency Models Considering Qualifications-Based Learning	165
Ramona Srbecky, Marcus Frangenberg, Benjamin Wallenborn, Matthias Then, Iván-José Pérez-Colado, Cristina Alonso-Fernandez, Baltasar Fernandez-Manjon, Matthias Hemmje	
Examinations in the Context of Curriculum Content: Case Study of a 1926 Irish Mathematics Exam Paper	181
Hazel Murray, David Malone	
Splitfed Learning Without Client-Side Synchronization: Analyzing Client-Side Split Network Portion Size to Overall Performance	188
Praveen Joshi, Chandra Thapa, Seyit Camtepe, Mohammed Hasanuzzaman, Ted Scully, Haithem Afli	

Chapter 5

Engineering and Society

Trust and Transparency in Data Protection in Online-Marketing — Differences Between Different Generations	197
Louis Kerker, Ingo Stengel, Stefanie Regier	
Three Source Electron-Beam Co-deposited Thermoelectric Bisbte Thin Films	210
Philipp Lorenz, Gabriel Zieger, Heidemarie Schmidt	
Temperature Determination During Flash Lamp Annealing	214
Lars Rebohle, Viktor Begeza, Thomas Schumann	

Workshop »Future of Education«

The Future of Education — Lessons learned from CERC Workshop »Future of Education« and an Online Course on »Digital Basics«	216
Tanja Kranawetleitner, Heike Krebs, Nina Kuhn, Robert Loew	

List of Authors

Afli	127, 188	Alonso-Fernandez	165
Andrade	12	Begeza	214
Bhutani	50	Bornschlegl	15
Bottacini	152	Browne	68
Bruchhaus	15, 143	Buckley	152
Camtepe	188	Carracedo	99
Cummins	38	D'Arco	116
Dockrell	38	Etumusei	99
Fernandez-Manjon	165	Fischman	154
Frangenberg	165	Freund	15
Hadjar	29	Hasanuzzaman	188
Hemmje	15, 29, 143, 154, 165	Jolkver	143
Joshi	188	Kerker	197
Kramer	143	Kranawetleitner	216
Krause	143	Krebs	216
Kuhn	216	Lan	116
Loew	216	Lohar	38
Lorenz	210	Lothary	154
Maier	29	Malone	181
Mayer	29	McCalmont	116
McClellan	99	Mc Kevitt	29
Murray	181	Pietruska	86
Pérez-Colado	165	Rebohle	214
Regier	197	Reidy	38
Reis	15	Schmidt	210
Schumann	214	Scully	188
Srbecky	154, 165	Stengel	86, 197
Sullivan	68	Sweeney	38
Thapa	188	Then	154, 165
Toosi	50	van Sinderen	152
Vu	29	Wallenborn	154, 165
Wang	68, 116	Way	38
Winterhagen	154	Wölfel	86
Yuan	127	Zheng	13, 68, 116
Zieger	210		

Keynotes

Microbiome, Health and Data Science: Current applications and future prospects

Bruno Gabriel Nascimento Andrade^a

^a *Munster Technological University, Department of Computer Science, Ireland*

Abstract

It's a known fact that humans are not made entirely of human cells, as your cells are outnumbered by a multitude of microorganisms living inside you in a symbiotic state, the so-called microbiome. The microbiome is an interesting and complex data science problem, a single project can generate terabytes of data, requiring the expertise of data scientists and powerful computers for data processing and data analysis. Although complex, unlocking the potential of the microbiome for Health, biotechnology and Agritech can change the world forever for the better.

Speaker Biography

Dr. Bruno Gabriel Nascimento Andrade is a data scientist with a career focused on biological dataset, with experience with both Public Health and Agritech fields. In order to continue this research line, he is currently working on his post-doctoral research, whose subject is the characterization and association of microbiome data with host phenotypes, using machine learning models.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ bruno.andrade@adaptcentre.ie (B. Andrade)

📞 0000-0003-2481-401X (B. Andrade)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Toward digital enabled connected health and well-being

Huiru Zheng^a

^a *Ulster University, School of Computing and Mathematics, Northern-Ireland, UK*

Abstract

Population ageing has become a global phenomenon and traditional health systems are under pressure from rising costs, growing consumer expectations in caring for their elderly. With the rapid advance of digital technology, it is increasingly possible to drive early interventions to better manage health outcomes and risks and to provide an effective solution for people with lower healthcare cost. In this talk, Prof. Zheng will highlight recent research work in digital health. The applications to well-being and self-management of chronic conditions will be presented. The presentation will conclude with the discussion of challenges and opportunities in connected health.

Speaker Biography

Prof. Huiru Zheng is a Professor of Computer Science with School of Computing. Her research interests include machine learning, integrative data analysis, complex network analysis, and assistive technology and their application in healthcare informatics and bioinformatics. She has a successful track record of winning research funding as a Principal Investigator and has been a grant holder of research projects funded by EPSRC, TSB, DEL, NHS, Invest NI and European Commission including SMART Self Management, NOCTURNAL, CLARCH COPD Self Management, Self Management Platform for Connected Health, CardioWorkBench, SenseCare, MetaPlat, STOP and MENHIRE. She is co-leader of WG4 in the COST ACTION – OpenMultiMed. The products of her research have been reflected in her over 330 peer reviewed journal and conference publications. Prof. Zheng is a Senior Member of IEEE. She serves on the editorial board of several international journals and serves as co-chairs and program committees of a number of international conferences.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ h.zheng@ulster.ac.uk (H. Zheng)

📞 0000-0001-7648-8709 (H. Zheng)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Chapter 1

Visual Computing

AI-based User Empowering Use Cases for Visual Big Data Analysis

Thoralf Reis^a, Sebastian Bruchhaus^a, Florian Freund^a, Marco X. Bornschlegl^a and Matthias L. Hemmje^a

^aUniversity of Hagen, Faculty of Mathematics and Computer Science, 58097 Hagen, Germany

Abstract

This paper aims to assess scenarios where AI empowers experts and end users carrying out visual Big Data Analyses by systematically deriving and classifying use cases that interconnect these trending research and application areas. Thereby, and to utilize a unified terminology, the results align with the AI2VIS4BigData Reference Model and its Service-Oriented Architecture, a conceptual framework for visual Big Data Analysis in combination with AI. The modeling of use cases with practical relevance within this paper follows a two step approach: An existing reference model for visual Big Data Analysis is reviewed by conducting a cognitive walk-through and the revealed challenges are utilized to define a set of use cases that drive existing research forward. These use cases are subsequently validated utilizing the result of an exploratory survey by the aid of a group of international scientists.

Keywords

User Empowerment, AI, Big Data, Visualization, Big Data Analysis, AI2VIS4BigData, Use Cases

1. Introduction and Motivation

The concept of user empowerment in **Information Systems (IS)** comprises methods and principles that aim at increasing the users' motivation and self-confidence to utilize as many of the system's capabilities as possible [1]. Thereby, the users can tailor the interface or adapt the usage of the IS to be more goal-oriented and efficient. Kim et al. summarize IS user empowerment to be strongly related to four psychological aspects: the "*individual's belief in his or her capability to use the system*" [1], a clear understanding and prioritization of the system activities, the awareness which decisions can be made to influence the IS, and the knowledge of the "*degree to which an individual can influence task outcomes based on the use of system*" [1].

Artificial Intelligence (AI) is a collective term for methods and techniques like, e.g., **Machine Learning (ML)** [2] and becomes more and more relevant for practical applications in health care, driverless cars, or humanoid robots [3]. AI and Big Data Analysis are closely connected to each other [4] as Big Data Analysis enables deriving, validating, applying, and improving AI models while AI-driven algorithms support the exploration of Big Data [4].

Although Big Data is popular in both science and industry, definitions of its terminology remain rather unclear. From a global perspective, Big Data "*refers to the explosion of available information*" [5]. A more formal definition can be derived from Doug Laney's data management challenges [6]

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ thoralf.reis@fernuni-hagen.de (T. Reis); sebastian.bruchhaus@fernuni-hagen.de (S. Bruchhaus); florian.freund@fernuni-hagen.de (F. Freund); marco-xaver.bornschlegl@fernuni-hagen.de (M.X. Bornschlegl); matthias.hemmje@fernuni-hagen.de (M.L. Hemmje)

ORCID 0000-0003-1100-2645 (T. Reis); 0000-0002-7783-2636 (S. Bruchhaus); 0000-0002-7344-6869 (F. Freund); 0000-0003-3789-5285 (M.X. Bornschlegl); 0000-0001-8293-2802 (M.L. Hemmje)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

which comprise three dimensions (the three v's); variety ("high dimensionality" [5], ambiguous data manifestations), volume ("massive sample size" [5]), and velocity (high data rates) [6].

Since "human brain tends to find pattern more efficiently when data is represented visually" [7], visualization is an important link between AI and Big Data Analysis [8] that "enhances comprehension and decreases entry barriers for new users" [9]. Its purpose ranges from simply presenting information, confirming assumptions to exploring new insights [8].

In 2020, the authors of this paper introduced AI2VIS4BigData, a reference model for scientific and industrial applications that apply AI, Big Data Analysis, and visualization [4]. The reference model's objective is to establish a common terminology, specify relations unambiguously, and enable the derivation [10] for Big Data Analysis systems that utilize visualization and AI [4]. It furthermore presents three use cases that interconnect Big Data Analysis, AI, and visualization. The visualization of these high level use cases in Figure 1 reveals that all use cases are somehow relevant for Big Data Analysis as well as AI user stereotypes. However, there exists neither a detailed description of these use cases nor an assessment to which extent they support user empowerment.

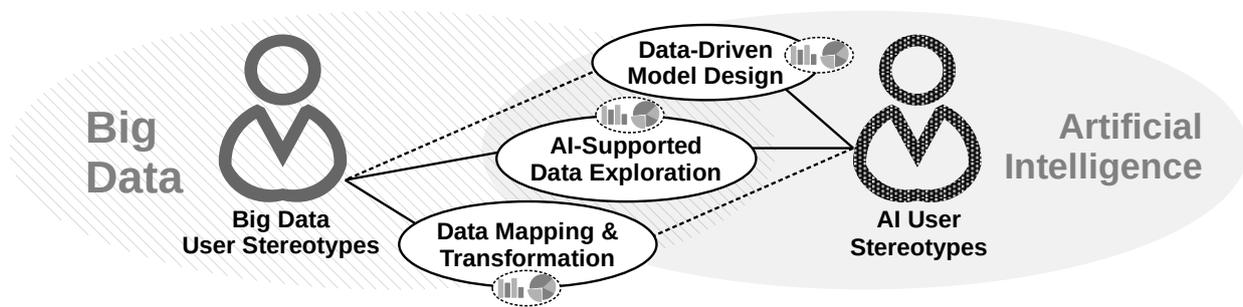


Figure 1: Use Cases interconnecting AI, Big Data Analysis, and Visualization [4]

In this way, this paper intends to provide both detailed descriptions and a classification of use cases with special focus on user empowerment. This shall be accomplished through specifying the use cases and creating a use case taxonomy. For achieving this objective, this paper follows the research approach introduced in [11] to assess manual activities that exist for visual Big Data Analysis over all AI2VIS4BigData processing steps from data integration to data analysis, data visualization, and data exploration as well as review existing challenges in literature to examine the three use cases from Figure 1 in more detail. All manual activities and challenges will then be utilized to derive use cases that solve the underlying problems or mitigate the negative impact of the challenges. Thereby, these use cases empower the users. Finally, these use cases will be clustered and hierarchically categorized within a use case taxonomy. The remainder of this paper contains a presentation of the state of the art (Section 2) with a detailed introduction of the AI2VIS4BigData Reference Model as well as a model review of it based on a cognitive walk-through. Section 3 introduces a derivation of detailed use cases as well as their relationships within a use case taxonomy. Section 4 validates the use cases initially before Section 5 summarizes the results and outlines future research directions.

2. State of the Art

This section outlines selected state of the art with relation to user empowerment and lays the foundation to identify potential use cases of AI for user empowerment to improve visual Big Data Analysis

in the AI2VIS4BigData Reference Model through conducting an cognitive walk-through based model review. The cognitive walk-through highlights potential problems as well as manual activities that can consume a significant amount of time.

2.1. AI2VIS4BigData Reference Model

AI2VIS4BigData Reference Model [4] is founded on Bornschlegl’s IVIS4BigData Reference Model for visual Big Data Analysis [12]. It extends the visual Big Data Analysis reference model on all aspects of AI and machine learning for data analytics. Its objective is to establish a common terminology, specify relations unambiguously, and enable the derivation [10] for Big Data Analysis systems that utilize visualization and AI [4]. This reference model has been successfully validated in an expert survey as well as an expert workshop with scientists and researchers from six different countries [13]. Further work includes the introduction of a conceptual architecture that is based on this reference model [9]. This paper addresses two shortcomings in the state of the art: Use cases for user empowerment have not been analyzed methodically in the context of Big Data visualization and AI so far. Furthermore the concept of user empowerment in AI2VIS4BigData itself is not comprehensively described as of yet. The reference model is visualized in Figure 2.

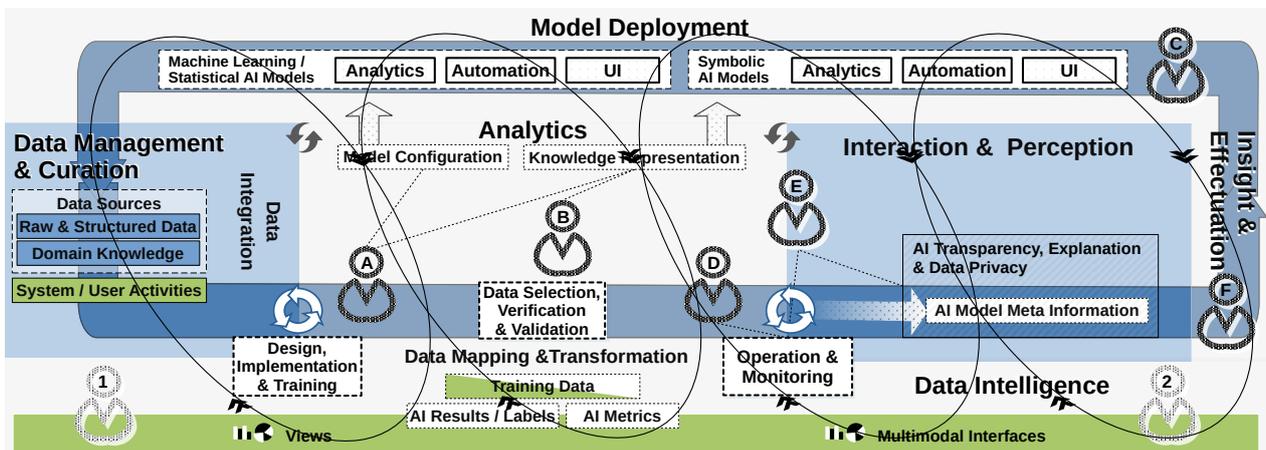


Figure 2: AI2VIS4BigData Reference Model [4] for AI-based support of visual Big Data Analysis

Value generation in AI2VIS4BigData is separated into three dimensions; firstly the transformation of raw data into visualizations of analyzed data, and secondly the transformation from data into information, knowledge, and wisdom [14]. The first dimension is represented through the four consecutive process stages of AI2VIS4BigData [4]. The second dimension is represented through a data intelligence layer that interconnects all process stages and enables the different expert and end user stereotypes that are involved in visual Big Data Analysis to interact with the system and its processing results [14]. Within this reference model, *“end users know the application domain”* [14] while expert users *“are able to configure technical details”* [12]. In regards to end users, it empowers them *“to configure, simulate, optimize, and run each phase of the IVIS pipeline in an interactive way”* [14]. The third dimension is the application of AI within the model deployment layer that lays the foundation for AI-based user empowerment [4]. As visualized in Figure 2, the view on each processing step, the processing step itself, and the processed artifact are interconnected with the respective end

user stereotype through four SECI¹ cycles. These cycles represent the consecutive transformation of "*explicit knowledge to tacit knowledge*" [14] and the other way round in order to generate novel knowledge.

2.2. Fischer and Nakakoji's Multifaceted Architecture

User empowerment in context of AI2VIS4BigData comprises user interface configurability, simulations, continuous improvements, and knowledge creation from IVIS4BigData [14]. Hence it is closely related [12] with Fischer and Nakakoji's multifaceted architecture [15]. This architecture consists five elements sorted into three layers [15]: a domain knowledge layer that contains knowledge in form of a catalog of rules and patterns, their meaning for the current application domain as well as the users' comprehension of their significance [15]; a design creation layer which contains a specification, a description of the required characteristics of a design, and a construction, the actual implementation of the design [15]; a feedback layer connects domain knowledge and design creation through critical reflections, case-based reasoning, and simulation [15]. Instead of trying to create an expert system for a well-understood problem that can be fully specified, Fischer and Nakakoji propose to utilize this domain knowledge to empower users to solve the problem by themselves [15]. Their architecture enables end user with lower expertise in the application domain to benefit from the existing knowledge base whereas expert users like, e.g., "*experienced designers*" [15] are empowered to increase the collection within this knowledge base [15].

2.3. Discussion and Remaining Challenges

AI2VIS4BigData and its foundation IVIS4BigData incorporate principles like continuous improvements and the utilization of knowledge [14] to improve the system's usability in a manual form depending on knowledge input and system adaptations by expert users. An application of Fischer and Nakakoji's multifaceted architecture and its feedback layer to empower users based on a knowledge base without requiring expert users to manually adapt the system's user interface remains a challenge. This paper targets to address this challenge by systematically deriving use cases for applications of Fischer and Nakakoji's multifaceted architecture on AI2VIS4BigData.

3. Conceptual Modeling

Section 2 introduced the AI2VIS4BigData Reference Model, Fischer and Nakakoji's multifaceted architecture, and emphasized on combining their underlying concepts. This section reviews the manual activities within AI2VIS4BigData for the purpose of visual Big Data Analysis in order to reveal manual activities and content-related problems which a consequent application of Fischer and Nakakoji's multifaceted architecture can address. These weak spots then serve as basis for conceptually modeling specific use cases based on the three high-level use cases from Figure 1. The objective is an overview as well as a taxonomy of user empowering use cases.

In order to establish a clear terminology in regards to the application areas' relationship like, e.g., is AI applied for the purpose of visual Big Data Analysis or vice versa, the use cases shown in Figure 1 are renamed: Data-driven model design is renamed to Big Data-supported AI model design as Big Data Analyses are applied for the purpose of designing an AI model; data mapping and transformation is redefined to AI-based Big Data transformation since AI is applied to extract relevant information

¹Socialization, Externalization, Combination, and Internalization

from Big Data; and in order to emphasize on the importance of user empowerment, AI-supported data exploration is renamed to AI-based Big Data user empowerment.

3.1. Model Review based on a Cognitive Walk-Through

Fischer and Nakakoji's user empowering architecture targets to equip users with relevant insights and knowledge to solve the problem themselves instead of trying to utilize AI to replace them [15]. They mention the challenges of *"inform[ing] and support[ing] the judgment"* [15] of the user as well as of *"automat[ing] tasks that people consider tedious or uninteresting"* [15] to be crucial. To properly support the users, it is crucial to be aware of potential problem types that can occur during visual Big Data Analysis. Consequently, this model review looking into manual, repetitive activities that need to be carried out by human experts and end users looks like a promising starting point for deriving valid user empowerment use cases as well as typical problems that occur thereby.

The model review of AI2VIS4BigData is structured alongside the four Big Data processing steps of its foundation IVIS4BigData since they comprise all Big Data Analysis related activities that can be supported through the application of AI. Beginning with data integration, the expert and end users are required to manually configure the system to adapt to certain data models, data schema, as well as semantics [12]. They then have to configure appropriate wrapper and mediator components before they can simulate and perform the actual data integration [12]. Relevant problems that can thereby occur comprise data level inconsistencies and a high heterogeneity [16] as well as the handling of the enormous amount high-dimensional data [5] which leads to *"scalability"* [5] issues as well as *"storage bottleneck[s]"* [5].

During the data transformation processing step, the users configure and select an appropriate Big Data analysis method as well the data they want to transform [12]. Finally, they configure, simulate, and finally execute the workflow [12]. Relevant problems in this context result on the one hand from bad data quality: *"noise, outliers, low precision, missing values"* [8], errors in measurement [8, 5], duplicates [8] and *"not maintained attribute[s]"* [17]; On the other hand, the massive amounts of data and the dynamic nature of Big Data with *"streams of time related or real time data"* [8] challenge data transformation significantly [5] since *"linear pass of the whole dataset [is] unaffordable"* [5] and these huge data amounts call for *"immense parallelization"* [7].

The third processing step, visual mapping, consists of configuring and selection activities for a visual representation and a visualization library [12]. Selecting the data that shall be visualized, configuring and simulating the visualization workflow conclude the manual visual mapping activities [12]. Problems for this processing step are closely linked to data quality since *"visual noise"* [7], *"spurious correlation, incidental endogeneity"* [5], or *"incidental homogeneity"* [5] make comprehensible visualizations hard to achieve. This applies especially, as the human receptivity is challenged and often overstrained by Big Data's data amount and data velocity [7].

The last processing step view transformation comprises the configuration of IVIS techniques, the selection of visualizations as well as the configuration and simulation of views [12]. All derived manual activities as well as relevant problems in visual Big Data Analysis per AI2VIS4BigData transformation are summarized as results of this cognitive walk-through in Table 1.

3.2. AI-based Big Data Transformation

The AI2VIS4BigData SOA presented in [9] contains AI-based services for every visual Big Data Analysis processing step from data integration to data exploration. Figure 3 visualizes these services together with the introduced SECI cycles and associates them with the use case AI-based Big Data Trans-

Table 1

Manual Activities and Problems in Visual Big Data Analysis derived from an AI2VIS4BigData [4] Model Review

Big Data Transformation	Manual Activity	Problem
Data Integration	Wrapper Configuration Mediator Configuration Data Schema Configuration Data Model Configuration Semantic Resource Configuration Semantic Resource Selection Data Integration Configuration and Simulation	Inconsistency in Data Levels [16] High Data Rate [5]
Data Transformation	Analysis Method Configuration Analysis Method Selection Raw Data Selection Analysis Method Workflow Configuration and Simulation	High Data Dimensionality [5] Measurement Errors & Noise [8, 5] Outliers [8] Missing Values [8, 5] Duplicate Records [8, 17]
Visual Mapping	Visual Representation Configuration Visual Representation Selection Visualization Library Configuration Visualization Library Selection Structured Data Selection Visualization Workflow Configuration and Simulation	Visual Noise [7] Spurious Correlation [5] Incidental Endogeneity & Homogeneity [5]
View Transformation	IVIS Technique Configuration Visualization Selection View Configuration and Simulation	

formation (B). In analogy to the model review (Section 3.1), the four vertical pillars from IVIS4BigData's transformation pipeline [14] can be utilized to detail the use case further on into sub use cases.

Use Case AI-based Data Integration

The first of these sub use cases is *AI-based Data Integration*. Manual activities and present challenges for this transformation step are utilized to detail this use case further on into the following third level use cases: *Wrapper Detection* (manual activity of wrapper configuration), *Mediator Detection* (manual activity of mediator configuration), *Data Schema Detection* (manual activity of data schema configuration and challenge of inconsistency in data levels), *Data Model Detection* (manual activity of data model configuration), *Data Semantics Detection* (manual activities of semantic resource configuration and selection), and *Data Inflow Prediction* (challenge of high data rates).

Figure 4 visualizes the relationship between AI2VIS4BigData and the IVIS4BigData services for data integration: the AI2VIS4BigData data integration services derive metadata and information on semantic representations of the raw data from the data source systems whereas the IVIS4BigData

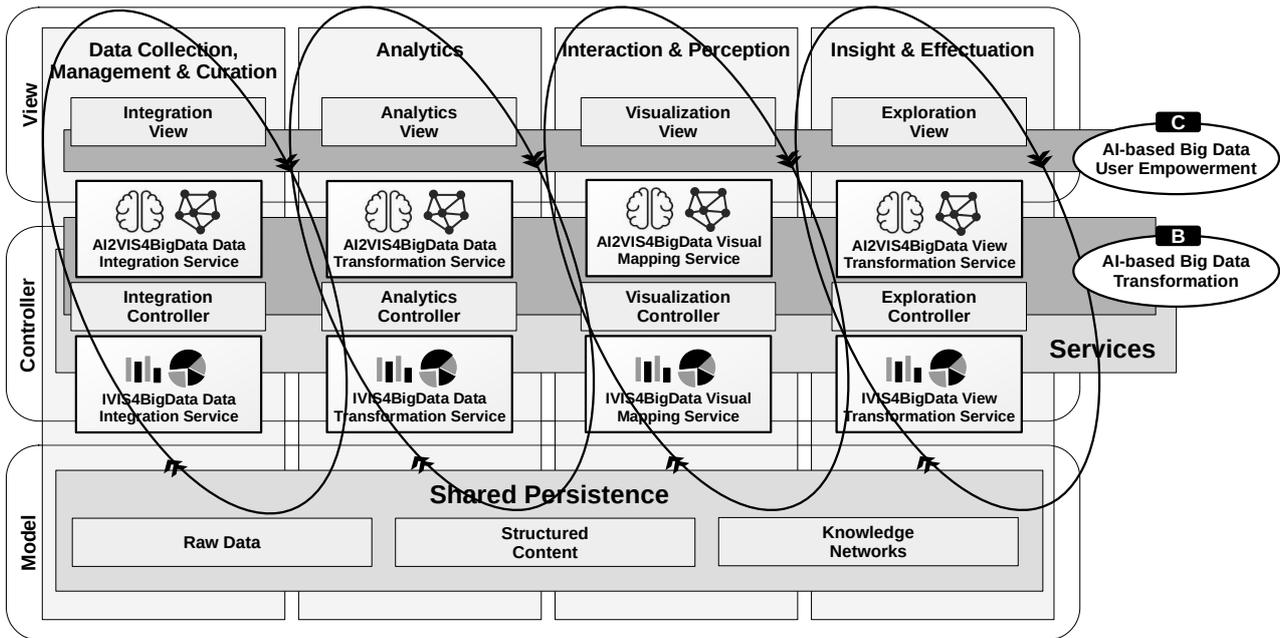


Figure 3: User Empowering Use Cases embedded in the AI2VIS4BigData SOA

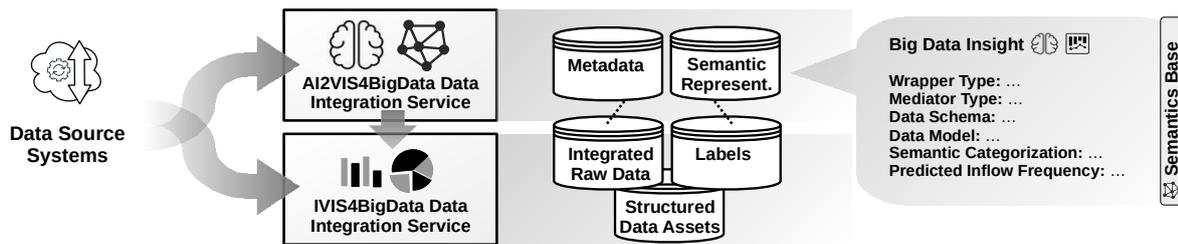


Figure 4: Data Integration Use Case Sketch as Application of AI2VIS4BigData Services

services actually integrate the raw data into the system. The metadata and information on semantic representations are thereby utilized to either assist the human user or to automatize parts of the data integration process.

Use Case B2: AI-based Data Transformation

The data transformation use case for the IVIS4BigData analysis processing phase is visualized as application of an AI2VIS4BigData data transformation service in Figure 5. It consists of the following third level use cases: *Analysis Method Prediction* (manual activities of analysis method configuration and selection), *Raw Data Hotspot Detection* (manual activity of raw data selection, challenge of outliers), *Dimensionality Reduction* (challenge of high data dimensionality, challenge of measurement errors and noise), *Missing Value Detection* (challenge of missing values), and *Duplicate Detection* (challenge of duplicate records).

Figure 5 shows the information flow of the formerly integrated data asses (refer to Figure 4) through both, AI2VIS4BigData and IVIS4BigData services. The different data transformation use cases thereby

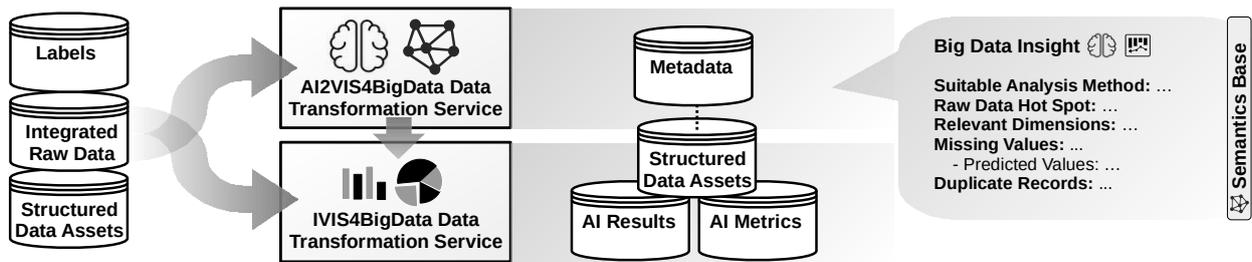


Figure 5: Data Transformation Use Case Sketch as Application of AI2VIS4BigData Services

support the IVIS4BigData transformation through generation metadata.

Use Case B3: AI-based Visual Mapping

The third use case, *AI-based Visual Mapping* (Figure 6), summarizes all application scenarios of AI in order to support the third transformation of IVIS4BigData from analyzed, structured data into visual representations. Consequently, its third level use cases are also derived from manual activities like, e.g. the configuration and selection of suiting visual representations or visualization libraries as well as from challenges like, e.g., visual noise or spurious correlations: *Visual Representation Prediction* (manual activities of visual representation configuration and selection), *Visualization Library Prediction* (manual activities of visualization library configuration and selection), *Structured Data Hotspot Detection* (manual activity of structured data selection and the challenge of incidental endogeneity and homogeneity), *Visual Noise Detection* (challenges of visual noise), and *Correlation Detection* (challenge of spurious correlation).

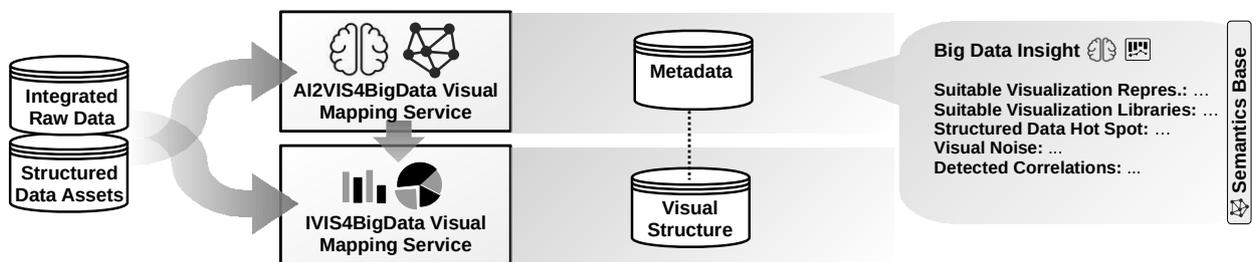


Figure 6: Visual Mapping Use Case Sketch as Application of AI2VIS4BigData Services

Use Case B4: AI-based View Transformation

The fourth and final second level use case derived from IVIS4BigData's data transformation steps is *AI-based View Transformation*. It is visualized in Figure 7 and comprises the transformation of visual structures under application of the selected and configured visualization algorithm into actual views and dashboards that can be perceived by non-expert user stereotypes [12]. Its third level use cases are *IVIS Technique Prediction* (manual activity of IVIS technique configuration), *Visualization Hotspot Detection* (manual activity of visualization selection).

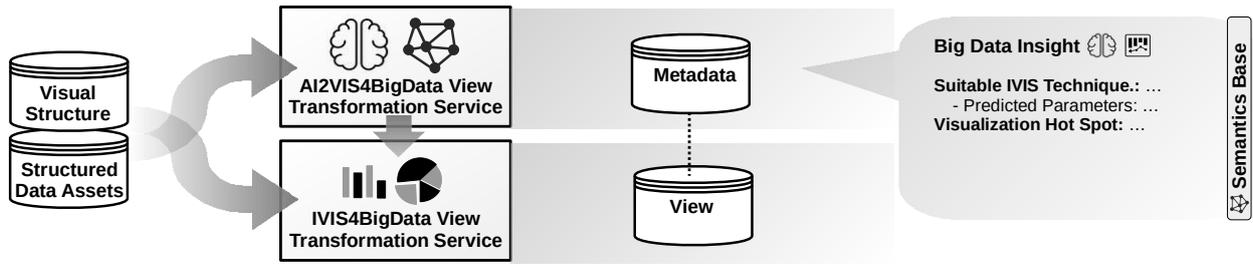


Figure 7: View Transformation Use Case Sketch as Application of AI2VIS4BigData Services

3.3. AI-based Big Data User Empowerment

The authors of this paper propose in [11] that AI-based user empowerment closely follows the principles that Fischer and Nakakoji introduced in their multifaceted architecture [15] with *Design Creation* being the expert and end user stereotype’s utilization of the Big Data analysis system with the objective to gain as much insight as possible. Following that interpretation, a *Specification* can be seen as the users’ mental model of how he or she thinks that the Big Data exploration objective can be reached, while a *Construction* is the actual articulation of this plan to the system like, e.g., executing a data query or the usage of an UI element. The likelihood that a user decides to follow more sophisticated plans increases with the users self-confidence regarding a holistic system understanding. Positive experiences with successful Big Data Analysis system usage results can be stored as best practices or rules within the *Catalog Base* of the domain knowledge layer while negative experiences can be stored within the *Argumentation Base* of the domain knowledge layer. Expert knowledge regarding Big Data Analysis workflows, useful patterns, and anti-patterns can be integrated in both. The actual meaning of the system and data state can be stored within the *Semantics Base*. The application of case-based reasoning, simulation, and critical challenging can improve the users’ self-confidence and thus increase the probability of fruitful Big Data exploration. Figure 8 visualizes the proposed interpretation of Fischer and Nakakoji’s multifaceted architecture for visual Big Data Analyses.

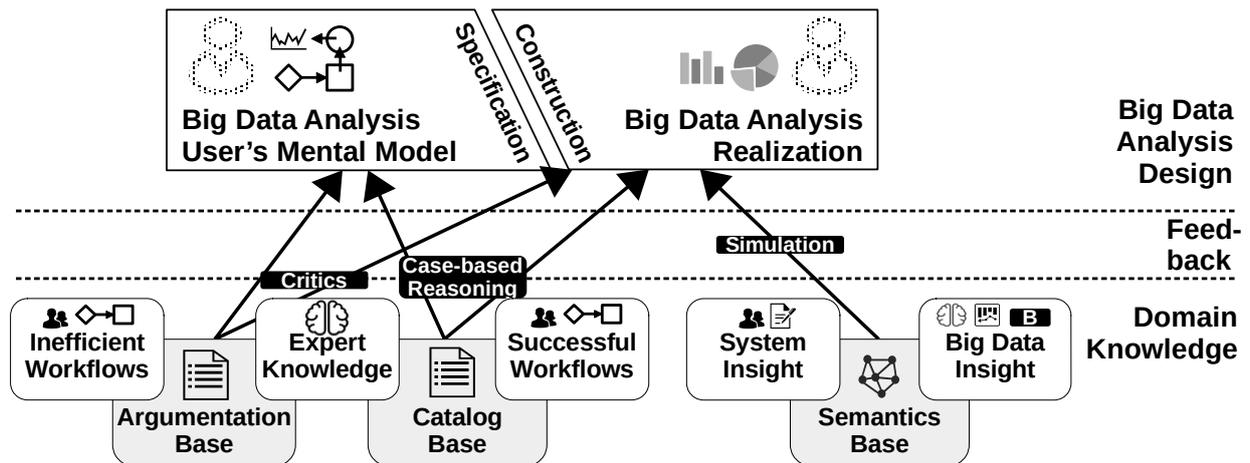


Figure 8: Multifaceted Architecture [15] Interpretation for Visual Big Data Analysis [11]

Following the interpretation in Figure 8, user empowerment for visual Big Data Analysis is closely related to the outcome of use case B: use case B applies AI in order to identify user-relevant information (*Big Data Insight*). Use case (C) then utilizes this insight and applies formalized rules, experiences, or expert knowledge to transport this information to the Big Data Analysis user stereotypes. The form of this information transport is utilized to derive two second-level use cases: *Interaction Guidance* and *Content Guidance*.

These use cases have the objective to address the four psychological principles introduced by Kim et al. [1]: strengthen the individual user's belief in his / her capabilities, sharpen the user's system understanding, raise the awareness which capabilities the system offers, and to clarify the understanding of the users influence on the system [1]. In addition, these use cases aim at minimizing the effect of the presented challenges and thereby follow the **Human-Centered Design (HCD)** approach that intends to improve system usability through "*applying human factors, ergonomics and usability knowledge and techniques*" [18]. According to the ISO9241 standard, the scientific discipline of **Human Factors and Ergonomics (HF/E)** is defined to comprise the research of human interaction with systems as well as all design activities "*to optimize human well-being and overall system performance*" [18]. Hence, use case C and its sub use cases apply methodology and technology from HCD and HF/E.

Use Case Interaction Guidance

Since "*human background knowledge, intuition, and decision-making either cannot be automated or serve as input for the future development of automated processes*" [8], the first use case *Interaction Guidance* focuses on improving the human user's interaction with the system. In more detail, the use case summarizes two different types of guiding the user interaction to make it simpler (addresses the lack of skilled personnel) or more efficient (addresses the challenge of expensive resources). The use case can be implemented as an interaction proposal (e.g. propose the action to fill data sample's missing values if missing values exist or to remove duplicates if duplicate records exist), to prioritize interaction capabilities (e.g. a user interface "*that over time automatically minimizes or even eliminates infrequently used features or menu items*" [19]), or to propose whole workflows (a set of single interactions carried out in a certain order).

Use Case Content Guidance

The second use case *Content Guidance* is closely related to the former one. Its focus lies on informing the user about the content of the Big Data Analysis itself and thereby enable Big Data Analysis user stereotypes to perceive information that might have only been accessible for very skilled and experienced data scientists. Examples range from simple tool-tips (e.g. hints that the current set of data contains missing values or duplicate records) and help dialogues that support the user [16], "*automatic warning messages*" [17] to more complex scenarios such as automatic adaption of visualizations [7] or "*systems that automate the data exploration process by discovering data objects*" [20] steering users "*towards interesting data*" [20].

3.4. Use Case Taxonomy

All derived use cases implicitly created a use case taxonomy through their ordering and through the relationships between each other. This taxonomy is visualized in Table 2. It consists of three use cases on the first level, six on the second level and 18 on the third level.

Table 2
Taxonomy for Use Cases interconnecting AI and Visual Big Data Analysis

Use Cases		
1 st Level	2 nd Level	3 rd Level
Big Data-based AI Model Design	AI-based Data Integration	Wrapper Detection
		Mediator Detection
		Data Schema Detection
		Data Model Detection
		Data Semantics Detection
		Data Inflow Prediction
		Analysis Method Prediction
	AI-based Data Transformation	Raw Data Hotspot Detection
		Dimensionality Reduction
		Missing Value Detection
		Duplicate Detection
	AI-based Visual Mapping	Visual Representation Prediction
		Visualization Library Prediction
		Structured Data Hotspot Detection
Visual Noise Detection		
AI-based View Transformation	Correlation Detection	
	IVIS Technique Prediction	
	Visualization Hotspot Detection	
AI-based Big Data User Empowerment	Content Guidance	
	Interaction Guidance	

Although the formal modeling of all use cases in 2 remains a challenge for future research, a validation of their practical relevance is required. For this purpose, this paper revisits the introduced AI2VIS4BigData expert survey [13] to compare the experts feedback with the derived use cases.

4. Initial Validation of Identified Use Cases

The derived use cases A to C will be initially validated in this Section. For this purpose, two questions from the expert survey in preparation for the AVI 2020 satellite workshop [13] will be reviewed and the experts' answers are associated to the different use cases (upon the second use case level). The result of this initial validation is summarized in Table 3. Validated use cases are visualized with a filled circle (●), non-validated use cases are visualized with an empty circle (○). The following questions have been assessed in the survey:

1. Question 1: What is the practical relevance of a given set of application scenarios for AI, Big Data Analysis, and visualization activities within the survey participant's research? (Table 1 in [13])

2. Question 2: What is the practical relevance of three different AI model types within the survey participant's research? (Figure 10 in [13])

Use case A (*Big Data-based AI Model Design*) is validated through approval of the survey participant's majority for the application scenario "*Applying Big Data to design AI models (e.g. adjust weights of a neural network)*" [13] as well as the high approval rate for the scenario "*Applying visualization to design AI models*" [13].

The survey results for use case B (*AI-based Big Data Transformation*) support only the validation use case the *AI-based Data Integration* (B1) and *AI-based Data Transformation* (B2). B1 is validated through the approval of the workshop participants to "*applying AI to integrate different sources of Big Data*" [13] while use case B2 received the participant's approval for the application scenario "*Applying AI to ease Big Data exploration (e.g. to programmatically identify outliers)*" [13]. The further sub use cases of use case B were not addressed in the survey.

The validation of use case C (*AI-based Big Data User Empowerment*) consists of the combined validation of its sub use cases. Both are validated through the very high approval rate for the application scenario "*Applying AI and Big Data to ease visualization and UI comprehension (e.g. through intelligent UI that explains and highlights useful tools)*" [13] with the explanatory aspect pointing to *Content Guidance* and the tool highlighting pointing to *Interaction Guidance*. In addition, the majority of survey participants state at least sometimes to utilize "*UI models*" [13] which strengthens the decision.

Table 3

Summary of Expert Survey [13] based Use Case Validation

Validation Method	Use Case A	Use Case B				Use Case C	
		B1	B2	B3	B4	C1	C2
Expert Survey [13]				○	○		

As Table 3 summarizes, the expert survey [13] validates the use cases A, B1, B2, C1, and C2. A validation of use cases B3 and B4 remains a challenge for future research. Further research directions include a formal UML modeling as well as an extensive validation of all use cases, the introduction of an use case framework, the technical specification of the use cases within the AI2VIS4BigData conceptual architecture (e.g. APIs and data models) as well as a prototypical implementation of them.

5. Conclusion and Outlook

This paper summarizes existing challenges in the application field AI, Big Data Analysis, and visualization, clusters the different user stereotypes involved in these application areas into experts and end users. It introduces 28 AI2VIS4BigData use cases that enable empowering experts and end users involved in the domain of visual Big Data Analysis. Furthermore, these use cases are ordered into a use case taxonomy and validated through analysis of a survey result with answers from international scientists and researchers [13]. Intensive work within one or multiple of the derived use cases, formal modeling, a comprehensive evaluation, a detailed specification, and a proof of concept implementation remain as open challenges and are future research goals.

References

- [1] H.-W. Kim, S. Gupta, A user empowerment approach to information systems infusion, *IEEE Transactions on Engineering Management* 61 (2014) 656–668.
- [2] ISO, ISO/IEC JTC 1/SC 42 Artificial Intelligence, 2018. URL: <https://isotc.iso.org/livelink/livelink/open/jtc1sc42>.
- [3] R. Bond, F. Engel, M. Fuchs, M. Hemmje, P. M. Kevitt, M. McTear, M. Mulvenna, P. Walsh, H. J. Zheng, Digital empathy secures Frankenstein’s monster, *CEUR Workshop Proceedings* 2348 (2019) 335–349.
- [4] T. Reis, M. X. Bornschlegl, M. L. Hemmje, Towards a Reference Model for Artificial Intelligence Supporting Big Data Analysis, To appear in: *Advances in Data Science and Information Engineering -Proceedings of the 2020 International Conference on Data Science (ICDATA’20)* (2021).
- [5] J. Fan, F. Han, H. Liu, Challenges of big data analysis, *National science review* 1 (2014) 293–314.
- [6] D. Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, Technical Report, META Group, 2001.
- [7] S. M. Ali, N. Gupta, G. K. Nayak, R. K. Lenka, Big data visualization: Tools and challenges, in: *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, 2016, pp. 656–660.
- [8] D. A. Keim, F. Mansmann, J. Schneidewind, H. Ziegler, Challenges in visual data analysis, in: *Tenth International Conference on Information Visualisation (IV’06)*, IEEE, 2006, pp. 9–16.
- [9] T. Reis, T. Krause, M. X. Bornschlegl, M. L. Hemmje, A Conceptual Architecture for AI-based Big Data Analysis and Visualization Supporting Metagenomics Research, *Proceedings of the Collaborative European Research Conference (CERC) (2020)* 264–272.
- [10] O. Thomas, Understanding the Term Reference Model in Information Systems Research: History, Literature Analysis and Explanation, *Proceedings of the Workshop on Business Process Reference Model (2005)* 16–30.
- [11] T. Reis, S. Bruchhaus, B. Vu, M. X. Bornschlegl, M. L. Hemmje, Towards Modeling AI-based User Empowerment for Visual Big Data Analysis, *Proceedings of the Second Workshop on Bridging the Gap between Information Science, Information Retrieval and Data Science (BIRDS 2021)* (2021) 67–75. URL: <http://ceur-ws.org/Vol-2863/{#}paper-07>.
- [12] M. X. Bornschlegl, K. Berwind, M. L. Hemmje, Modeling end user empowerment in big data applications, *26th International Conference on Software Engineering and Data Engineering, SEDE 2017* (2017) 47–53.
- [13] T. Reis, M. X. Bornschlegl, M. L. Hemmje, AI2VIS4BigData: Qualitative Evaluation of a Big Data Analysis, AI, and Visualization Reference Model, To appear in: *Lecture Notes in Computer Science LNCS 12585* (2021).
- [14] M. X. Bornschlegl, *Advanced Visual Interfaces Supporting Distributed Cloud-Based Big Data Analysis*, Dissertation, University of Hagen, 2019.
- [15] G. Fischer, K. Nakakoji, Beyond the macho approach of artificial intelligence: empower human designers—do not replace them, *Knowledge-Based Systems* 5 (1992) 15–30.
- [16] A. Kadadi, R. Agrawal, C. Nyamful, R. Atiq, Challenges of data integration and interoperability in big data, in: *2014 IEEE international conference on big data (big data)*, IEEE, 2014, pp. 38–40.
- [17] O. Azeroual, Data Wrangling in Database Systems: Purging of Dirty Data, *Data* 5 (2020) 50.
- [18] International Organization for Standardization, *ISO 9241–210: 2019 (en) Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems* (2019).
- [19] A. Blair-Early, M. Zender, User interface design principles for interaction design, *Design Issues* 24 (2008) 85–107.

- [20] S. Idreos, O. Papaemmanouil, S. Chaudhuri, Overview of data exploration techniques, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 277–281.

Video-based Emotion Detection Analyzing Facial Expressions and Contactless Vital Signs for Psychosomatic Monitoring

Hayette Hadjar^a, Binh Vu^a, Dennis Maier^a, Gwendolyn Mayer^b, Paul Mc Kevitt^c, and Matthias Hemmje^d

- a. *University of Hagen, Faculty of Mathematics and Computer Science, Hagen, Germany.*
- b. *Heidelberg University, Department of Internal Medicine II, General Internal Medicine and Psychosomatics, Heidelberg, Germany.*
- c. *Ulster University, Derry/Londonderry, Northern Ireland.*
- d. *Research Institute for Telecommunication and Cooperation, Dortmund, Germany.*

Abstract

Recently, automatic emotion recognition research is gaining attention in computer science. This paper outlines research on Affective Computing in the context of the Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) project for remote home healthcare applications. This includes: (1) developing the system architecture for monitoring emotions and vital signs using a simple camera, and (2) recognition and visualization of emotions and of physiological signals to synthesize patients' psychosomatic data for health-care providers. In this exemplar prototypical implementation we employ Convolutional Neural Networks (CNNs) and remote PhotoPlethysmoGraphy (rPPG) methods for recognition of psychosomatic states during patient monitoring.

Keywords

Continuous Emotion Recognition, Facial Expressions, Remote Photoplethysmography (rPPG), Contactless Physiological Signals, Affective Computing

1. Introduction and Motivation

Psychosomatic medicine is becoming a key medical specialty that targets a deeper understanding of the physical, emotional and social causes for a disease [1]. Relevance of the corresponding treatment of patients has been significantly magnified by the coronavirus pandemic [2] and it has boosted the practice of telemedicine [3]. It appears tele-home health care settings give a rich landscape for research and application of affective computing to improve the quality of patient care [4]. This work has been developed in the context of the Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) project [5]. SenseCare is a research and innovation project which was initially funded by the European Union [6] [7]. SenseCare focuses on developing a number of input interfaces for specific sensory devices, e.g., cameras and wearable sensors from the Internet of Things. Several analysis methods for emotion recognition within the web platform of SenseCare have already been developed. Three analysis methods are currently available on the SenseCare KM-EP (Knowledge Management Ecosystem Platform): (1) analysis based on Support Vector Machines according (see Healy et al. [8]), (2) Artificial Neural Networks (see Maier [9]), and (3) Convolutional Neural Networks with TensorFlow (see Hadjar et al. [10]). In the world of Big Data, data visualization tools and technologies are essential to analyze large sets of information and to make data-driven decisions. The facial appearance of patients may indeed give diagnostic clues to maladies, severity of diseases, and their vital parameters [11]. When it comes to very large and complex

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ hayette.hadjar@fernuni-hagen.de (H. Hadjar); binh.vu@fernuni-hagen.de (B. Vu); dennis.maier@fernuni-hagen.de (D. Maier); gwendolyn.mayer@med.uni-heidelberg.de (G. Mayer); p.mckevitt@ulster.ac.uk (P. McKevitt); mhemmje@ftk.de (M. Hemmje);

🌐 <http://www.paulmckevitt.com/> (P. McKevitt);

📄 000-0001-9540-6473 (H. Hadjar); 0000-0001-9715-1590 (P. McKevitt);



© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

datasets, the functionality of the data visualization library at play is also important for effective insights. In this preliminary study, we are primarily interested in the following two goals:

- Visualization and perception of all collected audio, video, and emotion feature annotations or graphical representation of emotions over time, in order to make optimal healthcare decisions.
- Extraction of physiological signals (pulse rate, respiratory rate) from the camera, detection of potential changes of various emotions, and characterization of these changes.

The remainder of this paper is organized as follows. Section 2 discusses the state of the art of video-based facial expression emotion recognition. Section 3 details the conceptual design of an architecture for recognition and visualization of facial expressions and contactless vital signs. Section 4 describes the implementation of the proposed system, and finally we conclude and discuss future work in Section 5.

2. State of the Art and Related Work

Machine learning algorithms are applied to a wide variety of areas, such as spam filtering, speech recognition, face recognition, document classification and the processing of natural language. Classification is one of the most common areas of machine learning application. Recently, Video Facial Emotion Recognition research has received focus of attention in the computer vision community. In the task of emotion recognition, varied types of input data are employed such as: facial expressions, speech, physiological parameters, and body gestures. There are many approaches for detection of emotion from facial expressions, e.g., the work of Michel Healy [8], where an emotion detection system is described based on a video feed in real-time, and employs a machine learning Support Vector Machine (SVM) to provide quick and reliable classification. Features employed in [8] are 68-point facial landmarks. The application has been trained to detect 6 different emotions by monitoring changes in facial expressions. The work of Dennis Maier [9] currently uses neural networks via TensorFlow to train image features and then achieves classification through fully connected neural layers. The advantage of image features over facial landmarks is the larger information space, where the spatial formation of landmarks gives a viable method for analyzing facial expressions. However, this is also accompanied by a higher computing power requirement. The structure provides for an outsourced classification service that runs on a server with a GPU. Images of faces are brought to the service in real time, which can perform a classification within a few milliseconds. In future, this approach will be extended to include text and audio features and conversation context to boost accuracy. Another approach uses Convolutional Neural Networks with TensorFlow [12]. An example for using TensorFlow.js with the SenseCare KM-EP is discussed in [8], which deploys a web browser and Node Server [13].

Furthermore, Speech Emotion Recognition is an important field employed by numerous multimodal sentiment analysis systems, e.g., robotics, security, automated identification, and language translation [14]. The MENHIR project [15] aims to investigate conversational technologies to promote mental health and assist people with mental ill health in managing their conditions. In the context of MENHIR, a new emotion-recognition system based on speech is being developed [16]. Physiological signs are detected such as facial electromyography, facial color patterns, blood volume pulse, and galvanic skin response. The research project “EEG-based Emotion Recognition” [17] leads in recognizing emotion from brain signals measured with the Bra Inquiry EEG PET device. Furthermore, the work in [18] implements a wearable sensing system for effective recognition of user emotional states, by employing heart rate, body temperature, and galvanic skin response sensors. The proposed system achieves up to 97% recognition accuracy when it adopts the k-nearest neighbor (KNN) classifier. For contactless vital signs, several methods are available for camera-based Heart Rate (HR), Heart Rate Variability (HRV), and Respiration Rate (RR) detection such as Remote PhotoPlethysmography (rPPG) where [19] describes a popular technique of non-contact measure HR from facial videos, and [20] describes how RR works. Recently, studies on deep learning based rPPG methods have been introduced such as [21] and [22]. Eulerian Video Magnification [23] applies spatial decomposition to the input standard video sequences, where amplification reveals hidden information in resulting signals. Other research employs Convolutional Neural Networks for remote pulse rate measurement and mapping from facial video, such as the DeepPhys Convolutional Neural

Network [24] and 3D Convolutional Neural Network [25]. Furthermore, several approaches exist for body gesture recognition, e.g., Noroozi et al. [26]. This survey classifies body language into 6 basic emotions: fear, anger, sadness, surprise, happiness, and disgust. Another study [27] introduces deep learning models for multivariate time series, from vehicle control to gesture recognition and generation.

3. Conceptual Architecture

The processing of large data streams from audio-video recordings (real-time and offline) to data visualization follows the architecture of the Affective Computing Strata (AC-Strata) model [28] where the S-Strata relates to an individual mode of affective monitoring. The S-Strata Information model indicates that under any AC analytical method, one, or any number of specific sensors of embedded devices, may participate. The S-Strata model describes the time aspects of the sensors that are integrated in a specific manner and also stipulates contextual and manual representations dealing with both internal and external features. Accordingly, the conceptual architecture of our SenseCare affective computing platform based on audio- and video-based emotion recognition within the SenseCare KM-EP is detailed in Figure 1.

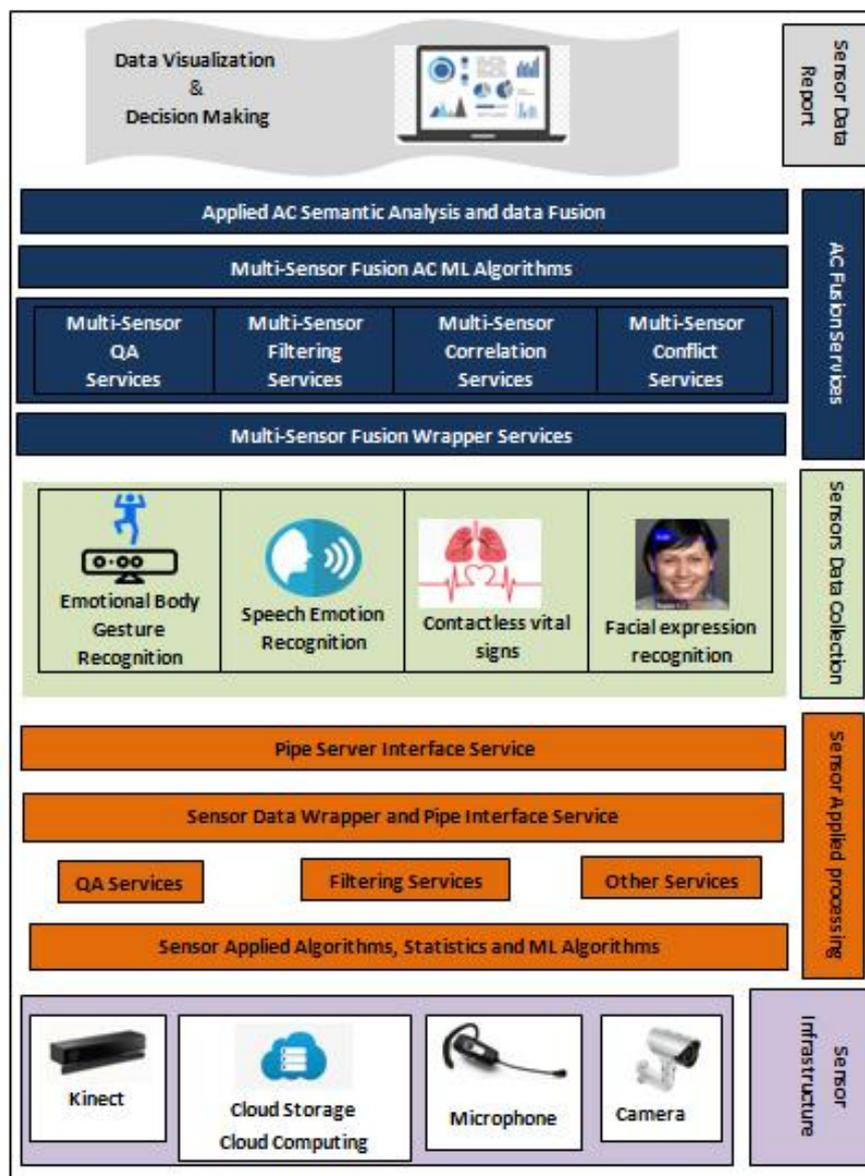


Figure 1: Conceptual Architecture of the SenseCare Affective Computing Platform

The conceptual architecture design is structured as follows. First of all, in the sensor & infrastructure area, sensors used are: (1) Camera (for Facial Expression Recognition, and Contactless

Vital Signs [HR, RR, HRV]), (2) Microphone (for Speech-based Emotion Recognition), and (3) Kinect (for Body Gesture emotion recognition). Furthermore, cloud storage & cloud computing is employed here as an infrastructure for data processing and storage. Secondly, processing applied to sensor data includes facial expression-based emotion recognition employing analysis methods discussed above, i.e., Support Vector Machines (Healy et al. [8]), Artificial Neural Networks (Maier [9]), and Convolutional Neural Networks with TensorFlow (Hadjar et al. [10]), deployed in a web browser and Node Server [13]. Finally, sensor data feature collection includes the following 4 types of Affective Computing data features: (1) Facial expression-based emotion recognition, (2) Speech-based emotion recognition, (3) Contactless Vital Signs (HR, RR, HRV), and (4) Body-Gesture based emotion recognition. On the higher level of the architecture, the AC Fusion services are responsible for the fusion of the collected AC data features. On top of this data fusion layer the architecture user interface layer generates advanced services for sensor data reporting including data visualization and decision making support available to end users.

4. Implementation

The prototypical implementation of this research concentrates on extracting emotion-related features from images of the human face, ideally in real-time, from a single sensor which is the camera. Hence, the current state of implementation of AC sensor data collection is the facial-expression based emotion recognition and the vital signs recognition. An implementation of a prototypical module that collects patients' facial expression and corresponding emotion data in real-time was already described in detail in our previous work [10]. The API allowed monitoring patients during treatment sessions, or at home for cases of patients with or at risk for a mental disorder. The software classifies emotions into 7 categories: happy, sad, angry, disgusted, fearful, neutral, and surprised, as determined by Paul Ekman [29]. The prototype employs deep learning in browsers using JavaScript. In the case of real-time video-based emotion analysis, the SenseCare KM-EP's Emotion Detection API stores the most significant emotion detected from the video every 500 milliseconds into a database. The face expression recognition model employs depthwise separable convolutions and densely connected blocks. The Perceptron model is a linear transformation, where the activation function makes it possible for the model to approximate non-linear functions (e.g. Softmax, Sigmoid). We have chosen line chart graphs based on Chart.js [30] to visualize stored emotions in the database during time periods of, e.g., the last 24h, the last 3 days, the last week, and the last month. Graphs showing the score distributions of emotions over time are given in Figures 2-5 below.



Figure 2: Emotions detected (last 24 hours)



Figure 3: Emotions detected (last 3 days)



Figure 4: Emotions detected (last week)

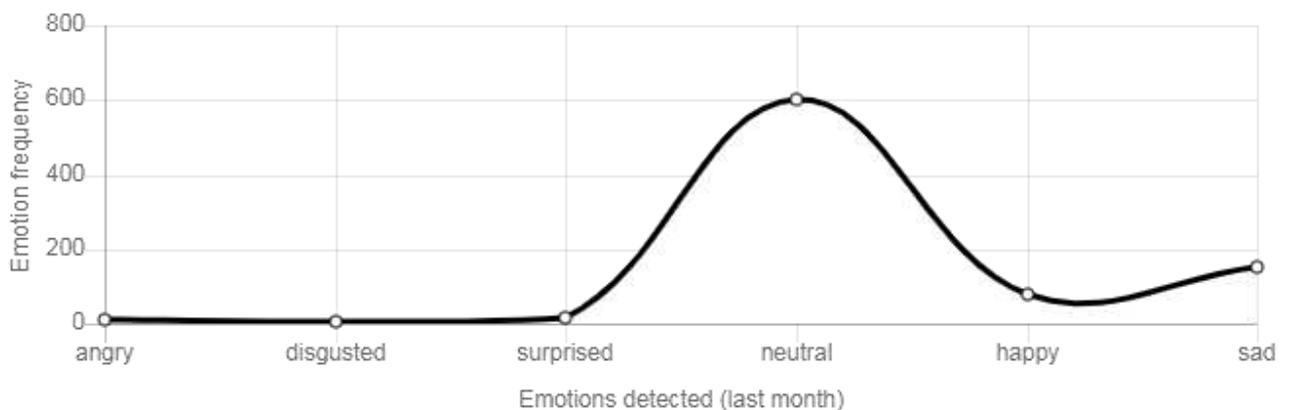


Figure 5: Emotions detected (last month)

Data acquired over several days (last 30 days) makes it possible to build a comprehensive picture compared to that acquired over a single day (last 24h). This data is potentially useful for analyzing psychosomatic states.

The development of video-based contactless sensing of vital signs yields opportunities for scalable physiological monitoring. For the measurement of stress, Heart Rate Variability is the focus of much work. Experimentation with a video-based pulse rate monitoring implementation in browser-based rPPG employing the Viola-Jones algorithm [31] for defining the Region of Interest and Haar-like features [32] for face detection was tested as shown at the following URL: <https://studev4.fernuni-hagen.de:20286/>

In respect of Data Fusion, visual facial expression analysis may in many cases be insufficient for emotion recognition [33]. Hence, we suggest in this approach the fusion of vital factors with the analysis of facial expressions to give greater precision of predicted results. The multimodal approach has been proposed for enabling more reliable recognition with minimal ambiguities that can arise

whilst using a single data channel. Bimodal emotion analysis of camera-based vital sign and facial expression recognition is shown in Figure 6.

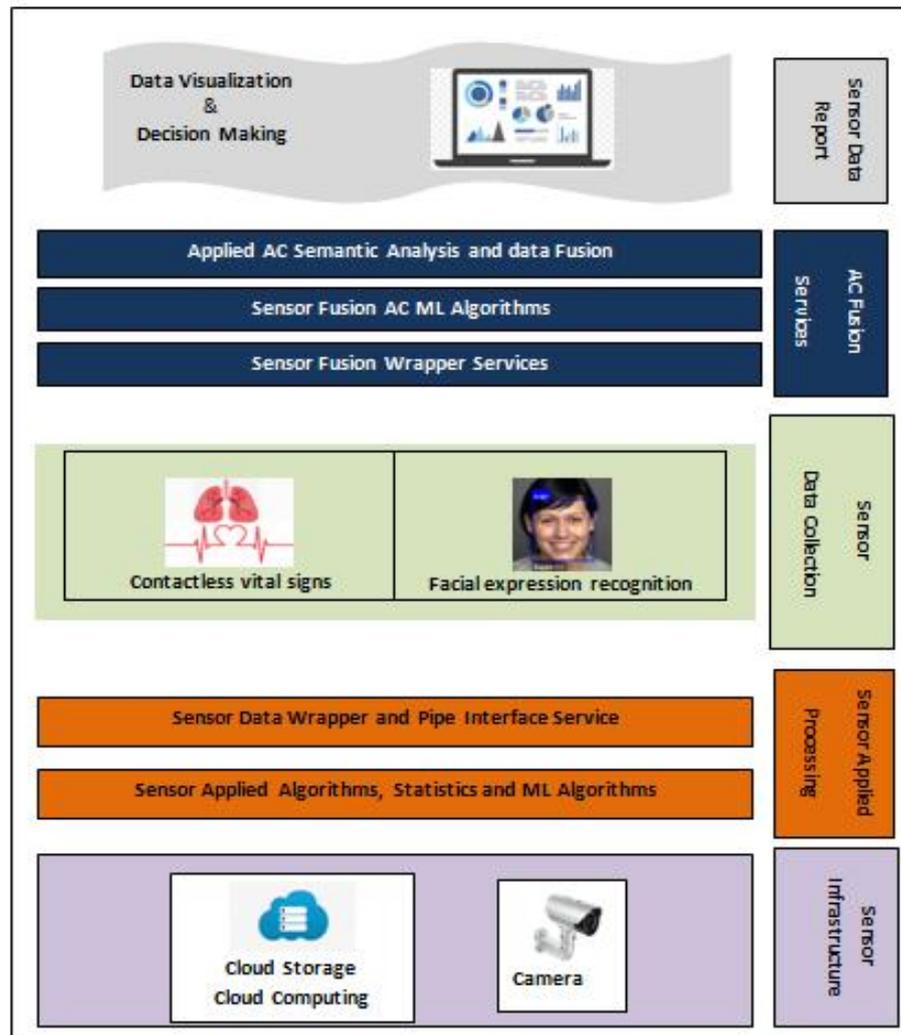


Figure 6: Multimodal Emotion Recognition based on Video

In this context, we must also handle the real-time tasks of acquiring and merging information and system performance.

5. Conclusion and Future Work

In this study, an approach of detection and recognition of emotions from the camera is proposed; bimodal systems will analyze facial emotions using deep learning and other algorithms by merging the two collected emotions. The current system recognizes emotions from facial expressions using Convolutional Neural Networks with Tensorflow.js, and heart pulse using the rPPG technique with the Viola-Jones algorithm, and finally displays results in the web browser. The advantages of the proposed system are:

- All the processing of the AI facial emotion analysis is done on the client machine, and no data such as real-time video is sent to the server.
- It is low cost with more comfort for the patient (no contact with skin).
- The system is accessible from any device with a camera.
- The system enables displaying a graphical result in a variety of devices, such as a Smartphone.
- The Node development environment supports a variety of popular libraries such as D3.js and Ggplot2, which can be included later, and enables collaborative visualization.

In this paper, we introduced a conceptual architecture of an affective computing platform supporting audio- and video-based emotion recognition within the SenseCare KM-EP. A new approach is proposed which includes facial expression recognition and contactless vital signs to provide an initial psychosomatic monitoring for the SenseCare KM-EP. The proposed system offers computer diagnostics and evaluation of emotions supporting diagnosis and treatment of psychosomatic illnesses. Our next steps are: (1) fusion of AC data and global visualization of emotions collected based on video and audio, (2) investigate the treatment of anxiety using eye-tracking and heart rate variability, and test and build different datasets and models, (3) develop a mobile application to offer a variety of choices for our users and improve the accessibility of our system, (4) investigate user data privacy, and (5) develop other systems with ability to detect different types of emotions and potential diseases. In future work, it will be important to evaluate the bimodal system with real patients to improve their performance.

6. References

- [1] K. Fritzsche, S. H. McDaniel, and M. Wirsching, Eds., *Psychosomatic Medicine*. 2020.
- [2] Yu-Tao Xiang, “Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed, *Lancet Psychiatry*, vol. 7, no. 3, pp. 228–229, 2020.
- [3] J. Torous, K. Jän Myrick, N. Rauseo-Ricupero, and J. Firth, *Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow*, *JMIR Ment. Heal.*, vol. 7, no. 3, p. e18848, Mar. 2020, doi: 10.2196/18848.
- [4] C. L. Lisetti and C. LeRouge, *Affective computing in tele-home health: design science possibilities in recognition of adoption and diffusion issues*, in *HICSS 2004, 37th IEEE Hawaii International Conference on System Sciences*, January 5-8, 2004, Hawaii, USA, 2004.
- [5] F. Engel et al., *SenseCare: Towards an Experimental Platform for Home-Based, Visualisation of Emotional States of People with Dementia*, in *Advanced Visual Interfaces. Supporting Big Data Applications*, 2016, pp. 63–74.
- [6] *Sensor Enabled Affective Computing for Enhancing Medical Care | SenseCare Project | H2020 | CORDIS | European Commission*. URL: <https://cordis.europa.eu/project/id/690862/fr>.
- [7] *SenseCare: Sensor Enabled Affective Computing for Enhancing Medical Care | FTK – Research Institute for Telecommunication and Cooperation*. URL: <https://www.ftk.de/en/projects/senscare>.
- [8] M. Healy, R. Donovan, P. Walsh, and H. Zheng, *A Machine Learning Emotion Detection Platform to Support Affective Well Being*, in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2694–2700, doi: 10.1109/BIBM.2018.8621562.
- [9] D. Maier, *Analysis of technical drawings by using deep learning*, Master’s thesis, Department of Computer Science, Hochschule Mannheim University, Germany, 2019.
- [10] H. Hadjar, *Video-based automated emotional monitoring in mental health care supported by a generic patient data management system*, 2020.
- [11] M. Leo and G. M. Farinella, *Preface in Computer Vision for Assistive Healthcare*, M. Leo and G. M. Farinella, Eds. Academic Press, 2018, pp. xxi–xxiii.
- [12] M. N. A. Wisal Hashim Abdulsalam, Rafah Shihab Alhamdani, *Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks*, *Int. J. Mach. Learn. Comput.*, vol. 9, no. 1, pp. 14–19, 2019.
- [13] Node.js, URL: <https://nodejs.org/en/>
- [14] M. El Ayadi, M. S. Kamel, and F. Karray, *Survey on speech emotion recognition: Features, classification schemes, and databases*, *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: <https://doi.org/10.1016/j.patcog.2010.09.020>.
- [15] *Mental health monitoring through interactive conversations | MENHIR Project | H2020 | CORDIS | European Commission*. URL: <https://cordis.europa.eu/project/id/823907>.
- [16] B. Vu., *A Content and Knowledge Management System Supporting Emotion Detection from Speech BT - Conversational Dialogue Systems for the Next Decade*, L. F. D’Haro, Z. Callejas, and S. Nakamura, Eds. Singapore: Springer Singapore, 2021, pp. 369–378.

- [17] D. Bos, EEG-based Emotion Recognition The Influence of Visual and Auditory Stimuli, 2007.
- [18] B. Myroniv, C.-W. Wu, Y. Ren, and Y.-C. Tseng, Analysis of Users' Emotions Through Physiology, in *Genetic and Evolutionary Computing*, 2018, pp. 136–143.
- [19] X. Ma, D. P. Tobón, and A. El Saddik, Remote Photoplethysmography (rPPG) for Contactless Heart Rate Monitoring Using a Single Monochrome and Color Camera, in *Smart Multimedia*, 2020, pp. 248–262.
- [20] M. Chen, Q. Zhu, H. Zhang, M. Wu, and Q. Wang, Respiratory Rate Estimation from Face Videos, in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4, doi: 10.1109/BHI.2019.8834499.
- [21] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, Heart Rate Estimation From Facial Videos Using a Spatiotemporal Representation With Convolutional Neural Networks, *IEEE Trans. Instrum. Meas.*, vol. 69, no. 10, pp. 7411–7421, 2020, doi: 10.1109/TIM.2020.2984168.
- [22] Q. Zhan, W. Wang, and G. de Haan, Analysis of CNN-based remote-PPG to understand limitations and sensitivities, *Biomed. Opt. Express*, vol. 11, no. 3, pp. 1268–1283, Mar. 2020, doi: 10.1364/BOE.382637.
- [23] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, Eulerian Video Magnification for Revealing Subtle Changes in the World, *ACM Trans. Graph. - TOG*, vol. 31, 2012, doi: 10.1145/2185520.2185561.
- [24] W. Chen and D. McDuff, DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks, in *Computer Vision -- ECCV 2018*, 2018, pp. 356–373.
- [25] F. Bousefsaf, A. Pruski, and C. Maaoui, 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video, *Appl. Sci.*, vol. 9, p. 4364, 2019, doi: 10.3390/app9204364.
- [26] F. Noroozi, C. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, Survey on Emotional Body Gesture Recognition, *IEEE Trans. Affect. Comput.*, vol. 12, no. 02, pp. 505–523, 2021, doi: 10.1109/TAFFC.2018.2874986.
- [27] G. Devineau, Deep learning for multivariate time series : from vehicle control to gesture recognition and generation, *Université Paris sciences et lettres*, 2020.
- [28] A. Keary, *Affective Computing for Emotion Detection using Vision and Wearable Sensors*, 2018.
- [29] P. Ekman and G. Yamey, Emotions revealed: recognising facial expressions: in the first of two articles on how recognising faces and feelings can help you communicate, Paul Ekman discusses how recognising emotions can benefit you in your professional life, *Student BMJ*, vol. 12, pp. 140–142, 2004.
- [30] Chart.js | Open source HTML5 Charts for your website.URL: <https://www.chartjs.org/>
- [31] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1, pp. I–I, doi: 10.1109/CVPR.2001.990517.
- [32] T. Mita, T. Kaneko, and O. Hori, Joint Haar-like features for face detection, in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2005, vol. 2, pp. 1619–1626 Vol. 2, doi: 10.1109/ICCV.2005.129.
- [33] A. Kwasniewska, J. Ruminski, M. Szankin, and K. Czuszynski, Remote Estimation of Video-Based Vital Signs in Emotion Invocation Studies, in *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2018, vol. 2018, pp. 4872–4876, doi: 10.1109/EMBC.2018.8513423.

Chapter 2

Data Processing and Machine Learning

Improving Machine Translation Quality Estimation Using Named-Entity Masking and Assessment Scores

Anthony Reidy^a, Sean Cummins^a, Kian Sweeney^a, George Dockrell^a, Pintu Lohar^b and Andy Way^b

^a*School of Computing, Dublin City University, Dublin 9, Ireland*

^b*ADAPT Centre, Dublin City University, Dublin 9, Ireland*

Abstract

This paper reports our findings in quality estimation (QE) in machine translation (MT) using the data set of WMT-2020 shared task. We perform sentence-level direct assessment (DA) and focus on the English–Chinese, Romanian–English, and English–German language pairs. We build on the single XLM-R transformer model within the state-of-the-art *Transquest* system [1] through named entity (NE) masking and analysis of quality assessment scores. Our methodologies result in the improvement in *Transquest* system for all of our chosen language pairs by achieving a higher Pearson correlation. We also obtain a reduction in error for all of these language pairs.

Keywords

quality estimation, machine translation, TransQuest, named-entity masking

1. Introduction

The MT quality estimation frameworks attempt to estimate the quality of translation outputs at varying levels of granularity: word, phrase, sentence, and document, without access to ‘gold-standard’ human-generated reference translations [2]. It can greatly reduce the cost associated with this evaluation process and also has the added ability of determining whether a machine-generated translation can be published as is, or whether it requires human post-editing efforts [3]. In this work, we use the framework developed by the TransQuest team [4] that achieved the best results in the WMT-2020 shared task of quality estimation. However, they mentioned in their error analysis that the presence of NEs causes the largest number of errors between their predicted scores and expected scores within their system. Considering this problem, we propose an approach of NE masking (discussed in details later in Section 5.1) with an aim of improving the performance of the QE system.

In addition, we utilise the following assessment metrics in our experiments to further extend our contribution.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ anthony.reidy3@mail.dcu.ie (A. Reidy); sean.cummins26@mail.dcu.ie (S. Cummins);

kian.sweeney27@mail.dcu.ie (K. Sweeney); george.dockrell2@mail.dcu.ie (G. Dockrell);

pintu.lohar@adaptcentre.ie (P. Lohar); andrew.way@dcu.ie (A. Way)

🌐 <https://github.com/reidya3> (A. Reidy)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- Root mean square error (RMSE)
- Mean absolute error (MAE)
- Spearman correlation coefficient
- Pearson correlation coefficient

We place an emphasis on the Pearson correlation metric as this is the metric used to compare models in the WMT-2020 shared task.

For the purpose of this research, we focus on the first task which is sentence-level direct assessment. We examine the (i) English–German, (ii) Romanian–English, and (iii) English–Chinese language pairs because we are interested in investigating how our approach performs with language pairs of the same and different scripts.

The remainder of this paper is organised as follows. We discuss the related works in this field in Section 2. The description of the DA dataset is provided in Section 3. Section 4 outlines our system architecture. Section 5 explains our experiment where we discuss the methodologies we use in this work to improve the performance of *MonoTransQuest* (the name of the system developed by the *TransQuest* team). We provide our results in Section 6. Finally, we conclude our work and provide some possible directions for the future in Section 7.

2. Related Work

A considerable amount of work has been done in the area of MT quality estimation. Fomicheva et al. [5] apply both a glass box and black box approach that achieves impressive results across a number of language pairs. This glass box approach uses features extracted from the NMT model and is very cost effective. However, it is their black box model, which looks at pre-trained representations using source and target text, that tied for the winning submission in four of the seven language pairs. In a similar manner, Moura et al. [6] uses this newly available features of the NMT model to further extend the OpenKiwi system [7] by using a Kiwi glass box ensemble alongside an OpenKiwi-based submission. This glass box extracts these features and feeds them into the OpenKiwi system.

Nakamachi et al. [8] makes use of an ensemble model of four regression models based on XLM-R [9], adding a language token for each sentence while Hu et al. [10] also uses an ensemble model with transfer learning and multilingual pretrained models. Zhou et al. [11] exposes explicit cross lingual patterns to zero-shot models in order to augment BERT scores. Ranasinghe et al. [1], the winning submission for WMT-2020 uses crosslingual embeddings to remove the dependency on parallel data using a pre-trained XLM-R large transformer model. This simplifies the complex neural network architecture and hence reduces the computational cost. However, the authors of this winning team mentioned that one of the main problems in their system was caused by the presence of NEs. The proper handling of NEs in quality estimation task is still less explored. In this work, we address this problem by NE masking in combination with the analysis of quality assessment scores.

Table 1
English–German training set record

Column	Value
Original	The burning propellant generates inert gas which rapidly inflates the airbag in approximately 20 to 30 milliseconds.
Translation	Das brennende Treibmittel erzeugt inertes Gas, das den Airbag in etwa 20 bis 30 Millisekunden schnell aufblasen lässt.
Scores	[55, 62, 85]
Mean	67.33333333333333
Z_scores	[-1.5259978765984137, -0.3605398796887971, -0.4917681450113435]
Z_mean	-0.792768633766184
Model_scores	-0.374686628580093

3. Description of the DA Dataset

The organizers of the WMT-2020 shared task on quality estimation provided the participants with the data sets extracted from Wikipedia for six language pairs, including our language pairs of interest. The training sets consists of 7,000 sentence pairs for each language pair whereas both the development and test sets contain 1,000 sentence pairs each. Preliminary analysis of both the training and the test data using latent dirichlet allocation (LDA) and hierarchical latent dirichlet allocation (HLDA) produced inconclusive results. The data comes from a variety of Wikipedia articles and as such, there are no general themes. Table 1 shows the example of an English–German training set record with the following entities.

- **Original:** The source sentence from a Wikipedia article.
- **Translation:** The translation of the source sentence produced by the state-of-the-art transformer-based NMT models, built using the fairseq toolkit [12].
- **Scores:** A score denoting the perceived quality of a translation, ranging from 0-100. Professional translators followed the FLORES guidelines [13] when manually annotating the DA scores. The number of annotators range from three to six.
- **Mean:** The collective scores of the raters for the translation are averaged to obtain this value.
- **Z_scores:** A list representing the standardised scores.
- **Z_mean:** The mean of the z-scores. Our architecture seeks to predict the mean DA z-scores of the test sentence pairs.
- **Model_scores:** The baseline QE system for the shared task is an LSTM-based Predictor-Estimator approach [14], implemented in OpenKiwi [3, 15]. This feature is provided to foster improvements over the described baseline.

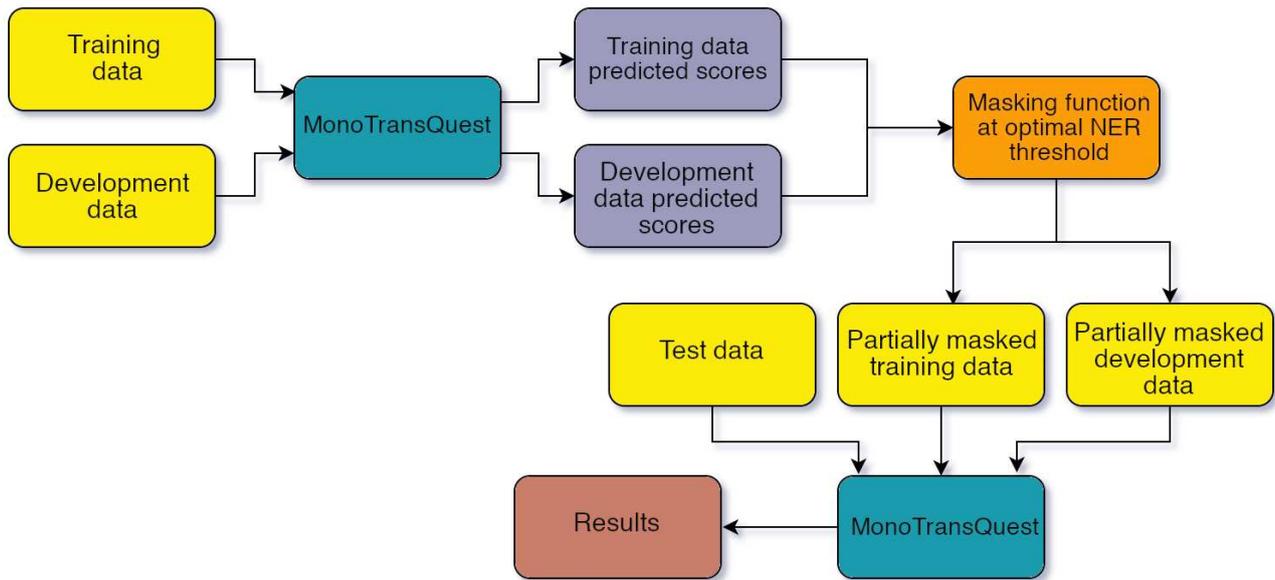


Figure 1: Architecture of our MTQE system

4. System Architecture

Our system architecture, shown in Figure 1, focuses on improving the results from the baseline *MonoTransQuest* system by incorporating our proposed approach.

Firstly, both the training and development data for a chosen language pair are passed through a standard run of *MonoTransQuest* where each tuple is assigned a predicted z-mean score. We calculate the absolute error (AE) for each tuple as the absolute difference between the predicted and actual z-mean scores.

We experiment with various AE thresholds (discussed later in Section 5). Once the masking process is complete for both the training and development data sets, the data is passed through *MonoTransQuest* again to obtain the final results. The improvements resulting from this architecture are highlighted and discussed in Section 6.

5. Experiments

In this section we present the methods we used in our experiments in chronological order.

5.1. NE masking

The *TransQuest* team documented their system’s difficulties with NEs in their paper. The occurrence of NEs in the source and target language sentences causes a large proportion of errors. The source-language sentences containing NEs regularly have translations that contain slight misspellings. *TransQuest* seems to penalise such occurrences greatly. In order to address this problem, we use *spaCy* [16], an open-source software for advanced natural language processing (NLP). In addition to *spaCy*, we use Stanford NLP’s *Stanza* [17], the Python equivalent of the originally Java-based Stanford NLP. *spaCy* and *Stanza* offer different language models

Table 2

An example of a Romanian–English instance after NE masking

Language	Input	Output
Romanian	În urmă explorărilor Căpitanului James Cook, Australia și Noua Zeelandă au devenit ținte ale colonialismului britanic.	În urmă explorărilor Căpitanului NE8734, NE27 și NE4612 au devenit ținte ale colonialismului britanic
English	Following the explorations of Captain James Hook, Australia and New Zealand became targets of British colonialism.	Following the explorations Captain NE5123, NE78113 and NE892 became targets of British colonialism.

meaning our choice of toolkit is dependant on the language pairs of our interest. We apply the following approaches for NE masking.

1. *The Regex¹ method* allows us to specifically target equivalent entities between the source- and target-language texts. This method is only practical for the English–German language pair for the reasons outlined in the ‘*The Above MAE method*’.

The Regex package provides a fuzzy-match functionality that allows users to find similarities between texts that may be less than 100% perfect. For example, in the German sentence ‘*Er regierte unterdrückerisch und fast bankrott Mali*’, a search for ‘*bankrupt*’ will return ‘*bankrott*’. Regex does this by allowing a certain amount of letter substitutions, deletions, and insertions. These operations contribute to an overall edit-distance.

For every NE found in an English sentence, the corresponding German translation is searched for sub-strings that are similar to the English NE within a specified amount of text alterations. We then mask the NE with the same ‘**NEXXX**’ value that is used for that NE in the source sentence. After experimentation, we deduce that allowing our matching function a total of 5 letter deletions and 5 letter replacements produces the best quality matches for this language pair. We also experiment with masking corresponding NEs with different values as opposed to the same value method mentioned previously. Masking with different values achieves better results.

2. *The double model method* involves using separate language models for the source- and the target-language texts. We use the source-language NE model to detect NEs within the source-language text, which once found, are replaced by an ‘**NEXXX**’ string, where ‘**XXX**’ is a uniformly random distributed number between 0 and 8,000 as rarely did we identify more than 8,000 NEs. An example of NE masking can be found in Table 2.

The target-language NE model is used to mask entities in the target-language text in an identical fashion. *spaCy* is used for the Romanian–English language pair whilst *Stanza* is used for the English–Chinese and English–German language pairs.

¹Python Regex - <https://pypi.org/project/regex/>

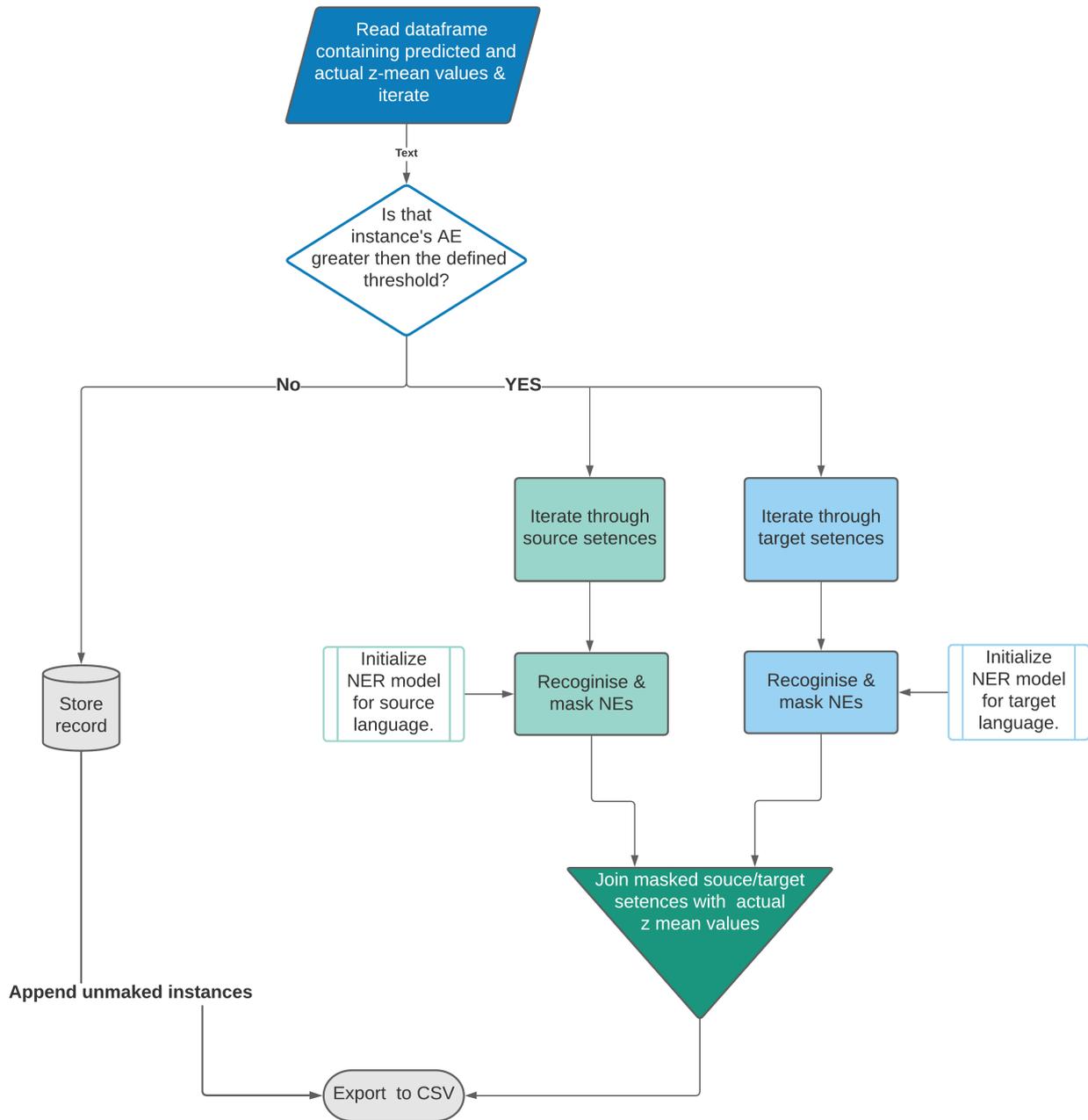


Figure 2: Workflow of the NE masking system

3. *The Above MAE method* is a less naive adaptation of the previous Regex method for the English–German language pair and the double model method for the Romanian–English and English–Chinese language pairs. It is impractical to use the Regex method for the two aforementioned language pairs due to large fundamental differences between the languages. English and Romanian are not part of the same language family whilst English and Chinese have different alphabets.

5.2. Analysis of the assessment scores

The TransQuest team indicates that they performed error analysis where the difference between the predicted score and actual score is the largest. We propose a feedback mechanism in order to retrieve the predicted z-mean scores of both the training and development data. We then calculate the value of AE between the expected and predicted z-mean value for each record in the data set. This is shown in Equation (1)

$$AE_i = |y_i - x_i| \quad (1)$$

where AE_i = the absolute error of an instance, y_i = the predicted z-mean value by *MonoTransQuest* of that instance, and x_i = the actual z-mean value of that instance calculated from the professional translators' scores. All the found NEs, in sentence pairs that have an AE greater than the MAE of standard *MonoTransquest* for that language pair are masked. Instances less than the MAE are left unchanged.

MAE is chosen as RMSE gives a higher weight to large errors. In addition, RMSE does not just simply describe the average errors. It also describes other characteristics which are often difficult to understand and comprehend. Willmott and Matsuura [18] suggest that the RMSE is not a good indicator of average model performance, and might be a misleading indicator of average error. Thus, the MAE is a better metric for this purpose. This method negates the possibility of our masking function reducing the score of sentence pairs that *MonoTransQuest* can predict relatively well.

4. *The Specific Threshold method* is an improvement of the above MAE method. Instead of masking all instances that have an AE value above their respective MAE value, we experiment with different thresholds. In this method, the training and development data are sorted by decreasing AE values.

In order to find the optimal setting for this method, we experiment with different threshold values. Through rigorous testing in the range of 5% to 1% (i.e 5% of training and development sentence pairs with the highest AE values are masked), we find the value that results in the highest increase in Pearson correlation.

6. Results

We first evaluate our framework with *the Above MAE method* followed by the *the Specific Threshold method*. Although Pearson correlation is the most commonly used evaluation metric in WMT quality estimation shared tasks, we decide to incorporate other metrics in our evaluation in order to gain a complete understanding of our architecture's performance and limitations. We choose *MonoTransQuest* as our baseline as this was the strongest system in WMT-2020 and what our architecture is built on. The results of our methods are shown in Figure 3 and also in Tables 3, 4, and 5.

Table 3 shows the results for the English–Chinese language pair. We achieve our highest Pearson correlation boost of 0.2204 (this is a 40% improvement on the *MonoTransquest* system) using *the Above MAE method*.

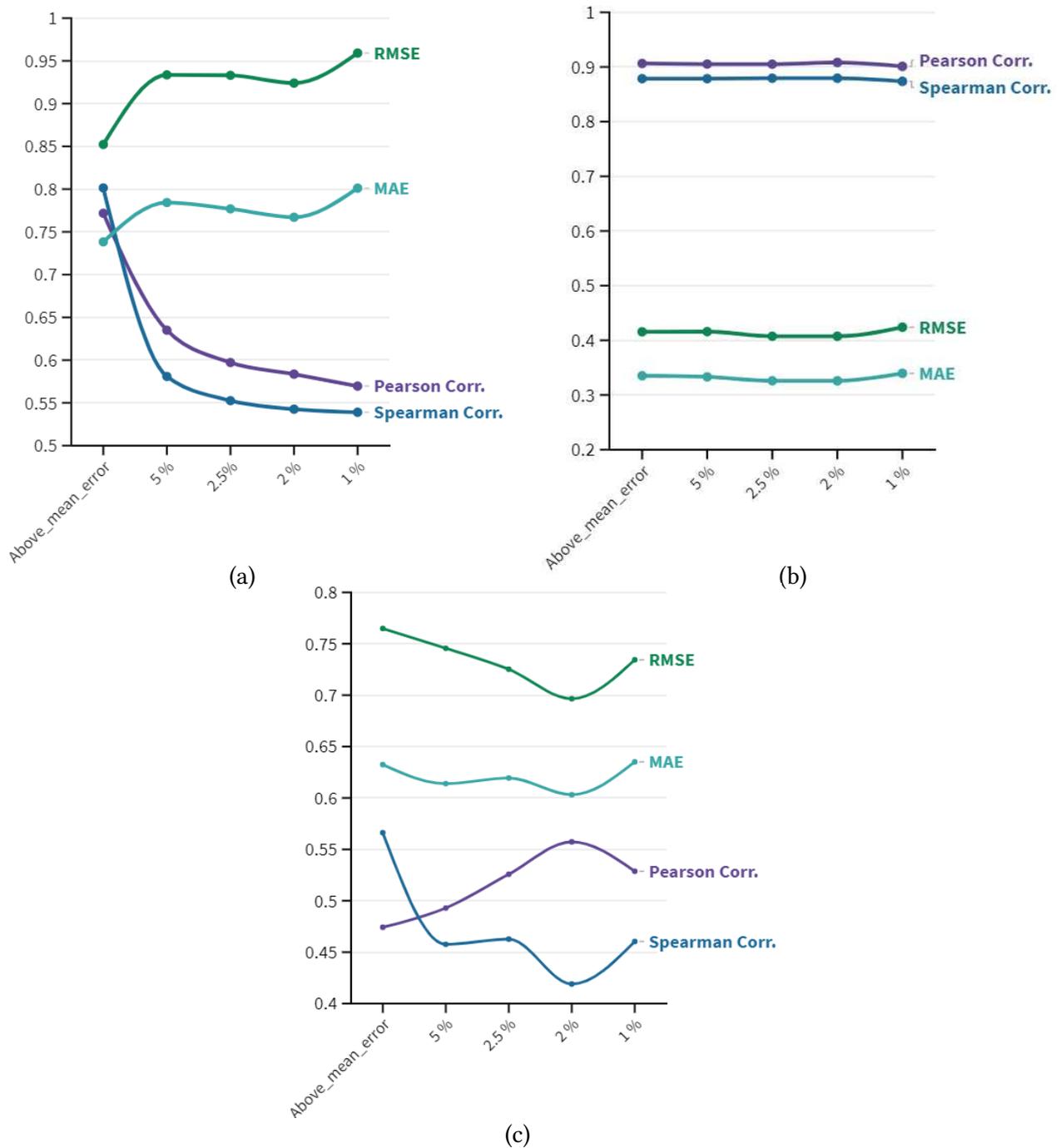


Figure 3: Figures displaying results for the (a) English–Chinese, (b) Romanian–English and (c) English–German language pairs.

This is also our highest Pearson correlation boost of any language pair. In addition, we achieve a 0.2525 Spearman correlation boost (an improvement of 46% on *MonoTransQuest*) as well as a reduction in both RMSE and MAE. This suggests that the presence of NEs in the English–Chinese language pair hinders *MonoTransQuest*'s performance significantly.

The results for Romanian–English are highlighted in Table 4. As can be seen in this table, we achieved our highest Pearson correlation boost (0.0061) over *MonoTransQuest* when using *the Specific Threshold method* at the 2% level.

Table 3
Results for English–Chinese

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.9511	0.7804	0.5489	0.5514
Above MAE	0.8522	0.7383	0.8014	0.7718
Worst 1% Masked	0.9591	0.8011	0.539	0.5696
Worst 2% Masked	0.924	0.7671	0.5426	0.5836
Worst 2.5% Masked	0.9331	0.7769	0.5526	0.5972
Worst 5% Masked	0.9336	0.7843	0.5809	0.635

Table 4
Results for Romanian–English

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.4209	0.3375	0.872	0.9021
Above MAE	0.4156	0.3352	0.8786	0.9064
Worst 1% Masked	0.424	0.3396	0.8737	0.9011
Worst 2% Masked	0.4074	0.3259	0.8794	0.9082
Worst 2.5% Masked	0.4131	0.333	0.8758	0.9051
Worst 5% Masked	0.4159	0.3331	0.8777	0.9053

Table 5
Results for English–German

Experiments	RMSE	MAE	Spearman	Pearson
MonoTransQuest	0.7757	0.6444	0.4807	0.461
Above MAE	0.7495	0.6259	0.3695	0.4309
Worst 1% Masked	0.7683	0.6588	0.3603	0.5085
Worst 2% Masked	0.6965	0.6031	0.4419	0.5573
Worst 2.5% Masked	0.7252	0.6194	0.4627	0.5258
Worst 5% Masked	0.7563	0.6365	0.3856	0.4271

This setting also achieves an increased Spearman value and a reduction in RMSE and MAE. Table 5 highlights the results for the English–German language pair. The *Specific Threshold method* produces the largest improvement in Pearson correlation (0.0963) at the 2% level. It also results in improvements in the RMSE and MAE over the baseline *MonoTransQuest* system.

7. Conclusions and Future Work

In this paper, we discussed the background of the WMT-2020 shared task, detailing the scientific topics that this benchmark intends to progress. We focused on the task of sentence-level direct assessment from WMT-2020. We explained our methodologies behind planned improvements on the *TransQuest* baseline system (*MonoTransQuest*), showing both our successful and less successful approaches.

The experimental results revealed that our proposed system outperformed the state-of-the-art *TransQuest* team. However, there are several possibilities to extend this work in future. One of them is to apply the NE masking system to other language pairs. Another possibility is to explore other medium and high resource language pairs in order to further test the robustness of our system. In addition, it is also possible to conduct experiments with data augmentation by extending the training data by using additional data that is similar to the test data set. It can be done by using a text similarity method based on word embeddings and other state-of-the-art sentence similarity methods in order to extract data that are semantically equivalent to the test data set.

Acknowledgements

This research has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106).

References

- [1] T. Ranasinghe, C. Orasan, R. Mitkov, *TransQuest at WMT2020: Sentence-level direct assessment*, in: *Proceedings of the Fifth Conference on Machine Translation*, Online, 2020, pp. 1049–1055. URL: <https://www.aclweb.org/anthology/2020.wmt-1.122>.
- [2] L. Specia, F. Blain, V. Logacheva, R. F. Astudillo, A. F. T. Martins, *Findings of the WMT 2018 shared task on quality estimation*, in: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, 2018, pp. 689–709. URL: <https://www.aclweb.org/anthology/W18-6451>. doi:10.18653/v1/W18-6451.
- [3] F. Kepler, J. Trénous, M. Treviso, M. Vera, A. F. T. Martins, *OpenKiwi: An Open Source Framework for Quality Estimation*, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Florence, Italy, 2019, pp. 117–122. doi:10.18653/v1/P19-3020.
- [4] T. Ranasinghe, C. Orasan, R. Mitkov, *TransQuest: Translation quality estimation with cross-lingual transformers*, in: *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online), 2020, pp. 5070–5081. URL: <https://www.aclweb.org/anthology/2020.coling-main.445>. doi:10.18653/v1/2020.coling-main.445.
- [5] M. Fomicheva, S. Sun, L. Yankovskaya, F. Blain, V. Chaudhary, M. Fishel, F. Guzmán, L. Specia, *Bergamot-latte submissions for the wmt 20 quality estimation shared task*, in: *Proceedings of the Fifth Conference on Machine Translation*, Online, 2020, pp. 1008–1015.
- [6] J. Moura, M. Vera, D. van Stigt, F. Kepler, A. F. T. Martins, *IST-unbabel participation in the WMT20 quality estimation shared task*, in: *Proceedings of the Fifth Conference on Machine Translation*, Online, 2020, pp. 1029–1036. URL: <https://www.aclweb.org/anthology/2020.wmt-1.119>.
- [7] F. Kepler, J. Trénous, M. Treviso, M. Vera, A. F. T. Martins, *OpenKiwi: An open source framework for quality estimation*, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, Association for Com-

- putational Linguistics, Florence, Italy, 2019, pp. 117–122. URL: <https://www.aclweb.org/anthology/P19-3020>.
- [8] A. Nakamachi, H. Shimanaka, T. Kajiwara, M. Komachi, Tmuou submission for wmt20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1035–1039.
- [9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.
- [10] C. Hu, H. Liu, K. Feng, C. Xu, N. Xu, Z. Zhou, S. Yan, Y. Luo, C. Wang, X. Meng, T. Xiao, J. Zhu, The niutrans system for the wmt20 quality estimation shared task, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1016–1021.
- [11] L. Zhou, L. Ding, K. Takeda, Zero-shot translation quality estimation with explicit cross-lingual patterns, in: Proceedings of the Fifth Conference on Machine Translation, Online, 2020, pp. 1066–1072.
- [12] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, fairseq: A fast, extensible toolkit for sequence modeling, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 48–53. URL: <https://www.aclweb.org/anthology/N19-4009>. doi:10.18653/v1/N19-4009.
- [13] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, M. Ranzato, The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6098–6111. URL: <https://www.aclweb.org/anthology/D19-1632>. doi:10.18653/v1/D19-1632.
- [14] H. Kim, J.-H. Lee, S.-H. Na, Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 562–568. URL: <https://www.aclweb.org/anthology/W17-4763>. doi:10.18653/v1/W17-4763.
- [15] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, A. F. T. Martins, Findings of the WMT 2020 shared task on quality estimation, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 743–764. URL: <https://www.aclweb.org/anthology/2020.wmt-1.79>.
- [16] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. URL: <https://doi.org/10.5281/zenodo.1212303>. doi:10.5281/zenodo.1212303.
- [17] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 2020, pp. 101–108. URL: <https://www.aclweb.org/anthology/2020.acl-demos.14>. doi:10.18653/v1/2020.acl-demos.14.

- [18] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate research* 30 (2005) 79–82.

An Empirical Comparison Analysis on the Evolution of RNN Models using Multiple European Languages

Vikram Bhutani^a, Farshad Ghassemi Toosi^a

^a*Munster Technological University, Cork, Ireland*

Abstract

Humans are becoming globally connected more than ever. Communication is the key of this connection and languages are the means of this communication. We are witnessing globalization in the world that potentially cause some languages to dominate other languages and consequently the extinction of some languages. To be part of the globalized world and still to be able to communicate with own local language, language translation agents may be employed. Language Translation (LT) as an (Natural Language Processing) NLP application is a short answer to the aforementioned problem. In the last few decades we have seen the development of multiple Machine Learning techniques in language translation. In this work, we are presenting a new Recurrent Neural Network (RNN) architecture/model and experimenting a set of comparison analyses between our model and four existing RNN models (Simple RNN, Bidirectional RNN, Embedded RNN and Seq2Seq). The experiment is performed on an open-source repository for four pairs of languages: 1) *English to Irish/Gaelic*, 2) *English to Spanish*, 3) *Irish/Gaelic to English* and 4) *Spanish to English*. Our result indicates that on average our proposed model outperforms all the other models for all four pairs of languages. The result also indicates that models, on average, favour the pairs of languages where English is the source language. *English to Spanish*, on average, has the highest performance compared to other pairs and *Irish/Gaelic to English*, on average, has the lowest performance.

Keywords

RNN, Bidirectional, Seq2Seq, Irish Language, Language Translation.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ vikram.bhutani@intel.com (V. Bhutani); Farshad.Toosi@mtu.ie (F.G. Toosi)

🆔 0000-0002-0877-7063 (V. Bhutani); 0000-0001-7116-9338 (F.G. Toosi)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

There are nearly 100 alive languages in the world. Each language has a set of different features (e.g., grammar, phonetic, structure and alphabet). The differences between these languages and their unique structures, make the task of automated translation, specific to the source and target languages. Therefore, developing and training a specific model with a known set of hyper parameters that can be generalized for several different languages if not impossible, is a complicated task. Even when for a fixed pair of languages (e.g., English and Arabic), it is still not easy to come up with a unique model and a fixed set of hyper parameters that can be used for every type of content (e.g., literature translation, technical translation, speech translation and etc) [1]. Deep learning techniques [2] as a subset of Machine Learning techniques are the most commonly used family of techniques in the area of machine translation. Most of the current translation technologies make use of Recurrent Neural Networks (RNN) for the task of language translation.

Recurrent Neural Network (RNN) is a sub-class of Artificial Neural Network that is comprised of a directed graph where the output is used again as a new input. Traditional neural networks (non-RNN models) have an input layer with a certain number of neurons that makes the network to be compatible with input with a certain size. This can be seen as a limitation when the input data has no pre-defined size. Examples of such cases is textual input where each input is a sentence with a number of words and sentences may have different size.

One of the objectives of the Recurrent Neural Network idea was to address the mentioned issue so that inputs with different sizes may be fed to the network [3]. The initial success of RNN made researchers to develop different variations of RNN model to deal with specific types of problems.

In this work, a new hybrid RNN-based model is presented. This new model takes advantage of existing RNN models and its performance is compared against other well-known RNN variations. Three European languages (English, Spanish and Irish/Gaelic) are employed for the comparison analysis task:

We mainly aim to address the following three questions:

- Although all these three languages are European languages, is there a noticeable performance difference among these pairs of languages under the same model?
- How model's performance is affected by reversing the source and target languages e.g., (English to Spanish) vs (Spanish to English).
- Each RNN model has a particular architecture with a set of layers, how to take advantage of different models' architecture in one single model (**Hybrid Model**, the main contribution of this work).

2. Literature Review

One of the limitations of basic neural networks is the requirement for the input to have a fixed or predetermined size. Recurrent Neural Networks (RNNs) advanced the basic neural networks by introducing a new structure where inputs do not have to have a fixed or predetermined size. Applications (e.g., language translation) where the input data does not or can not have a fixed size would greatly benefit from this type of architecture.

In the last few decades, there have been a number of RNN models developed for language translation such as *Sequence2Sequence* [4] and yet the research in this area is still active. For instance, [5], discussed the problem of multi language translation using encoder-decoder and allocating multiple decoders to multiple target languages. This way, they tried to have a simultaneous translation from language A (encoder) to languages B, C and so on (decoders). In another work, [6] proposed two novel models based on Recurrent Neural Networks for three different pairs of languages: German-English, Arabic-English and Chinese-English. Their novel approaches in Recurrent Neural Networks has two variations: a word-based model and the other one is phrase-based. The models indicate an improvement compared to the base models based on two metrics (BLEU and TER). Hu et al [7] also proposed a novel technique, MTU (Minimum Translation Unit) based approach against the classical n-gram back-off model on WMT 2012 French-English dataset. Their evaluation metric is BLEU and resulted in 0.8 improvement compared to traditional n-gram model.

Traditional RNN models like Seq2Seq are the model of choice for most of the NLP applications but training such models using big data can be very challenging. New technologies that use transformers with self and multi-head attention are proving to be state-of-the-art that can significantly reduce computational requirements [8]. The transformers use attention mechanism thus require less data computations and are less expensive as opposed to traditional RNN which uses LSTM (Long Short Term Memory) or GRU(Gated Recurrent Unit) [9].

There have been major developments in the field of NLP where RNN and attention algorithms are used together to achieve high accuracy and faster training times like in XLNet architectures [10]. Major difficulties when applying Transformer to language translation applications is that it requires more complex configurations(e.g., optimizer, network structure, data augmentation) than the conventional RNN based models. Recent studies shows that RNN models using global or local attention mechanism techniques can be used as the state-of-the-art solution [11].

Table 1
English-Irish-Spanish translation dataset

Details	English text	Irish text	Spanish text
Total Number of sentences	138460	138460	138460
Maximum sentence length	21	25	26
Total Number of words	1555241	2147273	1636051
Vocabulary size	679	1013	912

3. Dataset

In order to have a more comprehensive analysis, an open-source dataset is employed from *WMT14* repository published by *STATISTICAL MACHINE TRANSLATION* [12]. This dataset is currently available in many European languages such as *English, German, French* and *Czech*. The topic of the dataset is related to many widely used AI applications and the dataset has been mostly used for machine learning performance indicators and NLP benchmarks. Since the aim of this work was to work on Gaelic and Spanish languages, respective translations by services which include python APIs by Google [13] and Ai translate services [14] is performed.

The features of each language after the translation (English to Spanish using DeepL) are displayed in Table 1.

Like any other Machine Learning problems, the language translation requires pre-processing as well. Since languages have different types of symbols, features and structure, the task of pre-processing can be challenging and can affect the performance significantly. The textual data should be carefully examined, properly cleaned and transformed appropriately before feeding them into the RNN models. There are many stages involved in NLP data processing before it is fed into the translation models. All three data-sets are undergoing the following standard processes:

- Data cleaning: Below are some of the most frequent types of noise that is present in text data:
 1. Unicode and other symbols include removal of special characters such as "
 2. Removal of html tags.
 3. Removing numbers, generally numbers are not required to be translated as they are generic.
 4. Links: Links can be of many forms and most of them consist of strange symbols or short-codes that can present in the document.
- Tokenization: splitting our text into minimal meaningful units for ML algorithm to understand the data in the numerical form.

Table 2

The Details of all variants of RNN Models.

Model's Details	Simple RNN	Bidirectional RNN	Embedded RNN	Seq2Seq RNN	Hybrid RNN
Activation	relu	relu	relu	relu	relu
Final Layer Activation	softmax	softmax	softmax	softmax	softmax
Dropout	0.4	0.3	0.4	0.4	0.3
Layers	4	4	5	6	7
Optimiser	Adam	Adam	Adam	Adam	Adam
Learning rate	0.001	0.0001	0.001	0.001	0.001
Batch size	128	128	128	128	128
Recurrent Regularizer	0.000001	0.000001	0.000001	0.000001	0.000001
Epsilon	1e-08	1e-08	1e-08	1e-08	1e-08
Kernel Regularizer (decay rate)	1e-6	1e-6	1e-6	1e-6	1e-6

- **Normalisation:** Normalization is one of the important pre-processing steps and its attempt is to make a single representation of words with multiple representations. Stemming and lemmitising are key steps here.
- **Stop words removal:** Removing stop-words is another essential step. The main reason for stop-word removal is that these stop-words generally do not add new information to the text but are just a language construct.
- **Embeddings and Representations.** Once the dataset is cleaned, the text is converted into some kind of numerical representation to make them understandable for Machine Learning where they only understands numbers.
- **Sentences padding:** Proper padding is added to the sentences so that it will keep sentences to same size before the tensor multiplications is performed. This also helps in computational of high dimensional tensors. The <start> and <end> tokens are also added in each sentence to mark the start and end of sentences for tokenazation.

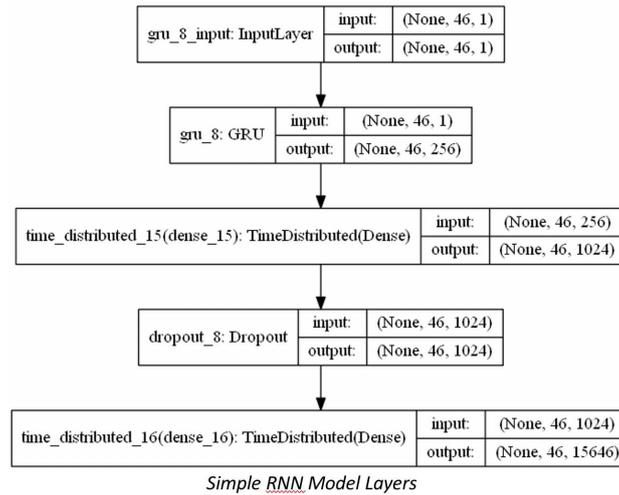


Figure 1: Simple Model.

4. Methodology

In this work, a comparative translation analysis is performed on different pairs of languages using five different models.

Four different pairs of languages using three European languages (English, Irish/Gaelic and Spanish) are employed in the experiment as follows:

- English to Gaelic language translation.
- Gaelic to English language translation.
- English to Spanish language translation.
- Spanish to English language translation.

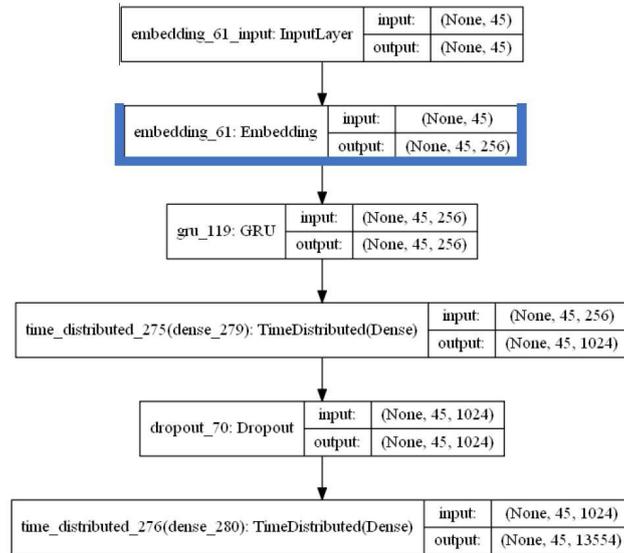
As part of this research, we also aim to find the optimal hyper parameters for the best performing model and final selection will be done based on the performance metric.

4.1. Models

All the employed models in this work are from the family of RNN architecture and the details of the models' architectures are illustrated in Figure 1. The hyper parameters of every single model were decided experimentally. The main contribution of this work is the Hybrid RNN model that is illustrated in Figure 1, the architecture has components/layers from three other models: Embedded, Bidirectional and Seq2Seq.

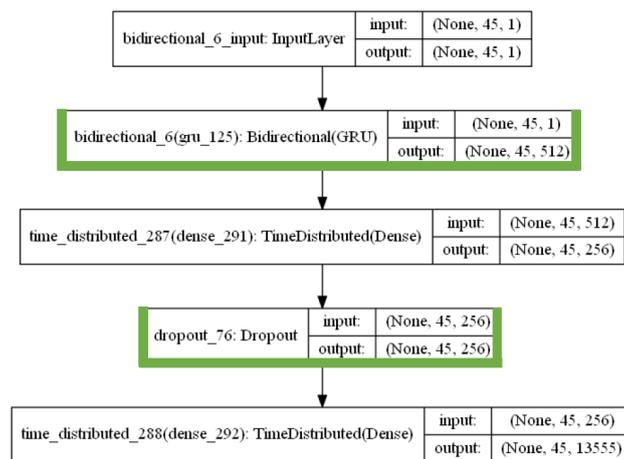
Each model has a set of hyper parameters such as learning rates, optimizer, dropout and etc; the details of the hyper parameters are displayed in Table 2.

The hyper parameters are decided experimentally. For Learning Rate, seven values are tested (0.01, 0.001, 0.009, 0.0001, 0.00001, 0.000001 and 0.0000001) and the winner is 0.001, See Table 2. A similar experimental approach is employed to decide an optimizer for the models. Six different optimizers (*adagrad*, *adam*, *SGD*, *adadelta*, *rmsprop* and *adamax*) are tested and based on



RNN model using Word Embedding Layers

Figure 2: Embedded Model.



RNN model using Word Bidirectional Layers

Figure 3: Bidirectional Model.

the accuracy performance, *adam* optimizer is selected. Although *adadelat*a optimiser performed well in certain language pairs, *adam* optimizer is decided for all models due to its consistency across all language pairs.

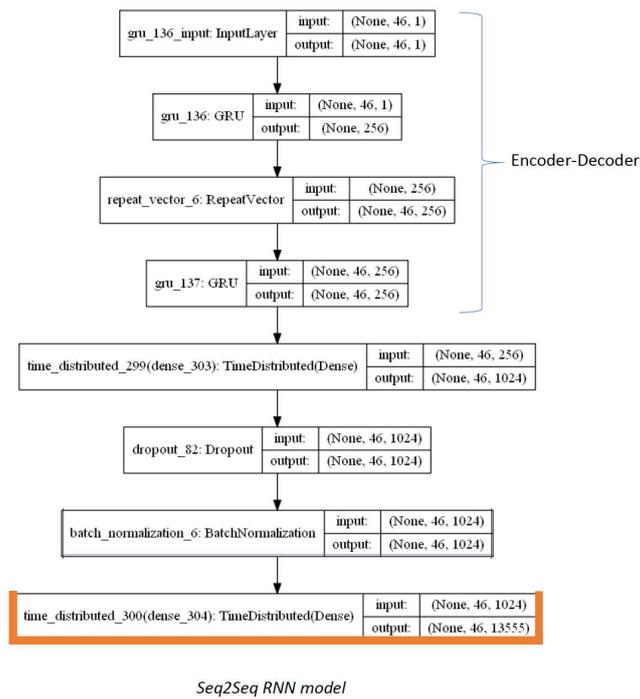


Figure 4: Seq2Seq Model.

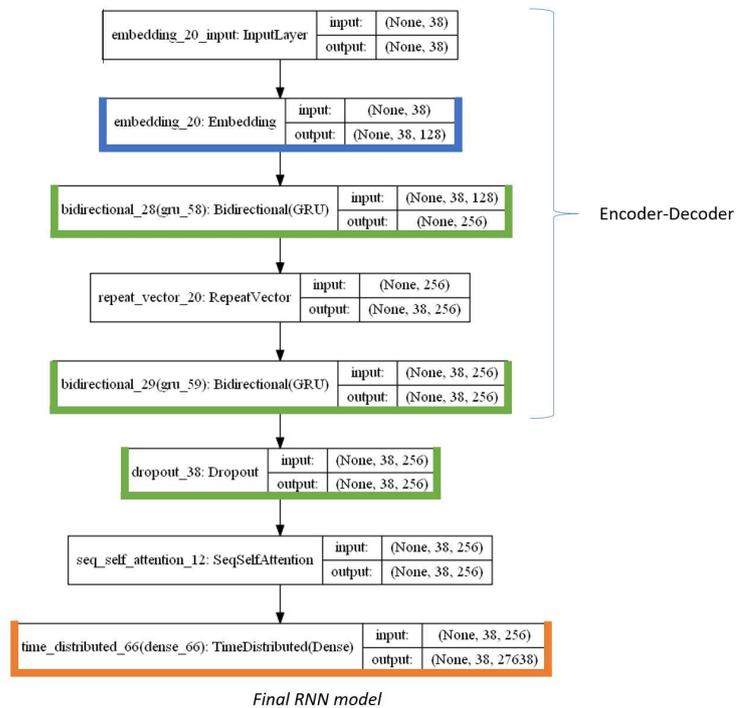


Figure 5: Hybrid Model.

5. Results

The results of the whole experiment is divided into five different sub-experiments using five different models as follows: 1) Simple RNN [15] 2) Embedded RNN [16] 3) Bidirectional RNN [17] 4) Seq2Seq [18] and 5) Hybrid RNN model. All the examined models in this work are variations of RNN (Recurrent Neural Network) on two pairs of languages reciprocally: 1) (*Gaelic/Irish* \iff *English*) 2) (*Spanish* \iff *English*). The following sections describe the details of each model and their results individually.

5.1. Simple RNN

The first experiment is carried out by a simple RNN [15] with 5 layers (details can be found in Table 2 and Figure 1. Simple RNN comprises of layer with GRU (Gated Recurrent Unit) followed by dense time distributed layer using softmax activation. This model is not deep enough and using minimal number of trainable parameters.

Figure 8 and 9 show the accuracy and loss distribution using simple RNN model on two pairs of languages reciprocally: *Gaelic to English*, *English to Gaelic*, *Spanish to English* and *English to Spanish*. The highest validation accuracy belongs to the *English to Spanish* pair followed by *Spanish to English* however the highest BLEU score belongs to *English to Gaelic* followed by *Gaelic to English*, See Table 3.

5.2. Embedding RNN

The first layer of the Embedding model [16] is word Embedding layer with the target vocabulary size. Word Embedding layer converts words into the dense vectors and helps in understanding the context of a word so that similar words have similar embeddings.

The highest validation accuracy for this model is resulted on *English to Spanish* pair followed by *English to Gaelic* and then *Spanish to English*. The highest BLEU score also belongs to *English to Spanish* and *Spanish to English* followed by (with relatively large gap) *English to Gaelic* and *Gaelic to English*

The performance of the validation and training sets are quite close to each other for *Spanish* \iff *English* pairs as opposed to *Gaelic* \iff *English* pairs, See Figures 12 and 13.

5.3. Bidirectional RNN

The Bidirectional model [17] includes Bidirectional Recurrent layers. The number of units in Bidirectional layer is doubled which resulted in more trainable parameters and thus increasing computational cost. While in simple RNN there is a single GRU layer, in Bidirectional model there are two LSTM or GRU cells activated to support forward and backward propagation. This can provide additional context to the network and result in faster learning of the model [17].

The highest validation accuracy by this model belongs to the *English to Spanish* followed by *Spanish to English* and *English to Gaelic*. *English to Spanish* is also the winner for BLEU score followed by *Spanish to English* and *English to Gaelic*.

Table 3
All Language Translations Performance Quality

Model	Simple RNN	Embedded	Bidirectional	Seq2Seq	Hybrid RNN Model
English to Gaelic					
Training Accuracy	80%	90%	85%	91%	95%
Validation Accuracy	78%	84%	82%	86%	89%
BLEU score	0.22	0.31	0.30	0.32	0.33
Epoch time (s)	140s	180s	220s	180s	190s
Gaelic to English					
Training Accuracy	77%	80%	81%	82%	83%
Validation Accuracy	76%	79%	79%	80%	82%
BLEU score	0.19	0.30	0.29	0.27	0.36
Epoch time (s)	120s	180s	260s	180s	210s
English to Spanish					
Training Accuracy	83%	88%	92%	95%	97%
Validation Accuracy	84%	87%	93%	93%	97%
BLEU score	0.16	0.31	0.33	0.27	0.42
Epoch time (s)	150s	220s	270s	200s	250s
Spanish to English					
Training Accuracy	81%	84%	82%	84%	85%
Validation Accuracy	81%	83%	82%	83%	84%
BLEU score	0.2	0.40	0.30	0.25	0.40
Epoch time (s)	150s	220s	290s	180s	260s

Spanish \leftrightarrow *English* pairs seem to have similar performance on the validation and training sets compared to *Gaelic* \leftrightarrow *English*, See Figures 10, 11 and Table 3.

5.4. Seq2Seq RNN

The Seq2Seq [18], also known as Encoder-Decoder, is the forth experimented model in this work. This model is the most popular model that is widely used in Autoencoders, Variational Autoencoders and in RNNs. We have also added self attention in this model to draw global dependencies between inputs and outputs [9].

While the highest validation accuracy under this model belongs to *English to Spanish* followed by *English to Gaelic*, the highest BLEU score belongs to *English to Gaelic* followed by *English to Spanish*.

Figures 14 and 15 show the accuracy and loss distributions using Seq2Seq model. A similar pattern to the previous models emerge here as well. The validation and training accuracy for *Spanish* \leftrightarrow *English* pairs seem to be so close to each other while there is a considerable difference between the validation and training accuracy for *Gaelic* \leftrightarrow *English* pairs.

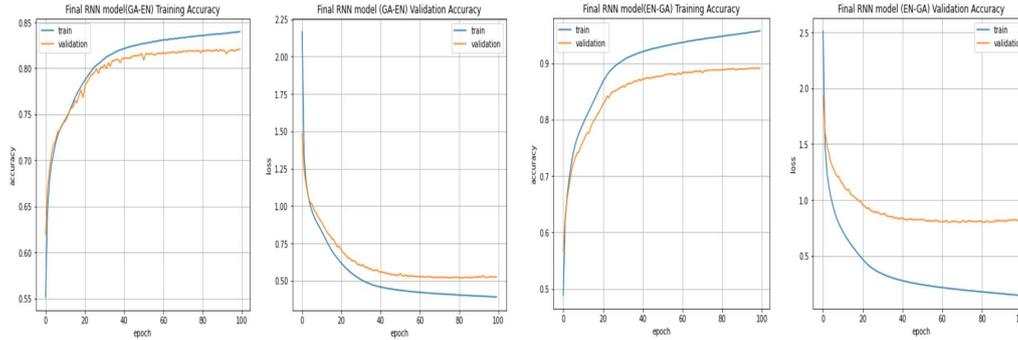


Figure 6: Hybrid RNN Model using Gaelic and English Pair

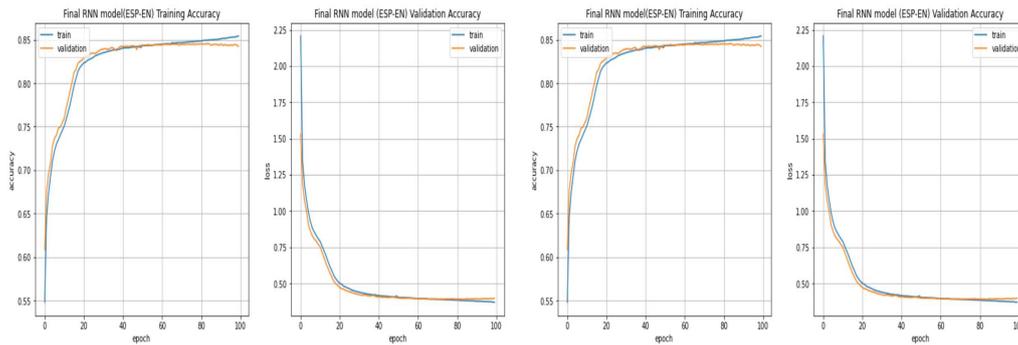


Figure 7: Hybrid RNN Model using Spanish and English Pair

5.5. Hybrid RNN Model

Hybrid RNN model is the main contribution of this work. Our proposed model contains features from three existing models (Embedding, Bidirectional and Seq2Seq). As shown in Table 3, Hybrid model is the winner in all the experiments and for all the metrics e.g., Validation and Training accuracy and BLEU score. Although Hybrid model has the highest performance compared to the previous models, its running time is computationally heavier as oppose to other models. As illustrated in Figures 6 and 7, the hybrid model converges faster for *Spanish* \iff *English* pairs as opposed to *Gaelic* \iff *English* pairs. The table 3 summarises the training and validation scores of all five employed neural networks for four different pairs of languages. The same table also shows the BLEU accuracy scores achieved by different models. The detailed comparison of variants of RNN models used in evaluation is shown in Figure 16. The result indicates that the performance of Hybrid model outperform all other RNN variants.

Figure 17 shows the different settings that are applied for model's hyperparameters and corresponding model performance. The model used in hyperparameter tuning is Hybrid RNN model. Based on the hypothesis and computational constraints, it is assumed that 100 epochs are enough to conclude the optimal hyperparameter values. The final distribution clearly shows the optimal learning rate, optimiser and regularisation that should be applied to the deep neural network. Based on the results, optimal settings of hyperparameter values are selected to maximize performance.

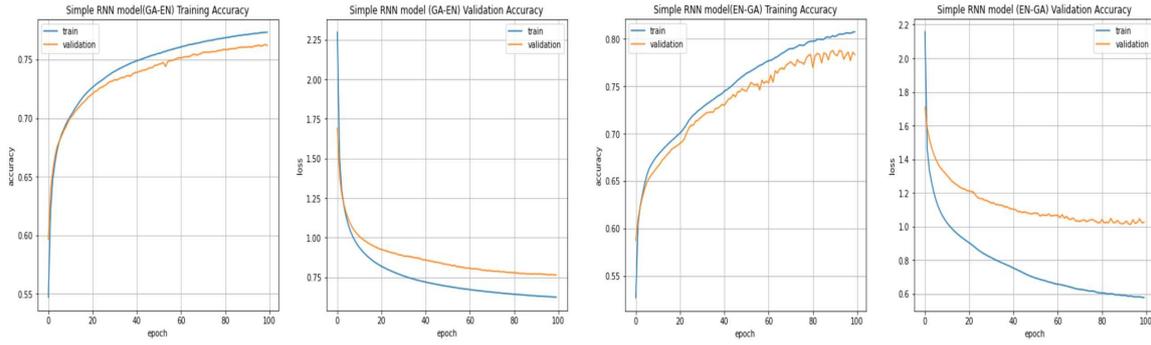


Figure 8: Simple RNN Model using Gaelic and English Pair

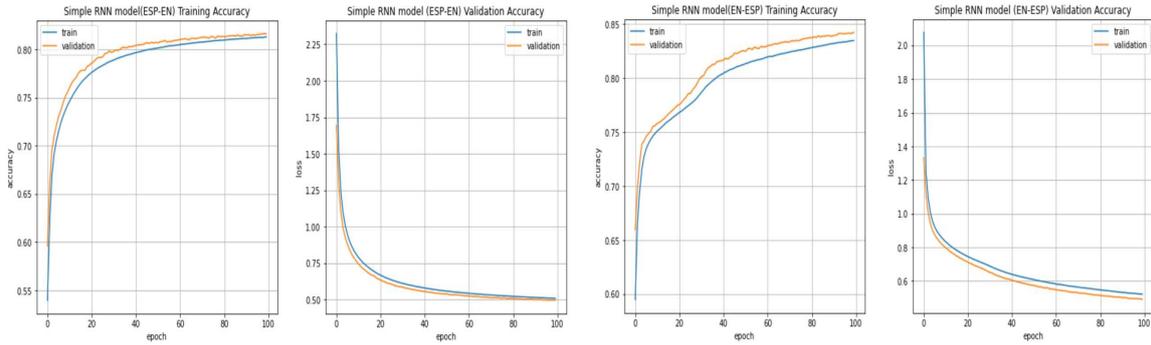


Figure 9: Simple RNN Model using Spanish and English Pair

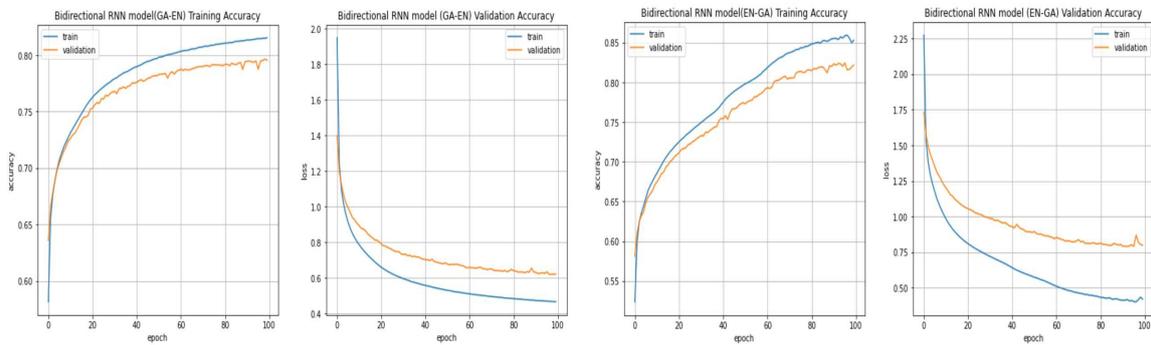


Figure 10: BiDirectional RNN Model using Gaelic and English Pair

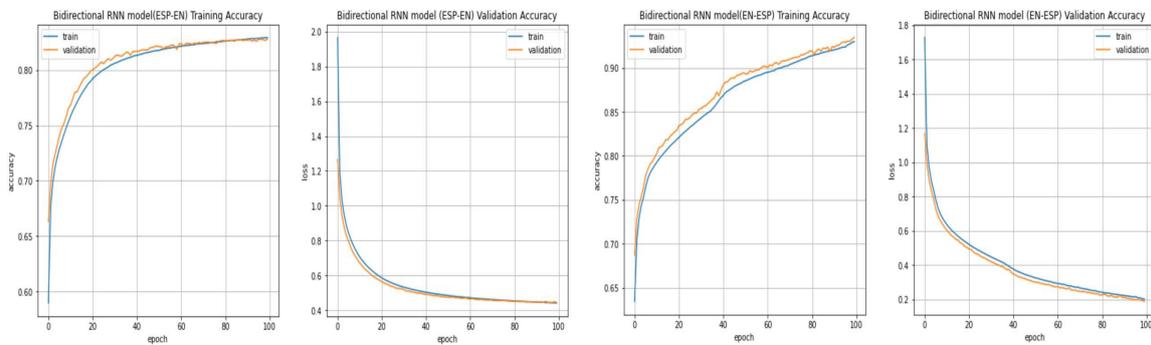


Figure 11: BiDirectional RNN Model using Spanish and English Pair

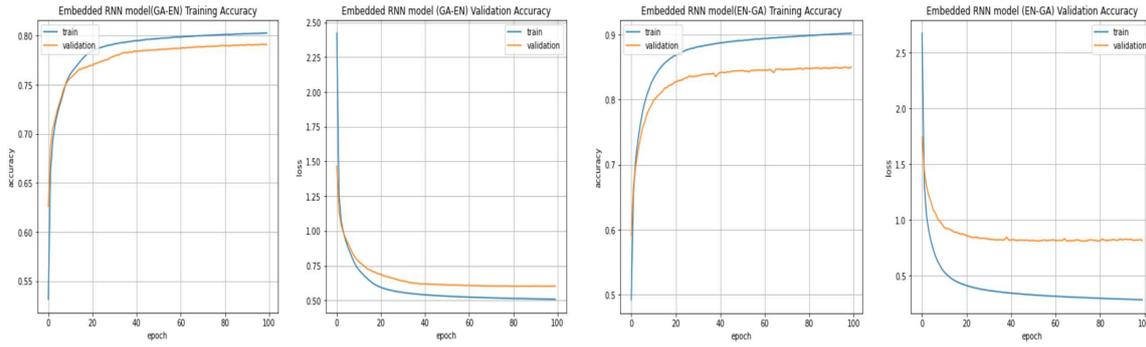


Figure 12: Embedded RNN Model using Gaelic and English Pair

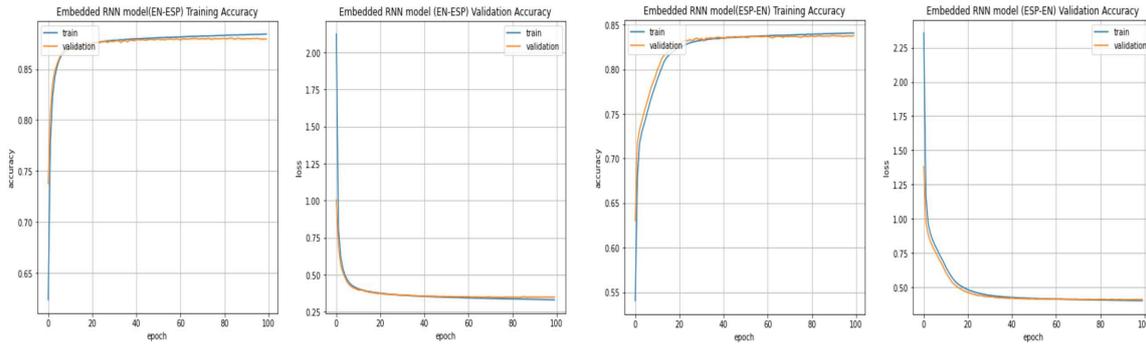


Figure 13: Embedded RNN Model using Spanish and English Pair

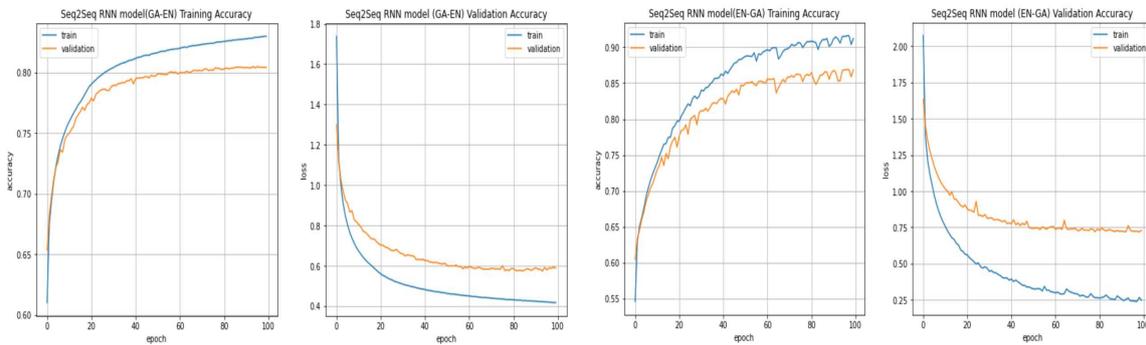


Figure 14: Seq2Seq RNN Model using Gaelic and English Pair

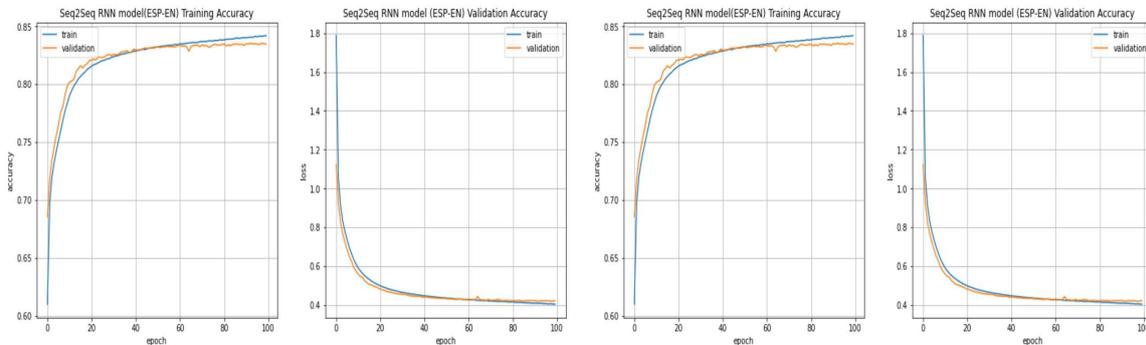


Figure 15: Seq2Seq RNN Model using Spanish and English Pair

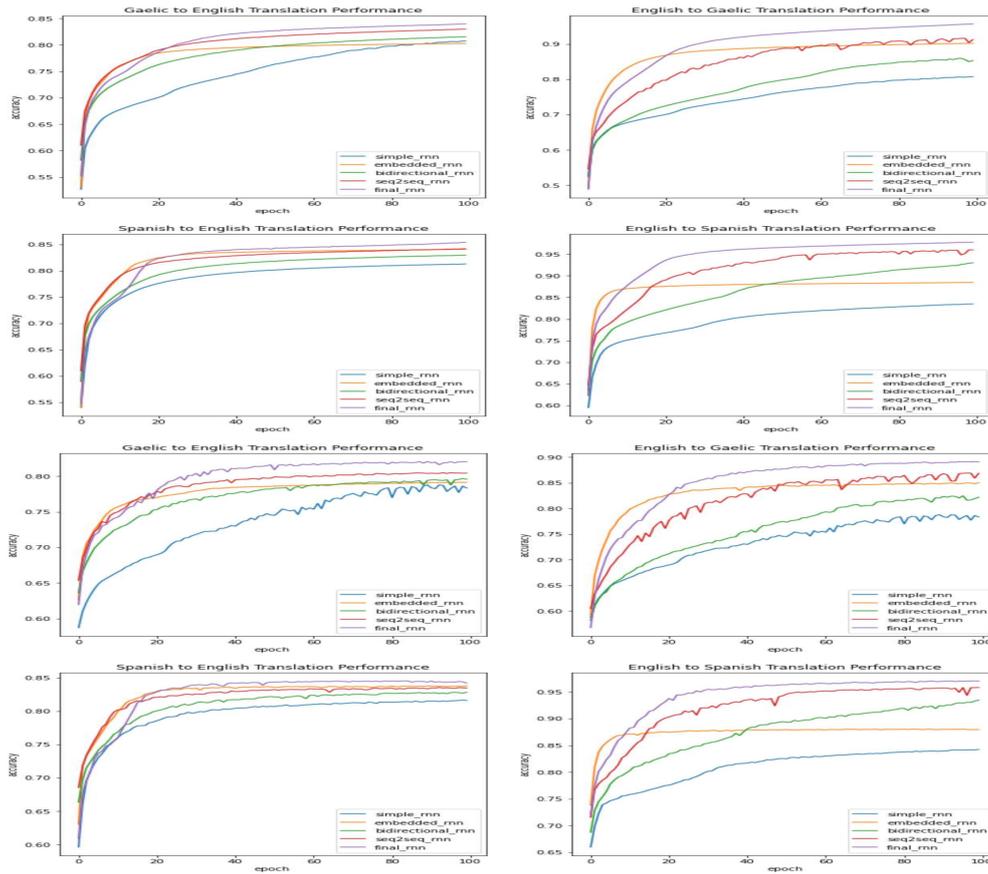


Figure 16: All RNN Model's Performance(Training and Validation)

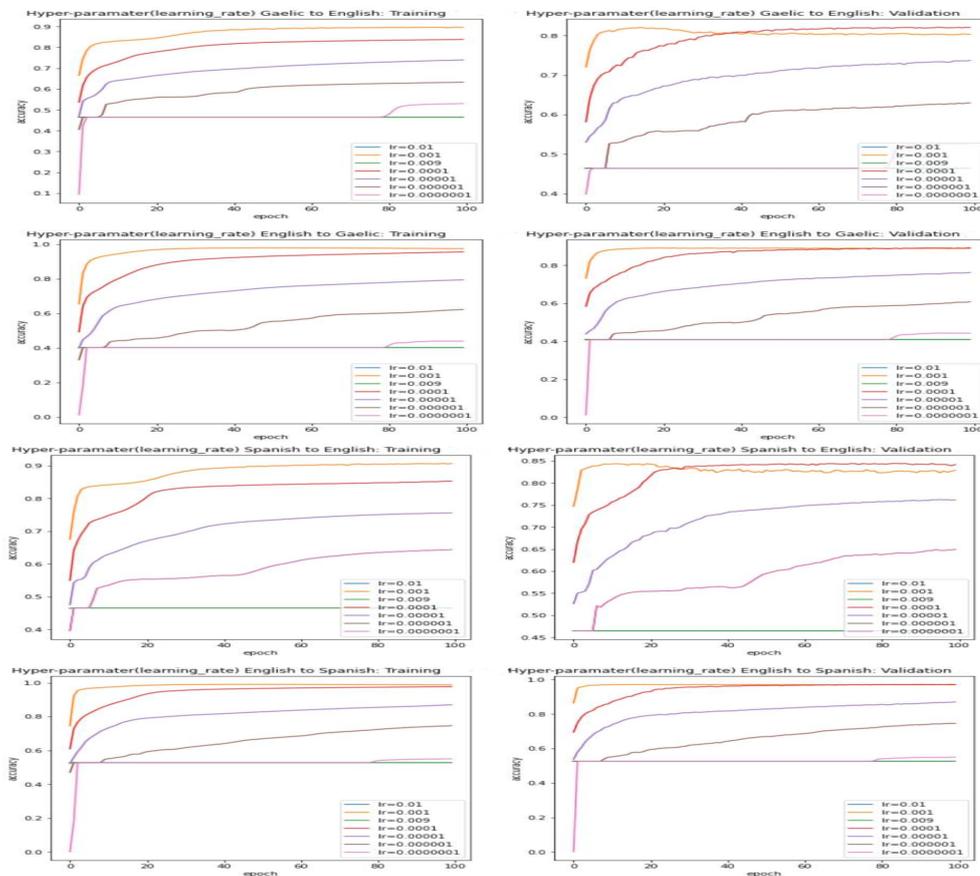


Figure 17: Hybrid RNN Model's Performance for Learning Rate Tuning

6. Discussion

In this work, five different RNN models are experimented under 2 pairs of languages in reciprocal way. A dataset from [12] is employed to perform the experiment. This dataset has been used in other studies and ML performance [19] and considered as a benchmark dataset. In total, 5×4 experiments are performed (5 models, 4 pairs of languages for each model). Four models out of these five models are stereotype models from literature while the last model is the main contribution of this work. As mentioned earlier, some of the layers of the hybrid RNN model are derived from other models (i.e., Bidirectional, Embedded and Seq2Seq models). As part of all experiments, a pre-processing step is carried out to make the data ready for neural networks. One of the observation from the pre-processing step is that stemmization and lemmitization techniques have very little impact on overall model performance.

Table 3 and Figures 8, 9, 10, 11, 12, 13, 14, 15, 6 and 7 show the performance details of all five models. Our proposed model (Hybrid model) has the best performance for training, validation accuracy and BLEU score. From performance point of view, the closest model to the Hybrid model is Seq2Seq. From BLEU score point of view, the second winner is Embedded model (except the English to Gaelic pair where Seq2Seq is the second winner). In all experiments and for all performances (e.g., training, validation and BLEU) simple RNN has the lowest performance.

Another interesting observation from all experiments is the behavior of languages under these models. On average, the highest validation accuracy for all models is for English to Spanish (90.8%) followed by English to Gaelic (83.8%) followed by Spanish to English (82.6%) and then at last Gaelic to English (79.2%). A similar behavior is observed for BLEU score where the BLEU score on average for all models is for English to Spanish (0.64) followed by English to Gaelic (0.60) followed by Spanish to English (0.588) and at last Gaelic to English (0.568).

The best validation accuracy from all models and all pairs of languages belongs to English to Spanish under our proposed model (97%) followed by Bidirectional and Seq2Seq models for the same pair of languages (English to Spanish) (93%). The best BLEU scores belong to Hybrid model (0.85 and 0.84 for English to Spanish and Spanish to English respectively) followed by Embedded model for English to Spanish and Spanish to English (0.80 and 0.80 respectively).

The first and obvious finding of this study shows that swapping the target and source languages has impact on the actual performance for each model. For example based on table 3 the performance metric for English to Spanish is better than Spanish to English. It shows that the language structure can be a contributing factor when training models thru RNN. Although the amount of data used for all experiment is the same, pairs of languages where Gaelic is the source language shows the worst performance. This can be an indication of differences in languages' structures.

From all the presented results here we can draw two main conclusions:

1. Hybrid model performs better than other four models on average for all quality metrics (e.g., validation and training accuracy and BLEU score). The running time per epoch for Hybrid model is the second highest (the highest one belongs to bidirectional model).
2. A quantitative comparison among these four pairs of languages indicates that English to Spanish favours the accuracy and BLEU metrics on average for all models. With almost

a large gap the second best pair of languages is English to Gaelic followed by Spanish to English and finally Gaelic to English. This indicates that pairs with English as the source language seem to have higher performance. On the other hand, among those pairs where English is the target language (i.e., Spanish to English and Gaelic to English), Spanish to English is the winner. This brings us to a new conclusion that reversing the source and target languages do not necessarily results in similar performance.

6.1. Future Work

The NLP data-sets from WMT releases [12] continue to evolve with the addition of more human languages to improve speech translation machine learning techniques, this work primarily focuses on pairs where English is either the source or target language. As a future work, other pairs of human languages e.g., Spanish to Gaelic or other European languages will be examined. We also aim to publish similar performance metric using bigger vocabulary sizes on newer WMT (e.g., WMT18 and WMT19) and other machine learning NLP data-sets.

Although the primary focus of this work is on analysis of RNN based models, as future work, more advanced structures such as transformers will be examined.

References

- [1] N. Yamashita, R. Inaba, H. Kuzuoka, T. Ishida, Difficulties in establishing common ground in multiparty groups using machine translation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 679–688.
- [2] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, S. Jain, Machine translation using deep learning: An overview, in: 2017 International Conference on Computer, Communications and Electronics (Comptelix), IEEE, 2017, pp. 162–167.
- [3] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, S. Khudanpur, Recurrent neural network based language model, in: Eleventh annual conference of the international speech communication association, 2010.
- [4] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112.
- [5] D. Dong, H. Wu, W. He, D. Yu, H. Wang, Multi-task learning for multiple language translation, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1723–1732.
- [6] M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney, Translation modeling with bidirectional recurrent neural networks, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 14–25.
- [7] Y. Hu, M. Auli, Q. Gao, J. Gao, Minimum translation modeling with recurrent neural networks, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 20–29.
- [8] C. Wang, M. Li, A. J. Smola, Language models with transformers, 2019.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762).
- [10] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2020. [arXiv:1906.08237](https://arxiv.org/abs/1906.08237).
- [11] Jing, Ran, Iop, A Self-attention Based LSTM Network for Text Classification, 2019, p. 012008. URL: <https://iopscience.iop.org/article/10.1088/1742-6596/1207/1/012008>. doi:10.1088/1742-6596/1207/1/012008.
- [12] O. Bojar, C. Buck, C. Federmann, A. s. Tamchyna, Findings of the 2014 workshop on statistical machine translation, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 12–58. URL: <http://www.aclweb.org/anthology/W/W14/W14-3302>.
- [13] Han, SuHun, Google, googletrans: Free Google Translate API for Python. Translates totally free of charge., 2018. URL: <https://github.com/ssut/py-googletrans>.
- [14] AiTranslate, LSI, Corportion, Languages | Ai Translate | 120+ Languages | 64 MT Language Pairs, 2020. URL: <https://aitranslate.com/languages/>.
- [15] Z. Francois Chollet, A. G. Tan, Konduit, J. Allaire, T. O’Malley, Keras documentation: SimpleRNN layer, 2014. URL: https://keras.io/api/layers/recurrent_layers/simple_rnn/.
- [16] X. Z. Geoffrey Irving, Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification, volume 541, 2015, pp. 123–288. URL: <http://www.sciencedirect.com/science/article/pii/S0378437119318436>. doi:10.

- 1016/j.physa.2019.123288.
- [17] Abadi, P. B. a. Ashish Agarwal, Bidirectional Layers in TensorFlow Core v2.4.0, 2015. URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Bidirectional.
 - [18] H. S. Y. B. Kyunghyun Cho, Fethi Bougares, Learning phrase representations using RNN encoder-decoder for statistical machine translation, volume abs/1406.1078, 2014. URL: <http://arxiv.org/abs/1406.1078>. arXiv:1406.1078.
 - [19] MLCommons, MLCommons - Home, 2021. URL: "<https://mlcommons.org/>".
 - [20] A. Einstein, Zur Elektrodynamik bewegter Körper, volume 322, 1905, pp. 891–921.
 - [21] M. Goossens, F. Mittelbach, A. Samarin, The companion, Addison-Wesley, 1993.
 - [22] J. Tenni, A. Lehtola, C. Bounsaythip, K. Jaaranen, Machine learning of language translation rules, in: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028), volume 5, IEEE, 1999, pp. 171–177.
 - [23] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 311–318. URL: <https://doi.org/10.3115/1073083.1073135>. doi:10.3115/1073083.1073135.
 - [24] M. o. J. EUPARL, Euparl corpus for irish-english data, in: Gaois bilingual corpus of English-Irish legislation (processed), 2014.
 - [25] D. Merckx, S. L. Frank, Comparing transformers and rnns on predicting human sentence processing data, 2020.
 - [26] D. J. Kutyłowski, DeepL Translate, 2020. URL: <https://www.DeepL.com/translator>.
 - [27] M. J. Varela-Salinas, R. Burbat, et al., Google translate and deepl: breaking taboos in translator training, 2018.
 - [28] M. Translation, Papers with Code - Machine Translation, 2021. URL: <https://paperswithcode.com/task/machine-translation>.
 - [29] R. Karim, Illustrated: Self-Attention, 2017. URL: <https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a>.
 - [30] E. Liu, Y. Chu, L. Luan, G. Li, Z. Wang, Mixing-rnn: a recommendation algorithm based on recurrent neural network, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2019, pp. 109–117.

Data Quality Improvement and Entity Alignment Optimization for Constructing Large-Scale Knowledge Graphs

Keaton Sullivan^{a,b}, Fiona Browne^a, Huiru Zheng^b and Haiying Wang^b

^a*Datactics, 1 Lanyon Quay, Belfast, Northern Ireland, UK*

^b*School of Computing, University of Ulster, Newtownabbey, Northern Ireland, UK*

Abstract

Poor data quality can have an impact on the accuracy and analysis of knowledge graphs. Remediating this involves maximizing the data quality of sources used in constructing knowledge graphs and aligning entities. By improving the underlying data quality, knowledge graphs and their analysis are subsequently improved. In this paper we propose and implement a parallelizable data quality driven pipeline. We compare the proposed approach against one utilizing common pre-processing actions. This involves the measurement of entities validated against an external comprehensive dataset. A higher percentage reduces the need for complex algorithms that scale with a polynomial degree. We then show how the validated entities resulting from the pipeline produces high quality nodes and relationships that can be modelled as a realistic knowledge graph.

Keywords

Data quality, Parallelisation, Entity alignment, Textual similarity, Open data

1. Introduction

A global trend of increased publicly accessible datasets has been observed in [1] and attributed to commitments to governmental transparency initiatives such as the 2011 Open Government Partnership and the 2013 G8 Open Data Charter [2]. In addition to the volume, the diversity of data has grown exponentially [3]. The decentralised nature of datasets requires heterogeneous solutions to integrating and representing how different sources of data relate to each other. Knowledge graphs have become an effective solution of modeling these relationships as they provide a more realistic representation of integrated heterogeneous datasets by standardising data to an ontology that is composed of entities and how they relate to each other [4]. The versatility in modeling data from the heterogeneous data sources has found knowledge graphs being used anywhere data sources need to be integrated to support decision making processes - from the original use in creating a semantic web of the internet [5] to developing publicly accessible knowledge graphs of governmental data product offerings [6].

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ keaton.sullivan@datactics.com (K. Sullivan); sullivan-k@ulster.ac.uk (K. Sullivan);

fiona.browne@datactics.com (F. Browne); h.zheng@ulster.ac.uk (H. Zheng); hy.wang@ulster.ac.uk (H. Wang)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The construction of knowledge graphs has been a well researched topic, and an architecture [4] has been established - however the performance of implementing these in real-world large datasets has been identified as a challenge in many studies [4] [7] [8]. Algorithms developed in academia provide very accurate results, however the time complexity of their operations are rarely considered [9]. As a consequence, their practical applications have been found to be limited due to their polynomial time complexity not scaling with realistic datasets, despite their high accuracy.

The ontology and the knowledge base are the two fundamental components to knowledge graph construction.

Datasets undergo a process called knowledge extraction to transform unstructured, semi-structured and external structured data into a knowledge base of structured information. This process transforms the data into entities, their relationships and attributes that describe them. However, being from heterogeneous sources, it contains multiple issues that prevent it from modeling the real scenario, such as: (1) attributes may be spilt between multiple extracted entities (2) there may be duplicated entities (3) there may be subsets of attributes that can be considered independent entities.

Because of this, the extracted knowledge base undergoes a process called knowledge fusion which ultimately constructs the ontology of the knowledge graph and populates it with the knowledge base that is a more realistic approximation of the scenario - this process is often iterative as improvements may only be identified after investigating the completed ontology.

The process of resolving these issues in knowledge fusion before ontology construction is called entity alignment - and the algorithms that carries this out are a focus of academic study. Before these algorithms can be applied, data pre-processing is required to standardise the diverse representations of data and the inconsistencies of how they are recorded. However, academic datasets typically select datasets with limited data quality issues and therefore requires very little data preprocessing actions to be carried out.

The time complexity issue is primarily the product of attempting to apply algorithms with a polynomial time complexity to all entity-relationship pairs - and as such is primarily a limitation in the entity alignment stage which involves those comparisons [9].

Motivated by these challenges facing the practical application of knowledge graphs and with suggestions for solutions provided by [9], [7] and [8], this paper describes how the introduction of a data quality improvement pipe line that extends the existing data pre-processing stage of knowledge fusion can significantly reduce the time complexity challenges that would typically limit the practical implementations of academic entity alignment algorithms. This paper makes the following contributions:

1. development of a knowledge graph with practical applications from Open Governmental Data (OGD) datasets with significant data quality issues that would limit academic algorithms
2. proposed data quality improvement pipeline for practical implementations of entity alignment algorithms
3. demonstrate the time complexity reduction of compared method to existing algorithms and a typical pre-processing approach

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 discusses considered datasets. Section 4 describes the approach to knowledge graph construction and includes the proposed data quality improvement pipeline. Section 5 discusses the results of implementing the proposed pipeline. Section 6 discusses future work. Section 7 concludes the paper.

2. Related Work

In this section we highlight key research in the areas of open data, data quality and the construction of knowledge graphs from these data.

2.1. Data Quality in Datasets

Data quality has been the subject of a number of studies which attempt to define data quality dimensions. However, there is no clear consensus on these dimensions with and across domains despite the research focus.

An attempt to standardize the heterogeneous nature of data quality was conducted in [10]. This research produced a model ontology of data quality characteristics, dimensions and domains providing first steps at standardizing data quality in literature. This research found that data quality assessment was too premature of a field for the ontology to be viable in every approach. It was clear that while objective measures could be made generic, there would always be subjective or business specific measures that required expert assistance in defining.

The approach to data quality evaluation has been investigated. This includes data quality rules executed against data with calculations performed to capture percentage of data that pass such rules. These are straightforward to calculate [11]. However [10] proposes that the fitness for purpose should define beneficial metrics then design rules as complex as required to address them. This complexity applies limits on the number of rules, but is regarded as more beneficial for the users of the data due to being specific. Rules are assigned to a relevant data quality dimension and individual rule results are aggregated to provide a broad overview in dashboards for business users.

Ultimately, data quality is defined as fitness for purpose. Generic objective measures can be carried out [12], however an expert is required for subjective and business specific measures. To provide structure, a framework similar to [6], based on the characteristics described in [10], is a comprehensive approach to measuring data quality.

Knowledge of purpose, descriptive metadata, familiarity of data, comprehensive profiling focusing on regular distribution and outlier values are some of the methods used to inform definitions of data quality rules. The fewer present when describing data quality rules, the less fit-for-use defined rules would be.

2.2. Knowledge Graphs and Entity Alignment

Construction of knowledge graphs to model entities and their relations is growing in popularity across diverse domains such as finance, medicine and social sciences. Key challenges in generating high quality robust knowledge graphs are in the areas of data quality defined above

and in knowledge fusion where datasets are integrated together. Zhao et al. [4] conducts a systematic review of knowledge graph construction and identified a typical architecture which is separated into knowledge extraction and knowledge fusion. Entity alignment algorithms are an important topic within knowledge graph construction as identified in [4] and [9] whereby entities from different datasets are matched and presented in a graph structure. We refer to this process as entity alignment. Single alignment algorithms for small-scale datasets have been making significant advancements in the field such as realising knowledge graph representation learning with neural network based models [13], [14]. Alignment in large-scale networks have been facing significant challenges due to the polynomial time complexity of alignment algorithms which have to overcome data quality issues such as consistency. As a consequence, the single algorithm approach for large scale datasets are often reduced to a combination of simple matching strategies [8]. These produce multiple similarity scores which are combined to judge which entities are duplicates. High confidence matched entities often go through further alignment to produce a similarity score vector which is time consuming and not as important as lower confidence matches.

Studies [8], [7] routinely call for the development of algorithms capable of parallelism to address the time complexity. For instance, using a multi threaded approach to increase speed. However, this approach is not feasible with single algorithms as there is little concurrent activity to partition. In addition, as all entities are compared together, there would be an immense amount of inter-partition communication - both of these issues need to be addressed to facilitate parallelism.

In addition, the same studies state that if the data has significant data quality issues, then no algorithm can be used. Surprisingly, this issue is discussed as critical to an algorithm's success but the data pre-processing which aims to reduce it is often not described in detail or addresses the minimum consistency and integrity data quality issues.

To address such limitations, we propose a data quality improvement pipeline in this paper. This pipeline aims to reduce the volume of considered entity pairs so that matching using complex alignment algorithms are only performed only on the least confident entity pairs. Furthermore, we focus in detail on the data quality aspect of this pipeline to improve the matching downstream.

3. Datasets

In order to demonstrate the impact of a proposed data quality improvement pipeline on a knowledge base, open governmental datasets were selected as they fulfilled the following requirements: (1) available to the general public, (2) used for practical commercial decision making, (3) no existing knowledge graph is accessible, (4) has measurable data quality issues to improve and (5) heterogeneous entities and relationships between datasets.

The OGD from the UK was selected as the author could act as the expert in decision making due to familiarity with data and the UK has a high ranking on the global open data index. Common datasets available in OGD initiatives include budget allocation, national statistics, environmental data, weather forecasts, location and company registers. Company register was selected due to its common use in financial sectors. The data collection for this register is self

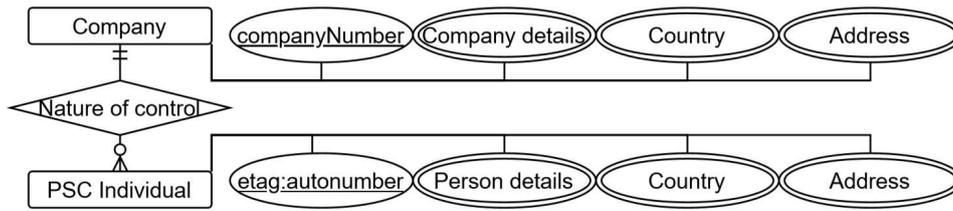


Figure 1: ERD of data sources with unrealistic 1:n cardinality

reported from users resulting in data quality issues. Multiple register datasets are available containing linking information. The specific OGD company register datasets selected in this study are the company data product and their persons of significant control (PSC) from March 2021.

The ontology represented by the data sources is summarised in Figure 1. Issues of why this is unrealistic is discussed in section 4.

Meta data for both companies data and PSC was minimal and only available from unstructured sources. The use of comprehensive profiling allowed the author to understand the content of the fields. A summary of these data along with data quality actions are detailed below along with additional corpora used in the data improvement pipeline for the entity alignment process.

3.1. Company Data Dataset

This set includes the basic name, registered address, operating status and classification of activities of all active UK companies. Updated monthly, available in CSV format and comprising of around 5 million records that have 60 fields. The majority of data quality issues related to companies house can be attributed to the lack of validation on addresses - particularly the lack of retrospectively correcting identified inaccuracies. We select all companies when building the knowledge graph - see Table 1 for summary of cleansing actions taken during pre-processing to ensure data quality was maximized before entity alignment.

3.2. Persons of Significant Control Dataset

This set includes information on the nature of control that individuals, companies and legal entities have over companies. It has user entered address and name information and is updated daily. The data is available in JSON format and comprises of more that 8 million records that have 36 properties. It includes the reference of the company and includes a list of 66 possible natures of control over a company which act as the direct relationship between PSC and companies dataset. PSC can be individuals with name and address information, companies who have address and registration information and legal entities who can have name, registration and address information. While discussed later, all fields except company number (excluding corporate number), kind (determines type of entity) and nature of control have data quality issues associated with them. It also includes companies that are exempt and reasons for non-compliance. We select only individuals as it comprises the majority of the dataset - see Table

Table 1

Table of Companies House company data cleansing and standardisation actions.

Column	Cleansing / Standardisation Process
Company Name	Extracts special character information along with standardising key values such as LTD, Limited.
	Standardise name based on rules from Companies House company search.
Country	List of UN Geoscheme countries was expanded to include UK regions. This was run against the country column to validate all the distinct countries in this column.
	Countries without PSC transparency were labelled as such through comparison to set of countries on the secrecy index dataset.
	As country of origin always contains a value, if country of registration was empty we inputted country of origin here.
Postal Town	When post town was not populated but postcode was, we inputted the post town value.
All fields	For dates – we ensure consistency by parsing all dates into a single format DD/MM/YYYY.
	Standard uppercase and punctuation.
	Cleansing and substitution of values to single values e.g. too, two, to.
	Replace often abbreviated words with their abbreviations.
	Extract characters and commas that could be used in SQL injections such as , “”, .

2 for summary of cleansing actions taken during pre-processing to ensure data quality was maximized before entity alignment.

4. Knowledge Graph Construction

Knowledge graphs construction can be classified as bottom-up when structures within data define the ontology - typically used in iterative implementation with minimal lead time to expanding functionality. Or top-down when well defined domain ontology and schema are considered first then the knowledge base is populated - typically used when interoperability of knowledge graphs is required.

A bottom-up approach was selected as the knowledge graph implemented heterogeneous data sets incrementally - being unable to define the resulting ontology prevented a top-down approach. The main limitation from this is limited interoperability with other knowledge graphs.

The construction of bottom-up knowledge graphs can be observed to follow a common architecture. This has been reviewed extensively by [4].

4.1. Knowledge Extraction

Heterogeneous datasets by their nature are provided in a variety of formats - knowledge extraction describes the approaches to identifying diverse entities, their relations and attributes within semi-structured and unstructured sources and transforming them into a more uniform format.

Table 2

Table of Companies House Persons of Significant Control cleansing and standardisation actions.

Column	Cleansing / Standardisation Process
Name	Titles were not expected in forename and surname yet were present. Applied clean titles rule to removed titles such as “Ms, Dr” from these columns.
	Performed profiling using string length, unexpected characters and numerals to identify outliers.
	Used a regular expression of common English words “to, the, and, at, a” against the name columns to profile non-name words to cleanse.
Country	Performed profiling and standardisation on all country columns.
	The three country columns were appended.
	Profiled the combined list to identify unique values (6.8K). This indicates data quality issues in this column. Shown and discussed in Figure 5
Postal town	Populated postal town based on associated postcode in corpus.
Postcode	Validating the structure of the postcode against valid postcodes from UK, US, India
All rows	For dates – we ensure consistency by parsing all dates into a single format DD/MM/YYYY.
	Standard uppercase and punctuation.
	Cleansing and substitution of values to single values e.g. too, two, to. Replace often abbreviated words with their abbreviations.
	Extract characters and commas that could be used in SQL injections such as , “”, , .

The diversity of entities and relationships dictates the available solutions. Natural Language Processing and Machine Learning methods are practical when handling diverse datasets, however they require significant investment in manually annotating training data to produce a model with acceptable accuracy. Rule-based methods are practical and provide high accuracy if all entities and relationships can be identified by matching with a set of manually developed predicates - it requires comprehensive implementations and a fixed structure which is only feasible in well defined datasets.

Ultimately, so long as the pipeline can differentiate the types of entities then either approach can be used, but a rules-based method was employed as the identified datasets came from semi-structured sources with clearly defined columns - making it easy to identify their type from the structure.

Implementation Figure 2: Entities in JSON were extracted using a series of rules that mapped all attributes to a csv row per entity - flattening the JSON to be in the same format as companies data. Each property had a corresponding column which had the heading of the fully qualified path of the property so no information would be lost. As nature of control was a limited list it was encoded in order to make it simplistic to identify the relationships.

Profiling was carried out to understand the content of each column. Investigating the structure of both datasets to identify candidate entities and the attributes involved in their conditional functional dependencies [7]. In general, the smallest subset of attributes that could be shared between entities were considered separate entities. For example the forename, surname, birth month and birth year was the smallest subset that could describe a person as their address could be considered a separate entity that people and companies share. In addition, if

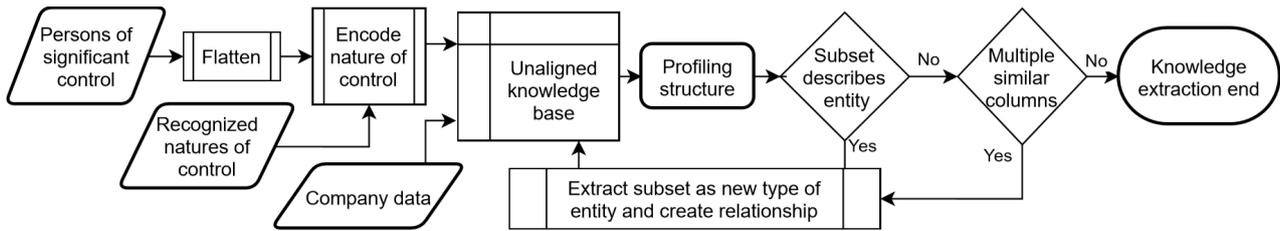


Figure 2: Process of knowledge extraction

the record could have multiple columns describing the same values then it would be best to be extracted as an entity and to create a relationship, such as 'country of origin' and 'registered country' in company data as well as 'country of residence', 'country registered' and 'country' in PSC.

4.2. Data Quality Improvement Pipeline

The focus of algorithms developed for entity alignment is the prediction accuracy and not the data itself so literature selects datasets with reasonably good data quality as they require minimal preprocessing actions. Even with reasonably good quality data, single algorithm approaches to entity alignment are too complex to be practical for large datasets - so a combination of simple matching strategies is typically used in the entity alignment algorithm. However, even the most advanced combination approaches cannot create knowledge graphs from sources with significant data quality issues [9], which is common place in open data. Therefore to have a practical implementation of knowledge graph construction from open data sources, data quality improvement is necessary regardless of algorithm.

The data quality pipeline encompasses the data preprocessing and entity alignment processes within knowledge fusion stage in the architecture of knowledge graph construction.

4.2.1. Data pre-processing

To address the challenge of parallelism, the pipeline partitions entities into concurrent specialised cleansing and alignment algorithms which align a certain type of entity via a series of matches.

A requirement of parallelism is that there needs to be no inter-partition communication. To satisfy this, entities were profiled then extracted to include any attributes involved in their conditional functional dependencies [7], as this resulted in disjointed partitions that contained all the attributes that would be involved in entity alignment (excluding those in external datasets).

Understanding conditional dependencies within the data goes beyond structural information. An expert or sufficient metadata is required to understand which values result in different entities or relationships being represented. Without this foresight, entities in a partition may have attributes in another - requiring inter-partition communication to complete tasks and therefore not be concurrent.

Comprehensive data cleansing was carried out within the pipeline to maximize data quality and has been summarised in Table 1 and Table 2.

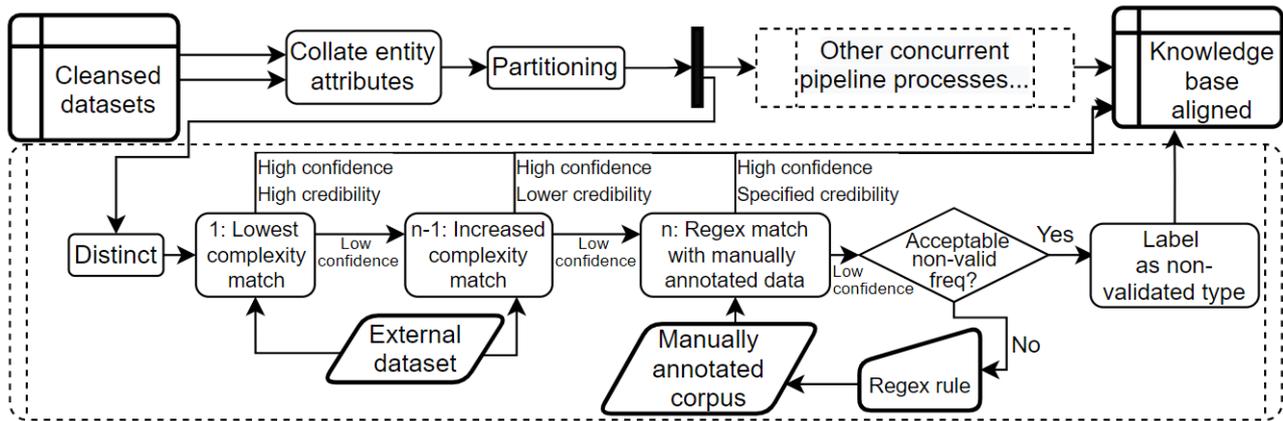


Figure 3: Generic entity alignment algorithm within pipeline

4.2.2. Entity alignment

The aim of entity alignment is to produce a set of unique entities that represent the real scenario. Duplication is expected from knowledge extraction but this leads to an unrealistic model in which relationships of a single real entity are spread across multiple duplicates that may not have any direct connections and therefore any network analysis would be inherently inaccurate.

For example, in the considered dataset this is taken to the extreme as multiple companies can be controlled by a single person and multiple people can control a single company in reality, however the data only represents the latter (shown in Figure 1) as no connections exist between people - network analysis only produces meaning results when the network represents reality, so would be ineffective. Entity alignment in this example takes all duplicated entities that had a single relationship and produces a set of unique entities with all relationships of their duplicates - representing the real many-to-many relationship.

According to [4], entity alignment in bottom-up approaches consists of textual similarity functions (like lexical similarity) producing values used in pairwise alignment algorithms (e.g. Levenshtein) then structural similarity functions being applied on collective alignment algorithms [8]. This process often relies on external datasets to validate entities. Due to the PSC dataset not containing enough distinguishing attributes to rely on the structure - this paper focus on textual similarity like those employed in [8] and demonstrated to be practical for OGD in [6].

The calculation of similarity metrics is used to indicate how likely two entities are to referring to the same entity. Multiple matching strategies may be applied within a single entity alignment algorithm and many exist for specific applications [9]. The resulting values are multidimensional so similarity combination is the process of evaluating all similarity scores and returning a single score. Alignment judgement is interpreting that score to specify which matched entities refer to a single entity then typically determine the winning attributes.

Character-based lexical similarity metrics: Lexical similarity calculates how similar a considered string is when compared to a defined corpus - the corpus is usually constructed from an external dataset of valid entities. Character-based lexical similarity measures how many

subtractions, updates and additions it takes for two strings to be equal.

An example is Levenshtein-ratio (occasionally known as fuzzy string matching), which considers all edit operations between two strings as the same weight [15]:

$$\text{lev}_{a,b}(i,j) \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \quad (1)$$

"where $1_{(a_i \neq b_j)}$ is the indicator function and it is equal to 0 if $a_i = b_j$, equal to 1 otherwise, and $\text{lev}_{a,b}(i,j)$ is the distance between the first i characters in a and the first j characters in b " [15]

The constructed knowledge graph utilizes a series of algorithms that returns lexical similarity - multiple algorithms were applied in series when the corpus that represented the allowable set of entity property values, didn't return satisfactory results.

Content-based category similarity metric: With content-based category metrics, similarity depends on the attributes shared by nodes - so similarity is calculated by "quantitatively evaluating the common information content of two categories" [8], matching-distance between compared content is used to evaluate the likelihood that comparisons refer to the same entity. However it assumes attributes are equally important, when in reality their importance actually depends on context. So improvements consider weights in the calculation.

The matching-distance similarity measure requires a selection of attributes that adequately differentiates the entity for others - attributes with without uniquely identifiable information need to consider many more attributes - considering additional attributes is adding another dimension in algorithms that have a polynomial time complexity so lead to severe performance issues. The comparison of time complexity is discussed in the results and the impact the data quality improvement pipeline provides, but common algorithms are defined below:

Simple matching coefficient compares how similar attributes are. It considered all attributes equally as important. It considers mutual absence in the nominator and denominator - which may not be realistic when comparing two subsets [15]:

$$SMC = \frac{M_{00} + M_{11}}{M_{01} + M_{10} + M_{00} + M_{11}} \quad (2)$$

Jaccard index is similar however it excludes mutual absences of both sets - so is effective when comparing subsets [15]:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{|X \cap Y|}{|X \cup Y|} \quad (3)$$

Overlap coefficient is the overlap of both sets [15]:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \quad (4)$$

Similarity combination and alignment judgement: The process of similarity combination takes the multidimensional results of those metrics and ultimately produces a single similarity score.

Alignment judgement is determining how those multidimensional results should be evaluated, producing a predicate that determines if entities are referring to the same entity, The typical approach to this is to unilaterally apply algorithms to all entities as this allows all entities to be matched to very high accuracy, however this results algorithms having the worst case of polynomial time complexity.

The chosen algorithms consider a subset of attributes that is the minimal amount to distinguish attributes. When attributes have high confidence in distinguishing entities (like company number), the algorithm requires few attributes - however entities that have only poor confidence in distinguishing attributes requires a larger number of attributes to be considered in algorithms for a similar result. Increasing attributes involves an exponential increase in time complexity of algorithms. Poor confidence attributes mainly exist because they contain non-unique values or that data quality issues have degraded confidence in the data. The implementation minimized the polynomial function in a number of ways:

1. Increasing the data quality before applying these algorithms boosted confidence in attributes so that entities could be matched in earlier algorithms (reducing size of entity pairs being considered), many new entities had enough confidence to match and entities without distinguishing attributes required fewer attributes.
2. Rather than consider all entity-pairs at once, the data quality pipeline selects only entity-pairs that would derive benefit from each algorithm. Rather than consider all metrics at once, the entity-pairs are filtered through a series of algorithms, where as if there is a confident match (e.g. exact country match) then it bypasses other similarity algorithms. Entities with confidence lower than the matching threshold underwent another algorithm to align as may as possible with the reduced set of unaligned entities. The time complexity of algorithms increased as the number of considered entities decreased - resulting in the algorithms with the highest polynomial function being applied on a minimal set of entity-pairs.

4.2.3. Ontology construction:

Bottom-up approaches like the one implemented have the data drive the formal definition of the ontology. Aligned entities and relationships are considered the knowledge base and are investigated to provide an ontology which would define the knowledge graph.

RDF and graph database are the two most widely implemented approaches to storage of knowledge graphs. Graph database was chosen as it performs faster when running queries and constructing graphs of a pre-defined structure. Neo4j is the most commonly used application for visualizing graph databases so was chosen.

The knowledge base was formatted to specify the ontology of entities and relationships according to neo4j syntax then the knowledge graph was constructed within minutes using their bulk import tool. The knowledge graph was visualized.

As expected, some PSC had such poor DQ that they couldn't be repaired and thus were missing relationships - but we discuss the improvement in results. Address and countries were shared by PSC and companies so were extracted to cluster entities around them. The data quality pipeline repaired data quality issues, derived new attributes and validated entities. The new ontology can be seen represented Figure 4.

5.2. Ontology

We developed an ontology to represent the control asserted by individual PSC.

The scenario represented in the original datasets looks like Figure 1. The lack of a many-to-many relationship between PSC and company meant it was unrealistic as the fundamental many-to-many control structure wasn't represented. Subsets of attributes like address were shared between entities but couldn't be clustered on, so entities at the same address were as closely related as entities in another country.

Whereas a realistic ontology Figure 4 produced from the data quality improvement pipeline Figure 5 has the following ontology. It models the missing many-to-many relationships, clusters entities that have similar attributes, validates country and address details and repairs dirty data that would typically degrade the performance of the entity alignment algorithms.

5.3. Boosting Data Quality to Reduce the Reliance on Complex Entity Alignment Algorithms

Time complexity in knowledge graph construction is centred around entity alignment. The presence of data quality issues that would degrade accuracy can be overcome with more complex entity alignment algorithms, however it is a polynomial cost function - addressing the data quality issues is a linear cost function.

We show how the cleansing, validation and repairing in data preprocessing (shown here 1) that addresses data quality issues in attributes that poorly distinguish entities can be sufficient to create a realistic network from the simplest match - where a complex algorithm would otherwise be required.

When there is no significantly distinguishing attributes such as phone number, then it isn't practical for large datasets. We show how an implementation of the data quality improvement pipeline on forename, surname, birth month and birth year can implement cleansing to simplify the required matches.

We select a reduced amount of attributes, profile attributes using string length, unexpected characters and numerals to identify outliers for data quality issues and use those findings to create a manually annotated corpus used to cleanse data. While the pseudonymised individual details in PSC prevents validating individuals that was possible with countries, the cleansing standardised name for consistency and removed irrelevant titles to improve accuracy. The example of this impact is cleansing the forename 'The Executors Of The Estate Of Geoffrey' to 'GEOFFREY' - however while cleansing was applied to all, 1,633 of the records had additional cleansing applied that would have required a more complicated match. We show how this cleansing is practical by implementing the simplest exact match and it creating a realistic structure.

If pre-processing was limited to those found in literature like then the scoring functions of entity alignment algorithm for syntactic matching would implement multiple matching strategies similar in complexity to 1 which is based on edit distance between two strings. As described in [9], even with these complex algorithms the poor data quality would result in missing matches that are significantly different like in the 'GEOFFREY' example, so while a complex algorithm would provide more matches than no cleansing, it would still miss matches with data

quality issues. A common approach in literature is to increase the amount of attributes to dilute the impact of data quality issues, but this results in an exponential increase in complexity.

We find that our data quality pipeline improves that data quality sufficiently so that a reduced amount of attributes need to be considered and the algorithm can be simplified and produce comparable results. The pseudonymised nature of PSC entities means they can't be distinguished further by textual similarity so incorporating structural similarity could be implemented in the future to distinguish individuals with the same details but disjoint egonets.

5.4. Reducing Time Complexity within Entity Alignment

In addition to boosting data quality to reduce the reliance on entity alignment algorithms – we demonstrate how a data quality pipeline can implement a series of increasingly complex matches with a minimal amount of entity-pairs – and how it provides the opportunity to prioritise manual labelling when validating against an expected list. We demonstrate this through aligning the 17,292,145 country values from 4 columns – of which 7,724 are distinct values to an external dataset of recognised countries under the UN Geoscheme [16], implementing ISO country codes [17].

We identify data quality issues originating from (1) column combination, (2) misspellings, (3) aggregated countries, (4) countries that no longer exist, (5) entering values that aren't countries and (6) values that have no logical reason to be entered like '12 July 2012'. All of these are considered valid by companies house but not fit for purpose – the data quality pipeline provided the opportunity to validate countries against a recognised list of countries and to implement multiple effective matching algorithms. While validating knowledge graph construction against rules derived from external dataset has been undertaken before in [18], the data quality pipeline provides the opportunity reduce the entity pairs applied in the alignment algorithms that makes it scale better.

Extract all company columns and partition from rest of the dataset – this allows each process to be carried out in parallel with other cleansing and alignments. Cleansing to standardise values was carried out to apply cleansing rules then profiling was carried out to identify data quality issues. As a limited list was expected, entities were made distinct to reduce the entity pairs. It then undertook a series of matches utilizing the external dataset as shown in Figure 5. The important aspects are that it begins with the simplest and most comprehensive matching then subsequent matches are more complex. Notice how matches with high confidence results bypass additional similarity metrics – this results in the entity being aligned with the simplest match possible – contrasted to approaches like [8] and [19] in literature which calculate multiple similarity scores for all entity then combines the score into a single value to judge if it should be aligned. While their match considering all similarity scores will result in more accurate matches, the polynomial time complexity makes it impractical for large datasets – this paper's implementation of the data quality pipeline provides significant dimensionality reduction to make large datasets feasible.

Non-validated entities can be prioritised based on frequency and therefore the number of entities to manually annotate can be prioritised. As experts are involved in manually annotating training datasets in entity alignment algorithms that utilize ML, this can be a reduced amount of work as only the low-confidence subset requires annotation in the pipeline. The

pipeline including the prioritised annotation resulted in successfully validating over 99% of the country entities - as only 12,000+ couldn't be repaired even with manual annotation. Principles for designing combination of matching strategies algorithm:

1. entity's expected to be from a limited set should be compared against externally recognised reference list
2. comparisons should begin with the lowest time complexity matches like exact matching
3. subsequent algorithms should be increasingly complex with minimal entity pairs
4. matching historic and aggregated data to valid entities may reduce accuracy which conceal malicious intentions, match/label with most likely value that retains suspicion
5. complex algorithms with a low return should be implemented if they can be selectively applied, such as extracting country from postcode only if it matches a postcode regex
6. values that don't match an external set of recognised values should be retained, but labeled as non-validated
7. manual annotation may be required for non-validated data, so integrate profiling to prioritise effort on most common values

We compare this to a typical approach shown in table 3. Each row corresponds to an additional stage that may be applied in pre-processing. A higher percentage of validated entities corresponds to requiring a less complex alignment strategy. Stages: (a) Initial values; (b) Distinct to reduce duplicates; (c) Case insensitive; (d) Standardize punctuation; (e) Remove punctuation.

The external dataset used for validation is the UN Geoscheme countries, it notes 249 countries. As UK is considered a single entity in this dataset, if used without any pre-processing it would result in only validating 3.421% of the dataset before alignment, therefore to provide a more realistic use case, the values of the relevantly clean company data were all mapped to the dataset. 129 matched exactly, 23 required modification to match, 29 were required to be inserted (mostly countries or variations of UK and Irish countries and some dissolved countries) and 97 had no matches so remained as the same value as in the initial dataset. In total, 278 countries were identified in this approach.

Within 3, " $n \subseteq$ " is used as shorthand to describe the number of distinct validated entities, which can be compared to the "Count" of entities after Stage (b). " $n \sum$ " refers to the percentage of total validated entities this validates in the selected column(s). The shaded cells

The typical approach achieves below 62% in validating all entities - this requires a more complex entity alignment algorithm to align entities with the same accuracy compared to our approach.

5.5. Derived Attributes

We additionally show that the data quality pipeline can be used to improve the completeness of the knowledge graph by imputing values from the validated values within the graph. While it has been carried out previously – incorporating it into partitions allows for actions to be completed concurrently. For example we use postcode and city to extract post town and nation for the dataset – we include additional country information like its region and if it can be trusted. All of this information provides tangible benefit to network analysis. For example, the report

Table 3

Table decomposing the number of distinct validated entities and resulting percentage of countries validated in a typical approach to pre-processing.

	id.country_registered			CH PSC country_of_residence			address.country			CH CD RegAddress.Country			CH All		
	Count	n	Σ	Count	n	Σ	Count	n	Σ	Count	n	Σ	Count	n	Σ
(a)	505714	149	47.783%	7084230	204	62.378%	6245555	200	58.063%	3456646	153	62.997%	17292145	217	60.517%
(b)	2934	149	47.783%	4882	204	62.378%	1361	200	58.063%	192	153	62.997%	7724	217	60.517%
(c)	2934	157	73.909%	4882	222	62.609%	1361	212	58.181%	192	153	62.997%	7543	233	61.418%
(d)	2884	174	73.928%	4724	311	62.615%	1322	239	58.208%	192	153	62.997%	7304	351	61.431%
(e)	2539	212	74.052%	4586	327	62.618%	1148	340	58.233%	192	154	63.006%	6691	483	61.446%

‘The companies we keep’ identifies that individuals who originate from countries high on the financial secrecy index, may not be relied on. They find that certain countries like Ukraine protect details of PSC making it appealing to money laundering. By creating a knowledge graph that is enriched by a data quality pipeline, we can conduct better network analysis by considering missing links and can cluster entities on more shared attributes. The drawback is internal data may need to be duplicated if used by multiple partitions to avoid dependencies, however as our example using countries uses an external list, it wasn’t required.

5.6. Limitations

The following are some of the major limitations in the implementation of the pipeline.

Partitions require disjoint datasets to avoid inter-partition communication, which is only possible if data sources have a well-defined structure – this means that it isn’t practical for unstructured data sources and data duplication may be required if attributes are shared between entities but can’t be extracted as an entity. The series of entity alignment algorithms excluding high confidence matches means that entity judgment is dependant on order of operations and that the most important accurate matches have the lowest time complexity.

Entities that didn’t match were excluded, when they should be retained and specified as non-validated. Also original values should be retained to comply with regulations of transparency.

Partitions are based on entity type and not size so parallel processing would never be balanced. This effect is reduced the greater the number of types of entities.

The entire pipeline requires a great deal of expert knowledge to implement. Particularly the cleansing, which becomes less feasible the more entities there are.

The bottom-up approach isn’t suitable for interoperability with other knowledge graphs and it hasn’t been considered for top-down approaches. Implementing the ontology following a comprehensive schema like one provided by schema.org [20] would allow the same approach but make it easier to integrate with other anthologies.

The datasets were semi-structured and conditional functional dependencies were easy to identify in knowledge extraction due to sample size - it is unlikely that disjoint sets of entities could be extracted from unstructured data sources using the pipeline. NLP approaches could be a solution but the sporadic extraction of attributes would degrade algorithms that rely on complete data.

6. Future Work

Knowledge graph construction is an iterative process, and therefore, additional work can expand its functionality. The constructed knowledge graph wasn't compared with knowledge graphs using different approaches. Differences in network analysis should be investigated to determine how effective the pipeline is at creating the graph from the same data.

The PSC dataset was limited to individuals which was the majority of entities, however, including other types would provide a more representative network and be more heterogeneous in nature – providing a more accurate comparison against other algorithms.

The current approach doesn't discuss selecting the winning attributes of aligned entities unless they are validated – realistically when referring to an individual entity there should be a golden record that describes it the most likely attributes.

7. Conclusion

As open data use becomes increasingly pervasive in supporting decision making, construction of the knowledge graphs that models and facilitates the network analysis increases in scale and data quality issues.

As described in [21], algorithms that focus on developing overly complicated models to dilute data quality issues or develop solutions with pristine data, degrade their usefulness in real-world scenarios. Academic papers on entity alignment are observed as following this as they propose algorithms with a polynomial time complexity which aren't feasible to create on large-scale open data.

We show how a use-case built from a typical approach to pre-processing in entity alignment can validate entities against external data sources to reduce the need for complex alignment algorithms and that it can be substantially improved with a dedicated pipeline such as what we propose.

The implementation of the data quality pipeline is a practical approach which attempts to overcome identified challenges of large-scale knowledge graph construction from dirty data but requires a great deal of expert knowledge to design initially. Maximizing data quality requires significant development time but it has more of an effect in terms of accurately reflecting entities than increasing the complexity of alignment algorithms.

References

- [1] J. Attard, F. Orlandi, S. Scerri, S. Auer, A systematic review of open government data initiatives, *Government Information Quarterly* 32 (2015) 399–418. doi:10.1016/j.giq.2015.07.006.
- [2] R. Matheus, M. Janssen, A systematic literature study to unravel transparency enabled by open government data: The window theory, *Public Performance & Management Review* 43 (2020) 503–534. doi:10.1080/15309576.2019.1691025.
- [3] Company Register, Open knowledge - company register 13 percent open, 2021. URL: <https://index.okfn.org/dataset/companies/>, accessed: 2021-06-25.

- [4] Z. Zhao, S.-K. Han, I.-M. So, Architecture of knowledge graph construction techniques, in: *International Journal of Pure and Applied Mathematics*, 2018, pp. 1869–1883.
- [5] W. Hu, J. Chen, Y. Qu, A self-training approach for resolving object coreference on the semantic web, in: *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, 2011, p. 87–96. doi:10.1145/1963405.1963421.
- [6] J.-w. Lee, J. Park, An approach to constructing a knowledge graph based on korean open-government data, *Applied Sciences* 9 (2019). doi:10.3390/app9194095.
- [7] W. Fan, Data quality: From theory to practice, *SIGMOD Rec.* 44 (2015) 7–18. doi:10.1145/2854006.2854008.
- [8] K. Sun, Y. Zhu, J. Song, Progress and challenges on entity alignment of geographic knowledge bases, *ISPRS International Journal of Geo-Information* 8 (2019) 77. doi:10.3390/ijgi8020077.
- [9] K. Zeng, C. Li, L. Hou, J. Li, L. Feng, A comprehensive survey of entity alignment for knowledge graphs, *AI Open* 2 (2021) 1–13. doi:10.1016/j.aiopen.2021.02.002.
- [10] C. Fürber, M. Hepp, Towards a vocabulary for data quality management in semantic web architectures, in: *Proceedings of the 1st International Workshop on Linked Web Data Management*, ACM, New York, USA, 2011, p. 1–8. doi:10.1145/1966901.1966903.
- [11] D. C. Corrales, A. Ledezma, J. C. Corrales, From theory to practice: A data quality framework for classification tasks, *Symmetry* 10 (2018). doi:10.3390/sym10070248.
- [12] H. Chen, G. Cao, J. Chen, J. Ding, A practical framework for evaluating the quality of knowledge graph, in: *Knowledge Graph and Semantic Computing*, Springer Singapore, 2019, pp. 111–122. doi:10.1007/978-981-15-1956-7_10.
- [13] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, D. Q. Phung, A novel embedding model for knowledge base completion based on convolutional neural network, *CoRR* (2017). doi:10.18653/v1/N18-2053.
- [14] D. Q. Nguyen, T. Vu, T. D. Nguyen, D. Q. Nguyen, D. Q. Phung, A capsule network-based embedding model for knowledge graph completion and search personalization, *CoRR* (2018). doi:10.18653/v1/N19-1226.
- [15] M. K. Vijaymeena, K. Kavitha, A survey on similarity measures in text mining, *Machine Learning and Applications: An International Journal* 3 (2016) 19–28. doi:10.5121/mlaij.2016.3103.
- [16] UN Department of Economic and Social Affairs, Countries or areas / geographical regions, 2020. URL: <https://unstats.un.org/unsd/methodology/m49/>, accessed: 2021-06-25.
- [17] International Organization for Standardization, Iso 3166 - country codes, 2020. URL: www.iso.org/iso-3166-country-codes.html, accessed: 2021-06-25.
- [18] S. Guo, Q. Wang, L. Wang, B. Wang, L. Guo, Knowledge graph embedding with iterative guidance from soft rules, *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (2018). URL: www.arxiv.org/abs/1711.11231.
- [19] Z. Wang, J. Li, J. Tang, Boosting cross-lingual knowledge linking via concept annotation, in: *IJCAI*, 2013, pp. 2733–2739.
- [20] Schema.org, Corporation, 2021. URL: www.schema.org/Corporation, accessed: 2021-6-25.
- [21] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. K. Paritosh, Lora, Aroyo, Everyone wants to do the model work not the data work: Data cascades in high-stakes ai, in: *2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

The Choice of Reference Channel in Channel Alignment and Channel Selection

Ingo Stengel^a, Karin Pietruska^a and Matthias Wölfel^a

^a Karlsruhe University of Applied Sciences, Moltkestraße 30, Karlsruhe, 76133, Germany

Abstract

Channel alignment based on the generalized cross correlation with phase transform (GCC-PHAT) is part of many multichannel speech processing procedures including the channel selection procedure based on multichannel cross-correlation coefficients (MCCC). Despite the wide application of the GCC-PHAT approach for channel alignment, little has been reported on how the choice of reference channel might affect alignment accuracy and subsequent processing steps when microphones are coarsely distributed. The present research investigates alignment accuracy with random selection of a reference channel in relation to heuristic selection of a reference channel using the GCC-PHAT approach for time difference of arrival (TDOA) estimation and subsequent MCCC based channel selection. Results show that the procedure for reference channel selection effects both: the accuracy of channel alignment as well as results of the subsequent channel selection procedure. Findings suggest that the choice of reference channel should not be left to chance in distributed microphone arrays in order to optimize processing steps following channel alignment.

Keywords

Channel selection, multichannel speech processing, microphone array

1. Introduction

In various contexts the auditory information of a scene is recorded by several spatially distributed microphones, so called microphone arrays. In film production or sports broadcasting, spaced microphone arrays are used to create an immersive audio experience and to separate sound sources of interest from ambient noise [1, 2]. Likewise, in conference rooms or lecturing halls microphone arrays have proven useful to enhance sound signals that emanate from the current speaker while reducing noise from spatially distinct locations.

The estimation of time differences of arrival (TDOA) of sound signals at different microphones forms a critical first step in many techniques employed in microphone array processing for noise reduction [3], speaker localization [4, 5], channel selection [6] or speech enhancement [7]. First introduced more than half a century ago [8], the generalized cross correlation technique remains a widely applied method for TDOA estimation in near field and far-field scenarios [9]. The cross-correlation technique for TDOA estimation takes two signals as input and finds the time lag between the two signals that maximizes the value of the cross-correlation function. In the generalized cross correlation technique [8], an additional weighting function, also referred to as filtering, is applied to the cross-correlation. This paper focuses on the PHAT-weighting function, a filtering approach that has proven particularly useful for TDOA estimation in indoor settings that are characterized by signals with different forms of reverb [10]. Throughout the years, research has dedicated much attention on expanding and optimizing the GCC-PHAT approach. Only recently, a subband analysis with GCC-PHAT has yielded improved accuracy in TDOA estimates in relation to the classic approach [11]. To date, the GCC-PHAT is widely applied in multichannel signal processing and often constitutes one of the first steps when combining multiple signals. Channel alignment based on TDOA estimation with

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ ingo.stengel@h-ka.de (I. Stengel); karin.pietruska@h-ka.de (K. Pietruska); matthias.woelfel@h-ka.de (M. Wölfel)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

GCC-PHAT constitutes also the first step of the multichannel cross-correlation coefficient (MCCC) procedure for channel selection first introduced by Kumatani et al. (2011) [6].

Selecting a subset of channels of a microphone array for further multichannel processing remains of key interest in present research, particularly regarding voice-based assistance systems or conferencing systems that integrate signals received by a dispersed microphone array. As research has shown, adding more channels, particularly when they have a low signal to noise ratio, may not improve but instead substantially decrease the performance of an automatic speech recognition system [12]. Throughout the last decades, different approaches for channel selection have been introduced including classic signal to noise ratio estimation [13], class separability of phonemes [14], multichannel cross-correlation coefficients [6], cepstral distance [15] and neural network posterior probability models [16]. Notably, channel alignment remains an integral processing step in many of these multichannel approaches also including end-to-end ASR models, particularly with masked based neural beamforming [17].

We focus on the MCCC approach for channel selection, due to a suggested decrease in computational complexity compared to the use of an automatic speech recognizer (ASR) for channel selection [6, 14] and its use of the GCC-PHAT approach for channel alignment [6]. The MCCC approach aims at discarding low quality, noisy channels by building on the assumption that noise is uncorrelated to the speech signal of interest. Setting information on the spatial correlation among signals in relation to their variance, the MCCC algorithm implements a channel selection procedure that in combination with beamforming has yielded similar word error rates compared to a close distance microphone while focusing on computational efficiency. Channel alignment is implemented in the first processing step of the MCCC algorithm in order to optimize the accuracy and computational efficiency of the subsequently computed spatial correlations among channels [6]. Despite the wide use of the GCC-PHAT in the first step of channel alignment, little is known on how the choice of reference channel affects alignment accuracy and subsequent processing steps. This is of particular interest for microphone arrays that are spaced coarsely across the recording room with variable inter-microphone distances. This coarse setting differs profoundly from much of previous research that focused on linear, evenly spaced microphone arrays [18].

The present research aimed at investigating the effect of the choice of reference channel on TDOA accuracy for channel alignment based on the GCC-PHAT approach. Moreover, follow-up effects of the choice of reference channel for alignment on the overall results of the MCCC based channel selection procedure were examined. Random choice of reference channel was compared to a choice of reference channel based on a delay heuristic. Short-Time Objective Intelligibility (STOI) scores for each channel served as an independent speech intelligibility measure for the achieved channel rankings [19]. The effects were examined on data recorded in indoor settings with microphones distributed on a table or stand. The microphone locations were similar to a distribution that can be expected in ad-hoc microphone arrays when recording business meetings or seminars with the smartphones of meeting participants. The use of real data was critical in the approach as synthetic data are known to generate results that are often not replicable in realistic indoor settings.

2. Methods

2.1. Data

In order to investigate the present research questions, we aimed at using data recorded in indoor settings with unobstructed microphones spaced across a table or located on a stand. The VOICES corpus provides recording conditions that fulfill these requirements along with spatial information that allowed the approximation of inter-microphone distances of a subset of the microphones used in the recordings [20]. We constrained the analysis to data recorded in room 3 (size: 7.6 by 7.6 m) with a foreground loudspeaker angle of 90 degrees. In the 90 degrees position, the loudspeaker is in line with the microphones of interest (azimuth angle = -90 degrees), causing a maximal time delay between subsequent microphones. None of the distractor noise loudspeakers were active. Pre-recorded speech from LibriVox recordings was played by the foreground loudspeaker in room 3 equipped with basic furniture including a table, chairs, a shelf as well as a refrigerator. 20 microphones of different type (studio microphones, lavalier microphones, MEM microphones) were placed at different locations within the room. In the present paper, we constrained the microphones included to microphones that

were unobstructed, positioned either on a table or stand and located in front of the foreground speaker box. Inter-microphone distances were approximated by taking the difference scores of the indicated distance of each microphone to the foreground speaker box. Table 1 lists the included 7 microphones with information on the type of microphone, their location with respect to the foreground speaker box. Data were recorded with a PreSonus StudioLive RML32AI digital mixer and PreSonus Capture recording software and all channels were sampled synchronously with a sampling frequency of 16 kHz [20]. Table 2 and Table 3 depict the approximated inter-microphone distances as well as the expected differences in TDOA values in terms of samples given a sampling frequency of 16 kHz.

Table 1

Type, model and location of microphones included in the analysis. Microphone model and location descriptions are based on the documentation of the VOICES corpus [20]

ID	Type	Model	Location
01	studio	SHURE SM58	close on table
02	lavalier	AKG 417L	close on table
03	studio	SHURE SM58	mid distance on table
04	lavalier	AKG 417L	mid distance table
05	studio	SHURE SM58	far distance on stand
06	lavalier	AKG 417L	far distance on stand
16	bar	ATR4697	mid distance on table

Table 2

Approximated inter-microphone distances in centimeters based on the given distance information of each microphone to the foreground speaker box. Height differences are not adequately represented in the calculated inter-microphone distances. Distance values in the original VOICES corpus are indicated in inches without positions after the decimal point. Indicated inter-microphone distances are therefore broad approximations.

	ID 01	ID 02	ID 03	ID 04	ID 05	ID 06	ID 16
ID 01	0	0	201	201	544	544	104
ID 02	0	0	201	201	544	544	104
ID 03	201	201	0	0	343	343	97
ID 04	201	201	0	0	343	343	97
ID 05	544	544	343	343	0	0	439
ID 06	544	544	343	343	0	0	439
ID 16	104	104	97	97	439	439	0

Table 3

Expected inter-microphone delays in samples based on the approximated inter-microphone distances and a sampling frequency of 16 kHz. Positive difference values indicate that the respective channel in the column is delayed with respect to the reference microphone ID denoted by the row label. Conversely, negative values indicate that the channel denoted by the column label was located more closely to the sound and was therefore ahead in time compared to the channel denoted by the row label.

	ID 01	ID 02	ID 03	ID 04	ID 05	ID 06	ID 16
ID 01	0	0	94	94	253	253	49
ID 02	0	0	94	94	253	253	49
ID 03	-94	-94	0	0	160	160	-45
ID 04	-94	-94	0	0	160	160	-45
ID 05	-253	-253	-160	-160	0	0	-205
ID 06	-253	-253	-160	-160	0	0	-205
ID 16	-49	-49	45	45	205	205	0

2.2. Analysis 1: Reference Channel on TDOA

This first analysis investigated the effect of the choice of reference channel on the accuracy of channel alignment. Time differences of arrival (TDOA) for each channel with respect to a chosen reference channel were estimated with the generalized cross-correlation with PHAT weighting (GCC-PHAT). First introduced by Knapp and Karter (1976) [8], the generalized cross-correlation function denoted by $R_{km}(\tau)$ takes two microphone signals k, m as an input and computes the cross-correlation of the filtered versions of these two input signals. When applying the PHAT weighting, these filters consist of the phat weighting function denoted by $\psi_{km}(\omega)$ as described by the following equations:

$$R_{km}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_{km}(\omega) X_k(\omega) X_m^*(\omega) e^{j\omega\tau} d\omega \quad (1)$$

$$\psi_{km}(\omega) = \frac{1}{|X_k(\omega) X_m^*(\omega)|} \quad (2)$$

The maximum of $R_{km}(\tau)$ is the lag value estimated by the GCC-PHAT function that corresponds to the relative delay between the input signals k, m .

$$\widehat{\tau}_{km} = \arg \max_{\tau} R_{km}(\tau) \quad (3)$$

In order to examine the effect of the choice of reference channel on alignment accuracy, random selection of one reference channel was compared to the selection of a reference channel based on a delay heuristic. Let k be the total number of channels. The delay heuristic attempts to estimate the channel located closest to the sound source by taking each channel k_r and computing the relative delays of all remaining $k-1$ channels with regard to channel k_r . The result is a delay matrix \mathbf{D} of dimension $k \times k$, whereby k is equal to the total number of channels. Each row of \mathbf{D} contains the GCC-PHAT delay estimates $\widehat{\tau}_{k_r, k_i}$ in relation to one specific reference channel k_r . The delay matrix \mathbf{D} is a hollow matrix in which the diagonal values are all zeros as the relative delay of a channel k_r to itself is always zero.

$$D_{k,k} = \begin{pmatrix} \widehat{\tau}_{1,1} & \widehat{\tau}_{1,2} & \cdots & \widehat{\tau}_{1,k} \\ \widehat{\tau}_{2,1} & \widehat{\tau}_{2,2} & \cdots & \widehat{\tau}_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\tau}_{k,1} & \widehat{\tau}_{k,2} & \cdots & \widehat{\tau}_{k,k} \end{pmatrix} \quad (4)$$

The first row of the delay matrix \mathbf{D} contains the estimated time delays $\hat{\tau}$ of each channel with respect to channel 1. The second row of the delay matrix contains the delays of each channel with respect to channel 2 up until row k with reference channel k . The value of $\hat{\tau}$ is of positive sign when the respective channel is delayed to the reference channel of the respective row and negative if it is ahead of the reference channel of the respective row. More precisely, if the value of $\hat{\tau}_{i,2}$ is of positive sign, the signal of channel 2 was delayed with respect to channel 1. The heuristic aims at choosing the channel as the reference for the alignment procedure to which all other channels are delayed. Due to the possibility of maxima of the generalized cross correlation function $R_{km}(\tau)$ that might result from signal reflections or noise, the TDOA estimate may in some cases fail to reflect the ground truth time difference between two channels k,m . The heuristic therefore adopts the channel as reference channel for alignment with the maximal number of $\hat{\tau}$ values of positive sign within the respective delay matrix row. This means, based on the GCC-PHAT TDOA estimates, the maximal number of channels are delayed with respect to the reference channel.

Delay Heuristic:

Take each row d_{i*} of delay matrix \mathbf{D} and compute the sum of the outputs of the sign function for each row entry.

$$g(d_{i*}) = \sum_{j=1}^J \text{sgn}(d_{ij}) \quad (5)$$

Take the index i of the row that maximizes the output of $g(d_{i*})$. This index i denotes the row of delay matrix \mathbf{D} that contains the delays with respect to the reference channel k_{ref} selected by the delay heuristic for channel alignment:

$$k_{ref} = \underset{d_{i*} \in \mathcal{D}}{\text{argmax}} g(d_{i*}) \quad (6)$$

2.3. Analysis 2: Reference Channel on MCCC Channel Selection

The second analysis investigates potential follow-up effects of the choice of reference channel for alignment on the results of the channel selection procedure based on the MCCC algorithm [6]. As described below, channel alignment based on TDOA estimates by GCC-PHAT constitutes the first step of the MCCC channel selection algorithm. Following the alignment procedure, the covariance matrix \mathbf{S} is computed for each sample to capture the spatial correlations among channels.

$$\mathbf{S}_{k,k} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_k^2 \end{pmatrix} \quad (7)$$

The MCCC score ρ_k for a specific sample is calculated by computing the determinant of the covariance matrix \mathbf{S} and dividing it by the product of the diagonal elements s_i^2 of \mathbf{S} , which denote the variance of the signal within the respective channel i .

(8)

$$\rho_k = \frac{\det[\mathbf{S}_k]}{\prod_{i=1}^k s_i^2}$$

For detailed mathematical notations, please refer to the paper by Kumatani et al. (2011) [6]. In the following, the channel selection algorithm based on MCCC values is briefly summarized:

1. TDOA estimation based on GCC-PHAT
2. Channel Alignment of the k signals
3. Denote all the channels in the search space as K_c
4. Find set K_s of K_c-1 signals with highest MCCC value
5. Remove k_i that was not included in set of K_c-1 with highest MCCC value from search space
6. Go to step (3.) if MCCC value of K_c is larger than the MCCC value of the subset K_s , $K_c > K_s$

The smallest set K_s of channels that can be retained by the algorithm consists of two channels. At least two channels are needed to compute a spatial correlation and thus the MCCC value. By saving the channels that are excluded in subsequent rounds of the algorithm, we receive a channel ranking from worst quality to the two best quality channels as ranked by the MCCC algorithm.

In the present analysis, for each of the $n=128$ utterances of the VOICES corpus, a channel ranking based on the MCCC algorithm was computed and compared to the ranking of the channels based on the Short-Time Objective Intelligibility (STOI) scores [19]. The STOI score has a range from 0 to 1 with higher scores representing increased speech intelligibility. Table 4 displays the STOI Scores for the first sample utterance for each of the $k=7$ channels as well as their distance to the foreground speaker-box. The LibriSpeech source recordings of the respective utterance served as the non-degraded signal for the STOI-score computation.

Table 4

STOI scores for each channel (=microphone) of sample 1 along with distance of each microphone to the foreground speaker. The original LibriVox signal served as the undegraded reference for computation of STOI scores.

Mic ID	Distance to Speaker (cm)	STOI Score
01	170	0.59
02	170	0.49
03	371	0.37
04	371	0.31
05	714	0.26
06	714	0.23
16	274	0.39

3. Results

3.1. Analysis 1: Reference Channel on TDOA

The present analysis investigated the effect of a random choice of the reference channel in relation to a heuristic choice of reference channel on the accuracy of TDOA estimates based on GCC-PHAT. As a measure of accuracy, the difference scores of the computed TDOA values in relation to the approximated ground truth values were computed. Ground truth values are defined as the expected TDOA in samples given the a priori approximated distances between microphones. Table 5 depicts the mean and standard deviation of the difference scores for each channel of the 128 samples in the random and heuristic condition of reference channel selection. Results show an increased mean difference score and variance in the random condition in relation to the heuristic condition.

Table 5

Difference scores of the TDOA values in relation to approximated ground truth values for channel alignment based on randomly selected reference channel and reference channel based on heuristic. Displayed are the mean and standard deviations for each channel of the n=128 samples.

MIC ID	Random Selection		Heuristic	
	Mean	SD	Mean	SD
01	5.4	10.7	1.0	0.0
02	1.7	1.9	0.0	0.0
03	5.4	10.3	1.4	1.1
04	12.3	28.5	0.0	0.0
05	3.4	1.9	4.0	0.0
06	27.4	36.8	5.2	2.2
16	5.7	10.5	0.0	0.0

3.2. Analysis 2: Reference Channel on MCCC Channel Selection

The second analysis investigated potential follow-up effects of the choice of reference channel during the alignment procedure on the channel selection and ranking based on the MCCC algorithm. A channel ranking based on the MCCC algorithm was computed for all 128 samples with a reference channel randomly chosen during the alignment procedure and a reference channel chosen based on the delay heuristic. The resulting MCCC based channel rankings from worst to best quality channels for both conditions were compared to the rankings based on STOI scores. As previously described, the MCCC algorithm can retain a minimum set of 2 channels during the selection procedure. These two channels are ranked as the signals of best quality according to the MCCC algorithm. For the random selection of reference channel condition, the number of samples in which the set of the two selected channels was identical to the set of the two best quality channels based on the STOI scores was decreased with n=33 samples compared to the heuristic condition with n=116 samples. As indicated in table 4, STOI scores decreased with increasing distance to the foreground speaker box indicating that STOI score rankings adequately captured the effects of signal attenuation and reverb. The supplementary material shows the channel rankings for the first 10 samples in the random and heuristic condition based on the MCCC algorithm along with the reference channel used for each sample and condition.

4. Discussion

Present findings reveal that the choice of reference channel for alignment effects TDOA accuracy based on GCC-PHAT. More specifically, random selection of the reference channel was associated with increased deviation from ground truth values as well as with increased between sample variability of TDOA estimates. In addition, the choice of reference channel for the alignment procedure affected results of the subsequent channel selection approach based on MCCC. Selection of the reference channel based on a delay heuristic yielded channel selection results that were congruent with STOI scores for the majority of the utterances. In contrast, random selection of a reference channel was associated with only 26% of the samples in line with STOI scores. Findings suggest that deviations from the ground truth in the alignment procedure as well as the selected reference channel per se might affect subsequent spatial covariance computations involved in the MCCC channel ranking approach and thus yield selection results that are not optimal for subsequent speech recognition steps in terms of speech quality.

In contrast to previous research that used a linear microphone array with $N=64$ microphones and equal inter-microphone spacings of 2 cm [6, 18], the present research employed only a small subset of microphones and these were distributed with inter-microphone distances up to 5 meters. Therefore it is to be expected that channel differences between microphones are more pronounced in the present data set due increased inter-microphone effects of reverb and sound attenuation. Consequently, when optimizing alignment to a remote channel with substantial reverb effects and a decayed source signal, the spatial correlation of channels with similar reverb shaded degradations could be enhanced and thus confound overall results of the channel ranking.

Notably, the microphones included in the present work were not of the same kind, but differed in terms of their operating principle: dynamic microphones as well as condenser microphones were included in the analysis. Although the present number of channels is very limited, findings show a trend towards increased variability in TDOA estimates in microphones not only as a function of distance to the sound source but also as a function of microphone type in the random selection condition. The condenser microphones were associated with increased variability, particularly when they were located more remotely from the sound source. It remains up to future research to further investigate these tentative findings and also if subband calculations of GCC-PHAT may decrease these effects [11].

Recent research focusing on far field speech recognition in noisy and reverberant conditions with coarsely distributed microphone arrays has focused increased attention on the choice of reference channel. Maximization of cross-correlation coefficients [12] as well as attention-based approaches [17, 21] have been suggested as strategies for reference channel selection. This is in line with the present findings, implying that the choice of reference channel should not be left to chance in environments where microphones are more widely distributed and thus record signals that differ more profoundly with regard to reverb and attenuation.

4.1. Limitations

The spatial accuracy of present calculations was limited by the distance information provided by the VOICES corpus documentation of the corresponding website [20]. Inter-microphone distances were broadly approximated by building the difference scores between the given distance information of each microphone to the foreground speaker box. Consequently, differences in height were not adequately represented in the derived ground truth distances. In addition, distances were indicated in inches without decimal points which also limits the accuracy of the present conversions to centimeters. Consequently, the present TDOA results of channel alignment were compared to broadly approximated ground truth values. Despite this limitation, present results on channel alignment are meaningful in that they do not only show an increased deviation from ground truth values when selecting a reference channel randomly, but they also show an increased variability in this deviation as compared to a heuristic selection of a reference channel.

Speech recordings of the present data were based on prerecorded LibriVox utterances played by a speaker box positioned in the room. This needs to be taken into consideration as spectrograms between

recorded speech and real human speakers may differ depending on the recording conditions and thus may be distinguishable based on spectral features.

The number of microphones included in the present calculations was constrained to unobstructed microphones that were in line with the speaker box yielding a maximal time delay between subsequent microphones. Furthermore, we did not report the results of a third operating type of microphone included in the corpus, so called MEM microphones. Initial results with MEM microphone recordings used in the corpus could not be related back to the approximated ground truth values and we therefore did not include them in the present paper. This was confirmed by written correspondence with one of the authors of the VOICES corpus stating that some of the MEM microphones had a short delay prior to the signal output.

Finally, the MCCC algorithm was introduced as a channel selection method with suggested decreased computational complexity compared to ASR based channel selection approaches [6]. The applicability in real-time settings and computational efficiency of this method when combined with a heuristic for reference channel selection still remains to be explored.

5. Acknowledgements

This research was funded by the Federal Ministry of Education and Research (Germany).

6. References

- [1] H. Riaz, M. Stiles, C. Armstrong, A. Chadwick, H. Lee, G. Kearney, Multichannel microphone array recording for popular music production in virtual reality, in: Proceedings of the AES 143rd Convention, New York, NY, 2017, Article Number: eBrief384.
- [2] A. Farina, A. Capra, L. Chiesi, L. Scopece, A Spherical Microphone Array for Synthesizing Virtual Directive Microphones in Live Broadcasting and in Post Production, in: Proceedings of the AES 40th Conference, Tokyo, Japan, 2010.
- [3] R. Martin, Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction, in: M. Brandstein, D. Ward (eds), *Microphone Arrays, Digital Signal Processing*, Springer, Berlin, Heidelberg, 2001, pp. 255–279. https://doi.org/10.1007/978-3-662-04619-7_12.
- [4] N. K. Chaudhary, S. Verma, A. Aditya, Sound Source Localization using GCC-PHAT with TDOA Estimation, *Journal of Basic and Applied Engineering Research*, 1.11 (2014): 54–58. Online ISSN: 2350-0255.
- [5] R. Lee, M. Kang, B. Kim, K. Park, S. Q. Lee and H. Park, Sound Source Localization Based on GCC-PHAT With Diffuseness Mask in Noisy and Reverberant Environments, *IEEE Access* 8 (2020) 7373–7382. doi: 10.1109/ACCESS.2019.2963768.
- [6] K. Kumatani, J. McDonough, J. F. Lehman, B. Raj, Channel selection based on multichannel cross-correlation coefficients for distant speech recognition, in: *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Edinburgh, UK, 2011, pp.1-6. doi: 10.1109/HSCMA.2011.5942398.
- [7] Z.-Q. Wang, D. Wang, All-Neural Multi-Channel Speech Enhancement, in: *Proc. Interspeech 2018*, Hyderabad, 2018, pp. 3234-3238. doi: 10.21437/Interspeech.2018-1664
- [8] C. Knapp, G. Carter, The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24.4 (1976): 320-327. doi: 10.1109/TASSP.1976.1162830.
- [9] W. Li, Y. Zhang, P. Zhang, and F. Ge, Multichannel ASR with Knowledge Distillation and Generalized Cross Correlation Feature, in: *2018 IEEE Spoken Language Technology Workshop*, Athens, Greece, 2018, pp.463-469.
- [10] C. Zhang, D. Florencio, and Z. Zhang, Why does PHAT work well in lownoise, reverberative environments?, in: *IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, NV, 2008, pp. 2565-2568. doi: 10.1109/ICASSP.2008.4518172.

- [11] M. Cobos, F. Antonacci, L. Comanducci, A. Sarti, Frequency-Sliding Generalized Cross-Correlation: A Sub-Band Time Delay Estimation Approach, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 (2020) 1270–1281. doi: 10.1109/TASLP.2020.2983589.
- [12] J. Dennis, T. H. Dat, Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASPIRE challenge, in: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Arizona, USA, 2015, pp. 518-524.
- [13] M. Wölfel, C. Fuegen, S. Ikbal, J. McDonough, Multi-source far-distance microphone selection and combination for automatic transcription of lectures, in: *Proc. Interspeech 2006*, Pittsburgh, USA, 2006, paper 1253-Mon2BuP5.
- [14] M. Wölfel, Channel selection by class separability measures for automatic transcriptions on distant microphones, in: *Proc. Interspeech, 2007*, Antwerp, Belgium, 2007, pp. 582-585.
- [15] C. G. Flores, G. Tryfou, M. Omologo, Cepstral distance based channel selection for distant speech recognition, *Computer Speech & Language*, 47 (2018) 314–332. doi: 10.1016/j.csl.2017.08.003.
- [16] F. Xiong, J. Zhang, B. Meyer, H. Christensen, J. Barker, Channel Selection using Neural Network Posterior Probability for Speech Recognition with Distributed Microphone Arrays in Everyday Environments, in: *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, Hyderabad, 2018, 19-24. doi: 10.21437/CHiME.2018-5.
- [17] T. Ochiai, S. Watanabe, T. Hori, J. Hershey, Multichannel End-to-end Speech Recognition, 2017. URL: <https://www.merl.com/publications/docs/TR2017-035.pdf>.
- [18] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, I. Tashev, Microphone array processing for distant speech recognition: Towards real-world deployment, in: *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, 2012*, pp. 1-10.
- [19] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, 19.7 (2011): 2125–2136.
- [20] C. Richey, M.A. Barrios, Z. Armstrong., C. Bartels, H. Franco, M. Graciarena, A. Lawson, M.K. Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, K. Ni, Voices Obscured in Complex Environmental Settings (VOiCES) Corpus, in: *Proc. Interspeech 2018*, Hyderabad, 2018, 1566-1570. doi: 10.21437/Interspeech.2018-1454.
- [21] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, Multi-channel Attention for End-to-End Speech Recognition, in: *Proc. Interspeech 2018*, Hyderabad, 2018, 17-21, doi: 10.21437/Interspeech.2018-1301.

7. Supplementary Material

Table S1

Channel rankings based on MCCC for condition with heuristic selection of reference channel for alignment. Channels are sorted upwards from worst quality to the two best quality channels based on MCCC. Columns 6 and 7 denote the two channels that were ranked as best quality channels based on the MCCC algorithm.

Channel ranking: Heuristic Reference Channel								Channel Alignment
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Reference Channel
Sample 1	16	4	3	6	5	1	2	2
Sample 2	16	5	6	4	3	1	2	2
Sample 3	3	4	16	6	5	1	2	2
Sample 4	16	4	3	6	5	1	2	2
Sample 5	16	4	3	6	5	1	2	2
Sample 6	16	4	3	6	5	1	2	2
Sample 7	16	6	5	4	3	1	2	2
Sample 8	16	5	6	4	3	1	2	2
Sample 9	16	4	3	6	5	1	2	2
Sample 10	3	4	16	6	5	1	2	2

Table S2

Channel rankings based on MCCC for condition with random selection of reference channel for alignment. Channels are sorted upwards from worst quality to the two best quality channels based on MCCC. Columns 6 and 7 denote the two channels that were ranked as best quality channels based on the MCCC algorithm.

Channel ranking: Random Reference Channel								Channel Alignment
	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Reference Channel
Sample 1	4	2	3	16	1	5	6	6
Sample 2	5	6	16	4	3	1	2	1
Sample 3	2	16	1	6	5	3	4	5
Sample 4	4	2	3	16	1	5	6	6
Sample 5	2	6	5	16	1	3	4	4
Sample 6	16	4	3	6	5	1	2	2
Sample 7	16	6	5	4	3	1	2	2
Sample 8	2	16	1	4	3	5	6	6
Sample 9	16	4	3	6	5	1	2	2
Sample 10	6	5	16	4	3	1	2	1

Chapter 3

E-Healthcare and Smart Diagnostics

A Robust Martingale Approach for Detecting Abnormalities in Human Heartbeat Rhythm

Jonathan Etumusei^a, Jorge Martinez Carracedo^a and Sally McClean^a

^aUlster University, Jordanstown, United Kingdom

Abstract

The analysis of electrocardiogram data is vital to the healthcare system to improve and monitor health conditions. Existing algorithms are effective in discovering abnormalities in electrocardiogram data streams but most of these approaches do not focus on the intensity and duration of these anomalies. In this paper, we propose a new method called alignment of the martingale sequence (AMS) that improves previous approaches using dynamic time warping and particle swarm optimisation to obtain the optimal parameter that maximises F1. Our proposed method can also estimate the severity and extent of an abnormal heartbeat rate. Experimental results show that the proposed approach makes some improvements over the traditional method.

Keywords

Heart rate, dynamic time warping, ECG sequence, martingales

1. Introduction

Persistent heart failure (PHF) is a dynamic, crippling condition that can lead to cardiac disorders and hospitalisation. Common symptoms of PHF include fatigue, shortness of breath and peripheral oedema. These symptoms can cause several effects to human health and disruption in daily life activities [1]. The predicaments caused due to PHF have not only negatively impacted patients and their loved ones but also the health care system and society generally. PHF diagnosis is around 2% of the general population in developed nations [2]. The British heart foundation reckons that heart failure affects around 2% of the UK population [3]. The ageing demographic, in the developed world, are more inclined to be affected by this disease. HF increases from 1% among those within the age group of 45 – 55 years old to over 5% in the

CERC 2021: Collaborative European Research Conference, September 09-10, 2021, Cork, Ireland

✉ etumusei-j@ulster.ac.uk (J. Etumusei); j.martinez-carracedo@ulster.ac.uk (J.M. Carracedo);
si.mcclean@ulster.ac.uk (S. McClean)

 0000-0002-1337-1471 (J. Etumusei); 0000-0001-8017-2598 (J.M. Carracedo); 0000-0002-6871-3504 (S. McClean)

 © 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

age group from 80 years old above [4].

It is assumed that heartbeat rate plays a key role in the risk of a heart attack. Heart diseases such as PHF, coronary heart disease, congenital heart disease and congestive heart failure are the main cause of mortality for men and women in many countries [5]. PHF patients will need to constantly monitor their heart rate for a sign of irregular behaviour which might be a sign of potential heart failure. There are several ways by which we can handle or manage the threat of PHF as stipulated in the guidelines from the European Society of Cardiology [6]. These procedures include:

- Monitoring of symptoms associated with PHF.
- Self-management of multiple chronic illnesses.
- Educating patients to observe their health conditions and be able to identify illness seriousness.
- Consistent exercises and physical activities.
- Consistent monitoring of the heartbeat for irregular signs.

The recent evolution of microelectronics and sensor technology has led to the development of many wireless sensor applications which can measure the heartbeat for signs of cardiovascular diseases [6]. The heart's rhythm can be measured using signals that are recorded by specialised devices to identify the normal functioning of the heart through heartbeat or heart rate. Heart rate is the number of periods the heart beats per minute while heartbeat is one complete pulsation of the heart. A normal heart rate for resting adults is within the range of 60 to 100 beats per minute [7]. A lesser heart rate suggests a more dynamic heart functionality.

Arrhythmia is an abnormal heart rate or rhythm that happens when electric impulses that originate from hearth beats do not function properly. Arrhythmia might cause concern and can be life-threatening exhibiting symptoms like shortness of breath, palpitations, fatigue, feeling dizzy, fainting. The most efficient avenue to diagnose an arrhythmia is through an electrical recording of the heart rhythm called an electrocardiogram (ECG)[8]. The use of ECG time series to identify heartbeat makes it possible to intervene in the situation of PHF. Abnormal heart rate intensity (AHI) is the extent or degree to which the heart rate becomes too slow or fast in an uncertain way. Abnormal heart rate duration (AHD) is the time or period when the heart becomes too slow or fast in an uncertain way.

An ECG is a recorded signal that can be used to check your heart's rhythm and electrical activity for the diagnosis process. ECGs are one of the primary diagnostic tests to detect cardiovascular abnormalities. Using ECG, it is also possible to estimate the dimension and position of the heart chambers to discover any form of damage in the heart [9].

An anomaly can be informally defined as anything that deviates from a normal standard.

Mathematically, an anomaly is often perceived as a point that is far away from the mean of a sequence. Anomaly detection is the discovery of irregularities that are different from the rest of the time series such as ECG data. ECG time series consist of real quasi-periodic signals and current approaches will need to learn about the relationship of the sequence before they can discover anomalies [10]. Identifying abnormalities in the ECG time series is very crucial in the medical and health area.

Anomaly detection in ECG can be very challenging as a result of heartbeat variation in patients. Consequently, some heart rate variations might not be life-threatening and this can produce misleading information that can affect the interpretation of ECG readings. This situation motivates us to propose a method that uses machine learning techniques to detect severe heart rates for quick intervention to save lives.

Recently, diverse methods have been created for analysing ECG signals. However, the complexity of these techniques has limited the performance of identifying anomalies in heart rate [11]. Most of these techniques can detect irregularities in heart rate but are also unable to completely isolate noise interference in ECG signal [12]. This situation gives rise to a false alarm rate (FAR). In this work, we are developing an algorithm that will discover abnormalities in ECG signals. Unlike most successful change detection approaches, our proposed approach can detect change intensity and duration in heart rate that occurs in ECG sequence. Our method uses the martingale frame to reduce the impact of noise interference in the ECG data set. The rationale behind our suggested technique is that it allows PHF patients and medical staff to monitor their heart rate intensity and abnormal duration to diagnose early signs of arrhythmia. Also, our method can be useful in measuring the heart rate to evaluate effort between several exercises or workout sessions.

To handle any challenges of similarity measures of ECG signal points produced by our algorithm, we implement dynamic time warping (DTW). DTW [13] is a popular technique that locates the optimal alignment between two sequences under certain conditions. The optimisation and DTW concepts are further discussed in Section 3 respectively.

To obtain the optimal parameters that improve the performance of the algorithm, we use particle swarm optimisation (PSO). PSO [14] is a stochastic optimisation technique that is motivated by the intuitive mutual (swarm) behaviour of animals such as a swarm of bees, a flock of birds and schools of fish. We use the PSO approach to identify the parameter values that maximise $F1$. The method explores or searches simultaneously through a group of individuals or particles to obtain the optimum value in a swarm whose trajectories are modified by stochastic and a deterministic component [15]. For this study, we use PSO rather than genetic algorithm (GA) [16] due to the following reasons:

- PSO can be adjusted to handle complex problems.
- PSO is computationally more efficient concerning speed and memory requirements [17].

PSO will be further explained in Section 3. In this work, we use the PSO to obtain the optimal

fitness function value using F1. We use F1 instead of accuracy as it takes into consideration both false negatives and false positives. F1 is a better metric to evaluate sequences where imbalance classes exist. We benchmark our proposed technique with traditional methods and obtain competitive prediction outcomes.

The paper structure is as follows: In section 2, we review the latest work done on identifying changes in ECG data. In Section 3, we introduce our novel approaches. In section 4, we show our experimental results and compare them with the existing Martingale algorithm. We finish the paper in section 5 discussing the results that we got and the next steps that we will take in the research.

2. Related work

In the last decades, new change detection techniques have been developed to discover transitions in a human heartbeat using ECG data. For instance, Varon et al. [18] proposed a methodology for the instinctive discovery of sleep apnea from an ECG sequence. The approach uses two novel well-known features common in heart variability analysis: standard deviation and serial correlation coefficients of the interval between heartbeats. The first feature utilises the main components of QRS complexes (the spread of impulses through the ventricles of the heart) that represent abnormalities in their structure as a result of increased sympathetic activity during sleep apnea conditions. The second novel feature captures the information distributed between the respiration system and heart rate using orthogonal subspace projections. The respiratory information is obtained using the ECG signal through three robust algorithms. The features use the radial basis function (RBF) kernel implemented as input to the least-square support machine classifier. Two independent ECG data sets which include hypopneas and apnea points were analysed. The algorithm can achieve a comparable result of 100% accuracy rate in classifying sleep apnea and also able to determine the contamination level of each ECG timing.

The rise in electronic medical observation and sensors applications such as electrocardiograms are becoming available as a result of the big data revolution. However, most of these signal recorded remains unlabelled thereby making anomaly detection challenging. This situation motivates Pereira et al. [19] to introduce an unsupervised method that uses a technique to learn about the features of the ECG sequence to discover any abnormality using numerous detection strategies. Experimental result shows that the suggested method can learn demonstrative representations of ECG time series to discover divergence with scores that outperform conventional supervised and unsupervised approaches respectively.

High false alarm rates (FAR) occur in ECG signals as a result of the inability to distinguish between actual ECG signals and ECG artefacts (electromagnetic alterations that are unrelated to cardiac impulse activities) as both signals are similar in terms of structure and frequency. These characteristics lead to a misconception of ECG readings. Sivaraks et al. [20] were motivated by this fact to propose a robust approach that can discover abnormalities while minimising FAR

in ECG data. The method design takes into consideration the cardiologist and motif identification approaches. Every step of the algorithm complies with the review of a cardiologist. The approach can make use of both single-lead and multi-lead ECGs respectively. Experimental results show that the algorithm can achieve 100% on accuracy on detection, specificity, sensitivity and positive predictive with 0% FAR. The outcome depicts that the suggested method performs better compared to conventional anomaly detection techniques.

Conventional change point detection approaches can discover changes in electromagnetic (EM) signals but are often limited by the issue of noise interference. This situation motivates Etumusei et al. [21] to propose two approaches that utilise the martingale framework to discover abnormalities in EM signals. The methods can isolate noise and makes use of cross-validation to optimise its parameters. Experimental result shows the proposed algorithm makes improvements over the previous technique within the martingale framework.

In this paper, we present a novel method based on enhancing the performance of our previous work [21] known as the moving median of the martingale sequence. The proposed method is known as the alignment of the moving median of the martingale sequence (AMS). The method improves the previous moving median of the martingale sequence (MMMS) by applying time warping and optimisation techniques to enhance performance for discovering abnormality in ECG time series. This technique is applied to analyse ECG sequences. Our unsupervised method uses previous methods such as randomised power martingale to distinguish between normal and anomalous data points by learning the ECG sequences. To make this anomalous data point outstanding, our method uses the moving median approach to isolate ECG artefacts (noise). We apply dynamic time warping (DTW) to align any displaced points and PSO to optimise the algorithm parameters. We compare the suggested approach results with the original methods called the randomised power martingale (RPM) and running average of the martingale sequence (RAS) respectively. These methods will be discussed in detail in Section 3. We have summarised the proposed system in Figure 1. The next paragraph explains the heart rate model.

3. Heart rate Model

This Section discusses the pre-processing approach used and the model for discovering anomalies in ECG sequence.

3.1. Data pre-processing

The first step to implement the proposed algorithm involves obtaining accelerometry data from a Shimmer wireless sensor platform (SWSP) [22] attached to healthy participants. The participants perform activities in different scenarios within a home environment (for more details about this data set see [23]). For each scenario, ECG and accelerometer signals were captured

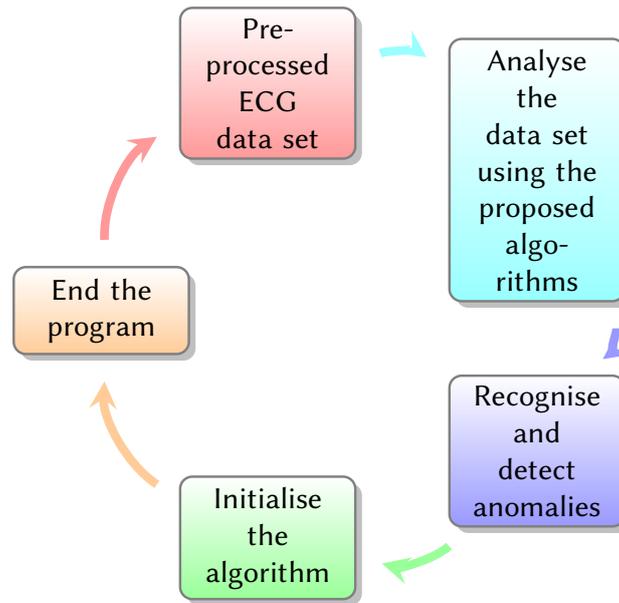


Figure 1: Proposed system

before, during and after each activity of the participants. The data obtained are streamed to a computer via the IEEE 802.15.1 Bluetooth communication protocol using the BioMOBIUS windows-based application development platform[23][24]. To process the ECG data captured, a fast Fourier transform (FFT) [25][23] is used to transform the ECG signal to determine its frequency components. Secondly, we use R-peak filtering techniques to remove and filter low-frequency noise. Finally, the average per interval heart rate for every activity is computed obtained after the filtering process. The labelled ECG data obtained from the sit to stand scenario can be seen in Figure 5 which also shows the changes, their duration and intensity.

3.2. Martingale technique

The following techniques aim to discover abnormal changes, change intensity and duration in ECG time series. In the next paragraphs, we shall focus on the martingale concept.

A martingale is a succession of a stochastic process, for which, at a specific time, the conditional expectation of the next value given all previous points is equal to the present value.

Definition 1: [26] A sequence of random variables $\{M_i : 0 \leq i < \infty\}$ is a martingale regarding the sequence of random variables $\{X_i : 0 \leq i < \infty\}$, if for all $i \geq 0$, the following conditions hold:

- The martingale M_i is a function that is measurable of X_0, X_1, \dots, X_i ,
- $E(|M_i|) < \infty$ and

- $E(M_{n+1}|X_0, \dots, X_n) = M_n$.

Ho and Wechsler [26] suggested a fundamental unit of the martingale framework by defining a metric called strangeness. Strangeness measures how much a new data point diverges from the previous one in a time series.

Let us consider a sequence $Z = \{z_1, \dots, z_{i-1}\}$, where there is a newly recorded point z_i . Let us also consider that the data points in Z have been clustered into k disjoint sets Y_1, \dots, Y_k , ($k \leq i - 1$) [27].

Definition 2: The strangeness of z_i is defined as

$$s_i = s(Z, z_i) = \| z_i - C_r \| , \quad (1)$$

where C_r is the centroid of the cluster Y_r , for some $r \in \{1, \dots, k\}$ such that $z_i \in Y_r$. $\| \cdot \|$ denotes the chosen distance.

The strangeness of z_i is used to compute a "probability" time series where its points are named \hat{p}_i . If for $j = 1, 2, \dots, i$, s_j is the strangeness of z_j and θ_i is a fixed value in $[0, 1]$ [26][28], \hat{p}_i is computed as follows:

$$\hat{p}_i(Z \cup z_i, \theta_i) = \frac{\#\{j : s_j > s_i\} + \theta_i \#\{j : s_j = s_i\}}{i}. \quad (2)$$

Intuitively, \hat{p}_i measures the probability of being more estranged than z_i . It should be noted that \hat{p}_i can be seen as an unusual case of the statistical notion of p-value [26]. The set of \hat{p}_i can be used to compute a new random variable that will create a new sequence known as the randomised power martingale.

Definition 3: [26] The randomised power martingale (RPM) is indexed by $\epsilon \in [0, 1]$ defined at each time-point as

$$M_n^{(\epsilon)} = \prod_{i=1}^n (\epsilon \hat{p}_i^{\epsilon-1}). \quad (3)$$

Fixed an $\epsilon \in [0, 1]$, once we have computed $M_n^{(\epsilon)}$, the model will detect a change in the $n - th$ timepoint if

$$M_n^{(\epsilon)} > t, \quad (4)$$

where the threshold t is chosen in a probabilistic way based on Dobb's Inequality [26].

In the following section, we introduce an approach that aims to improve the overall performance of the previously described martingale approach.

3.3. Moving Median of a martingale sequence (MMMS)

A moving median (MM) is a robust and effective smoothing technique to detect a transition in a data stream [29]. The moving median computes the median of a sequence using a sliding window. For our method, once we have computed the martingale sequence, we apply MM over that sequence to smooth the martingale points. We use median, rather than a mean for the analysis of ECG time series because it is more robust against extreme values as it is not determined by the individual points of the ECG sequence, but only by their order. This behaviour suggests that the median tends to smooth the time series, thereby, reducing the effect of noise.

Let us consider a martingale succession $\mathcal{M} = \{M_i : 0 \leq i < \infty\}$ and fix a window length $l > 0$. We define D_k as the MM of the k -th window in martingale sequence \mathcal{M} . Although we could use this technique to look for anomalies in the data, we will refine the sequence D_k before using it for that purpose. In the following Section, we will describe dynamic time warping.

3.4. Moving Average of the Martingale Sequence (MAS)

For our baseline method, we implement the moving average technique on the martingale sequence. The moving average computes the mean of a sequence using a sliding window. The moving average is given as;

$$MA_n = \frac{M_{n-k+1} + M_{n-k+2} + \dots + M_n}{k}, \quad (5)$$

$$MA_n = \frac{1}{k} \sum_{i=n-k+1}^n M_i, \quad (6)$$

where M_i is the martingale point and n is the length of the data sequence. In a later stage, we will apply PSO to the MA_n sequence to obtain the optimal parameter that maximises F1.

3.5. Dynamic time warping (DTW)

Heart rate undergoes minor changes between successive heartbeats and thereby produces a linear functional to ECG readings. One way of handling this challenge is to use dynamic time warping to specify a nonlinear heart rhythm [30][31]. Dynamic time warping (DTW) measures the affinity between the original ECG data set and our proposed algorithm output. DTW discovers the minimum path by producing a non-linear alignment between the two sequences[32]. DTW computes the optimal match between the two sequences with certain rules and conditions:

- Each one of the indexes from the initial sequence must correspond with one or more indices of the other succession and vice versa
- The first index of the initial time series must match with the first index of the other sequence
- The last index from the first time series must be similar to the last index of the other succession
- The aligning of the indices from the first sequence to indices from the other sequence must always be increasing and not becoming constant or decreasing vice versa. For instance, if $j > i$ are indices obtained from the first time-series then there should never be any two indices $l > N$ in such a way that i corresponds with index l and index j is matched to index N and vice versa

Given two sequences X and Y , we will say that each tuple (i,j) is the alignment between $X[i]$ and $Y[j]$. We define the mapping path DS as the map that minimises the distance between the sequences X and Y . We implement DTW on $MMMS$ and the original ECG sequence to obtain a new succession known as the aligned moving median of the martingale sequence (AMS). The process is repeated using the MAS and the original ECG sequence to obtain a new sequence known as the aligned moving average of the martingale sequence ($AMAS$) respectively. In the next Section, we discuss the implementation of the PSO on our new sequences.

3.6. PSO optimisation

PSO algorithm is an optimisation technique that uses a search process based on swarm exploration. In this type of exploration, each individual retains the optimal location in the swarm. For each generation, the information accumulated by the particle is then used to adapt the new location of the particle. The particles are constantly evolving in a multi-dimensional search capacity until an optimal condition is found. Each particle adjusts its position depending on its present velocity, its preceding best location (P_{best}) and the global best location (G_{best}) of the whole swarm.

To tune and explore the direction of the swarm, the velocity and location of the particles at iteration k are accomplished using the following steps:

1. Particles are initialised with arbitrary location and velocities according to the search range or space
2. Estimation of each particle using the fitness function
3. Particles are updated with individual best and global best fitness values and their location
4. The candidate solution's location and velocity are renewed
5. If the convergence criterion is satisfied then the algorithm is halted and the final solution output is presented otherwise the process will progress to step 2

Table 1
PSO component values

Parameters	value
InertiaRange	[0.10000, 1.1000]
InitialSwarmSpan	200
MaxIterations	200 * NumberOfVariables
MaxStallIterations	20
MinNeighboursFraction	0.250
SwarmSize:	100
SelfAdjustmentWeight	1.4900
SocialAdjustmentWeights	1.4900

PSO locates the optimal parameters (ϵ , window size) using fitness function (FF) to maximize F1. The fitness function is given as:

$$F1_{max} = \max_{(\epsilon, window\ size)}(F1_{(AMS)}), \quad (7)$$

where ϵ ranges from 0 to 1 and window size from 2 to 20 for each activity.

PSO components implemented to maximise the fitness function are shown in Table 1. PSO implementations on the methods RPM, AMS, AMAS will be called RPM(PSO), AMS(PSO) and AMAS(POS) respectively. Furthermore, the proposed approach is illustrated in Figure 2.

3.7. Threshold computation

While Ho and Weschler [26] proposed a probabilistic way of computing threshold, we suggest a threshold based on the median absolute deviation (MAD) of the martingale sequence. MAD is a robust technique for analysing ECG time series because it measures the variability of the univariate ECG points. For a univariate sequence, $S = \{S_1, S_2, \dots, S_n\}$ (in our case, we use the new PSO sequences) MAD is the median of the absolute deviations of the sequence. MAD is given as follows:

$$MAD(S) = \text{median}(\{|S_i - \tilde{S}|, i = 1, \dots, n\}), \quad (8)$$

where $\tilde{S} = \text{median}(S)$. MAD shows how spread out the data is. Ley et al. [33] proposed, based on outlier detection, a threshold for change detection of $ME \pm MeAD$, where ME is the mean of the data points and $MeAD$ is the mean absolute deviation. We used a similar approach using the median to compute a threshold t for our methods. Therefore, this model will detect a change when :

$$H_k \geq t. \quad (9)$$

```

Data: Input (F): ECG univariate data set
Result: Output: AMS(PSO) points
1 Initialise:  $M(0) = 1; i = 1; F = \{ \}$ ;
2 Set values for cluster group  $k$ ,  $\epsilon$  value, window size;
3 while do
4   A new example of normalised  $z_i$  is discovered;
5   if  $F \neq \{ \}$  then
6     Set the strangeness of  $z_i := 0$ 
7   else
8     Compute the strangeness of  $z_i$  and the data points in  $F$ 
9     Compute the  $\hat{p}_i$  of  $z_i$ ;
10    Compute the RPM points using equation (4);
11    Compute the MMMS points;
12    Compute the AMS points;
13    Compute the AMS (PSO);
14    Compute the threshold  $t$ 
15  end
16  if  $MMS \geq t$  then
17    Discover abnormalities
18    Estimate the AHI
19    Estimate the AHD
20    Re-initiate  $M_i = 1$ 
21  else
22    Add  $z_i$  into  $F$ ;
23  end
24  if  $i = i + 1$ ; then
25  end
26 end

```

Figure 2: The algorithm

where H_k can represent RPM(PSO), AMS(PSO) and AMAS(PSO). If H_k exceeds the given threshold t , a change has been detected. When the analysis of data of D_k is finalised, the algorithm is restarted.

The evaluation performance for the approaches is measured using robust evaluation metrics (EM)[34] such as accuracy, precision, recall(sensitivity), harmonic mean (F_1).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} * 100, \quad Precision = \frac{TP}{TP + FP} * 100,$$

$$Recall = \frac{TP}{TP + FN} * 100, \quad F1score = \frac{2 * Recall * Precision}{Recall + Precision} * 100,$$

$$Specificity = \frac{TN}{TN + FP} * 100,$$

where TP, TN, FP, FN are true positive, true negative, false positive and false negative respectively. The performance metric provides a proper estimation of the suggested method, especially on imbalanced time series [35].

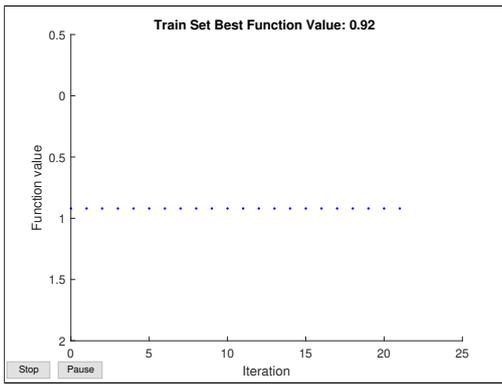
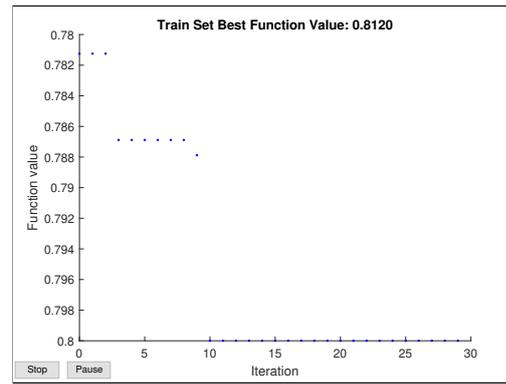

Figure 3: PSO iteration using AMS method

Figure 4: PSO iteration using AMAS method

Table 2

Confusion matrix using optimised parameters for training set

Approach	Training set	ϵ	window size	TP	TN	FP	FN
RPM(PSO)	N1	0.9	-	10.0(63.0%)	35.0(83.3%)	17.0(63.0%)	7.0(16.7%)
AMAS(PSO)	N1	0.9285	19.0	26.0(96.3%)	30.0(71.4%)	1.0(3.7%)	12(28.6%)
AMS(PSO)	N1	0.5741	19.0	23.0(85.2%)	42.0(100.0%)	04(14.8%)	0.0(0.0%)

3.8. Abnormal heart rate intensity and duration

Our proposed method can measure the abnormal heart rate intensity (AHI) and abnormal heart rate duration (AHD). To compute the AHI and AHD, we first find the threshold and then subtract it from the highest algorithm output point. The HAI is given as:

$$AHI = M - T, \quad (10)$$

where $M = \max\{H_k \mid k = 1, \dots, n\}$, being H_k the output time series of the used algorithm and T as the used threshold. Furthermore, we can also compute the duration of the abnormality in heart rate data by computing the time (sec) of the changes. AHD is the length of time in seconds, the changes take place. AHD is given as:

$$AHD = \text{Seconds}(TP), \quad (11)$$

In the next Section, we shall discuss the experimental results of our proposed algorithm.

4. Experimental results

The section exposes an overview of the different approaches adopted to detect anomalies in the ECG data set. The ECG data used for this experiment was discussed in Section 3.

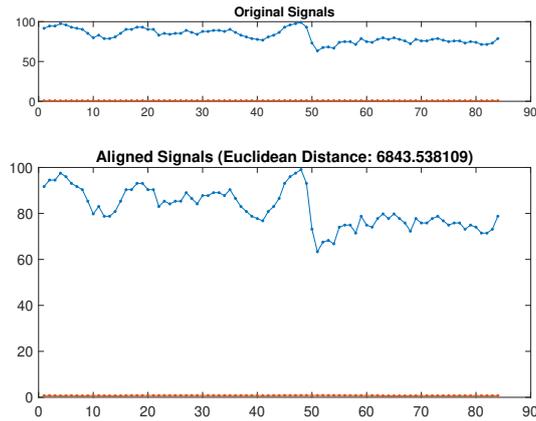


Figure 5: Original and aligned Signals

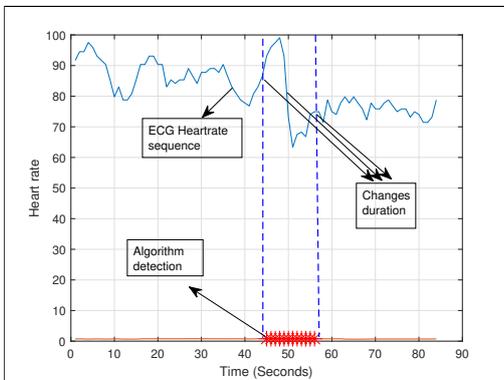


Figure 6: Test data change detection

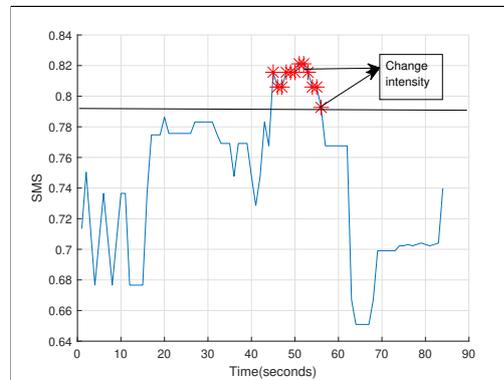


Figure 7: AMS(PSO) output

Table 3

Confusion matrix using optimised parameters for test set

Approach	Test set	ϵ	window size	TP	TN	FP	FN
RPM(PSO)	N2	0.9	-	05(45.5%)	62(84.9%)	06(54.5%)	11(15.1%)
AMAS(PSO)	N2	0.9285	19.0	11(100.0%)	53(72.6%)	00(0.0%)	20(27.4%)
AMS(PSO)	N2	0.5741	19.0	11(100.0%)	72(98.6%)	00(0.0%)	01(1.4%)

4.1. Cross validation technique

To show the rational potential, we evaluate the proposed method using different training ($N1$) and test set ($N2$). Both datasets are obtained from an individual performing similar activities. The dataset captured the participant heart rates. Our objective is to detect when there is a change in these heart rates and the actual duration of this change.

We implement our proposed method to obtain the optimal parameters. In this specific work, we used ϵ and the window size as the parameters to be optimised. This technique is used to

Table 4
Evaluation metrics of the proposed and previous approaches

Approaches	N1	N2	ϵ	window size	Sensitivity(%)	Specificity(%)	Accuracy(%)	Precision(%)	F1-Score(%)
RPM(PSO)	70	84	0.9	-	41.3[37.04, 45.5]	84.1[83.3, 84.9]	72.5[65.2, 79.8]	45.1[58.8, 31.3]	41.3[45.5, 37.0]
AMAS(PSO)	70	84	0.9285	19.0	97.7[96.3, 99.0]	72.0[71.4, 72.6]	78.7[81.2, 76.2]	52.0[68.4, 35.5]	66.0[80.0, 52.0]
AMS(PSO)	70	84	0.5741	19.0	92.1[85.2, 99.0]	98.8[99.0, 98.6]	96.6[94.2, 98.9]	95.4[99.0, 91.7]	93.9[92.0, 95.7]

Table 5
Estimation of algorithm performance

Approach	Ave. iteration time	AHI	AHD
RPM(PSO)	0.3801	0.2716 RPM	5(sec)
AMAS(PSO)	0.5236	0.0349 RAS	11(sec)
AMS(PSO)	0.4329	1.5275 AMS	11(sec)

Table 6
Evaluation performance comparison

Approach	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1(%)
RPM(PSO)	72.5	41.3	84.1	45.1	41.3
AMAS(PSO)	78.7	97.1	72.0	52.0	66.0
AMS(PSO)	96.6	92.1	98.8	95.4	93.9

analyse both the previous and newly proposed algorithms, that is RPM, AMS, AMAS.

Once we found optimal parameters for our data set N1, we used that configuration to compute the evaluation metrics for N2. Both confusion matrices can be found in TABLE 2 and 3. In addition, the information from the confusion matrices is then used to compute the performance metrics. This information can be seen in TABLE 4.

4.2. Performance of ECG detection algorithm

Our proposed algorithm (TABLE 4) produces better results compared to conventional approaches such as RPM(PSO) and AMAS(PSO) respectively. These comparison are summarised (TABLE 6) for better evaluation. The performance comparison (TABLE 6) shows that the suggested approach gives an accuracy rate of over 15% compared that of AMAS(PSO) and RPM(PSO) independently. Also, our proposed approach gives a specificity of over 10% compared to the AMAS(PSO) and RPM(PSO) independently. Overall our suggested technique produces a preferable output of over 40% in terms of precision and F1-score. However, the AMAS(PSO) method is slightly sensitive compared to our proposed approach. This can be attributed to slightly higher TP detected by the AMAS(PSO). This might not be a major issue at the moment, but we aim to address it further in future work. Our proposed algorithm (TABLE 5) also produces a better AHI (1.5275 AMS) and a lower average iteration run time of 0.4329 seconds. The AHD time for our proposed AMS method is 11 seconds greater than that of the RPM method.

5. Conclusion and future work

This paper discusses persistent heart failure and its impacts on humanity. This predicament inspires us to introduce a method (that uses the martingale framework) to detect abnormality in heart rate measuring devices such as ECG. Experimental results show that our proposed technique outperforms previous martingale approaches. Furthermore, the proposed algorithm can measure the heart rate intensity and duration of abnormality in ECG sequence. Future work is required to confirm this hypothesis using big data streams in specific populations areas and age categories especially the elderly who are more prone to heart failure and attack.

Acknowledgments

This work was supported by a University of Ulster Vice-Chancellor's Research Studentship. The authors would like to thank anonymous reviewers for their constructive suggestions.

References

- [1] N. Albert, K. Trochelman, J. Li, S. Lin, Signs and symptoms of heart failure: are you asking the right questions?, *American Journal of Critical Care* 19 (2010) 443–452.
- [2] A. Bundkirchen, R. H. Schwinger, Epidemiology and economic burden of chronic heart failure, *European Heart Journal Supplements* 6 (2004) D57–D60.
- [3] K. Sutherland, Bridging the quality gap: heart failure, Health Foundation, 2010.
- [4] A. J. S. Coats, Ageing, demographics, and heart failure, *European Heart Journal Supplements* 21 (2019) L4–L7.
- [5] W. M. Jubadi, S. F. A. M. Sahak, Heartbeat monitoring alert via sms, in: 2009 IEEE Symposium on Industrial Electronics & Applications, volume 1, IEEE, 2009, pp. 1–5.
- [6] K. Swedberg, J. Cleland, H. Dargie, H. Drexler, F. Follath, M. Komajda, L. Tavazzi, O. A. Smiseth, A. Gavazzi, A. Haverich, et al., Guidelines for the diagnosis and treatment of chronic heart failure: executive summary (update 2005) the task force for the diagnosis and treatment of chronic heart failure of the european society of cardiology, *European heart journal* 26 (2005) 1115–1140.
- [7] Y. Ostchega, Resting pulse rate reference data for children, adolescents, and adults: United States, 1999-2008, 41, US Department of Health and Human Services, Centers for Disease Control and . . . , 2011.
- [8] R. D. White, G. Flaker, Smartphone-based arrhythmia detection: Should we encourage patients to use the ecg in their pocket?, *Journal of atrial fibrillation* 9 (2017).
- [9] S. Chauhan, L. Vig, Anomaly detection in ecg time signals via deep long short-term memory networks, in: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2015, pp. 1–7.
- [10] M. Thill, S. Däubener, W. Konen, T. Bäck, P. Barancikova, M. Holena, T. Horvat, M. Pleva,

- R. Rosa, Anomaly detection in electrocardiogram readings with stacked lstm networks, in: Proceedings of the 19th Conference Information Technologies-Applications and Theory (ITAT 2019), CEUR-WS, 2019, pp. 17–25.
- [11] M. I. Owis, A. H. Abou-Zied, A.-B. Youssef, Y. M. Kadah, Study of features based on nonlinear dynamical modeling in ecg arrhythmia detection and classification, *IEEE transactions on Biomedical Engineering* 49 (2002) 733–736.
- [12] G. D. Clifford, F. Azuaje, P. Mcsharry, Ecg statistics, noise, artifacts, and missing data, *Advanced methods and tools for ECG data analysis* 6 (2006) 18.
- [13] D. J. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series., in: *KDD workshop*, volume 10, Seattle, WA, USA., 1994, pp. 359–370.
- [14] D. Wang, D. Tan, L. Liu, Particle swarm optimization algorithm: an overview, *Soft Computing* 22 (2018) 387–408.
- [15] H. Lu, J. Chen, L. Guo, *Energy quality management* (2018).
- [16] S. Shabir, R. Singla, A comparative study of genetic algorithm and the particle swarm optimization, *Int. J. Electr. Eng* 9 (2016) 215–223.
- [17] C. M. Martínez, D. Cao, Integrated energy management for electrified vehicles, *Ihorizon-Enabled Energy Management for Electrified Vehicles* (2019) 15–75.
- [18] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, S. Van Huffel, A novel algorithm for the automatic detection of sleep apnea from single-lead ecg, *IEEE Transactions on Biomedical Engineering* 62 (2015) 2269–2278.
- [19] J. Pereira, M. Silveira, Unsupervised representation learning and anomaly detection in ecg sequences, *International Journal of Data Mining and Bioinformatics* 22 (2019) 389–407.
- [20] H. Sivaraks, C. A. Ratanamahatana, Robust and accurate anomaly detection in ecg artifacts using time series motif discovery, *Computational and mathematical methods in medicine* 2015 (2015).
- [21] J. Etumusei, J. M. Carracedo, S. McClean, Novel martingale approaches for change point detection, in: *International Conference on Intelligent Systems Design and Applications*, Springer, 2020, pp. 457–467.
- [22] A. Burns, B. R. Greene, M. J. McGrath, T. J. O’Shea, B. Kuris, S. M. Ayer, F. Stroiescu, V. Cionca, *Shimmer™—a wireless sensor platform for noninvasive biomedical research*, *IEEE Sensors Journal* 10 (2010) 1527–1534.
- [23] S. Zhang, L. Galway, S. McClean, B. Scotney, D. Finlay, C. D. Nugent, Deriving relationships between physiological change and activities of daily living using wearable sensors, in: *International Conference on Sensor Systems and Software*, Springer, 2010, pp. 235–250.
- [24] M. J. McGrath, T. J. Dishongh, A common personal health research platform-shimmer and biomobius., *Intel technology journal* 13 (2009).
- [25] M. Garrido, M. L. López-Vallejo, S.-G. Chen, Guest editorial: Special section on fast fourier transform (fft) hardware implementations, *Journal of Signal Processing Systems* 90 (2018) 1581–1582.
- [26] S.-S. Ho, H. Wechsler, A martingale framework for detecting changes in data streams by testing exchangeability, *IEEE transactions on pattern analysis and machine intelligence* 32 (2010) 2113–2127.
- [27] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, An

- efficient k-means clustering algorithm: Analysis and implementation, *IEEE transactions on pattern analysis and machine intelligence* 24 (2002) 881–892.
- [28] V. Fedorova, A. Gammerman, I. Nourtdinov, V. Vovk, Plug-in martingales for testing exchangeability on-line, *arXiv preprint arXiv:1204.3251* (2012).
- [29] A. Kawala-Sterniuk, M. Podpora, M. Pelc, M. Blaszczyzyn, E. J. Gorzelanczyk, R. Martinek, S. Ozana, Comparison of smoothing filters in analysis of eeg data for the medical diagnostics purposes, *Sensors* 20 (2020) 807.
- [30] B. Vahabzadeh, R. Sameni, The notion of cardiac phase and its applications in electrophysiological studies, *Biomedical Engineering* 9 (2012).
- [31] R. Sameni, C. Jutten, M. B. Shamsollahi, Multichannel electrocardiogram decomposition using periodic component analysis, *IEEE transactions on biomedical engineering* 55 (2008) 1935–1940.
- [32] M. Müller, Dynamic time warping, *Information retrieval for music and motion* (2007) 69–84.
- [33] C. Leys, C. Ley, O. Klein, P. Bernard, L. Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology* 49 (2013) 764–766.
- [34] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, in: *European Conference on Information Retrieval*, Springer, Meylan, France, 2005, pp. 345–359.
- [35] M. S. Orendurff, J. A. Schoen, G. C. Bernatz, A. D. Segal, G. K. Klute, How humans walk: bout duration, steps per bout, and rest duration., *Journal of Rehabilitation Research & Development* 45 (2008).

Artificial Neural Network for Human Activity Recognition by Use of Smart Insoles

Luigi D'Arco^a, Haiying Wang^a, Graham McCalmont^a, XianQi Lan^b and Huiru Zheng^a

^a*School of Computing, Ulster University, Belfast, Antrim, UK*

^b*Beitto Ltd., Fuzhou, China*

^c*Corresponding author*

Abstract

Human Activity Recognition (HAR) is an area with high interest. It can be used, for example, for medical purposes, for rehabilitation, for monitoring in sports as well as for prevention and monitoring of the elderly. The most used devices for this purpose are wearable devices. These devices have small sizes and can integrate different sensors inside. The only problem with these devices is that when using multiple devices at the same time, it can create an annoyance for the user. In addition, processing and understanding the data provided by such devices are sometimes not immediate and there is no standard rule that can be followed. In this study, a solution to HAR is presented by integrating a pair of smart insoles as a non-hindering device for the user and as a secondary purpose, there is the creation of a pipeline that can be simply recreated and extended to multiple subjects as needed. Smart insoles integrate pressure and inertia sensors inside. Alongside this device, an Artificial Neural Network model is developed to autonomously extract salient information directly from the raw data. Three subjects were included in the study. Each of them completed a series of activities among a well-defined set (fast walking, normal walking, slow walking, sitting, standing, downstairs, upstairs, and sit to stand). The results obtained with this solution achieved an average accuracy of 99.47%.

Keywords

Artificial Neural Network, Human Activity Recognition, Smart insole

1. Introduction

The Human Activity Recognition (HAR) has been a theme that has always fascinated the scientific community, which has spent many years looking for solutions suitable for everyday use. One of the objectives of HAR is to reveal information about a user's behaviour so that computing systems can help them with their daily tasks more effectively. Initially, the leading technology involved for HAR was computer vision. Computer vision made it possible, through the use of capture devices or video devices, to analyse the fundamental patterns and gestures of some actions. Different works can be found in literature [1, 2, 3], however, they all share the constraint of being prepared for a given environment, be it small or large, and cannot be easily adapted to different scenarios. Returning to the main goal of integrating this technology into

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ darco-l@ulster.ac.uk (L. D'Arco); hy.wang@ulster.ac.uk (H. Wang); mccalmont-g@ulster.ac.uk (G. McCalmont); allan@beitto.com (X. Lan); h.zheng@ulster.ac.uk (H. Zheng)

ORCID 0000-0001-7179-8281 (L. D'Arco); 0000-0001-8358-9065 (H. Wang); 0000-0001-7648-8709 (H. Zheng)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

everyday life settings, efforts to find new solutions have been shifted from computer vision to systems that could be worn by users (wearable systems). Wearable systems allow not only to be integrated into multiple environments without having to worry about external constraints but at the same time reduce the overall dimensions and work continuity factors over time.

The study of HAR using wearable devices has made it possible to extend the possible applications of HAR to scenarios that may previously have been unthinkable [4] and by the spread of ubiquitous devices, the integration of HAR in daily life has been greatly simplified so that today we can find several studies in the literature that address it [5, 6, 7]. HAR has been found to be useful in a variety of real-world situations, such as rehabilitation [8], sport monitoring [9] as well as prevention for the elderly [10]. Despite the numerous applications and benefits of wearable systems, there is always the area covered by these devices that is too large for pervasive use. In addition to the aforementioned problem, it must be considered that the processing of sensor's data is not as immediate as images video because they are difficult to understand for the human and there is no technique that could be applied for all the sensor's data. Indeed, the majority of the studies use feature extraction techniques or even statistical analysis to make the data more readable and meaningful before processing. With the growth of Machine Learning and Neural Networks, the processing of data is becoming more simple and it does not require so depth knowledge of mathematics to process that data. Neural Networks provides an intrinsic function of auto discovering of the most important features providing the scientist with a way to process directly the raw data.

The purpose of this work is to develop a HAR system that can be used by users without any restrictions, as well as to establish a pipeline that can be easily replicated and expanded to include as many subjects as desired. A smart insole is employed as a device for the HAR to achieve this goal because it decreases the user's encumbrance as well as the work necessary for installation in a new settlement. Pressure sensors and inertia sensors have been integrated into the smart insole. An Artificial Neural Network (ANN) is the engine that recognises the activities and allows raw data to be processed directly.

The rest of the paper is structured as follow: in Section 2 the existing solutions are analysed, the study details and the approach chosen are presented in Section 3 followed by the results obtained and the discussion in Section 4, and in Section 5 the overall paper will be summarised and the future work highlighted.

2. Related Work

Using smart insoles for HAR is certainly not new, and multiple studies can be found in the literature which can be divided into two main categories according to the way the HAR is carried out: threshold algorithms, machine learning algorithms.

Threshold algorithms base their functioning on the study of the data collected by the subjects and on the basis of observations on the change of such data in relation to the activity carried out, minimum thresholds are defined which, if exceeded, indicate the activity performed. Mofawad el Achkar et al. [11] introduce instrumented shoes that can collect movement and foot loading data unobtrusively during daily living. The instrumented shoes include a 3D accelerometer, a 3D gyroscope, a 3D magnetometer, a barometric sensor and a force sensing

insole comprised of 8 pressure cells. Ten elderly subjects were recruited for data collection, and asked to complete a predefined track to mimic physical activities of daily life (level walking, sit-to-stand, sitting, standing, uphill, downhill, upstairs, downstairs and elevator use). The data were sampled at 200 Hz, then segmented at 5 seconds with 2.5 seconds overlap. The classification relies on an expert-based hypothesis, the foot loading, orientation, and elevation can be used to classify postural transitions, locomotion, and walking type. The resulting classifier has sensitivity and precision for sitting, standing, and walking that are greater than 95%. Stair ascending had the lowest precision (89%) and sensitivity (79%) whereas elevator up and down had the lowest precision (89%) and sensitivity (79%) respectively (78%). The overall accuracy rate was 97.41%.

Machine learning algorithms try to imitate human learning and consequently process data incrementally, improving its accuracy during execution through the use of statistical methods uncovering key insights within data. De Pinho et al. [12] propose a machine learning HAR classifier based on a foot-based wearable device. The wearable device in question comprises two components: a plantar with pressure sensors and a microcontroller equipped with a 3D accelerometer, 3D gyroscope, 3D magnetometer and a barometric sensor. Eleven volunteers participated in the experiment from which 2 hours of feet posture and movement data were gathered. The activities that the classifier identifies are: walking straight, walking slope up and down, ascending and descending stairs and sitting. The data collected were segmented at 0.3 seconds and instead of processing sensors raw data, the features were extracted, in particular the standard deviation, variance, minimum, maximum and average values are used. Random Forest was used as classified and validated by the means of a Leave-one-out Cross-Validation strategy. The average accuracy of the classifier was 93.34%.

Identifying the most important features to extract requires in-depth knowledge of the dataset and requires several statistical studies on the data so that the identified features can best represent the data collection. To overcome these problems, several studies use deep learning techniques instead of machine learning as they differ from the latter for their intrinsic ability to filter data and then select only the most important information, thus avoiding the feature extraction phase. Pham et al. [13] propose a deep convolutional neural network (CNN) for detecting a set of low-level activities including running, walking, standing, jumping, kicking, cycling. They used a 3D accelerometer sensor integrated into a pair of shoes. The data were sampled at a frequency of 50 Hz and segmented with a sliding window approach of 2 seconds and 50% overlap between two consecutive windows. Ten subjects were included in the study which were requested to perform each activity from 10 to 30 minutes. The CNN was evaluated using a 10 Fold Cross Validation reaching relatively high performances, on average 93.41% precision and 93.16% recall. Zhang et al. [14] design a HAR system based on an insole-based IMU sensor driven by a subject-independent algorithm. The IMU sensors involved are a 3D acceleration and a 3D gyroscope, and the data are sampled at a frequency of 100 Hz. In the data preprocessing section, raw IMU sensor data is separated into two parts: one part is manually labelled with activity labels, while the other half is left unlabelled, The data are segmented in slices of 2 seconds and then the subject-specific information are removed from them. The deep learning model uses the labelled data to train but also can take advantage of unlabelled data in the training process which can improve the recognition performance. Eight volunteers are asked to perform 5 basic daily activities: standing, laying, walking downstairs, walking, running. The

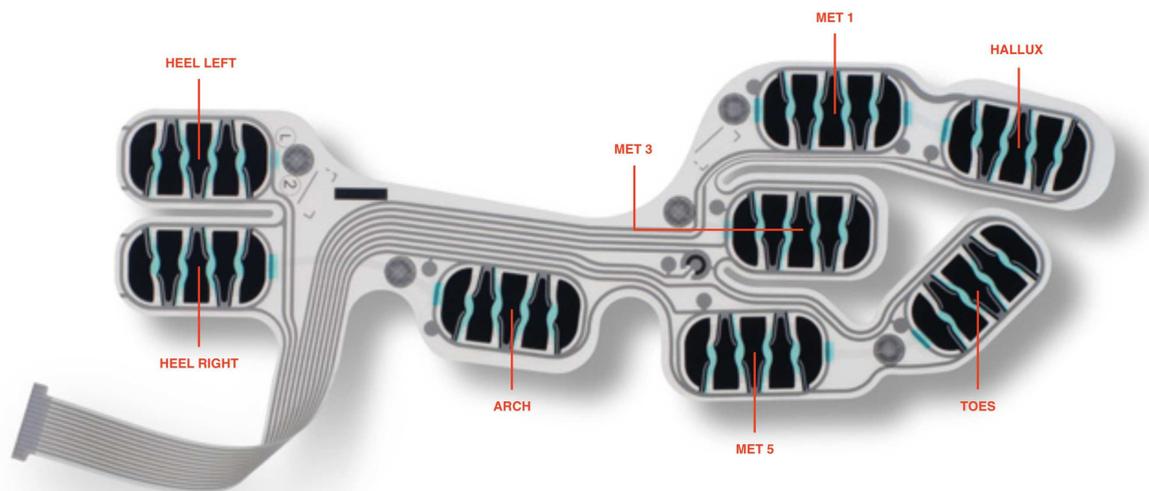


Figure 1: Pressure cells on the IEE Smart Foot Sensor.

model was evaluated with a Leave-One-Out approach reaching an overall accuracy of 98.92%. Paydarfar et al. [15] present a HAR system based on piezoresistor-based instrumented shoes and a Recurrent Neural Network (RNN). The hardware consists of a pair of sneakers, with an onboard microcontroller that is connected to 3 piezoresistor sensors located at the calcaneus, metatarsals, and phalanges. Twenty healthy subjects are engaged in the experiment. The activities collected are: walking, standing, balancing on the left foot, balancing on the right foot, toe-up, and ascending stairs. Each activity is performed by each subject for 45 to 120 seconds and the sampling frequency was set to 50 Hz. The data are segmented in slices of 1 second but in a way that each slice differ from the previous just by 1 timestep. The RNN was validate using a Leave One Out strategy reaching an accuracy of $87.0 \pm 8.9\%$.

3. Method

3.1. Sensors and Data Collection

In this study a smart insole kit has been used, the ActiSense Kit provided by the IEE Luxembourg S.A. Two IEE Smart Foot Sensors and two ActiSense electronics (ActiSense ECU) are included in the ActiSense kit; the former is a pressure-cell insole and the latter is a device made up of various IMU sensors. The IEE Smart Footwear sensor is available in eight different sizes, ranging from 28 to 47 (EU) in which are located eight high dynamic pressure cells, at positions: left heel, right heel, arch, met 1, met 3, met 5, hallux, and toes (see Fig. 1). The ActiSense electronic includes a 3-axis accelerometer, a 3-axis gyroscope and a 3-axis magnetometer.

The data was collected on three healthy volunteers (3 males, ages 24-45). The IEE ActiSense Kit was worn by each participant, and two feet measurements were taken. The data was col-

lected via an Android software provided by the equipment maker, which took advantage of a Bluetooth link between the smartphone and the kit. The sampling frequency was set to 200 Hz, and each participant performed various activities in a predefined set (Downstairs, Fast Walking, Normal Walking, Sit to Stand, Sitting, Slow Walking, Standing, Upstairs) without supervision or restriction. In total, 120 minutes of recording were collected.

3.2. Data Distribution and Over-Sampling Technique

The data collected from the subjects are in form of multivariate time series. Due to the multiple sensors involved and the capability of the same sensors to provide multiple values is possible to summarise the data collected as described in Eq. 1.

$$\sum_{i=1}^k s_i = (d^1, d^2, \dots, d^t) \quad (1)$$

where k represents the number of sensors involved, d^i the multiple values (such as for accelerometer that typically has three axes).

A single measurement at time t (sample) is not sufficient for the representation of an activity performed by a subject, so we need to group together multiple consecutive samples that are likely to contain information about one activity (segment). A segment, hence, can be represented through the Eq. 2.

$$w_i = (t_1, t_2) \quad (2)$$

where t_1 and t_2 represent, respectively, the starting and ending time within the time series. The segment usually are referred as "windows" and the difference between t_1 and t_2 as "window size".

In this study, the window size is set to 2 seconds, as it was noted that this number is repeatedly chosen in the literature when there are conditions that agree with ours, for instance, sample rate, type of data and subjects included.

After data segmentation one of the most important analysis to perform is to assess the data distribution in relation to the classes. If the data show that one or more classes have a large number of samples in contrast to other classes that have few, what is encountered is a class imbalance problem. The problem of class imbalance is relevant since it has been demonstrated to produce a major bottleneck in the performance that can be achieved using traditional learning methods that presume a balanced class distribution [16]. As shown in Fig. 2 the number of elements in the classes: downstairs, sit_to_stand, sitting, standing and upstairs, are much less than the other classes resulting in an unbalanced dataset. To overcome this problem two possible approaches are available: assign distinct costs to train examples [17], re-sample the original dataset, either by over-sampling the minority class and/or under-sampling the majority class. The approach chosen is the latter, in particular, the Synthetic Minority Over-sampling Technique (SMOTE) [18] was used. Starting with the minority classes, SMOTE develops new synthetic examples. A representative from the minority class is picked at random. The k closest neighbours of the example are picked at random depending on the quantity of over-sampling

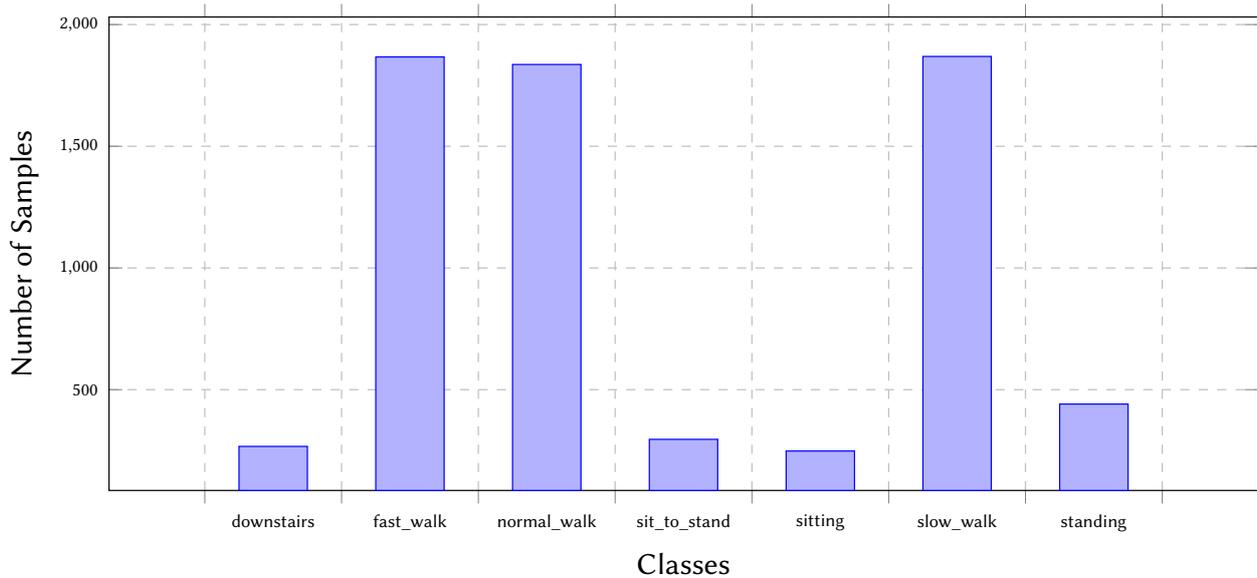


Figure 2: Samples distribution of the dataset after data segmentation with window size of 2 seconds.

Table 1

Layers that compose the architecture of the ANN.

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 13600)	0
dense (Fully Connected)	(None, 256)	3481856
dropout (Dropout)	(None, 256)	0
relu (Activation)	(None, 256)	0
dense_1 (Fully Connected)	(None, 128)	32896
dropout_1 (Dropout)	(None, 128)	0
relu (Activation)	(None, 128)	0
dense_2 (Fully Connected)	(None, 8)	1032
softmax (Activation)	(None, 8)	0

necessary. Multiply the difference between the example and the neighbour by a random number between 0 and 1, then add the result to the example causing the selection of a random point along the line segment between two specific examples. This strategy forces the minority class's decision-making region to become more general.

3.3. Model Architecture

Human activities by their nature are difficult to recognise as they can be influenced by several factors. Each person differs from others as well as from himself in different circumstances. The large number of devices that can be used also affects the recognition itself. All this has led to a missing common definition of Human Activity and therefore having to analyse the data of each study in a different way [19]. Data analysis and core information extraction (feature extraction) take time and effort. Neural Networks has become a popular choice because they

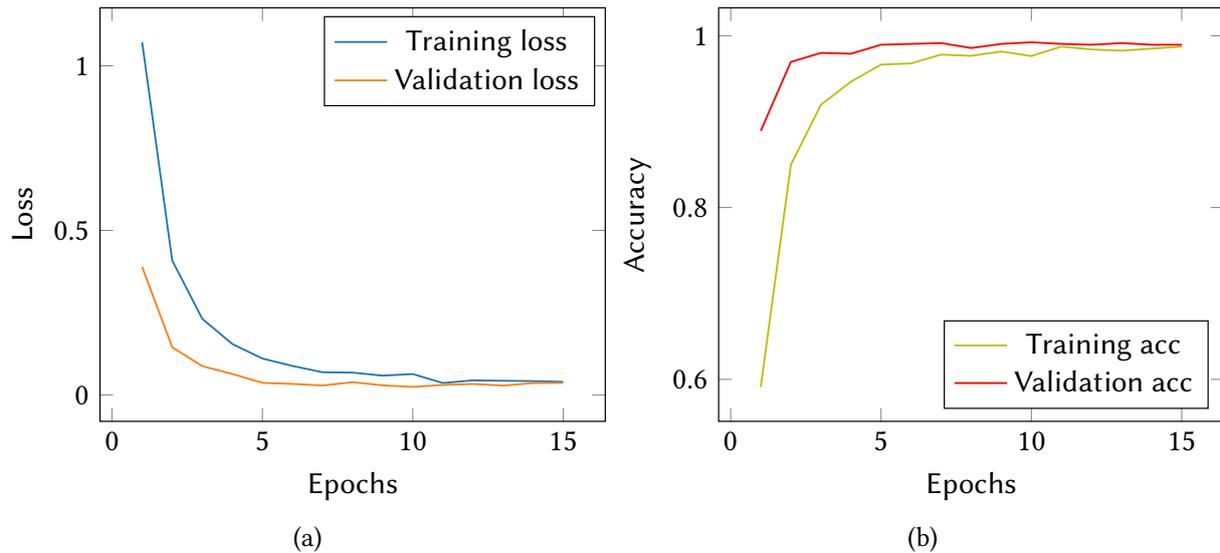


Figure 3: ANN model training history according to training and validation datasets. (a) Loss (b) Accuracy.

can overcome this problem.

In this study an Artificial Neural Network (ANN) is used to discriminate activities based only on raw data provided by the smart insole. The ANN tries to mimic the behaviour of the human brain. An ANN is therefore composed of three or more layers that are interconnected. The first layer consists of input neurons. Those neurons send data to the deeper layers, which in turn will send the final output data to the last output layer.

The layers that build up the ANN are shown in Table 1. The ANN takes as input the previously separated windows which have a size of (400, 34). Since the layers of our architecture cannot process a two-dimensional sample, the first layer of the architecture is a Flatten. A Flatten layer converts the input data in a column-wise shape to feed into the next layers. The Fully Connected layer consists of the weights and biases along with neurons and all the inputs are connected to every activation unit of the next layer. As a neural network learns, the weights of neurons are tuned providing some specialisation. Neighbouring neurons, on the other hand, begin to rely on this specialisation, which, if carried too far, can result in a weak model that is overly specialised to the training data. For this reason, a Dropout Layer is required to prevent overfitting by dropping out units in the neural network. Each neuron can be dropped with a probability p or kept with a probability $1 - p$ [20]. The Activation functions define how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network. Two types of activation functions are used: Rectified Linear Activation (ReLU) and Softmax.

The dataset was split for the training purpose, train and test sets were created. The ANN was trained for 15 epochs and the batch size set to 32. Moreover, the train set was split into two parts (train and validation, respectively 80% and 20%) for the objective of evaluating if during training overfitting or underfitting occurred. The training history of the model is shown in Fig. 3.

Table 2

Performances of the ANN model against the testing dataset.

	Precision	Recall	F1-Score
downstairs	0.9930	0.9965	0.9947
fast_walk	0.9811	0.9924	0.9867
normal_walk	0.9894	0.9824	0.9859
sit_to_stand	0.9966	1.0000	0.9983
sitting	1.0000	1.0000	1.0000
slow_walk	1.0000	0.9891	0.9945
standing	1.0000	1.0000	1.0000
upstairs	0.9963	0.9963	0.9963
Accuracy			0.9947
Macro avg	0.9945	0.9946	0.9946
Weighted avg	0.9947	0.9947	0.9947

4. Results and Discussion

The ANN model was trained with the goal of recognising 8 different basic activities: *downstairs*, *fast walk*, *normal walk*, *sit to stand*, *sitting*, *slow walk*, *standing* and *upstairs*.

To evaluate the performance of the classifier four metrics were used: accuracy, precision, recall and f1-score. The accuracy is the number of correctly predicted data points out of all the data points, in other words, it measures how often the algorithm classifies a data point correctly (see Eq. 3).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

The precision is the number of correct positive results divided by the number of positive results predicted by the classifier (see Eq. 4).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

The recall is the number of correct positive results divided by the number of all relevant samples (see Eq. 5).

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

The F1_Score is the Harmonic Mean between precision and recall. It tells you how precise your classifier is, as well as how robust it is (see Eq. 6).

$$F1_Score = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

The overall accuracy of the model on the test set is 99.47%. Analysing the Table 2 is possible to notice that the classes *sitting* and *standing* can be clearly identified against others with precision and recall that are 100%. The worst class recognised is the *normal_walk* with an F1-Score of 98.59%.

Table 3

Confusion matrix of the ANN model against the testing dataset.

Predicted → Reference ↓	downstairs	fast_walk	normal_walk	sit_to_stand	sitting	slow_walk	standing	upstairs
downstairs	284	0	0	1	0	0	0	0
fast_walk	1	260	1	0	0	0	0	0
normal_walk	0	5	279	0	0	0	0	0
sit_to_stand	0	0	0	293	0	0	0	0
sitting	0	0	0	0	278	0	0	0
slow_walk	0	0	2	0	0	273	0	1
standing	0	0	0	0	0	0	298	0
upstairs	1	0	0	0	0	0	0	267

Analyzing the performance of the system it seems that walking activities have a lower recognition rate. To learn more about what this result entails, a confusion matrix has been created. As shown in Table 3, the highest number of misclassifications is between the *normal_walk* and the *fast_walk*. This error is due to the nature of the dataset. The dataset was created by people at different times and places in a completely independent way and with few if almost no guidelines. So for instance one person can consider the *normal_walk* at the speed of 6 km/h, instead, another person can recognise the same speed as a *fast_walk*.

Overall, the produced solution outperforms the findings of the other studies examined; nevertheless, the number and kind of participants are insufficient to cover the entire population, which could lead to a loss of performance if the system is employed on a subject that differs a lot from those examined.

5. Conclusion

In this paper a Human Activity Recognition solution that may be non-invasive for the user has been discussed. A smart insole consisting of pressure sensors and inertial sensors is used as unique device able to capture activity information from the user. Moreover, with the aim of creating a replicable pipeline that is not bound to the subjects studied so as to be able to expand the number of subjects if necessary, the smart insole was supported by an Artificial Neural Network which allows the processing of raw data and auto extraction of meaningful information from them. The proposed system can resolve the human activity recognition with an overall accuracy of 99.47%. The results obtained exceed the existing solutions analysed, however it turned out that the lack of guidelines for collecting data has resulted in problems in the recognition of some activities such as normal and fast walking, since not having defined a standard speed, each user has decided their own.

Inasmuch as, not all types of subjects were treated, the number of people assessed narrows the prospective targets for everyday use. As a result, the number of subjects included in the study may be increased in the future, allowing the system to cover a larger number of users. In addition, to cover as many scenarios of daily life, it is possible to expand the number of activities involved.

Acknowledgments

Luigi D'Arco is supported by the Ulster-Beitto Collaboration Program, Graham McCalmont is supported by Department of Economy PhD Scholarship.

References

- [1] H. Zheng, H. Wang, N. Black, Human activity detection in smart home environment with self-adaptive neural networks, in: 2008 IEEE International Conference on Networking, Sensing and Control, IEEE, 2008, pp. 1505–1510.
- [2] W. Wolf, I. B. Ozer, A smart camera for real-time human activity recognition, in: 2001 IEEE Workshop on Signal Processing Systems. SiPS 2001. Design and Implementation (Cat. No. 01TH8578), IEEE, 2001, pp. 217–224.
- [3] M. Leo, T. D'Orazio, P. Spagnolo, Human activity recognition for automatic visual surveillance of wide areas, in: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks, 2004, pp. 124–130.
- [4] G. McCalmont, H. Zheng, H. Wang, S. McClean, M. Zallio, D. Berry, A lightweight classification algorithm for human activity recognition in outdoor spaces, in: Proceedings of the 32nd International BCS Human Computer Interaction Conference 32, 2018, pp. 1–5.
- [5] S. Irene, N. Shwetha, P. Haribabu, R. Pitchiah, Development of zigbee triaxial accelerometer based human activity recognition system, in: 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, IEEE, 2015, pp. 1460–1466.
- [6] S. Ashry, R. Elbasiony, W. Gomaa, An lstm-based descriptor for human activities recognition using imu sensors, in: Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics, ICINCO, volume 1, 2018, pp. 494–501.
- [7] W. Gomaa, Statistical and time series analysis of accelerometer signals for human activity recognition, in: 2019 14th International Conference on Computer Engineering and Systems (ICCES), IEEE, 2019, pp. 351–356.
- [8] E. Kańtoch, Human activity recognition for physical rehabilitation using wearable sensors fusion and artificial neural networks, in: 2017 Computing in Cardiology (CinC), IEEE, 2017, pp. 1–4.
- [9] F. Nurwanto, I. Ardiyanto, S. Wibirama, Light sport exercise detection based on smart-watch and smartphone using k-nearest neighbor and dynamic time warping algorithm, in: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), IEEE, 2016, pp. 1–5.
- [10] W. Ugulino, M. Ferreira, E. Velloso, H. Fuks, Virtual caregiver: a system for supporting collaboration in elderly monitoring, in: 2012 Brazilian Symposium on Collaborative Systems, IEEE, 2012, pp. 43–48.
- [11] C. Moufawad el Achkar, C. Lenoble-Hoskovec, A. Paraschiv-Ionescu, K. Major, C. Büla, K. Aminian, Instrumented shoes for activity classification in the elderly, *Gait & posture* 44 (2016) 12–17.

- [12] R. De Pinho André, P. H. F. Diniz, H. Fuks, Bottom-up investigation: Human activity recognition based on feet movement and posture information, in: Proceedings of the 4th international Workshop on Sensor-based Activity Recognition and Interaction, 2017, pp. 1–6.
- [13] C. Pham, N. N. Diep, T. M. Phuong, e-shoes: Smart shoes for unobtrusive human activity recognition, in: 2017 9th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2017, pp. 269–274.
- [14] X. Zhang, J. Zhang, Subject independent human activity recognition with foot imu data, in: 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), IEEE, 2019, pp. 240–246.
- [15] A. J. Paydarfar, A. Prado, S. K. Agrawal, Human activity recognition using recurrent neural network classifiers on raw signals from insole piezoresistors, in: 2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob), IEEE, 2020, pp. 916–921.
- [16] M. Kubat, R. C. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, *Machine learning* 30 (1998) 195–215.
- [17] P. Domingos, Metacost: A general method for making classifiers cost-sensitive, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 155–164.
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* 16 (2002) 321–357.
- [19] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, *ACM Computing Surveys (CSUR)* 46 (2014) 1–33.
- [20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research* 15 (2014) 1929–1958.

Investigation of the use of Deep Learning and emotion detection for the improvement of Text-based medical conversational agent

Bing Yuan^a, Haithem Afli^b

^aMunster Technological University of Ireland, Bishopstown Campus Cork, Ireland.T12 P928

^bMunster Technological University of Ireland, Bishopstown Campus Cork, Ireland.T12 P928

Abstract

In the last decade, there was an explosion in Natural Language Process applications led by the progress of Deep Learning architectures. Recently, Google researchers published BERT (Bidirectional Encoder Representations from Transformers) [1], a deep bidirectional language model based on the Transformer architecture, and advanced the state-of-the-art in many popular NLP tasks, including Machine Translation, Information Retrieval, and Sentiment Analysis. Emotion Detection is a sub-branch of Sentiment Analysis that focuses on detecting specific emotions. This project investigated the combination of emotion detection using Deep learning methods with rule-based Chatbot such as ELIZA. To improve the performance of conversation of agent focus on helping People with Obesity.

Keywords

emotion detection, BERT(Bidirectional Encoder Representations from Transformer), NLP(Natural Language Process), DL(Deep Learning), PwO(Person with Obesity), STOP(STOP Obesity Platform), LSTM,

1. Introduction

Conversational interfaces provide a different (and possibly easier) way to do things, e.g., engage in a conversation to ask a query or provide a response instead of navigating menus, forms, and drop-down boxes a traditional graphical user interface. As a result, chatbots have emerged as a new type of interface to serve on the web. It has also become possible in the e-health medical area due to the significant advances in artificial intelligence and speech and language technologies.

This project aims to design an intelligent conversational agent Smart-ELIZA to enable the STOP platform interactive communication with People with Obesity(PwO). The second main feature is the integration of the emotion detection model. When the user's emotion is negative, the Smart-ELIZA Chatbot will respond differently from what ELIZA system's original reply; for example, the new reply can be: "I am sorry to hear that, how can I help you?".

This project use the new State-art-of Deep Learning model: the BERT model and a fine-grained training dataset such as GoEmotions containing 58000 data and twenty eight emotions.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ hbingyuan@gmail.com (B. Yuan); Haithem.Afli@cit.ie (H. Afli)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Contribution

The primary contributions of this study are summarized below:

From the research aspect

1. This project is feasibility research regarding adding the emotion detection function to the ELIZA Chatbot with emotion detection function using Deep Learning technologies. ELIZA Chatbot is best known for acting in the manner of a psychotherapist, and the therapist "reflects" on questions by turning the questions back at the patient. By adding emotion detection within ELIZA, which empowers the ELIZA's function.
2. For the emotion detection model development: This project demonstrates the effect of using the GoEmotions dataset to train the Bert pre-trained model for emotion detection. To my knowledge, this is the first study to use Bert Model with the GoEmotions dataset combination to detect twenty eight different emotions.

From industry application aspect

1. The ELIZA Chatbot offer to the PwO is a novel approach that offers the ability to initiate a conversation any time and anywhere through a smartphone or laptop. It shows friendly companionship and also acts as an assistant to the PwO. By understanding emotions in textual dialogue to provide emotionally aware responses to e-health users enhances the healthcare-related platforms.
2. The first-hand chat information through the conversation with PwO provides insights into the common queries, concerns, and general psychological state of PwO. The captured chat data will be fused with further sensor data and knowledge resource database within the STOP platform and passed to an analysis pipelines service to provide insight for healthcare professionals.

3. Background

3.1. Introduction to Chatbot

Conversational agents are "systems that mimic human conversation" using communication channels such as speech, text, and facial expressions and gestures [2]. Conversational agents roughly consist of three main categories: Chatbots without embodiment, virtually embodied avatars, and physically embodied robots.

A Chatbot is a contraction of chat and robot, and it is a computer program that conducts conversations by using dialogue or text [3]. A Chatbot is an instant messaging application, an automated chat system; it is a web-based application that is easy to use and no need to install.

Chatbot application usage categories include commerce (e-commerce via chat), education, entertainment, finance, health, news, and productivity.[4]. With artificial intelligence (AI) technologies, including deep learning, machine learning (ML) algorithms, and natural language processing, Chatbot can proactively assist and predict the user or customer behaviors, requests, and needs that bring many domains.

3.1.1. Chatbot for HealthCare Overview

Chatbots have been used in many health-related practices, such as sports activities, psychological health, drug compliance, and depression and anxiety, as well as expressing sympathy and compassion [5]. In addition, some of the e-health Chatbot could accept consultation in the form of pictures, images, and even voice.

Semantic analysis technology allows Chatbots to autonomously converse with the patient to identify and respond to their needs. Thus Chatbots have become preferred by the HealthCare domain for communicating instantaneously with users in terms of efficiency and benefits. In addition, research shows that the deployment of conversational agents in service encounters is growing exponentially in many other sectors, including healthcare.[2]

3.2. Introduction to ELIZA

ELIZA was programmed in 1966 by Joseph Weizenbaum at MIT, and it was the first public known Chatbot. ELIZA was an early test case for the Turing Test. ELIZA does not understand the conversation; ELIZA uses pattern matching and returns the response by the selection scheme based on templates. ELIZA can only accept text format input; the input sentences are analyzed based on decomposition rules triggered by keywords appearing in the input text. Responses are generated by reassembly rules associated with selected decomposition rules.[6] ELIZA only contains 200 keywords and rules.[7] However, many users believe they were talking to a natural person when ELIZA was invented.

3.3. Introduction to Emotion Detection

Based on Merriam-Webster definition: Emotion which is a state of feeling, is the affective aspect of consciousness: Feeling is a conscious mental reaction (such as anger or fear) subjectively experienced as a strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body. The prominent work in understanding and categorizing emotions include Ekman's six class categorization and Plutchik's "Wheel of Emotion," which suggested eight primary bipolar emotions.

Emotion detection is the state or fact of emotion being detected. There are several ways to recognize human emotion, such as source from the text message, voice tones, facial expressions, gestures, etc. Myriam Munezero's research shows that current emotions detection is from text has focused on capturing emotion words based on three emotion models, that is, categories of basic emotions, emotion dimensions, or cognitive appraisal categories. Suppose emotion detection application can recognize emotions such as joy, sadness, anger with possible high accuracy. In that case, it will bring a wide range of impact for many emotion-based services, such as medical applications and public emotion sense services.

3.4. Introduction to Sentiment Analysis

Sentiments are "socially constructed patterns of sensations, expressive gestures, and cultural meanings organized around a relationship to a social object, usually another person or group such as a family." Examples of sentiments include romantic love or nostalgic feeling. Both emotions and sentiments refer to "experiences that result from the combined influences of the biological, the cognitive, and the social". However, sentiments are differentiated from emotions by the duration in which they

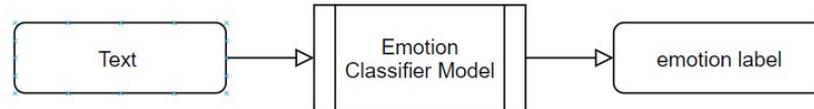


Figure 1: A high-level overview of an emotion detection classifier shows input and output

are experienced. Although emotions are brief episodes of brain, autonomic, and behavioral changes, sentiments have been found to form and be held for a more extended period. Sentiments, in addition, are more stable and dispositional than emotions.

Sentiment Analysis is a natural language processing (NLP) task. It has been defined as analyzing the sentiment according to the opinion expressed about a given subject is positive, negative, (but sometimes also neutral or vague). It aims to classify the text-based content, but sometimes it also includes audio and video. Sentiment analysis is a suitcase research problem that requires tackling many natural language processing (NLP) tasks. Emotion Detection is a branch of sentiment analysis within the NLP research domain.

3.5. Introduction to Natural Language Processing

Natural Language Processing (NLP) started in 1950, is an intersection of artificial intelligence and linguistics domain. NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications (Liddy, 2003). NLP is a large area of research and application, and the NLP tasks include automatic text summarization, machine translation, information search, question answering, and sentiment analysis, etc. Because NLP helps machines to "read" text by simulating the human ability to understand language; plus the availability of cloud computing, big data technologies, and deep learning algorithms, the NLP applications has been widely used in many industries such as linguistic analysis, speech recognition, and notes analyze in electronic health records, etc.

4. Using Deep Learning techniques to develop emotion detection models

Emotion detection and Sentiment Analysis attract diverse attention, including researchers from machine learning, natural language processing, and computational linguistics. The Emotion Detection classification within this project is a reference to a form of text classification in which text content will be classified into the pre-defined emotion classes. This section will present a state-of-the-art overview of various deep learning technologies and relevant datasets for text-based emotion detection.

4.1. Literature Review

From the dataset aspect, many pieces of research are done by using human physical reaction signals (Such as respiration signals, heartbeat and breathing signals) work together with a deep learn-

Table 1

Deep Learning Model and Datasets Summary for emotion detection

Deep Learning Model	Dataset	Detection Target	Author
LSTM+CNN	Public Twitter datasets of SemEval Task-1, Affect in Tweets dataset.	Text	Malak Abdullah [8]
LSTM with GloVe and SSWE as embedding layer	The Twitter dataset with 17.62 million tweet conversational pairs	Text	Ankush Chatterjee from Microsoft [9]
LSTM	electroencephalography (EEG) signals data from DEAP Dataset	Text	Alhagry S. [10]
multiple-fusion-layer based ensemble classifier of a stacked autoencoder	multimodal physiological signals from DEAP Dataset	Text	Zhong Yin [11]

ing model. Alhagry S. who use electroencephalography (EEG) signals data together with the LSTM model for emotion prediction and classified category for emotions. They used the Dataset for Emotion Analysis using Physiological signals (DEAP) data set in the study, and a classification process was performed for valence, arousal, and liking classes. Zhong Yin uses multimodal physiological signals and a multiple-fusion-layer-based ensemble classifier of stacked autoencoder (MESAE) for recognizing emotions. He also uses DEAP multimodal database for the study.

From the Deep learning model aspect, there are quite several pieces of research done by using the Long-short term memory (LSTM) model. LSTM is a particular type of Recurrent Neural Network(RNN) with DL family, its capability well knows of modeling the semantic relationships between words in the text, or combine of LSTM with other neuro network method. Such as Malak Abdullah use Convolutional Neural Networks CNN combine with LSTM on public Twitter datasets of SemEval Task-1, Affect in Tweets dataset. Ankush Chatterjee from Microsoft uses LSTM with GloVe and SSWE as an embedding layer on an extensive dataset: the Twitter dataset with 17.62 million tweet conversational pairs. The table 1 shows the comparison details :

1.

With the development of neural language models such as word vectors , paragraph vectors , and GloVe , the transfer learning (pre-training and fine-tuning) revolution started in NLP. Recently, Google researchers published BERT (Bidirectional Encoder Representations from Transformers) , a deep bidirectional language model based on the Transformer architecture [1] and advanced the state-of-the-art in many popular NLP tasks.

There have been many Sentiment Analysis pieces of research that used the BERT model. For example, Quoc Thai Nguyen [12] used the Fine-Tuning BERT together with LSTM/TextCNN/RcNN got an average 90% F1 score for negative and positive classification of Vietnamese Reviews. Manish Munikar [13] uses the BERT to classify five-category sentiments: very negative, negative, neutral, positive, and very positive. Another fine-grained BERTBase and BERTLarge model was performed on sentiment classification with the Stanford Sentiment Treebank dataset, which got 94% accuracy. Yun Xiang Zhang [14] also combine BERT with Deep neural networks DNN to propose a classification model applied to aspect-level sentiments classification.

The detection or recognition of emotions means the extraction of finer-grained sentiments. Text-based emotion recognition is a sub-branch of emotion recognition (ER) that focuses on extracting fine-grained emotions from texts. [15] Although BERT has been widely used in sentence-level sentiment classification, as shown above, BERT has not found widespread application in sentence-level emotion classification. However, only several research pieces have been done by using BERT for emo-

Table 2
Deep Learning Model and Datasets Summary for Sentiment Analysis

Deep Learning Model	Dataset	Emotions Categories	Author
CNN, biLSTM, BERT	Mass Shooting dataset (MS), the Terrorism dataset (TR)	7 emotions: fear, anger, joy, sad, contempt, disgust, and surprise	Jonathas G.D. Harb
BERT model+ bi-LSTM classifier	ISEAR dataset with 7666 sentences	7 emotions: joy, anger, sadness, shame, guilt, surprise, and fear.	Acheampong Francisca Adoma
BERT, RoBERTa, DistilBERT, and XLNet	ISEAR dataset with 7666 sentences	7 emotions: joy, anger, sadness, shame, guilt, surprise, and fear.	Acheampong Francisca Adoma
BERT Embeddings+GloVe word embeddings+BiLSTM neural network	Set of psycholinguistic features (e.g. from AffectiveTweets Weka-package.)	Four categories: Happy, Sad, Angry, and Other	Hani Al-Omari

tion detection; this is a very new research domain.

Jonathas G.D. Harb [16] design a framework using CNN, biLSTM, and BERT to perform the sentiment classification in terms of Ekman's seven basic emotions. The platform is to analyze the emotional reactions to mass violent events on Twitter. The research was done by applying three different DL technologies into different datasets Mass Shooting dataset (MS), the Terrorism dataset (TR), or MS+TR. CNN(MS+TR) produced the best results for three emotions and BERT(MS) for two emotions. biLSTM(MS) models produced statistically superior f-measure results for Anger only.

Acheampong Francisca Adoma [17] used the BERT model and bi-LSTM classifier to classify seven emotions from the text-based ISEAR dataset, which contains 7666 sentences. It attains an overall accuracy of 72.64%. That research shows that the BERT model performs an average much better performance than the machine learning-based classifier, including SVM, Naïve Bayer, and Randon Forest. They performed additional research to analyze the efficacy of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from texts [15]; the recorded model accuracies in decreasing order are 0.7431, 0.7299, 0.7009, 0.6693 for RoBERTa, XLNet, BERT, and DistilBERT, respectively.

Hani Al-Omari [18] then performed future research based on the above research with the combination of GloVe word embeddings, BERT Embeddings, and a set of psycholinguistic features (e.g., from AffectiveTweets Weka-package). The research combined a fully connected neural network architecture and BiLSTM neural, which achieve the result with F1-Score 0.748 for four emotions. The table 2 summarize the BERT model and the dataset.

5. Design

In order to design of the Smart-ELIZA Chatbot, this part reviews one of implemented ELIZA python version source code, summarizes how the ELIZA system was designed in detail with workflow. Also gives the overview of how the Smart-ELIZA will design based on the existing ELIZA system and narrows down the emotion detection working process within Smart-ELIZA then presents the detailed workflow design of Smart-ELIZA Chatbot. The final two parts are about the approaches for emotion modeling and the process for emotion detection.

```

doctor.txt x
1  initial: How do you do.  Please tell me your problem.
2  final: Goodbye.  Thank you for talking to me.
3  quit: bye|
4  quit: goodbye
5  quit: quit
6  pre: dont don't
7  pre: cant can't
8  pre: wont won't

doctor.txt x
278 key: can
279   decomp: * can you *
280     reasmb: You believe I can (2) don't you ?
281     reasmb: goto what
282     reasmb: You want me to be able to (2) .
283     reasmb: Perhaps you would like to be able to (2) yourself.
284   decomp: * can i *
285     reasmb: Whether or not you can (2) depends on you more than me.
286     reasmb: Do you want to be able to (2) ?
287     reasmb: Perhaps you don't want to (2) .
288     reasmb: goto what
289   key: what
290     decomp: *
291     reasmb: Why do you ask ?
292     reasmb: Does that question interest you ?
293     reasmb: What is it you really wanted to know ?
294     reasmb: Are such questions much on your mind ?
295     reasmb: What answer would please you most ?
296     reasmb: What do you think ?
297     reasmb: What comes to mind when you ask that ?
298     reasmb: Have you asked such questions before ?
299     reasmb: Have you asked anyone else ?
300   key: because
301     decomp: *

```

Figure 2: The screenshot of doctor.tx in ELIZA SourceCode

5.1. ELIZA Design Detail Review

With reference to Charles Hayden Source Code Readme document.[?], and a step-by-step debug of the Source Code, below is the summary of how ELIZA was designed at the code level.

A script file controls all the behavior of ELIZA with text format, which is only 359 lines. In this project, the file is named "doctor.txt." Every line of the script is prefaced by a tag that tells what list it is part of, the tags such as: initial, pre, post, key, etc. The script is used to construct the pre and post substitution lists, the keyword lists, and the decomposition and reassembly patterns, and a synonym matching facility.

The screenshot of the "doctor.txt" file shows in figure 2

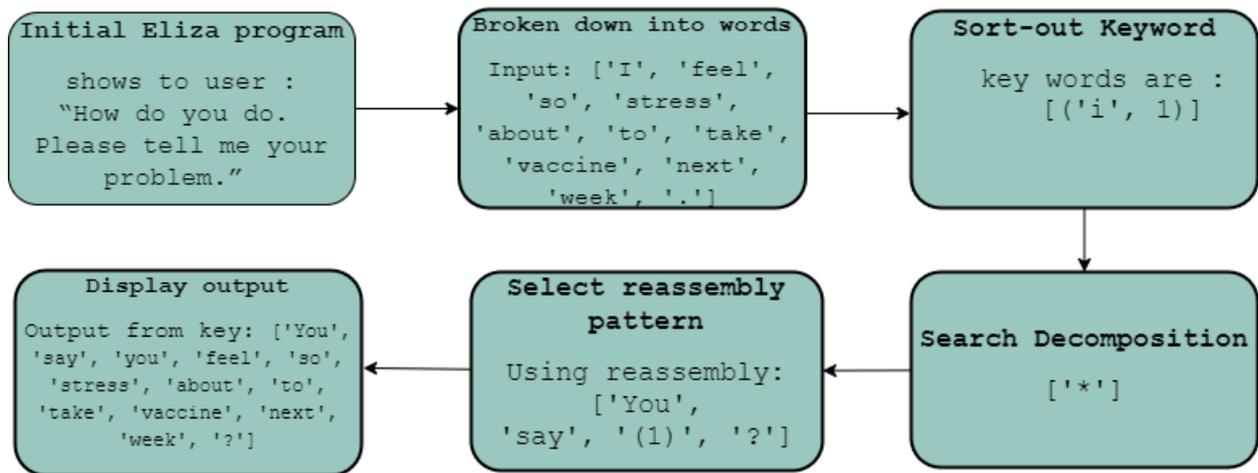


Figure 3: ELIZA Chatbot working process

5.1.1. ELIZA Chatbot Working Process

The ELIZA Chatbot working process consists of the following six main steps as show in figure 3

5.2. Smart-ELIZA Design Overview

For Smart-ELIZA, the design analyzes the whole input sentence’s emotion rather than word by word. So, the deep learning model will be engaged as long as the PwO returns a sentence to the Smart-ELIZA system. Reference the ELIZA original design explained in the above section. The Smart-ELIZA’s emotion detection process should be the extra step between the "Initial ELIZA system” and “Broken down reply into words” steps.

To perform the emotion detection for the user’s input sentence, the DL model plays an essential role by analyzing the PwO’s output and identifying the emotion category. Considering the sentence might contain negation (such as I am not feeling well these days during the lockdown.), it is essential to re-evaluate the sentiment in this situation. If the final emotion belongs to the negative list, then Smart-ELIZA will return a pre-defined text message to PwO; otherwise, continue the original ELIZA rest of the system’s function. The workflow details show as 4

5.3. Process for emotion detection

Regarding text-based emotion detection, generally, it contains five steps to perform the emotion detection. The steps are Data Source selection, Text Pre-processing, Emotion Detection Model Development, Process text using model, and Model evaluation. See 5

1. Data Source Selection: The first step is to select the data source to be analyzed or used as a training dataset. The data source could be any text format of data, for example, tweeter tweets, amazon customer reviews, email messages, or service request tickets.
2. Text Pre-processing: Text pre-processing step is trying to sanitize the text/source data and transform the text with a specific format to train the model file. Depending on what ML model is selected, there are a few sub-processes that will be included. For example: stop words detection, tokenization, part-of-speech (pos) tagging, parsing (syntactic analysis), stemming, and lemmatization.

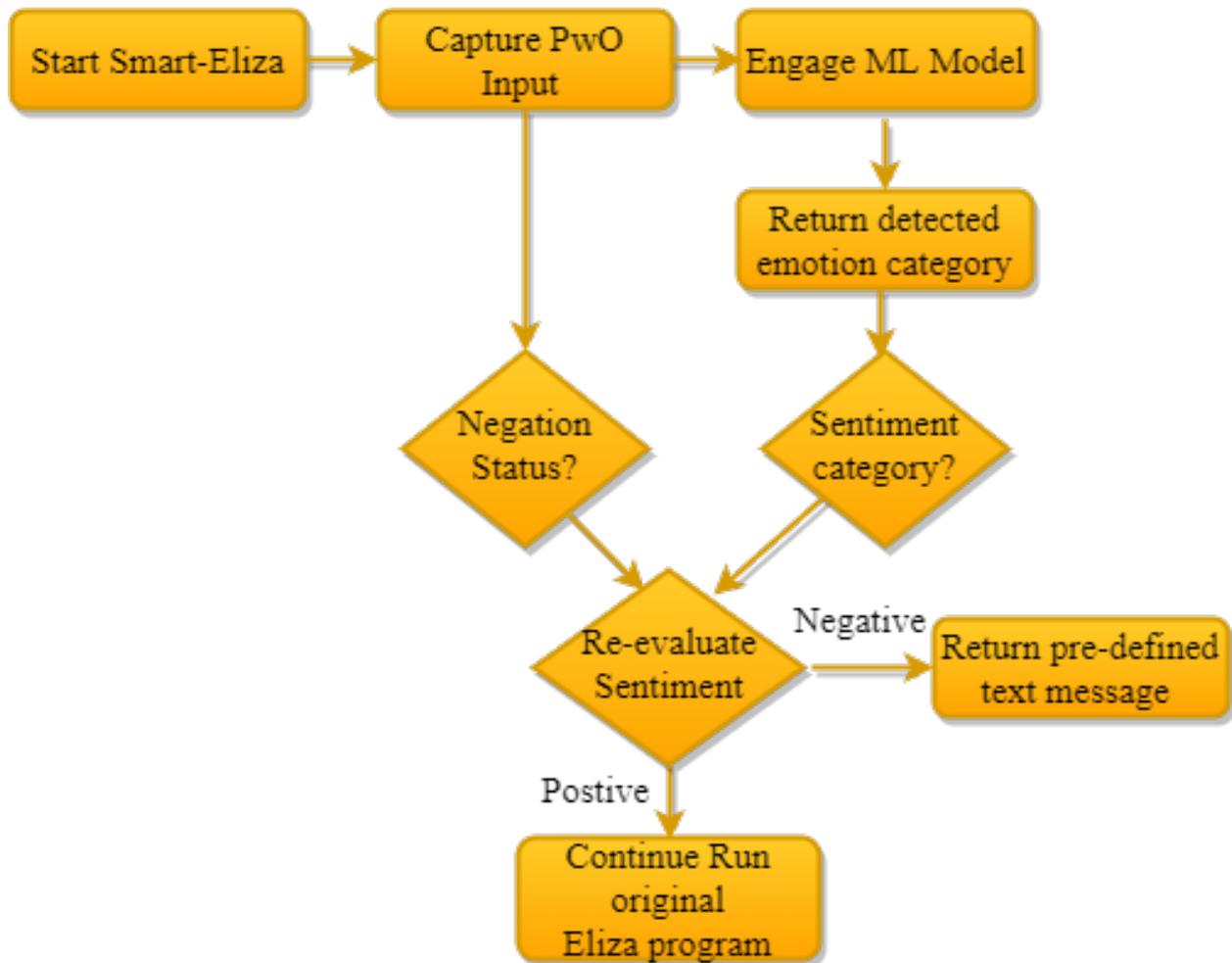


Figure 4: Smart-ELIZA Chatbot emotion detection working process

3. Emotion Detection Model Development: The emotion Detection Model is built using a method and works with several processes in which the emotions are detected in text source . They are corpus-based methods (e.g. lexical affinity) machine learning-based techniques (e.g. statistical methods), knowledge-based techniques (e.g. hand-crafted models, keyword spotting) and hybrid techniques. This project use state of the art deep learning model to implement emotion detection.
4. Process text using the model: Once the model file is generated, this step is to apply the text source with the model file and give the emotion detection result; the output emotion will be one of the emotion categories based on the source/training dataset's categories.
5. Model evaluation Three main evaluation metrics are mainly used in the research: Kappa Coefficient, Precision, and Recall, F-Score.

6. BERT Model implementation

BERT (Devlin et al., 2019) [1] Bidirectional Encoder Representations from Transformers is a method of pre-training language representations; it has been trained as a general-purpose "language understanding" model on a large text corpus (like Wikipedia). BERT model can be used for downstream NLP

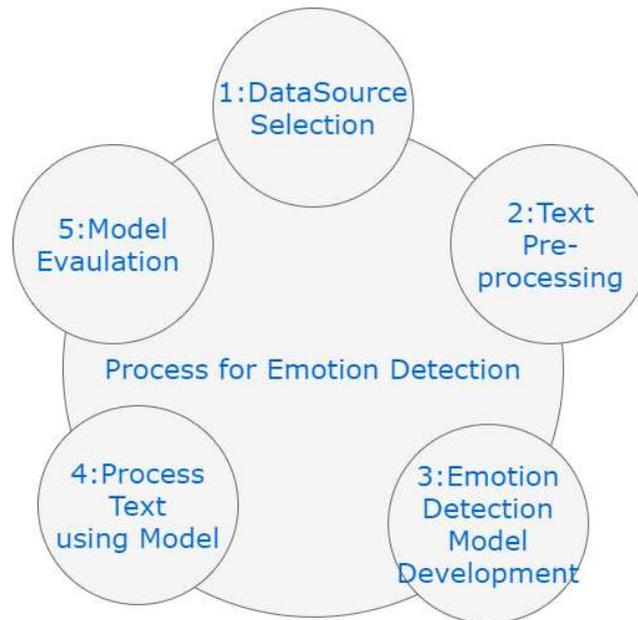


Figure 5: Five steps to perform the emotion detection

tasks (such as question answering, emotion detection). BERT is the first unsupervised, deeply bidirectional system for pre-training NLP. BERT provides a practitioner with a model with a significant amount of general knowledge of the language.

Google releasing several pre-trained models from the paper are available to download. Google release BERT-Base and BERT-Large when the paper first published; in March 2020, Google release 24 smaller BERT models.

Google proposed two main BERT models: BERT-Base contains 12 encoder blocks, 12 head attention, and 110 million parameters. BERT-large contains 24 encoder blocks, 16 head attention, and 340 million parameters.

6.1. Dataset selection

To get better accuracy of the emotion detection, the DL model implement firstly focus also on to find out the best suit emotion datasets for the deep learning BERT model.

Google "GoEmotions have performed another research: A Dataset of Fine-Grained Emotions," Dorottya Demszky introduce "the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral.", they also perform the "high quality of the annotations via Principal Preserved Component Analysis". The GoEmotions have also been tested on the BERTBase model and got the 46% across all the proposed taxonomy. This test shows the generalizability of the data across domains and taxonomies via transfer learning experiments.

6.1.1. GoEmotions dataset overview

GoEmotion GitHub website contains the entire dataset and Source Code that has been used for the research performed by Google. The complete datasets which are shows three CSV format files; the first few columns are text, id, author, subreddit, link_id, parent_id, created_utc, rater_id,

example_very_unclear. The rest of the columns are emotions, the value is either 0 or 1. By mapping emotions with sentiment, all the emotions are shown below:

- "positive": ["amusement", "excitement", "joy", "love", "desire", "optimism", "caring", "pride", "admiration", "gratitude", "relief", "approval"]
- "negative": ["fear", "nervousness", "remorse", "embarrassment", "disappointment", "sadness", "grief", "disgust", "anger", "annoyance", "disapproval"]
- "ambiguous": ["realization", "surprise", "curiosity", "confusion"]

6.2. Generate BERT Model

Feature extraction is one of the approaches to use BERT; this method uses BERT as a feature extraction model. The architecture of the BERT model is preserved, and its outputs are feature input vectors for subsequent classification models to solve the given problem.

This project uses BERT-Base, the pre-trained BERT model, and runs a further train on GoEmotions datasets without substantial task-specific architecture modifications.

Google's research mentioned that GoEmotions had been tested on BERTBase. However, they did not detail how the new Bert Model generated by the GoEmotions dataset and GitHub does not include the new BERT model for download.

Future research has been done regarding how to train BERTBase using GoEmotions to generate a new BERT model. The research is based on reading Google published paper [19] and debug the source code step by step. This part of the research is re-validate and investigating the procedure that has been performed in the Google GoEmotions research.

6.2.1. Detail Experiment Procedure

Below is the summarized the whole procedure based on the experiments locally:

1. Clone the GoEmotions code from [19] locally.
2. Clone the Bert code [19] from GitHub locally and put it under the GoEmotions folder.
3. Download the BERTBase model file from google storage, the folder contains three files:
 - A TensorFlow checkpoint (bert_model.ckpt) containing the pre-trained weights
 - A vocab file (vocab.txt) to map WordPiece to word id.
 - A config file (bert_config.json) specifies the hyperparameters of the model.
4. Install local running environment; the python libraries are listed in their requirements.txt file. There are more than 30 libraries. The TensorFlow version is straightly mentioned within the file.
5. bert_classifier.py is the main file to perform fine-tuning on the GoEmotions dataset on top of BERTBase,
 - Many parameters need to configure to make it work. To centralize the configuration, to enhance the extra code was added into configuration_constant.py, which specifies the variable: DATA_DIR, SENTIMENT_FILE, EMOTION_FILE, etc.

Table 3
BERTBase Model runs on GoEmotions Dataset

eval_accuracy = 0.4354155	eval_accuracy = 0.43209878
eval_loss = 2.545235	eval_loss = 2.5506651
global_step = 10852	global_step = 10852
loss = 2.5417128	loss = 2.5470078

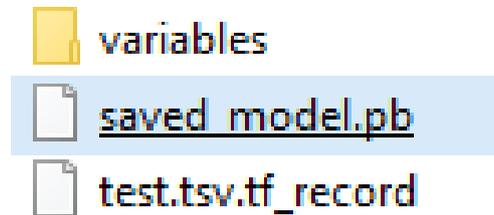


Figure 6: New BERT Model file folder structure

ant > output_dir > 1618244107 > variables

Name	Type	Size	Date modified
variables.data-00000-of-00001	DATA-00000-OF-00001 ...	427,750 KB	2021-04-12 05:15 PM
vocab.txt	Text Document	227 KB	2020-02-20 06:57 PM
variables.index	INDEX File	9 KB	2021-04-12 05:15 PM
config.json	JSON File	1 KB	2021-04-11 10:32 PM

Figure 7: New BERT Model file variables folder

- Parameter “do_export” within the “flags.DEFINE_bool” need to set as True to save the model file locally.
- Other “train_batch_size”, “learning_rate”, “num_train_epochs,” and the rest of the parameters are kept the same as the paper’s suggestion.

6. The finally working parameters are showing as below:

```
python bert_classifier.py -vocab_file=C:\BERT-Base-model\vocab.txt
-bert_config_file=C:\BERT-Base-model\bert_config.json -output_dir=C:\output_dir
```

6.2.2. The New BERT Model Running Result

The training model steps have been running twice; on average, it took more than 20 hours to complete the training on a local DELL laptop (CORE i7 8th Gen without GPU) with an accuracy of about 43.4%. This accuracy is very similar to Google Research paper’s result: 46%. The table 3 shows the running accuracy result.

The Saved Model contains a complete TensorFlow program, including trained parameters (i.e, tf.Variables) and computation.

The model files generated are under output_dir with two files and one variables folder from the implementation. The screenshot of all files are showing in FIGURE 7 and 6.

Major files' function are :

- The saved_model.pb file stores the actual TensorFlow program and a set of named signatures.
- The index file is a string-string immutable table.
- variables.data-00000-of-00001 is TensorBundle collection which stores the actual values of all the variables is about 427MB.

6.3. Apply New BERT model and Negation into ELIZA system

From the previous step, the New Bert model is generated. This section focuses on applying this new Bert model to make emotion detection for ELIZA chat input. A new “emotion_prediction_class” was created delicate to predict the chat context.

Several experiments have been performed to figure out how to use the above TensorFlow model file.

The first experiment uses “-do_predict=true” to run the model to predict classifier base on the BERT website. However, this parameter will engage rerunning training and evaluation once again, which is about 20 hours long, which would not apply to the Chatbot system’s requirement.

Second experiments is to investigation from GoEmotions bert_classifier.py and try to how the Model file is saved.

The final working prediction detailed experiment steps are:

1. Load Graph (tf.graph) into TensorFlow session
2. Load the exported model to a TensorFlow session with tag is [‘serve’] so it is available in the current context.
3. Pre-process the Chat text by adding extra “data_type ==‘chat_prediction’” into GoEmotions “DataProcessor” class so it will create examples base on chat content.
4. Perform from single example to tf.train.Features format with Feature Tensors structure required for next step.
5. Adjust the tf.train.Features into tf.train.Example and feed into prediction.
6. Figure out the prediction session run “feed_dic” structure from GoEmotions TensorFlow structure settings.

GoEmotions exist class like “DataProcessor” for this prediction; the extra step is to transfer the input text into CSV format and save it into local disk. The negation been applied to this sentiment result by reading the negative and positive list been defined from GoEmotions (see “GoEmotions dataset overview”).

7. Conclusions and Future Work

7.1. Evaluate BERT version of Smart-ELIZA function

For the the function evaluation,when the user inputs every sentence, the system will perform the emotion negation detection. Once the final sentiment is negative, the return will be using the pre-defined sentence “I am sorry to hear that, how can I help you?”. The system runs shows as in FIGURE ??

```
(base) C:\0.Data_Science_and_Data_analytic\0.CIT\AI_Master\4.FourthTerm\Code\Eliza\eliza_ml_goemotions_bert>python -W ignore eliza_with_ml_bert.py
How do you do. Please tell me your problem.
> I am doing great.
2021-05-08 17:07:21.285198: I tensorflow/core/platform/cpu_feature_guard.cc:142] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX AVX2
***-----DEBUG-----*** emotion detection result is : admiration
***-----DEBUG-----*** negation_status is : False
***-----DEBUG-----*** User is positive sentiment
Is it because you are doing great that you came to me ?
> I am worry about the exam.
***-----DEBUG-----*** emotion detection result is : approval
***-----DEBUG-----*** negation_status is : False
***-----DEBUG-----*** User is positive sentiment
How long have you been worry about the exam ?
> I am not feeling well about it.
***-----DEBUG-----*** emotion detection result is : neutral
***-----DEBUG-----*** negation_status is : True
***-----DEBUG-----*** User is positive sentiment
Do you believe it is normal to be not feeling well about it ?
> I am so worried about exam.
***-----DEBUG-----*** emotion detection result is : sadness
***-----DEBUG-----*** negation_status is : False
***-----DEBUG-----*** User is negative sentiment
I am sorry to hear that,how can I help you?
> I am worry about my exam
***-----DEBUG-----*** emotion detection result is : optimism
***-----DEBUG-----*** negation_status is : False
***-----DEBUG-----*** User is positive sentiment
Your exam ?
> I am worried about my exam
***-----DEBUG-----*** emotion detection result is : sadness
***-----DEBUG-----*** negation_status is : False
***-----DEBUG-----*** User is negative sentiment
I am sorry to hear that,how can I help you?
```

Figure 8: Smart-ELIZA function test using new BERT Model

Another finding is using a different way of expressing "worry": "I am worried about my exam," "I am so worried about my exam," "I am worried about my exam," it returns emotion detection with "approval," "sadness" and "optimism." The above test result leads the space to improve the model file. However, the answer from the original ELIZA system is still acceptable to the user. For example, the answers are "How long have you been worry about the exam?," "You exam?".

8. Conclusion

The project achieved a successful result with the state-of-art BERT model successfully detected ELIZA Chatbot input and returned a pretty high accuracy score considering this project using much more advanced emotion categories in the training datasets. In the meantime, the BERT version Smart_ELIZA returns the answer simultaneously just after the user input the sentence. The emotion detection steps works as seamless with ELIZA system. This also indicates BERT Model can be used in the STOP system for real users.

9. Future Work for this project

From the technical point of view, the future work that could be done is to improve BERT model accuracy; for example, another experiment could be done by combining the BERT model with the biLSTM model and testing the accuracy because these two models' combination has not been used in a few other sentiment analyses research so far. However, no relevant research has been done for

emotion detection.

Another potential work could be using this project's emotion detection and the user's chat history to predict the new message input and generate a personalized message to PwO.

10. Acknowledgments

This research was conducted with the financial support of the Horizon 2020 project STOP Obesity Platform under Grant Agreement No. 823978 and at the ADAPT SFI Research Centre at Cork Institute Of Technology. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

The project was partially supported by the Horizon 2020 project ITFLOWS under Grant Agreement No. 882986 and Dell-EMC (Ireland).

References

- [1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [2] M. M. Van Pinxteren, M. Pluymaekers, J. G. Lemmink, Human-like communication in conversational agents: a literature review and research agenda, *Journal of Service Management* (2020).
- [3] A. Deshpande, A. Shahane, D. Gadre, M. Deshpande, P. M. Joshi, A survey of various chatbot implementation techniques, 2017.
- [4] J. BARON, 2017 messenger bot landscape, a public spreadsheet gathering 1000+ messenger bots (2017).
- [5] B. Liu, S. S. Sundar, Should machines express sympathy and empathy experiments with a health advice chatbot, *Cyberpsychology, Behavior, and Social Networking* 21 (2018) 625–636.
- [6] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine, *Communications of the ACM* 9 (1966) 36–45.
- [7] B. Heller, M. Proctor, D. Mah, L. Jewell, B. Cheung, Freudbot: An investigation of chatbot technology in distance education, in: *EdMedia+ Innovate Learning*, Association for the Advancement of Computing in Education (AACE), 2005, pp. 3913–3918.
- [8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
- [9] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, P. Agrawal, Understanding emotions in text using deep learning and big data, *Computers in Human Behavior* 93 (2019) 309–317.
- [10] S. Alhagry, A. A. Fahmy, R. A. El-Khoribi, Emotion recognition based on eeg using lstm recurrent neural network, *Emotion* 8 (2017) 355–358.
- [11] Z. Yin, M. Zhao, Y. Wang, J. Yang, J. Zhang, Recognition of emotions using multimodal physiological signals and an ensemble deep learning model, *Computer methods and programs in biomedicine* 140 (2017) 93–110.
- [12] Q. T. Nguyen, T. L. Nguyen, N. H. Luong, Q. H. Ngo, Fine-tuning bert for sentiment analysis of vietnamese reviews, arXiv preprint arXiv:2011.10426 (2020).
- [13] M. Munikar, S. Shakya, A. Shrestha, Fine-grained sentiment classification using bert, in: 2019

- Artificial Intelligence for Transforming Business and Society (AITB), volume 1, IEEE, 2019, pp. 1–5.
- [14] Y. Zhang, Z. Rao, Deep neural networks with pre-train model bert for aspect-level sentiments classification, in: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), IEEE, 2020, pp. 923–927.
 - [15] A. F. Adoma, N.-M. Henry, W. Chen, Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition, in: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2020, pp. 117–121.
 - [16] J. G. Harb, R. Ebeling, K. Becker, A framework to analyze the emotional reactions to mass violent events on twitter and influential factors, *Information Processing & Management* 57 (2020) 102372.
 - [17] A. F. Adoma, N.-M. Henry, W. Chen, N. R. Andre, Recognizing emotions from texts using a bert-based approach, in: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2020, pp. 62–66.
 - [18] H. Al-Omari, M. A. Abdullah, S. Shaikh, Emodet2: Emotion detection in english textual dialogue using bert and bilstm models, in: 2020 11th International Conference on Information and Communication Systems (ICICS), IEEE, 2020, pp. 226–232.
 - [19] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, *arXiv preprint arXiv:2005.00547* (2020).

An RT-qPCR Data Analysis Platform

Thomas Krause^a, Elena Jolkver^a, Sebastian Bruchhaus^a, Michael Kramer^b and Matthias Hemmje^c

^aUniversity of Hagen, Faculty of Mathematics and Computer Science, Hagen, Germany

^bImmBioMed Business Consultants GmbH & Co. KG, Pfungstadt, Germany

^cResearch Institute for Telecommunication and Cooperation (FTK), Dortmund, Germany

Abstract

Gene expression data produced by RT-qPCR instruments is becoming increasingly important in laboratory diagnostics. The evaluation and management of this data often involve manual steps, as current software does not map the entire laboratory process. Workflow management systems (WMS) offer flexibility and expandability and can thus take into account individual requirements in laboratories. In addition to the mere evaluation of individual gene expression data, the requirements also include, for example, the linking with medical data, the generation of reports, the storage for meta-analyses, and the archiving of all relevant data. In addition, regulatory requirements such as those of the “In Vitro Diagnostic Regulation” (IVDR) must be taken into account for all of these requirements. This paper proposes a conceptual architecture consisting of a WMS for processing data, as well as an ecosystem of different components for the more advanced requirements. It also outlines a planned prestudy as the next step towards the implementation of such a system.

1. Introduction and Motivation

Genomics is an elementary branch of biology that deals with the genetic material of organisms (the “genome”). The subfield of genetics deals with individual sections of the genome, the “genes” and their interactions. Genes are information carriers that describe the structure of proteins, which are produced in a cell to carry out elementary functions.

“Gene expression” describes the mechanism, by which the information of a gene is transferred into protein. The extend and pattern of which genes (from the multitude of genes in the genome) are being “expressed” in a cell at a certain time point defines the pattern and degree of cellular protein production and thus the potential involvement of these proteins in cellular physiology and pathophysiology. The degree of the expression of a given gene depends on the cell type as well as its activation and differentiation status. Finally, gene expression is subject to external influences.

Therefore, the measurement of gene expression in cells can provide important diagnostic information and as such is being developed into a promising part of medical diagnostics being performed by appropriately equipped medical laboratories.

The processing of gene expression data in laboratories is supported by IT systems. Such IT systems aim to achieve a high degree of automation in order to reduce sources of error, increase throughput, reduce turnaround times, and allow for monitoring of the entire process under aspects of quality assurance. In this context, special regulatory challenges apply, which are specified, e.g. in the EU

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ thomas.krause@fernuni-hagen.de (T. Krause); elena.jolkver@studium.fernuni-hagen.de (E. Jolkver); sebastian.bruchhaus@fernuni-hagen.de (S. Bruchhaus); m.kramer@immbiomed.de (M. Kramer); mhemmje@ftk.de (M. Hemmje)

ORCID 0000-0003-4912-1703 (T. Krause); 0000-0003-4711-6708 (E. Jolkver); 0000-0002-7783-2636 (S. Bruchhaus); 0000-0001-8293-2802 (M. Hemmje)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Regulation for In Vitro Diagnostics (IVDR) [1], in specific DIN/ISO documents [2], or in the so-called “MIQE guidelines” [3].

1.1. Reverse-Transcription Quantitative PCR

Within a cell, the genes are defined sections of the DNA sequence. These can be “read” and transferred into messenger RNA (“mRNA”) by certain enzymes. This process is called “transcription.” Only in a second step, the mRNA sequence is being “translated,” into an amino acid sequence (the protein).

The extent to which genes are transcribed and translated varies from cell-type to cell-type as well as from the differentiation/activation status of the respective cell type. The latter not only depends on internal factors within the respective cell type but also a multitude of factors in the cellular micro-environment, whether they stem from endogenous processes within the body or environmental factors (like food ingredients, environmental toxins, or physical factors, like irradiation).

The strength of gene expression can be measured by determining how many transcription products are currently present as mRNA in a sample at a specific time point. Various methods have been established for this purpose. One of these methods is reverse-transcription quantitative PCR (“RT-qPCR”)¹.

For RT-qPCR, mRNA is first transcribed back into DNA (“reverse transcription”) and then amplified in several cycles by polymerase chain reaction (PCR). By using sequence-specific so-called “primers” the amplification process is restricted to a specific target sequence corresponding to the gene of interest. By using fluorescent indicators (“probes”), the increase of DNA can be measured quantitatively during PCR cycling, i.e. in “real-time.” This real-time approach is in contrast to so-called “end-point methods” where the DNA is only detected after completing a certain number of amplification rounds which makes them unsuitable for optimal expression quantification.

1.2. Delta Delta C_q method

RT-qPCR instruments process several samples and genes in parallel in different test tubes (“wells”). With each amplification cycle, the fluorescent signals of the wells are recorded. Plotted as a curve the fluorescent signal is exponential if a sample contains the target sequence as each cycle approximately doubles the amount of DNA. In the first cycles, the signal is still so weak that it is masked within the method-related background noise. Only when a certain threshold value is exceeded does the signal stand out clearly from the background noise. The corresponding cycle number at which the signal exceeds the threshold is known as the C_q value [5].² The higher the number of initial copies in a sample, the fewer cycles are needed to exceed the threshold and the lower the associated C_q value. The C_q value thus represents a measurable unit that correlates with gene expression.

Unfortunately, C_q values cannot be directly compared between different runs, samples, genes, instruments, or laboratories, as it is also dependent on many other factors. E.g. the measured value depends on how diluted the sample is or even how effective the amplification process is, which in turn depends on the equipment and reagents used. For better comparability, the measured C_q values can first be compared to a control gene for which the degree of activation is known and as constant as possible. Such genes are often termed “housekeeping genes.” Both genes can be observed in parallel for the same sample. Taking the difference between the C_q value calculated for the control gene and

¹Sometimes misunderstandings arise from the fact “rt” could stand for “real-time” or for “reverse transcription.” In fact, the method combines both. To avoid confusion the term “RT-qPCR” should be used [4] per MIQE guidelines. The term RT-PCR should be used to describe PCR after reverse transcription (i.e. starting from (m)RNA) and not real-time PCR. Adams has provided a recent educational article on the aspects of “RT-PCR, qPCR, and RT-qPCR” [4].

²The symbol C_t is also often found in the literature, the MIQE guidelines [3] recommend the use of C_q though

the C_q value for the target gene, the ΔC_q value is obtained, which expresses whether the target gene has stronger or weaker expression [5]. The value thus determined can be better compared across different samples and runs on the same instrument.

From a scientific and diagnostic perspective, it may be interesting to compare the ΔC_q values for the gene of interest with control samples to determine differences in gene expression. For example, to test whether a particular drug affects gene expression and to what extent it does so, a treated sample and an untreated sample can each be compared with a control sample. To do this, one again calculates a difference between the ΔC_q values of the samples under investigation for the target genes and the ΔC_q values in the control sample for the same target genes. The results are $\Delta\Delta C_q$ values expressing relative enhancement or attenuation of gene expression compared to the control [5]. This method is therefore referred to as the $\Delta\Delta C_q$ method.

Since the C_q values and the delta values are logarithmic (each increment by one corresponds to a halving of the gene sequence under study), it is often desirable in practice to bring these values back to a linear scale. For this purpose, the conversion $2^{-\Delta\Delta C_q}$ can be used to determine the so-called “fold change” value [5]. On this linear scale, a value of 1 then graphically corresponds to “no change in gene expression of the target gene” compared to the control sample, while, for example, a value of 2 would express a doubling compared to the control sample.

The fold change formula above assumes that the PCR efficiency of the genes of interest as well as the control gene is the same. A more accurate calculation can be performed, when the individual efficiencies are measured and taken into account. The formula also assumes a single control gene. The use of only one control gene is however insufficient for most use cases [6]. To address both shortcomings the formula

$$\frac{E_{goi}^{\Delta C_{q,goi}}}{\sqrt[n]{\prod_z^n E_{refz}^{\Delta C_{q,refz}}}}$$

should be used, where $\Delta C_{q,refz}$ are the ΔC_q values of the control genes and E refers to the PCR efficiencies of the target and control genes respectively. Several algorithms have been developed to select the best reference genes from a set of candidates [7].

1.3. Processing Pipelines

Data analysis for genomics can involve a multitude of computer programs written in different programming languages. A flow of data between these programs orders them, typically forming a directed acyclic graph (DAG) of processing steps. The input for each step is an output of either its antecedent step or the original data source. The data is funneled through a metaphorical processing pipeline. Standardized data formats accomplish a separation of concerns between the individual steps. Such steps may be regarded as functional units or black boxes. They are the building blocks of an abstract automated workflow.

For example, a simple pipeline for qPCR might consist of a step that reads in a file in a particular format and transforms it into a uniform schema. A second step could then analyze the data based on this uniform schema and produce a result. The first processing step might later be replaced with another supporting a different data source and input format.

The considerable effort of orchestrating and documenting such tasks has prompted the development of many Workflow Management Systems (WMS) [8]. They are used for scalable batch processing of recurring interdependent tasks on-premise and in the cloud. WMSs handle monitoring, logging, dependency resolution, and deployment of workflows. Furthermore the abstract, highly regular DAG

structure of workflows reduces complexity. This facilitates the transparency and reproducibility of analytical results. Workflow definition languages like CWL and OpenWDL [9] allow versioning, publishing, and sharing of workflows. Some WMSs are specifically intended for the analysis of biomedical data, e.g. the Galaxy platform [10] for which some RT-qPCR tasks have already been developed [11].

WMS can be used to establish traceability, accountability, and data integrity. They log different kinds of metadata for documentation and monitoring purposes. Monitoring, in particular, is necessary for corrective and preventive actions (CAPA). These are part of the quality management requirements in the IVDR and the EU's proposed rules on artificial intelligence [12].

1.4. Requirements for Analysis Software

To understand the field of qPCR diagnostics and the real world requirements for analysis software in laboratories, a board-certified physician for laboratory medicine ("clinical pathologist"), head of a clinical pathology lab active in indirect patient care, and head of a laboratory consulting firm, Professor Michael Kramer, was invited to co-author this paper and to provide expert knowledge on the field.

Based on this practical experience, an initial review of relevant literature, legal frameworks, and industry guidelines, several requirements were identified that are essential for efficient use of qPCR analysis software in a diagnostic setting:

- Where possible, tasks should be automated to save time, increase throughput, and reduce errors
- All basic user interactions with the system should happen through graphical user interfaces as laboratory staff cannot be assumed to be IT experts
- Data processing and analysis should be extendable as well as customizable to react to new needs
- Integration with existing Laboratory Information Management System (LIMS) should be possible
- Basic analysis of RT-qPCR data should support relative quantification analysis (e.g. $\Delta\Delta C_q$ method [5])
- Relative quantification analysis should support multiple control genes as well as variable PCR efficiencies
- Extensibility should allow additional analysis methods such as standard curve and copy-number-variation [13]
- Analysis runs need to be reproducible
- Results need to be stored in a way to allow subsequent additional analyses
- Quality control is essential [14] and required by legal frameworks
- All results should be easily archivable, data formats must remain readable even after many years
- Legal requirements and industry guidelines like the IVDR and MIQE should be followed where possible

The last point is especially important. To be able to use the solution for diagnostic purposes in the European Union the IVDR has to be taken into account starting next year. Among many other requirements, the IVDR mentions reproducibility, reliability, and performance as key targets for software. It also requires solutions to be developed according to the state of the art and taking into account the software lifecycle, risk management, IT security, verification, and validation. Besides technical requirements, it demands several administrative measures such as naming a person responsible for compliance with suitable professional experience.

2. State of Art

qPCR instruments are supplied with basic software packages that enable the measurement of fluorescent changes, calculations of quantification cycle (C_q) values, and relative quantity determination [15]. Not all devices support the researcher with additional features, like qPCR efficiency correction, normalization to multiple reference genes, averaging and statistical tests. Therefore, several instrument-independent tools have been developed, and most of them can roughly be categorized as software running under Windows, web-based tools, or packages for the R computing environment [16].

For the determination of relative expression levels, the C_q value is first calculated from the raw fluorescence values. For this, several pre-processing steps might be needed like noise reduction, curve smoothing, removal of outliers, normalization, and curve fitting. Based on C_q , the different gene expression levels need to be normalized on one or multiple references and the final $\Delta\Delta C_q$ value determined. The ability of the software to handle error propagation in these steps as well as the subsequent support in statistical analysis and graphical visualization is of high importance for lab staff lacking statistical experience. Ideally, the software supports all of the aforementioned steps and is compliant with MIQE guidelines. Such an all-rounder is the qpcR package for R [17], which, combined with R's rich statistical and graphical package infrastructure can provide all required analyses. However, the usage of R requires some proficiency in scripting, which can impose an initial barrier for less programming-savvy researchers. Out of the various tools developed for qPCR analyses, only a few support multiple of the aforementioned calculations.

QPCR [18] comprises a parser to import generated data from qPCR instruments and includes technical and biological replicate handling, incorporation of gene-specific efficiency, normalization using single or multiple reference genes, inter-run calibration, and fold change calculation. Moreover, the application supports the assessment of error propagation throughout all analysis steps and allows conducting statistical tests on biological replicates. Results can be visualized in customizable charts and exported for further investigation.

qBase [7] contains several algorithms for reference gene selection, normalization, quality control, inter-run-calibration, as well as copy number variation analysis. For reference gene selection and normalization, it offers the geNorm algorithm. geNorm ranks candidate genes by their stability and then picks a suitable number of genes to be used [6]. For normalization, a geometric averaging of these selected reference genes is then used. Other tools specifically for the selection of reference genes include BestKeeper [20] and NormFinder [21]. Both are available as plugins for Microsoft Excel.

While the analysis and normalization procedures offered by these tools are useful, they often only represent only a small part of the overall workflow from instrument setup to result visualization and beyond. Software included with the instruments is also usually proprietary so that they cannot be easily modified or used with other instruments [16]. Most tools seem to be tailored for research purposes and not for diagnostics, requiring several manual steps instead of automatization and integration. Usage for diagnostics in the European Union would also require compliance with IVDR which as of today is not available for the majority of tools.

3. Prestudy Design and Research Questions

Based on the perceived lack of freely available software that can be used for diagnostic purposes, considering especially the upcoming hardened IVDR legislation, it seems worthwhile to design, implement and certify such a system on an open-source base, to allow easier collaboration between

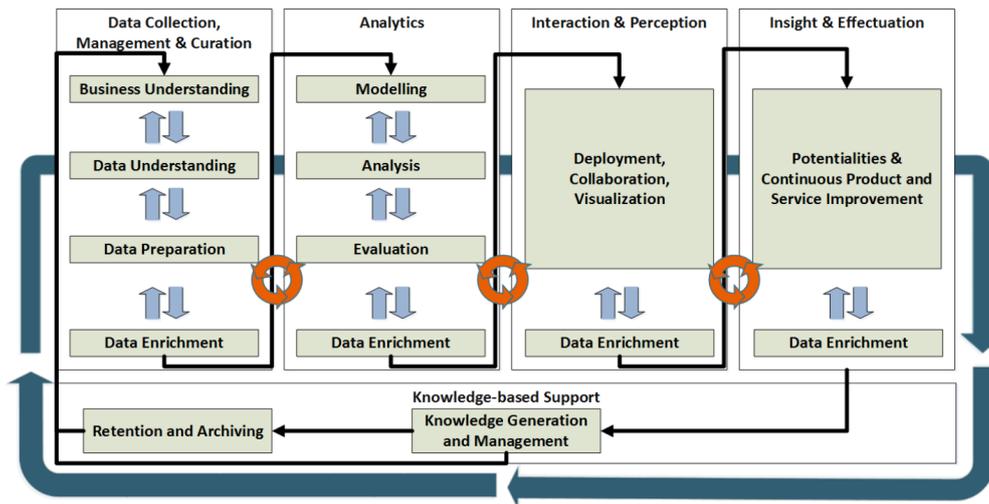


Figure 1: CRISP4BigData Reference Model

research teams and to reduce the time needed for new research to be used in diagnostics. While some basic requirements, existing software, and legal requirements have been outlined, further investigation into the current processes and resulting requirements seems necessary before implementation is possible. To support this, a prestudy is currently being planned that should answer the following research questions:

- What does a typical analysis process for RT-qPCR data look like in a laboratory?
- Which manual activities can be automated?
- What software already exists to automate the entire process or individual parts of it?
- What regulatory and technical requirements must be considered?
- How must a system be structured to meet the requirements and enable a high degree of automation?

4. Proposed Conceptual Architecture

While a detailed technical design and implementation require the prestudy that was outlined, it is possible to outline a conceptual architecture on a high level based on the known requirements so far, which can then be further detailed in the prestudy. For this goal the CRISP4BigData [22] reference model shown in fig. 1 was used as a basis. CRISP4BigData combines the Cross-industry Standard Process for Data Mining (CRISP-DM) with the IVIS4BigData model [23]. Figure 2 shows the conceptual architecture derived from the requirements and mapped to the reference model.

At the core of the architecture lies a WMS that allows a dynamic processing workflow where each step and the workflow as a whole can be customized. As described previously there are already well-known, good working tools for some parts of the qPCR workflow. Rebuilding these from scratch would be very time-consuming. Using a configurable workflow allows the reuse of these tools.

The processing workflow that is shown in fig. 2 has been separated into distinct phases. These phases map directly to the CRISP4BigData reference model and should be seen as a general template for a typical qPCR analysis process that can be customized by the laboratory. In practice, they will consist of one or more concrete processing tasks executed in sequence or parallel inside the WMS.

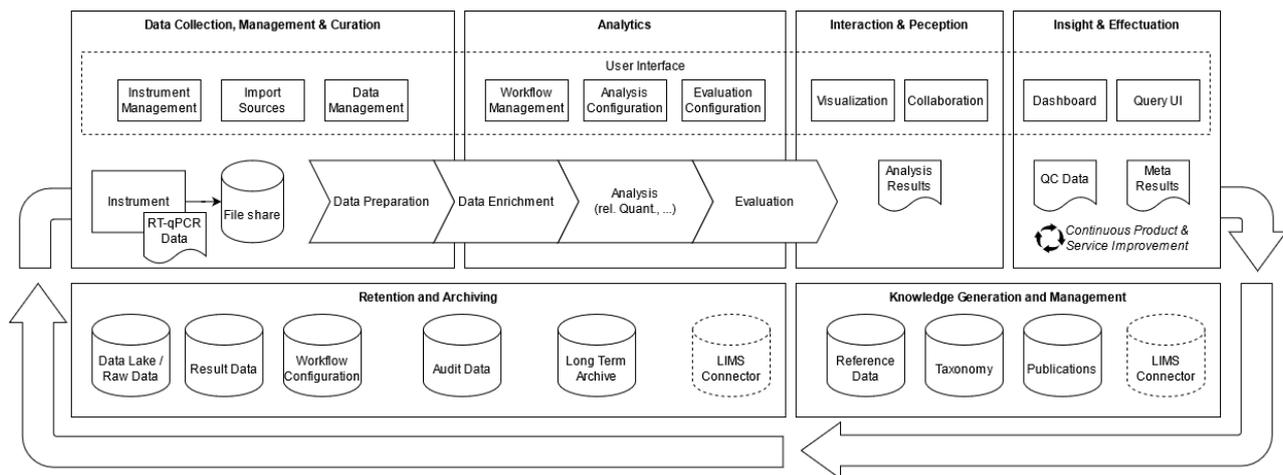


Figure 2: Conceptual Architecture

The workflow starts after data created by the instrument is read automatically from a file share or uploaded manually from the user interface. Analysis settings can be specified at upload time or pre-configured. During import, the data is transformed into an internal unified data structure that allows subsequent processes to work independently of the original data format (part of the “Data Preparation” phase). Data from previous runs is aggregated when required to allow for inter-run calibration and further augmented by loading reference data or computing intermediate results required by subsequent analysis in the “Data Enrichment” phase. The “Analysis” phase contains the analysis steps such as determining the relative quantification amounts. In the last phase (“Evaluation”) the analysis results are interpreted and transformed into desired outputs.

The artifacts created by the workflow are available for visualization and collaboration via a user interface. The same user interface is also available to administer the system as a whole, which can be seen in the upper part of fig. 2. Long-term insights from several runs as well as quality control measures and monitoring can be found in the “Insights & Effectuation” layer. New Knowledge discovered at this level, like suitable reference genes for analysis, can be stored for future use as part of the “Knowledge Generation and Management” layer.

Lastly, all input data, output data, intermediate results, and metadata such as the processing parameters used are persisted to enable reproducibility of the results at any time later. Legal requirements regarding storage and deletion periods are ensured by technical and organizational measures. These data stores are found at the bottom of the diagram as part of the “Retention and Archiving” layer. For long-term storage, suitable standards such as OAIS [24] are used to ensure future readability. For integration with existing LIMS, a connector is displayed in the same layer to accommodate the need for synchronization between the systems.

Some of the aspects mentioned above such as reproducibility are also part of IVDR compliance. Using a layered architecture with individual components can also help to provide other aspects such as reliability and performance because components can be used to parallelize workloads or to provide redundancy. The pre-study will aim to address additional points of the IVDR.

5. Evaluation

The conceptual architecture can be evaluated on a preliminary basis by comparing the identified requirements with the proposed solution. Further evaluation is planned at a later stage during the

prestudy.

The requirement of a graphical user interface has been addressed in the architecture by including it as an explicit component. The desire for flexible and configurable processing and integration with existing software is fulfilled by using a WMS as the core of the solution. This also partially addresses the subsequent points for the specific types of analysis to be used as the workflow approach allows any kind of analysis to run in principle. Furthermore, it was proposed to use existing tools and algorithms where possible inside the workflow which would allow the mentioned analysis methods to be executed. WMSs also facilitate reproducibility as the concrete settings used in a workflow are easily saved to allow re-execution. Storage for later analysis and for archiving is part of the architecture. For archival a specific format was suggested. Lastly, compliance with the IVDR was discussed. The specific requirements for compliance are a complex topic and need to be evaluated separately.

References

- [1] The European Parliament and the Council of the European Union, In vitro diagnostic regulation: IVDR, 2017. URL: <http://data.europa.eu/eli/reg/2017/746/2017-05-05>.
- [2] DIN, Medizinische laboratorien - anforderungen an die qualität und kompetenz, 2014.
- [3] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments, *Clinical Chemistry*, 55 2009 611–622. doi:10.1373/clinchem.2008.112797.
- [4] G. Adams, A beginner's guide to RT-PCR, qPCR and RT-qPCR, *The Biochemist*, 42 2020 48–53. doi:10.1042/BIO20200034.
- [5] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-(\Delta\Delta C_t)}$ method, *Methods*, 25 2001 402–408. doi:10.1006/meth.2001.1262.
- [6] J. Vandesompele, K. de Preter, F. Pattyn, B. Poppe, N. van Roy, A. de Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, *Genome Biology*, 3 2002 RESEARCH0034. doi:10.1186/gb-2002-3-7-research0034.
- [7] J. Hellemans, G. Mortier, A. de Paepe, F. Speleman, J. Vandesompele, qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, *Genome Biology*, 8 2007 R19. doi:10.1186/gb-2007-8-2-r19.
- [8] J.M. Perkel, Workflow systems turn raw data into scientific knowledge, *Nature*, 573 2019 149–150. doi:10.1038/d41586-019-02619-z.
- [9] J. Vivian, A.A. Rao, F.A. Nothhaft, C. Ketchum, J. Armstrong, A. Novak, J. Pfeil, J. Narkizian, A.D. Deran, A. Musselman-Brown, H. Schmidt, P. Amstutz, B. Craft, M. Goldman, K. Rosenbloom, M. Cline, B. O'Connor, M. Hanna, C. Birger, W.J. Kent, D.A. Patterson, A.D. Joseph, J. Zhu, S. Zaranek, G. Getz, D. Haussler, B. Paten, Toil enables reproducible, open source, big biomedical data analyses, *Nature Biotechnology*, 35 2017 314–316. doi:10.1038/nbt.3772.
- [10] B. Grüning, J. Chilton, J. Köster, R. Dale, N. Soranzo, M. van den Beek, J. Goecks, R. Backofen, A. Nekrutenko, J. Taylor, Practical computational reproducibility in the life sciences, *Cell Systems*, 6 2018 631–635. doi:10.1016/j.cels.2018.03.014.

- [11] N. Zanardi, M. Morini, M.A. Tangaro, F. Zambelli, M.C. Bosco, L. Varesio, A. Eva, D. Cangelosi, PIPE-t: A new galaxy tool for the analysis of RT-qPCR expression data, *Scientific Reports*, 9 2019 17550. doi:10.1038/s41598-019-53155-9.
- [12] European Parliament and Council, Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act), 2021. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>.
- [13] C.A. Heid, J. Stevens, K.J. Livak, P.M. Williams, Real time quantitative PCR, *Genome Research*, 6 1996 986–994. doi:10.1101/gr.6.10.986.
- [14] S.C. Taylor, K. Nadeau, M. Abbasi, C. Lachance, M. Nguyen, J. Fenrich, The ultimate qPCR experiment: Producing publication quality, reproducible data the first time, *Trends in Biotechnology*, 37 2019 761–774. doi:10.1016/j.tibtech.2018.12.002.
- [15] S. Bustin, A. Bergkvist, T. Nolan, In silico tools for qPCR assay design and data analysis, *Methods in Molecular Biology (Clifton, N.J.)*, 760 2011 283–306. doi:10.1007/978-1-61779-176-5_18.
- [16] S. Pabinger, S. Rödiger, A. Kriegner, K. Vierlinger, A. Weinhäusel, A survey of tools for the analysis of quantitative PCR (qPCR) data, *Biomolecular Detection and Quantification*, 1 2014 23–33. doi:10.1016/j.bdq.2014.08.002.
- [17] A.-N. Spiess, qpcR: Modelling and analysis of real-time PCR data, 2018. URL: <https://CRAN.R-project.org/package=qpcR>.
- [18] S. Pabinger, G.G. Thallinger, R. Snajder, H. Eichhorn, R. Rader, Z. Trajanoski, QPCR: Application for real-time PCR data management and analysis, *BMC Bioinformatics*, 10 2009 268. doi:10.1186/1471-2105-10-268.
- [19] What makes qbase+ unique?, 18.09.2018. URL: <https://www.qbaseplus.com/features>.
- [20] M.W. Pfaffl, A. Tichopad, C. Prgomet, T.P. Neuvians, Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–excel-based tool using pair-wise correlations, *Biotechnology Letters*, 26 2004 509–515. doi:10.1023/B:BILE.0000019559.84305.47.
- [21] C.L. Andersen, J.L. Jensen, T.F. Ørntoft, Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets, *Cancer Research*, 64 2004 5245–5250. doi:10.1158/0008-5472.CAN-04-0496.
- [22] K. Berwind, M.X. Bornschlegl, M.A. Kaufmann, M. Hemmje, Towards a cross industry standard process to support big data applications in virtual research environments, in: *Proceedings of the Collaborative European Research Conference (CERC) 2016*, 2016.
- [23] M.X. Bornschlegl, F.C. Engel, R. Bond, M.L. Hemmje, eds., *Advanced visual interfaces. Supporting big data applications: AVI 2016 workshop, AVI-BDA 2016, bari, italy, june 7-10, 2016, revised selected papers*, Springer International Publishing, Cham; s.l., 2016. doi:10.1007/978-3-319-50070-6.
- [24] CCSDS Secretariat, Reference model for an open archival information system (OAIS). Recommended practice, 2012. URL: <https://public.ccsds.org/pubs/650x0m2.pdf>.

Diversity of the Bifidobacterial Phageome in the Infant Gut

Darren Buckley^a, Douwe van Sinderen^b and Francesca Bottacini^c

^aCork University Maternity Hospital, Neonatal Department, Cork, Ireland

^bAPC Microbiome Ireland and School of Microbiology, University College Cork, Cork, Ireland

^cMunster Technological University, Department of Biological Sciences, Cork, Ireland

Extended Abstract

Members of Bifidobacterium play an important role in the development of the immature gut and are associated with positive long-term health outcomes for the host. It has previously been shown that intestinal bacteriophages are detected within hours of birth, and that induced prophages constitute a significant source of such intestinal phages. The gut phageome can be vertically transmitted from mother to newborn and is believed to exert considerable selective pressure on target prokaryotic hosts affecting abundance levels, microbiota composition, and host characteristics.

The objective of the current study was to investigate prophage-like elements and the CRISPR-Cas viral immune system of publicly available, human-associated Bifidobacterium genomes. Analysis of 585 fully sequenced bifidobacterial genomes identified 480 prophage-like elements with an occurrence of 0.82 prophages per genome. Interestingly, we also detected the presence of corresponding bifidobacterial prophages and CRISPR spacers across different strains, and species, thus providing an initial characterisation of the human-associated bifidobacterial phageome in early life. Our analyses show that closely related and likely functional prophages are commonly present across four species of human-associated Bifidobacterium.

Further comparative analysis of the CRISPR-Cas spacer arrays against the predicted prophages provided evidence of historical interactions between prophages and different strains at an intra- and inter-species level. Notably, a spacer representing a putative major capsid head protein was found on different genomes representing multiple strains across *B. adolescentis*, *B. breve*, and *B. bifidum*, suggesting that this gene may be a preferred target for bifidobacterial phage immunity. Overall, our analysis showed clear evidence of CRISPR-Cas acquired immunity to bifidobacterial prophages across several bifidobacterial strains and species.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

darren.buckley@ucc.ie (D. Buckley); francesca.bottacini@mtu.ie (F. Bottacini)

 0000-0002-0142-2956 (F. Bottacini)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Chapter 4

E-Learning and Competences

QBL – A Software-Technical Approach for Supporting Competence-Based Learning

Matthias Then, Benjamin Wallenborn, Felix Fischman, Sebastian Lothary, Ramona Srbecky, Michael Winterhagen, Matthias Hemmje

FernUniversität in Hagen, Germany

Abstract

Qualifications-Based Learning (QBL) is a software-technical approach that takes up former research & development activities in the area of *Competence-Based Learning (CBL)* and proposes improvements. The QBL domain model was conceived with the goal to support standardized qualifications frameworks as well as institution-specific catalogues designed for internal study or training programs. Associations between qualifications from different frameworks/catalogues advance the emergence of comprehensive qualifications networks and thereby improve the cross-institutional comparability of qualifications-based programs and learning content. Furthermore, the informative value of personal qualifications profiles is increased. In addition to the domain model, QBL proposes an architectural model for the higher education sector and introduces it on the example of an exemplary institution's (FernUniversität in Hagen) IT-landscape. The application scenarios described in this paper refer to that environment. Several proof-of-concept prototypes were developed, for example, a plugin for the learning management system Moodle, a management component for qualifications frameworks, and an authoring toolkit for qualifications-based programs and learning content. As a consequence of the positive evaluation results, subsequent QBL-related projects and theses were initiated. This paper gives an overview of QBL concepts, prototypes, and current projects.

Keywords 1

qualifications-based learning, competence-based learning, QBL, CBL, CBL software, competence model

1. Introduction

The development of the *Qualifications-Based Learning (QBL)* approach was triggered by the observation that the idea of *Competence-Based Learning (CBL)* is increasingly finding its way into the teaching/learning processes at educational institutions, which goes along with a growing demand for CBL software solutions. This demand is only partially met by existing approaches, models and software systems, there is still much need for improvements. QBL aims to contribute concepts and tools realizing CBL visions such as: creation of courses and study programs, design and implementation of teaching/learning scenarios, availability of user profiles, and cross-institutional comparison of competence-related information such as *learning goals* and *access requirements* for courses, modules, study programs, and learning content. QBL takes up former research & development activities, proposes improvements and provides extensions.

The term QBL was introduced, because in CBL the term competence is not always clearly distinguished from other qualification types such as skill or knowledge. In the following, the term *Competence/Qualification (CQ)* will be used as a generic notation for qualifications of any kind.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ [firstname.lastname]@fernuni-hagen.de



2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In 2014, when the development of QBL was initiated, CQ-based learning goals and access requirements of study programs, modules, courses, and learning content were usually described in form of free text. Regarding comparability, this is a serious problem, because such specifications leave too much room for interpretation. The usage of standardized *CQ Frameworks (CQF)* had not yet prevailed, not least because they were not sufficiently supported by learning platforms and other software systems. In addition, only a few universities and enterprises offered institution-specific CQ catalogues. Furthermore, the integrability of CQ-based concepts into higher education institution's IT landscapes and the data exchange between the involved software components was not yet sufficiently researched. From these observations, general *research goals* were derived:

RG1: investigation of CBL from a software developer's perspective. This includes: existing approaches, available software support, applicability in distributed architectures, data flows, and combinability with common exchange formats.

RG2: design of a general *QBL Model (QBLM)* that effectively supports CQFs and CQ-related data flows within a higher education institution's IT-landscape. The QBLM includes a domain model, an architectural model, and a service model.

RG3: development of proof-of-concept prototypes implementing the QBLM domain model.

RG4: realization of distributed QBL application scenarios within an exemplary institution's IT-landscape.

This paper outlines the current state of QBL concepts, prototypes, and currently running further developments. At the beginning (chapter 2), the results of the investigation required by RG1 are briefly summarized. After that (chapter 3), the modeling activities associated with RG2 are described, followed by (chapter 4) a summary of the prototypical solutions that emerged from RG3 and RG4. As a consequence of the positive evaluation results, subsequent QBL-related projects and theses were initiated, an overview is given in chapter 5.

2. State of the Art in Science and Technology

In the following, a few approaches, technologies, and software systems are described that can be regarded as basics and building blocks for QBL.

CQ-based comparison of study programs, modules, courses, and learning content demands an appropriate selection of available CQs. In this context, an overview of standardized CQFs, institution-specific CQ catalogues, and domain taxonomies in the area of information and communication technology has been obtained. The *European Qualifications Framework* [1], a widely recognized template for designing concrete, domain-specific CQFs, is of particular importance. Its proficiency levels can be regarded as an EU-wide basis for categorizing proficiency. Another milestone is the *European e-Competence Framework* [2], an implementation of the European Qualifications Framework focusing on the sector of information and communication technology.

A research & development project with similar objectives to QBL was *TENCompetence* (official title: Building the European Network for Lifelong Competence Development), it is described in [3], chapters 1, 18 und 19. Like QBL, the TEN-Competence approach recommends the usage of standardized formats, for example, its way of modeling study programs and learning content is strongly oriented towards *IMS Learning Design* [4]. CQ-related data then are added to these structures, an overview of the resulting domain model can be found in [5]. The concepts that emerged from TENCompetence, especially the mentioned domain model, are a suitable basis for QBL, so QBL is conceived as a further development of TENCompetence.

A useful means for conceiving distributed teaching/learning scenarios is the *IMS Learning Tools Interoperability (LTI)* specification [6] which can be applied to embed externally hosted, access-protected resources into courses provided by *Learning Management Systems (LMS)*. Such resources can be, for example, assignments, tests, videos, learning programs, or educational games. This is achieved on the basis of standardized services; the LTI basic services are: single sign on, tool launch, and return of outcomes.

In many cases, educational institutions use LMSs for designing and processing their teaching/learning scenarios, so a software solution is required that introduces QBL functionality into a common LMS. The decision on a suitable basis system was made in favor of *Moodle* [7] (Moodle: Modular Object-Oriented Dynamic Learning Environment), a popular, freely available open-source LMS. Moodle provides an easily extendable plugin architecture, capable APIs, and a large number of extension points. Furthermore, it already offers comprehensive support for CQ-based scenarios, even though there is still need for improvement. Another argument for Moodle is the fact that it is the default LMS at the *FernUniversität in Hagen (FUH)* [8], where most of our QBL application scenarios are located.

For research projects, FUH's productive IT environment is only available to a limited degree. For example, the central CQ management component has to be implemented separately from FUH's campus management system, so a temporary solution is required. Such prototypes can be implemented on the basis of the *Knowledge Management Ecosystem Portal (KM-EP)* [9], an educational ecosystem focusing on research and development activities in the areas of learning content creation and knowledge management. The further development of the KM-EP is accompanied by FUH's chair of *Multimedia and Internet Applications* [10].

In the following, concepts, models, and software solutions are presented that close the gap between state-of-the-art technologies and QBL requirements.

3. Conceptual Design of the QBL Model

This chapter introduces the QBLM and its components. As required by RG2, it includes a *Domain Model (QDM)*, an *architectural model*, and diverse *service distribution models*, each of them referring to a concrete application scenario.

3.1. QBLM Domain Model

The QDM is shown in the class diagram in Figure 1 (Fig. 54 in [11]). It can be regarded as a further development of the TENCompetence domain model [5]. The blue, red, orange, and gray colored elements have been derived from TENCompetence, modifications and extensions contributed by QBL are drawn in green color. Teaching/learning content is represented by blue classes, the orange ones stand for students' actions, goals, and received gradings. The major part of QBL-specific extensions refers to the competence model represented by the gray classes.

In compliance with TENCompetence and IMS Learning Design, QBL interprets online courses as *units of learning* composed of *learning activities* and *knowledge resources*. Each of these elements contributes to the course's CQ-based learning goal. Learning activities and knowledge resources are regarded as independent elements defining their own learning goals and access requirements, which facilitates both the modular design of CQ-based courses and the definition of processing sequences (i.e., CQ-based *learning paths*). For this reason, the QDM introduces the *Qualifications-Relevant Learning Element (QRLE)*. In Figure 1, it is visualized by an abstract class that is implemented by the blue classes, so a QRLE instance can de facto stand for a *personal development plan* (for example, a study program), a unit of learning, a learning activity, or a knowledge resource.

In QBL, CQ-based learning goals and access requirements are represented by *CQ profiles* consisting of *CQ instances* (the actual CQs). To achieve maximum comparability of CQ-related data, QBL recommends consequent application of standardized CQFs and institution- or domain-specific CQ catalogues. On the other hand, the creation of innovative didactic scenarios often demands additional course-, module-, or program-specific catalogues, so the QDM has to be flexible enough to handle both variants. This is achieved by the class *CQ framework* and its associated elements: from a software-technical point of view, both standardized and individually created CQ catalogues are CQFs. To enable relations between CQs from different CQFs and connections to elements of domain taxonomies, a tagging mechanism was introduced via the classes *simple tag* and *semantic tagging object*.

CQ profiles are not only applied for specifying QRLE-related learning goals and access requirements, but also for describing each student’s personal learning goal (*target profile*) and the current state of attested CQs (*actual profile*). After successful completion of a QRLE, the student’s actual profile is updated. The gap between the actual and the target profile has to be bridged by appropriate CQ programs, which include suitable QRLEs such as courses, learning activities, and knowledge resources. The personal learning goal is achieved as soon as the actual profile is equal to the target profile.

In [11], chapters 3.2, 4.1, and 5.2, the QDM, its elements, and their added value are explained in detail. This also includes the description of advanced concepts that have not been mentioned here, e.g., the distinction between *CQ scopes* and CQ instances, the revised proficiency concept, and the definition of completion criteria for attesting CQs to students.

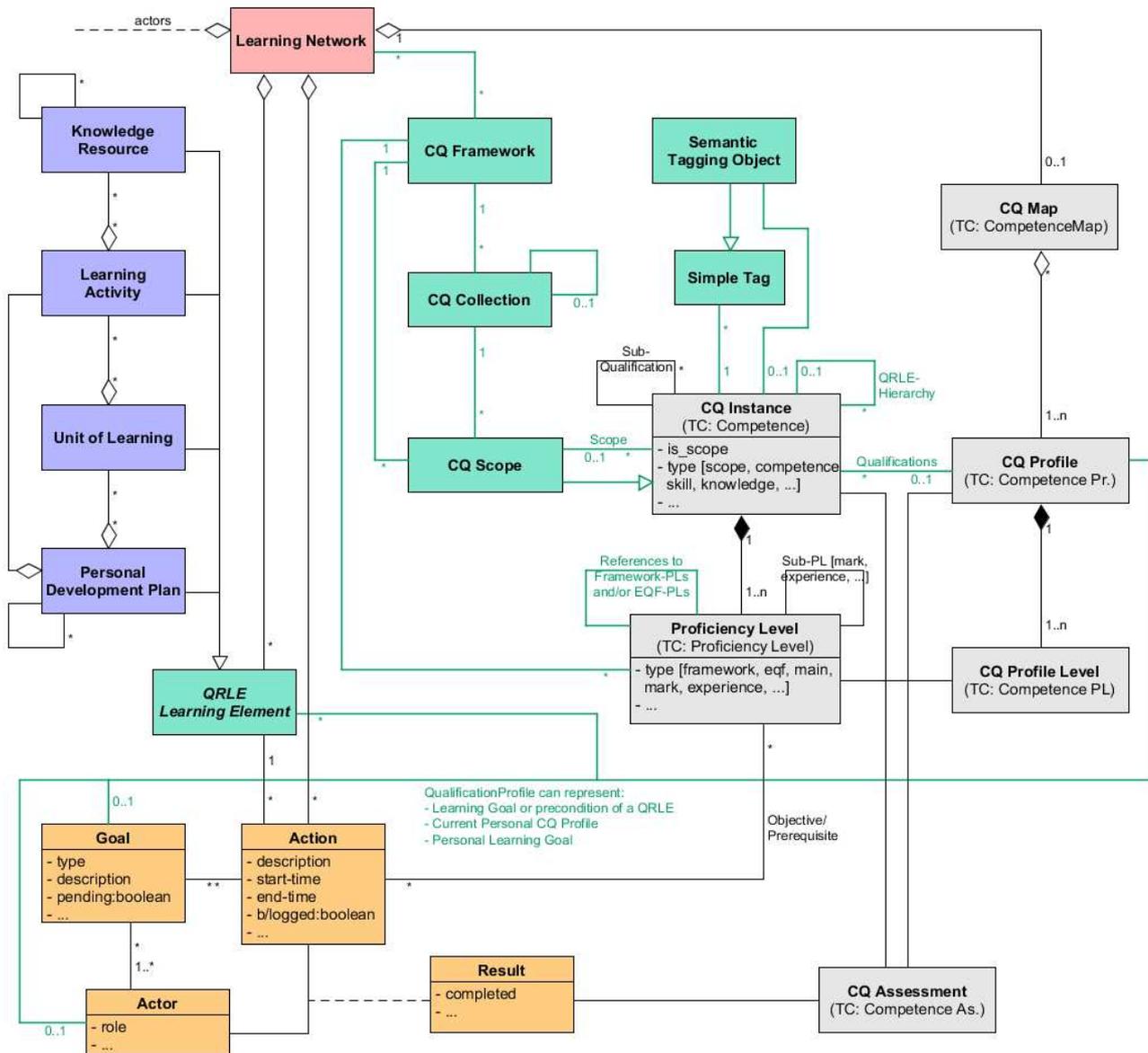


Figure 1: QBLM Domain Model (QDM, Fig. 54 in [11])

3.2. QBLM Architectural Model

The architectural model focuses on the introduction of QBL into a higher education institution’s IT-landscape and identifies involved *application areas* and *software components*. Currently, it consists of a general version and a concrete implementation representing the system landscape at the FUH. In both

cases, the as-is state was modeled at the beginning and later extended by additional components and relations required for QBL. The application scenarios outlined in the following refer to the FUH-specific version in the target state, which is displayed in Figure 2 (Fig. 64 in [11]). The yellow structures stand for application areas, software components are represented by green elements; the highlighted ones (light green, red border) and their interactions (red relations, dashed) are directly involved into our application scenarios. For more detailed information see [11], chapters 2.5 and 3.3.

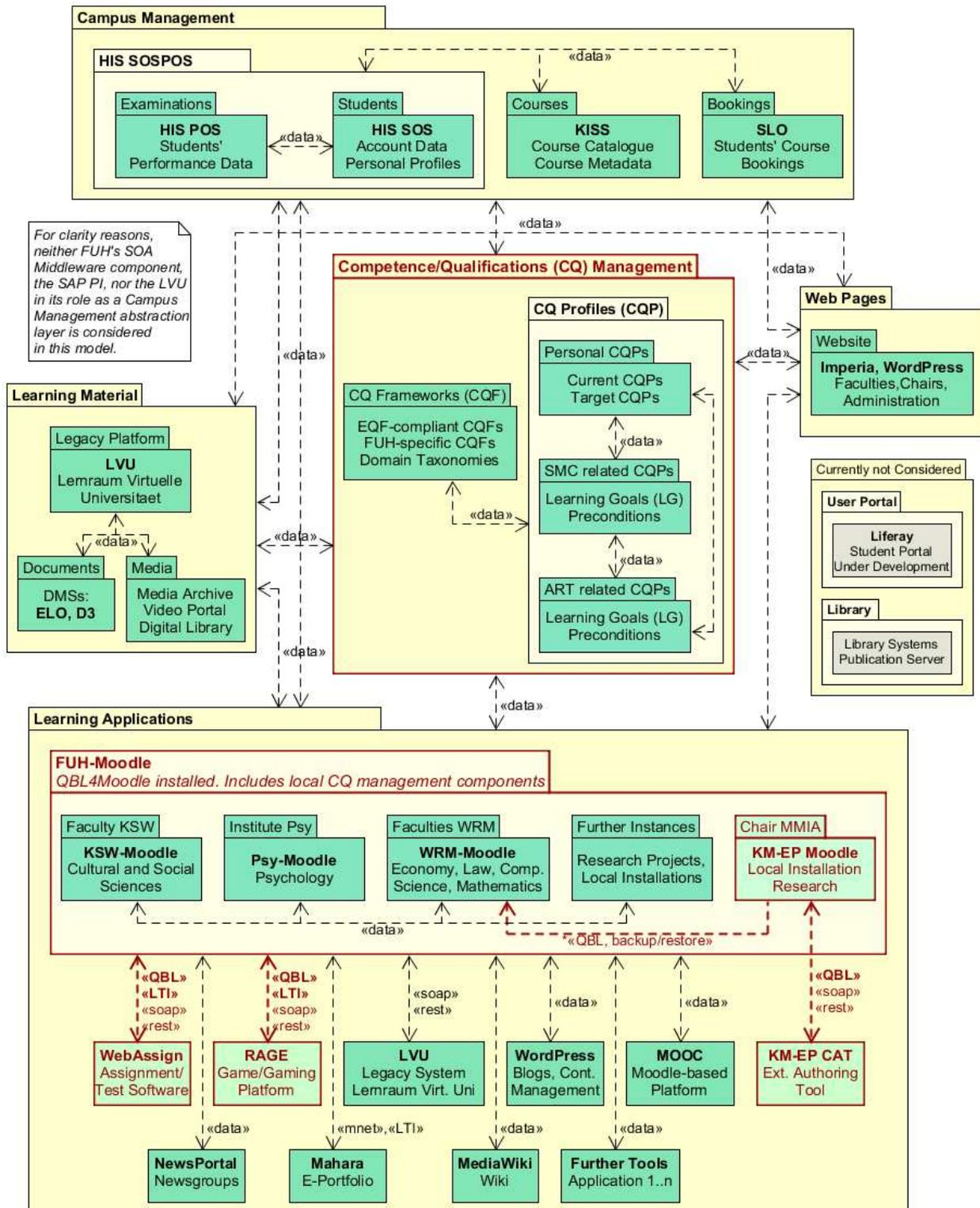


Figure 2: QBLM architectural model, FUH-specific version (Fig. 64 in [11])

In all visualizations of the architectural model, service-based interactions are indicated by relations labeled <data>. In some cases, especially when the relations are used in QBL application scenarios, they have been further concretized by service distribution models.

3.3. QBLM Application Scenarios and Service Distribution Models

In the following, these *Application Scenarios (AS)* are briefly outlined and links to the corresponding service distribution models and software prototypes are given.

AS1 is concerned with the creation of CQ-based, QDM-compliant courses. As demanded by RG3, the required QBL functionality is provided by an LMS extension, the LMS of choice is Moodle (see chapter 2). The resulting Moodle-plugin *QBL4Moodle* is documented in Then's PhD thesis [11], chapter 4.1, and in [12].

The idea of **AS2** is to extend the KM-EP with a CQ-based authoring toolkit that includes management components for QDM-compliant CQs/CQFs as well as tools for creating QRLEs with CQ-based learning goals and access requirements. Furthermore, an export mechanism for transferring such QRLEs to Moodle is required (this also demands QBL4Moodle extensions), as well as a study portal for students. AS2 is covered by Wallenborn's PhD thesis [9], for a service distribution model referring to Figure 2 see [11], chapter 3.3.5, Fig. 66.

AS3 focuses on the standardization of service-based interactions between e-learning systems, the goal is to connect Moodle and *WebAssign* [13], FUH's default application (besides Moodle) for exercises and tests, in an LTI- and QDM-compliant way. AS3 is called the *LTI-based Moodle-WebAssign-Integration*, detailed descriptions of this application scenario and the resulting software solution can be found in [11], chapters 3.3.6 and 4.2.

AS4 adapts the approach from AS3 for CQ-based *educational games*. Initial considerations and pre-projects for the *LTI-based Moodle-EduGame-Integration* are described in [11], chapters 3.3.7 and 4.3. Further developments are in progress.

4. Prototypical Implementations

In this chapter, a few prototypical implementations are outlined that emerged from the above-mentioned application scenarios.

The Moodle-plugin *QBL4Moodle* originally refers to AS1, but it also plays an important role in many other application scenarios. It is documented in [11], chap. 4.1, and in [12]. QBL4Moodle consequently uses the APIs and extension points offered by Moodle, modifications of the core code have been avoided as far as possible. The functional scope covers QBL basic functionalities such as:

- Support of standardized, institution-specific, and individually designed CQFs;
- Definition of CQ-based learning goals and access requirements for QRLEs;
- Assignment of achieved CQs to students;
- Realization of personal CQ profiles.

Subsequent application scenarios require extensions, for example, regarding the exchange of CQ-related data with the KM-EP, WebAssign, and educational games.

The *CQ-based Authoring Toolkit for the KM-EP (KM-EP-CAT)* was conceived and implemented in the context of application scenario AS2, for details see [9]. Figure 3 (Fig. 62 in [9]) gives an architectural overview of the KM-EP components that are involved into CQ-related processes and, therefore, were either introduced or modified in the context of AS2. In addition, the LMS Moodle is displayed, because it is the target system for course execution and therefore has to interact with the KM-EP-CAT. Before AS2 had been conceived, the KM-EP did not offer any support for CQ-based approaches, so a *CQ manager* for defining and modifying QDM-compliant CQs and CQFs was added, as well as a *CQ profile manager* for CQ profiles. The existing *course authoring tool* was extended with functionalities

enabling authors to assign CQ profiles to courses and use them as learning goals and access requirements. Via QBL-specific web services provided by QBL4Moodle, such courses can be exported to Moodle where the teaching/learning process takes place. The QRLE types study program and module are supported by the *study program authoring tool*. It offers functionalities for creating/managing study programs and their associated modules which are usually composed of courses. All mentioned QRLE types can be associated with CQ-based learning goals and access requirements that are represented by CQ profiles created with the CQ profile manager. Besides these QRLE-related tools, a *study portal* is provided. In this portal, students can create and manage their accounts, enroll into programs, modules, and courses, and monitor their learning progress.

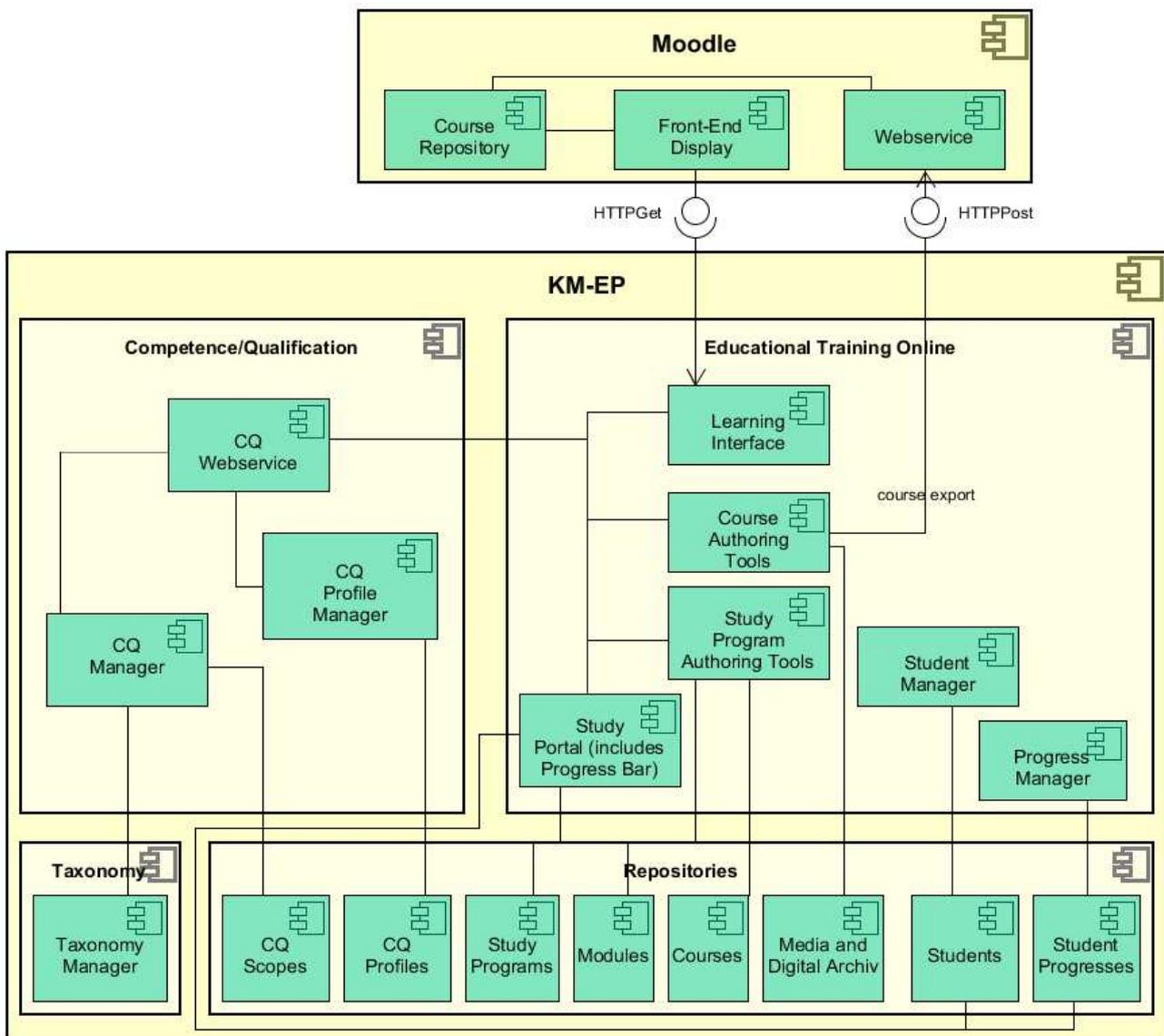


Figure 3: Application scenario AS2: CQ-related KM-EP components (Fig. 62 in [9])

The implementation of the *LTI-based Moodle-WebAssign-Integration* refers to application scenario AS3, Figure 4 (Fig. 68 in [11]) gives an overview of the involved components and their interactions. On Moodle-side, the activity type *external tool*, which is included by standard, enables course creators to seamlessly integrate access-protected exercises from WebAssign. Seamless means that before the requested resource is loaded, the logged in Moodle user is authenticated in WebAssign (LTI-service *single sign on*) and in the case of success, the resource is launched (LTI-service *tool launch*) within the Moodle course. If it is directly embedded or displayed in a separate window/tab, depends on Moodle-sided settings. Submitted attempts and received gradings (scores, comments, etc.) are stored in

WebAssign and via the LTI-service *return of outcomes*, the achieved scores can be transferred to Moodle where they are stored in the students' gradebooks.

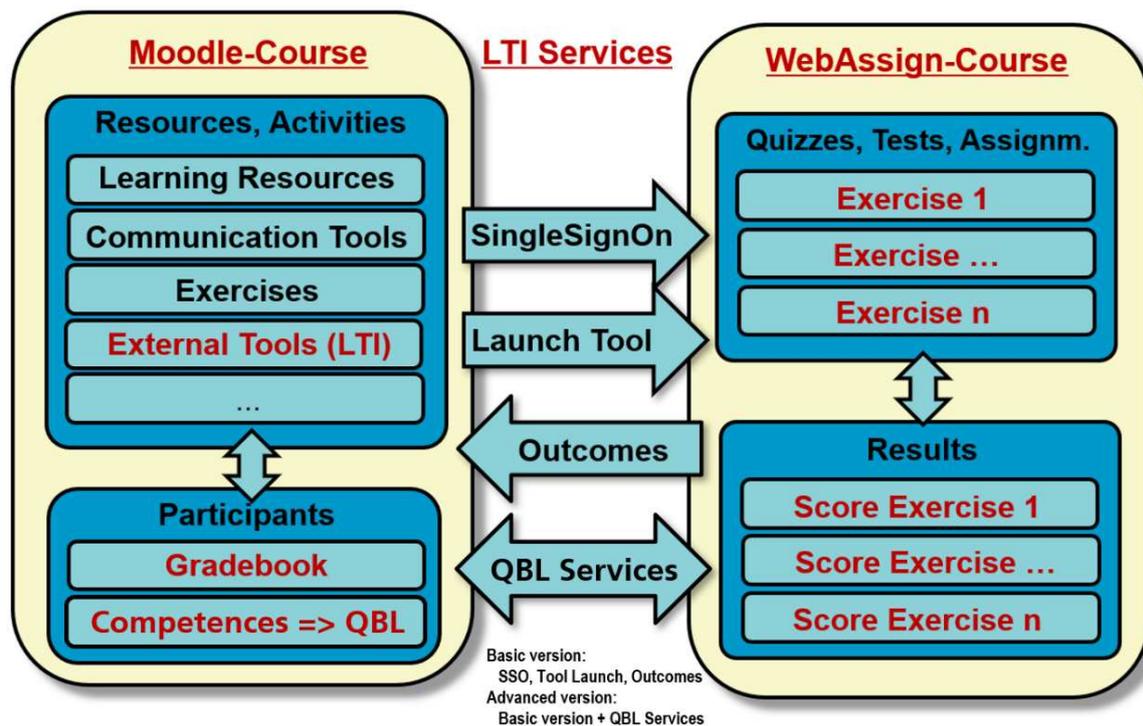


Figure 4: LTI-based Moodle-WebAssign-Integration (Fig. 68 in [11])

In this *basic version* of the LTI-based Moodle-WebAssign-Integration, achieved CQs can be attested to the students by configuring the involved external tools (more precisely: instances of activity type external tool) in a way that combines the configuration settings grade, activity completion, and competencies. As long as the scheduled CQ-based teaching/learning scenarios can be realized with the tools offered by standard Moodle, the basic version does not require any Moodle-sided extensions. QBL application scenarios can only be processed, if QBL4Moodle is installed on the involved Moodle instance. At the time AS3 was conceived, WebAssign did not yet support LTI, LTI-specific extensions were developed in the context of Then's PhD project, see [11] and [14]. The *advanced version* goes beyond the possibilities of the LTI outcomes service and allows a more extensive data exchange. To achieve this, QBL-specific extensions are required in both systems. Since winter semester 2017/18, the basic version of the LTI-based Moodle-WebAssign-Integration is an integral part of FUH's IT-landscape, it is applied for online exercises as well as for examinations. A user manual with the title "Einbinden von Übungssystem-Aufgaben als LTI-Tool in Moodle" can be found in [13], section "Handbücher für Kursbetreuer". A technical documentation describing the implementation is available in [15]. Furthermore, a QBL-related description is given in [11], chap. 4.2.

5. Evaluation, Outreach, and Impact

The QBLM, the QBL application scenarios, and the described software prototypes had been positively rated in several evaluation scenarios with participation of researchers, software developers, and course authors; see [11], chapter 5. As a consequence, subsequent PhD, master and bachelor theses were initiated that are concerned with improvements, extensions, and new application scenarios.

The LTI-based Moodle-EduGame-Integration (AS4) is continuously further developed, results are, for example, presented in [16,17]. Since then, a prototypical framework providing tools for different user groups and tasks was created. Developers are supported during game design and implementation; teachers can monitor and analyze students' gameplay performances. The latter is achieved by a gaming

platform-sided component that tracks game events and transfers the tracked data back to the embedding LMS. Visualization tools offer different types of diagrams and statistics about the collected performance data. A mechanism for analyzing the tracked data, extracting the achieved CQs, and adding them to successful players' personal CQ profiles is currently in progress. The described tracking concept must already be considered during game development, the events to track and the corresponding CQs have to be specified via an associated API.

The objective of **AS5** is to design a concept for *dynamic learning paths in CQ-based courses* (more precisely: QDM-based courses) and implement it on the basis of Moodle/QBL4Moodle and the KM-EP-CAT. Initial outcomes are presented in [18]. In a first step, a course consisting of activities/resources with separate, QDM-compliant learning goals and access requirements has to be designed. The course can be created either directly in Moodle or in the KM-EP-CAT, from where it is later exported to Moodle. Learning goals and access requirements are adjusted in a way that allows the derivation of reasonable, student-specific processing sequences for each student. Such a personal processing sequence is called dynamic personalized learning path; it depends on the student's current state of attested CQs, which results from prior CQs and the learning progress within the course. Besides CQ-based learning content, this approach requires a personal CQ profile for each student, which has to be automatically updated each time the student achieves a new CQ. Depending on the current CQ profile, the course's activities and resources are either accessible/visible for the student or not. A central evaluation scenario for this approach is a course that conveys basic knowledge and skills in object-oriented programming. Based on a student's current CQ profile, it is decided what is displayed and what is hidden. Some topics, activities (e.g., quiz, assignment, test), and resources can be skipped, others have to be processed/assessed, and still others remain hidden because the required access requirements are not yet fulfilled. So, for each participant, an individually customized course is generated and continuously adapted to his/her learning progress. This way, the learning process for object-oriented programming becomes more effective and, furthermore, all achieved CQs can be considered for follow-up courses that use personal CQ profiles in a similar way.

AS6 emerged from the idea of using *IMS Learning Design (IMS-LD)* compliant *Didactical Structural Templates (DST)* for representing, processing, and exchanging CQ-based QRLEs. Like in AS5, the basis systems for prototypical implementations are the KM-EP-CAT and Moodle/QBL4Moodle. The current state of AS6 is presented in [19]. A DST describes the didactical structure of, for example, a Moodle course and is based on the e-learning standard IMS-LD enriched with QBL-specific concepts. By combining the QBL-approach with IMS-LD, it is possible to define learning goals and access requirements on every IMS-LD element. In [19] it has been shown that it is possible to realize a DST as a "classical" Moodle course, a Moodle course with gaming content, and a standalone applied game. A benefit of DSTs is that learners can choose/switch between different realizations of a specific DST whenever they want. Gamers will usually choose the standalone applied game, others might prefer the traditional approach with a Moodle quiz, still others might want to try both variants. For details about educational games and their interaction mechanisms with gaming platforms and LMSs see AS4. The AS4-approach can also be applied for DSTs, which has been proven in evaluation scenarios covering both web-based and non-web-based games.

AS7 is concerned with the application of QBL for continuous professional education and lifelong learning. Based on the concept described in [20,21], an approach for a *Semantic Qualification Web (SQW)* is designed. This application scenario was motivated by the rapid change of job profiles caused by the ongoing web digitization and globalization. Employees continuously have to gain new and renew existing CQs in order to maintain or improve their employability and to adapt their personal CQ profiles to the requirements of the labor market. Therefore, a goal-oriented (re-)qualification concept for up- and reskilling is needed that helps people to continuously develop and maintain the demanded CQs for their targeted jobs and to react to future job market needs. We propose the creation of an SQW that includes a knowledge base consisting of defined CQs, job profiles, and CQ-based QRLEs (programs, courses, learning content). On this basis, suitable learning paths optimized for the needs of the individual employee can be automatically generated and suggested, the user can make his/her own choice. The term SQW stands for the idea of a semantic network in the qualification sector on the basis of the QDM.

The initial concept is described in [20,21], further developments and proof-of-concept implementations are in progress and will be published soon.

6. Summary and Outlook

In this paper, the QBL approach was introduced and an overview of the current state of development was given. This includes the achievements from 2014-2019, which are summarized in the PhD theses of Wallenborn [9] and Then [11] (short version: [14]), as well as the follow-up developments made by² Fischman, Lothary, Srbecky, and Winterhagen since 2018/19. The QBL approach, the QBLM, and the described software prototypes are continuously improved and extended; further QBL-related projects and theses are scheduled.

7. References

- [1] European Union, Recommendation of the European Parliament and of the Council of 23 April 2008 on the establishment of the European Qualifications Framework for lifelong learning, 2008, in: Official Journal of the European Union, pp. 1-7, 06 May 2008.
- [2] European Committee for Standardization CEN, European e-Competence Framework (e-CF) version 3.0 - a common European Framework for ICT Professionals in all industry sectors, CWA 16234:2014 Part 1, 2014.
- [3] R. Koper, Learning Network Services for Professional Development, Springer Berlin Heidelberg, 2009, ISBN 978-3-642-00977-8. DOI 10.1007/978-3-642-00978-5.
- [4] R. Koper, C. Tattersall, Learning Design - A Handbook on Modelling and Delivering Networked Education and Training, Springer Berlin Heidelberg, ISBN 978-3-540-27360-8, 2006.
- [5] R. Koper, The TENCompetence Domain Model - Version 1.1., 2008. URL: <https://core.ac.uk/download/pdf/55533686.pdf>, 18 June 2021.
- [6] IMS Global Learning Consortium, Learning Tools Interoperability, 2021. URL: <https://www.ims-global.org/activity/learning-tools-interoperability>, 18 June 2021.
- [7] Moodle.org, Moodle open-source learning platform, 2021. URL: <https://moodle.org>, 18 Jun 2021.
- [8] FernUniversität in Hagen, Germany's State Distance-Learning University, 2021. URL: <https://www.fernuni-hagen.de>, 18 June 2021.
- [9] B. Wallenborn, Entwicklung einer innovativen Autorenumgebung für die universitäre Fernlehre, 2018, Dissertation, FernUniversität in Hagen. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001428, 18 Jun 2021. DOI: 10.18445/20180911-091907-0.
- [10] Lehrgebiet MMIA, LGMMIA - Lehrgebiet Multimedia und Internetanwendungen, 2021. URL: <http://www.lgmmia.fernuni-hagen.de>, 18 June 2021.
- [11] M. Then, Supporting Qualifications-Based Learning (QBL) in a Higher Education Institution's IT-Infrastructure, 2020, Dissertation, FernUniversität in Hagen. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001608, 18 June 2021. DOI: 10.18445/20200309-141118-0.
- [12] M. Then, M. D. Hoang, M. Hemmje, A Moodle-based software solution for Qualifications-Based Learning (QBL), 2019, FernUniversität in Hagen. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001501, 18 June 2021. DOI: 10.18445/20190225-103757-0.
- [13] I. Schulz-Gerlach, Hilfe/Handbücher - Online-Übungssystem, 2021. URL: <https://online-uebungssystem.fernuni-hagen.de/hilfe/hilfe.html>, 18 June 2021.
- [14] M. Then, Ein Ansatz zur Softwaretechnischen Unterstützung des Qualifikationsbasierten Lernens (QBL) an Hochschulen, 2021, in: Hölldobler, S. (Ed.), Ausgezeichnete Informatikdissertationen 2020, Gesellschaft für Informatik. URL (preprint): https://www.researchgate.net/publication/349694936_Ein_Ansatz_zur_Softwaretechnischen_Unterstuetzung_des_Qualifikationsbasierten_Lernens_QBL_an_Hochschulen, 18 June 2021.
- [15] I. Schulz-Gerlach, Technische Doku zur LTI-Integration - Das Online-Übungssystem als LTI-Tool-Provider, 2021. URL:

² in alphabetical order

- https://online-uebungssystem.fernuni-hagen.de/download/LTI_Moodle/LTI_TechnicalDoc.html, 18 June 2021.
- [16] M. Winterhagen, M. Salman, M. Then, B. Wallenborn, T. Neuber, D. Heutelbeck, M. Fuchs, M. Hemmje, LTI-Connections Between Learning Management Systems and Gaming Platforms: Integrating a Serious-Game Prototype into Moodle Courses, 2020, in: Journal of Information Technology Research (JITR) 13(4), pp. 47-62, 2020, URL: https://www.researchgate.net/publication/346228285_LTI-Connections_Between_Learning_Management_Systems_and_Gaming_Platforms_Integrating_a_Serious-Game_Prototype_Into_Moodle_Courses. DOI: 10.4018/JITR.2020100104, 18 June 2021.
- [17] R. Srbecky, M. Then, B. Wallenborn, M. Hemmje, Towards learning analytics in higher educational platforms in consideration of Qualifications-Based Learning, 2021, Edulearn 2021 - 13th International Conference on Education and New Learning Technologies, July 2021. DOI: 10.21125/edulearn.2021.1046.
- [18] F. Fischman, H. Lersch, M. Winterhagen, B. Wallenborn, M. Fuchs, M. Then, M. Hemmje, Individualized Educational System Supporting Object Oriented Programming, 2020, in: Advances in Software Engineering, Education, and e-Learning - Proceedings from FECS'20, FCS'20, SERP'20, and EEE'20, Springer International Publishing, 2021. ISBN: 978-3-030-70872-6. DOI: 10.1007/978-3-030-70873-3.
- [19] M. Winterhagen, M. D. Hoang, B. Wallenborn, D. Heutelbeck, M. Hemmje, Supporting Qualification Based Didactical Structural Templates for Multiple Learning Platforms, 2020, EEE'20 - The 19th International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, Las Vegas, USA, July 2020. URL: https://www.researchgate.net/publication/343255892_Supporting_Qualification_Based_Didactical_Structural_Templates_for_Multiple_Learning_Platforms, 18 June 2021.
- [20] B. Humm, H. Bense, M. Fuchs, B. Gernhardt, M. Hemmje, T. Hoppe, L. Kaupp, S. Lothary, K.-U. Schäfer, B. Thull, T. Vogel, R. Wenning, Machine intelligence today: applications, methodology, and technology, 2021, in: Informatik Spektrum, 18. June 2021. DOI: 10.1007/s00287-021-01343-1.
- [21] S. Lothary, W. Krasowski, B. Wallenborn, M. Then, M. Fuchs, M. Hemmje, Supporting the Representation of Personal Learning Paths for Goal-Oriented Professional Qualification, 2021, Edulearn 2021 - 13th International Conference on Education and New Learning Technologies, July 2021. DOI: 10.21125/edulearn.2021.0813.

Supporting the Mapping of Educational Game Events with Competency Models considering Qualifications-Based Learning

Ramona Srbecky^a, Marcus Frangenberg^a, Benjamin Wallenborn^a, Matthias Then^a, Iván-José Pérez-Colado^b, Cristina Alonso-Fernandez^b, Baltasar Fernandez-Manjon^b and Matthias L. Hemmje^a

^a FernUniversität Hagen, Universitätsstraße 47, Hagen, Germany

^b Universidad Complutense de Madrid, Av. Séneca 2, Madrid, Spain

Abstract

Computer and video games have established themselves in society and are increasingly finding their way into learning with so-called Educational Games (EduGame). EduGames often provides a fundamental analysis of learners' learning results, but there is currently no existing approach to map the learning results automatically to digital standardized machine-readable Qualifications. In addition, different learning providers tend to use different approaches to describe Qualifications, often in the form of free text. Therefore, it is currently impossible to compare the reached Competencies and Qualifications (CQ) across different learning providers or only manually with high effort. In this paper, a prototypical implementation for automated mapping of learners' learning results in EduGames to standardized CQ will be introduced. The paper presents the conceptual work, the subsequent prototypical implementation with the chosen Analytics Environment, Game Engine, Learning Management System, and the evaluation results.

Keywords

Qualifications-Based Learning, Game-Based Learning, RAGE Analytics, CbKST, Learning Analytics

1. Introduction and Motivation

Computer and video games have established themselves in society and are increasingly finding their way into learning with so-called Educational Games (EduGame). Around 30,4 million Germans play at least occasionally, including 30 percent of 10- to 29-year-olds and 30 percent of 30- to 49-year-olds [19]. Schoolchildren and students are, therefore, very familiar with the medium. Learning content presented in game form builds on these experiences. An experiment [6] conducted at the Hult International Business School proves that game-based knowledge transfer has potential. In this experiment, teaching content through a video game was similarly successful as teaching directly by the teacher [6]. However, EduGames often provide only restricted analysis of learners' learning results. There is currently no existing approach to automatically map the learning results to digital standardized machine-readable Competencies and Qualifications (CQ). In the context of this research, the term CQ will be used for Qualifications. CQs consist of competencies, skills, and Proficiency Levels. If an assessment is performed and the CQ of the learners gets attested, it is a Qualification. Otherwise, learners acquire only Competency. To prove CQ in assessments, Learning Analytics should be used.

Therefore, it is currently impossible to compare the reached CQ across different learning providers or manually with high effort. To address this problem, the use of standardized CQ lends itself. The problem that there is no comparability between CQ was taken up by [36] and [32] with the Qualifications-Based Learning Model (QBLM). The QBLM allows the modeling of learning scenarios using digital (machine-readable) CQ [32] [36]. Based on this model, [32] designed the QBL4Moodle plugin for the Learning Management System (LMS) Moodle [25] to work with QBLM-based CQ in

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ {ramona.srbecky; benjamin.wallenborn; matthias.then; matthias.hemmje}@fernuni-hagen.de (R. Srbecky, B. Wallenborn, M. Then, M. Hemmje); marcus.frangenberg@studium.fernuni-hagen.de (M. Frangenberg); ivanjper@ucm.es (I. J. Perez-Colado); crisal03@ucm.es (C. Alonso-Fernandez); balta@fdi.ucm.es (B. Fernandez-Manjon)



Moodle. [32] also addresses the connection of EduGames with Qualification-based courses in Moodle, but they do not yet enable CQ-Profiles (CQPs) for individual users [32]. CQPs describe each student's personal Learning Goals (target Profile) and the Current State of attested CQs (actual Profile). As a future topic, [32] addresses game behavior analysis and the mapping to standardized machine-readable CQ and their assignment to learners CQPs. In this context, game behavior is the player's individual interaction in the game. This includes the way the player acts or conducts himself in the game [31]. This interaction causes various data that can be measured, stored, and analyzed. Since preliminary work has already been done with Moodle to support QBLM, this LMS is chosen as the technological basis for the present work. Through the LMS, students can access the study material and learning units for each study course [7]. The study courses are a subdivision of study programs from Higher Educational Institutions and are presented in different LMS formats. [28] has realized the first prototypical implementations for analyzing and evaluating the game behavior data in EduGames. This paper aims to track and analyze game behavior and subsequently enable mapping them to QBLM-based CQ.

Several Problem Statements (PS) can be derived from the motivation mentioned above. PS1 is that the individual evaluation of EduGame events is not possible. Various Analytics Engines have been established for the analysis and evaluation of game events. "An event is represented by any action or decision that is made by the player or the game itself." [24] However, [28] encountered the problem in that an evaluation per player was not possible with Unity Analytics [34], the chosen Analytics Engine in [28]. However, a separation by user profiles on the EduGame or Analytics Engine side and the LMS is a prerequisite for synchronizing results and the learner's personal CQP. PS2 is that EduGame results cannot be transferred to the CQP of an LMS. PS3 is that there is currently no mapping between game behavior and achieved CQ. This means a lack of criteria determines the achievement of CQ and Proficiency Level of CQ. The PSs mentioned above result in the following two Research Questions (RQ). RQ1: Can player individual evaluations of EduGame events be achieved via Analytics Engines be achieve? RQ2: Can a model be designed to map learning game events and CQ in a CQ model?

Based on the research methodology of [27], the following Research Objectives (ROs) were derived from the RQs. RO1 is assigned to the Observation Phase (OP). In this phase, a suitable CQ model is identified that can be mapped in an LMS to be determined. For this purpose, the QBLM will be investigated. Also, possible interfaces between Analytics Engine and LMS are considered. RO2 is assigned to the Theory Building Phase (TBP). A concept is designed that shows how the EduGame results are transferred into the CQP. However, also, a concept for mapping the game events with CQs will be designed. The System Development phase (SDP) moves the concept into a prototype and is assigned to RO3. The result of the SDP is evaluated in the Evaluation Phase (EP) in the context of a Cognitive Walkthrough (CW) [38]. Finally, the EP is assigned to RO4. In this phase, both RQs are evaluated. This paper is structured according to the ROs. This means that in the State of the Art (SOA) section, the OP is described. In the Conceptual Design section, the TBP is described, and the SDP phase is presented in this paper in the Proof of Concept (PoC) implementation section. In the Evaluation section, the EP is presented. The paper concludes with a summary and an indication of future developments.

2. State of the Art and Technology

Some research projects and software systems related to the research goals have already been mentioned in the previous section. In the following, the most important are described in more detail. The work in this section represents the results of the OP of the research methodology.

2.1. State of the Art

In this section the State of the Art of the QBLM model and the Competence-based Knowledge Space Theory (CbKST) will be briefly described.

To increase the comparability of learning objects based on Qualifications, the QBL-approach described by [33] was implemented at the University of Hagen [16]. The QBLM approach includes a

Domain Model (QDM), an Architectural Model, and diverse Service Distribution Models. The QDM introduces the Qualifications-Relevant Learning Element (QRLE). A QRLE instance can de facto stand for a Personal Development Plan (for example, a study program), a Unit of Learning, a Learning Activity, or a Knowledge Resource. In QBL, CQ-based Learning Goals and Access Requirements (ARs) are represented by CQPs consisting of CQ Instances (the actual CQs). CQPs are applied to specify QRLE-related Learning Goals and ARs and describe each student's personal Learning Goal (target profile) and the Current State of attested CQs (actual Profile). After successful completion of a QRLE, the student's actual Profile is updated. The gap between the actual and the target profile must be bridged by appropriate CQ programs, including suitable QRLEs such as courses, learning activities, and knowledge resources [32].

CbKST is an extension of the original Knowledge Space Theory (KST) [3]. The KST is a mathematical theoretical framework for representing the knowledge of learners in a certain learning domain. The domain is specified by a set of problems (in the following denoted with X), which the learner is able to solve [3]. The subset of problems that the learners can solve is defined as the knowledge state of the individual learner in the KST [20]. Because of the mutual psychological dependencies between the problems, not all possible subsets of X are plausible knowledge states. One possible approach to solving these dependencies is the concept of surmise systems. Here it is assumed that if a learner can solve the problem $x \in X$, with a family of subsets of X as clauses, the learner is also capable of solving all problems in one of the clauses [3]. CbKST involves psychological assumptions about underlying skills and competencies required to solve specific problems. In the CbKST approach, competencies are mapped to learning objects (taught competencies) and tested competencies [29]. A competency state can be defined similarly to a knowledge state. The competency state consists of a set of skills which the learner possesses. Based on a learner's current competency state and learning structure, personalized learning paths can be created [20].

2.2. State of Technology

In this section, the State of Technology will be addressed in more detail. First, the Course Authoring Tool (CAT) and the LMS will be introduced, followed by a short description of the data exchange formats for learning content. Afterward, a project for tracking students' behavior in gameplay and a Competence-based Knowledge Space Theory (CbKST) [3] based tool for CQ in-game assessment will be addressed. Finally, the section concludes with a short description of an EduGame prototype.

The Knowledge Management Educational Portal (KM-EP) [35] will be further developed within the project Realising an Applied Gaming Ecosystem (RAGE) [36]. The portal offers various web-based tools for Knowledge Management and user-friendly CAT to create Moodle courses without much previous knowledge [36]. In addition, the KM-EP enables QBLM-based work with CQs. For this purpose, there is a CQ administration, the possibility to create CQPs, and to assign learning units (modules, courses, and teaching resources within courses) to these profiles. Furthermore, courses with assigned CQs can be exported to Moodle and executed there [36].

Moodle stands for Modular Object-Oriented Dynamic Learning Environment [25] and is a freely available open-source software under the GNU Public License. It is software with which courses can be conducted and developed via the internet. With the Learning Tools Interoperability (LTI) [21] standard, further applications such as games can be integrated into an LMS like Moodle [21]. To transfer QBLM-based CQs to Moodle, the University of Hagen developed the plugin QBL4Moodle [32]. This plugin is the interface between Moodle and KM-EP. QBL4Moodle is used to work with QBLM in Moodle itself and map CQs created with it to the CQ approach of Moodle itself. The plugin also serves to import QBLM-based CQs, profiles, and frameworks from other systems. Currently, this is realized for the KM-EP [32]. The Learning Management System (LMS) used at the University of Hagen is Moodle. This LMS already offers digital learning content at the University of Hagen. Therefore, the already existing LMS will be used in this work.

To gather more information about learners' digital or non-digital learning experience, the Advanced Distributed Learning Initiative developed the Experience API (xAPI) [1]. To facilitate the compatible exchange of data about a learner's behavior and performance, the xAPI is designed as a specific standard [2]. In the context of this research, the xAPI will be used in the RAGE Analytics component to track the data from the learners [11].

The Learning Tool Interoperability (LTI) [21] standard, introduced by IMS Global Learning Consortium, integrates external learning tools and content into existing LMSs [21]. In order to use LTI, the following two systems are required to build an LTI ecosystem: A platform or LMS does not provide certain functionalities itself but delegates these functions to other tools. The second required component is the tools. These are software systems that offer digital learning games or other services that the platform does not provide but are integrated into the platform [21]. For this research, the LTI standard will be used to connect the LMS Moodle, which aims to be the platform, and the RAGE Analytics Environment, which will be the second required component, also called the tool.

The Realizing an Applied Gaming Ecosystem Analytics Environment (RAGE Analytics) has been developed at the Complutense University of Madrid, Department of Software Engineering and Artificial Intelligence [11], [18]. Besides software components that facilitate tracking students' behavior during gameplay, it offers tools for analyzing the traces and displaying the evaluation results in diagrams or graphics. Furthermore, RAGE Analytics will be used to generate and collect the corresponding game data analyzed regarding CQ.

The Player Competence Adaption Pack (P-CAP) was developed within the RAGE project [32]. The P-CAP includes a collection of assets that empower developers to define CQs, in-game situations for qualification assessment, and learning path in-game creation. The CQ model of P-CAP is derived from the CbKST [3]. A CQ State is given for every player who complies with the CQP in QBLM. P-CAP's CQ Assessment Asset provides the functionality to track and analyze the learner's gameplay in the respective task situations. When the player has completed the task, the corresponding CQ is stored in his own CQ State. There is no RAGE Analytics components usage in the P-CAP, but future development is considered [32]. The basic concept of P-CAP is shown in Figure 1.

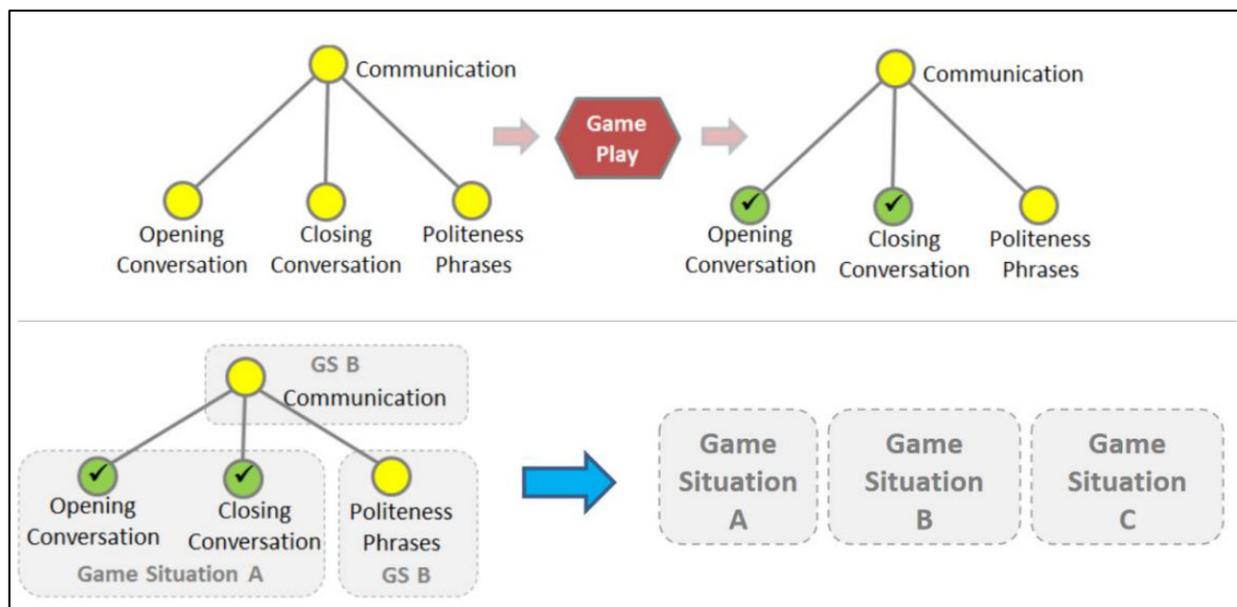


Figure 1: P-CAP Assets Overview [32]

The top-left graph in Figure 1 shows a simple CQ tree designed with the domain model asset included in the P-CAP. The CQ Communication is not considered to be achieved until the three sub-CQs, Opening Conversation, Closing Conversation, and Politeness Phrases, are satisfied. The upper right graph shows the CQ acquisition of a player after completing the game: The first two sub-CQs have

been achieved, but the third (Politeness Phrases) has not. This is recorded in a so-called CQ State, which is kept for each player. This functionality is provided by the CQ Assessment Asset provided by P-CAP. The lower left graph shows the Game Situation from the example. The Opening Conversation and Closing Conversation sub-CQs were taught through Game Situation A. To teach Politeness Phrases, Game Situation B follows. Game Situations are created using the CQ Assessment Asset, and the following appropriate Game Situation is determined using the CQ-based Adaption Asset algorithms. Currently, P-CAP does not use any components from the RAGE Analytics environment to track and store game data, but this is being considered for the future [36] [32].

The Serious Game Prototype (SGP) by [28] was developed based on the Unity Engine. The Unity Engine is a multiplatform development environment for games [5]. The engine supports Windows, Linux, Android, WebGL, and current game consoles, among others [10]. The manufacturer's goal is to provide a user-friendly tool for developing virtual 3D content. After starting the SGP in the Unity editor, the player gets to the MainMenu. This event is to be referred to as opening the game. By clicking on the start button, the game begins. It offers the possibility to construct different buildings in an area. For this, raw materials are subtracted from the raw material starting values. Then, by clicking on the constructed building, the player is asked four multiple-choice questions in random order. As soon as a question is answered correctly, the player continuously receives new raw materials at intervals of seconds. The SGP developed by [28] is to be used as an EduGame in the work context.

In this section, the to-be-used software systems and approaches had been introduced. The KM-EP is used to map and store CQ Frameworks (CQF) such as QBLM-based ones quickly. Since this is currently not possible in Moodle, the QBL4Moodle plugin links Moodle and the KM-EP. Moodle should only run the related courses and provide interfaces such as LTI connection for external learning content in this approach. xAPI and Unity Engine are used for the EduGame, which an LTI connection in Moodle will provide, and RAGE Analytics and the P-CAP to analyze the player individual data regarding CQ. Since the focus of the work is not on the development of the EduGame, the existing SGP will be used as EduGame.

3. Concept

In the following, the results of the conceptual and theory-building work will be described in more detail. First, the overall concept is presented. Afterward, the use cases and the concept for a mapping component between the game events and the CQ model will be described. This is followed by the system design decisions for a modified approach for the implementation due to limitations of the current system. Subsequently, the concept is explained according to the modified system design.

3.1. Overall Concept and Use Cases

To address the RQs, different software components and systems are required to be combined. Moodle is used as the LMS, RAGE Analytics as the Analytics Engine, and Unity Engine as the Game Engine. The overall system architecture is described in Figure 2.

Moodle is to be connected to the Unity Engine and RAGE Analytics via "External Tool" activities. This allows the Unity game to be displayed directly in Moodle and enables authentication of the Moodle user. Evaluations from RAGE Analytics can also be displayed directly in Moodle, and authentication can also automatically occur here. The technical realization of the mentioned interfaces is done via LTI connections. Q2 requires a mapping to take place between the EduGame outcomes and CQ. CQs, according to the QBLM and CQP, were Moodle via QBL4Moodle of the RAGE Analytics environment is to implement an extension that performs the mapping and stores the CQs achieved. An interface to Moodle needs to be created to access the Moodle CQ Frameworks. The acquired CQs should also be transferred from the extension to the respective Moodle CQPs. The Mapping extension is located within RAGE Analytics, as the transferred game data in the environment is already stored in a database and

can be accessed per player. Only information about the Competency progress must be transferred to Moodle.

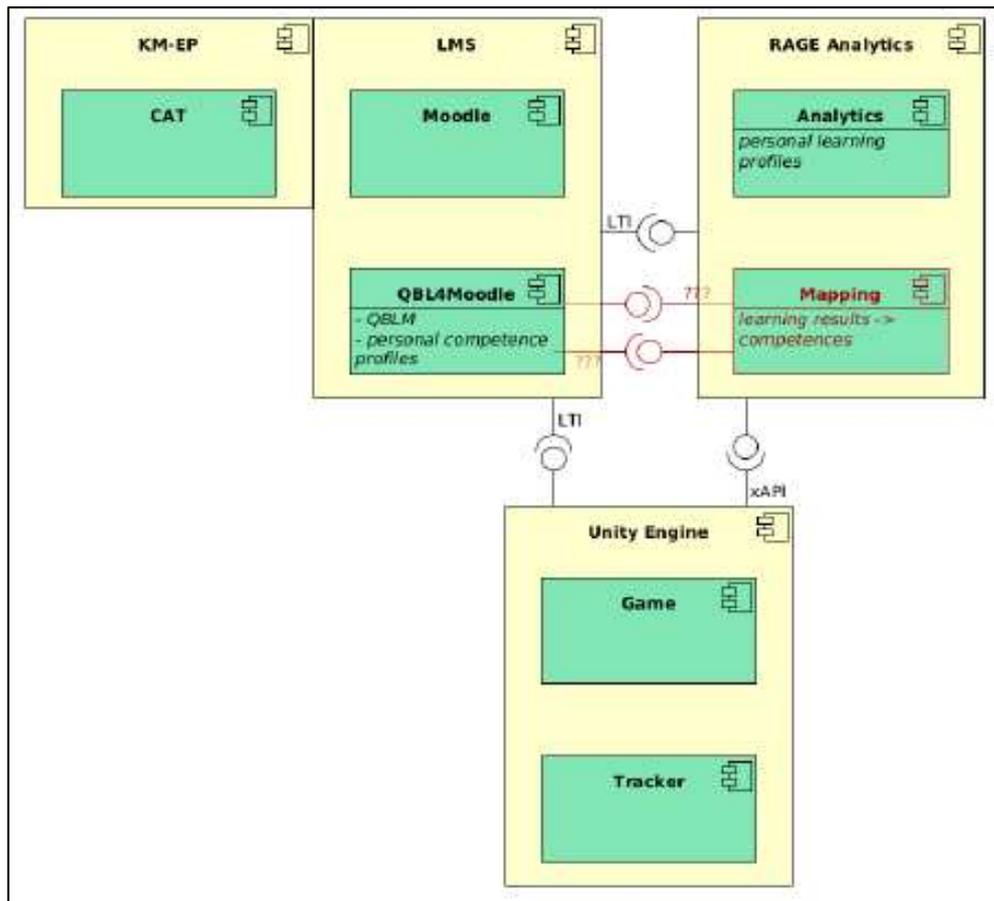


Figure 2: overall system architecture

In the following, the conceptual task assignments and use cases, as shown in Figure 3, will be addressed. This assignment is for the user roles administrator, developer, teacher, and learner. On the LMS and Analytics Engine side, the administrators are making configurational settings. This includes the setup of an External Tool [26] (via the LTI interface) on the side of the LMS to integrate the EduGame or the Analytics Engine. Also, the administrators shall perform the configuration of the Game Analytics on the Analytics Engine side. This includes assigning role permissions and managing Analytics Engine services. Developers shall create the EduGame or adapt an existing EduGame and then connect it to the Analytics Engine and from the Analytics Engine to the CQP of the LMS. Teachers should manage a course in the LMS. The creation of the course should be done in a simplified manner using a CAT and then export the course to the LMS. In the LMS, the teachers should be able to see the learners' CQP, which also considers the results of the EduGame. In the Analytics Engine, which they can call up directly from the LMS, they should be able to monitor the results of the learners. The learners should access the EduGame via a course in the LMS. The learners must be uniquely identified in the EduGame through the LMS. The learning results should be visible in the Analytics Engine or the individual CQP.

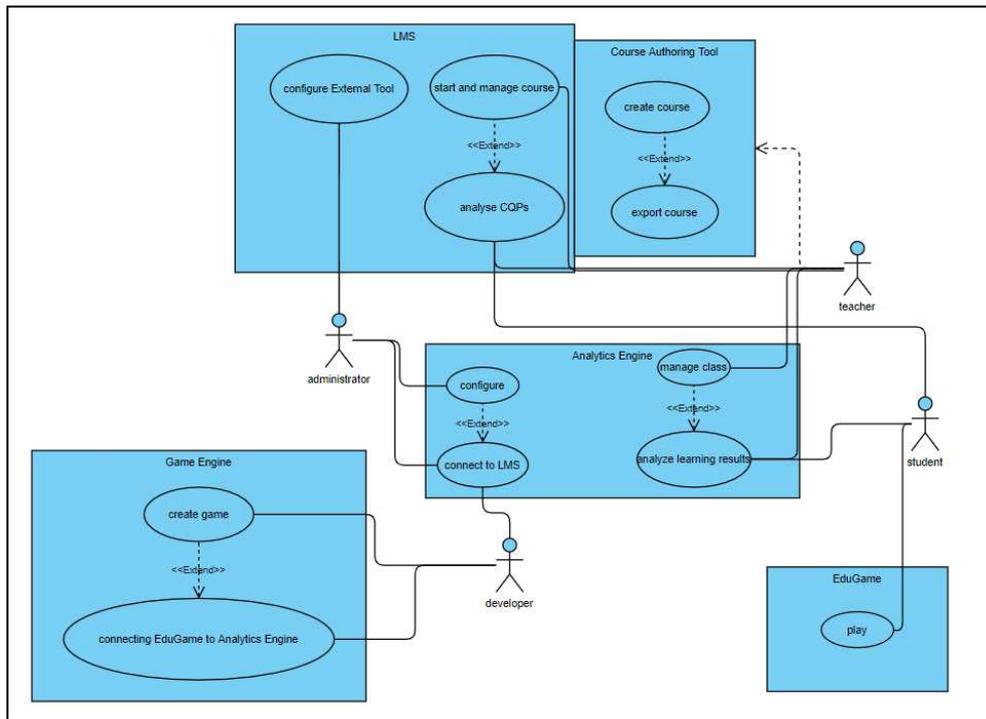


Figure 3: Use Cases

3.2. Concept for a Game Event Framework

With this Game Event Framework (GEF), game events should be defined and mapped (see Figure 2). A game event is subject to 0 to n conditions in the form of other game events. For example, event B could only occur if event A has passed before. In each game event, the player can choose from 1 to n game decisions. These decisions must be included in the dependencies of events to each other. Event B, for example, could only occur if the player had made decision A. A level is assigned to each decision in the model to capture the varying complexity of game decisions. To connect the GEF with the CQ Framework, game events must be set CQ (class Qualification in QBLM), and the game decisions resp. their levels with Proficiency Levels be linked. The GEF contains a game event. When the learner runs through this event, he can choose between three Choices. It is assumed that going through the Game Event extends CQ 1 from the CQ Framework. Therefore, the game decisions that can be made while going through the game event one should differ in their complexity and should each be assigned to one of the levels of CQ 1.

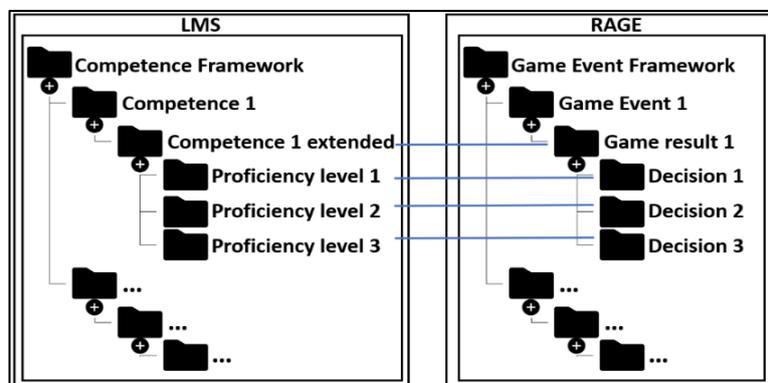


Figure 4: Mapping of the Competence Framework and GEF

Next, the GEF to be implemented within RAGE Analytics is classified in the QBLM Service Distribution Model (QSM) [32]. The QSM describes the communication between Moodle, augmented

by extensions for CQ-based learning, and the RAGE Ecosystem. All components of the model that already exist or are yet to be implemented are considered. Based on the State of development, the implementation of the GEF is assessed. A standard Moodle requires the extensions EduGameConnector and QBL4Moodle. The EduGameConnector has only been conceptually designed so far [32]. This extension is responsible for the interaction with the RAGE Analytics environment (also: Gaming Analytics Suite, GAS) and the P-CAP. The LTI interface is used to authenticate and call RAGE Analytics from Moodle. The tracker, which sends game data to the analytics server, should inform the EduGameConnector about starting games in the future. The EduGameConnector should also be able to retrieve input traces and player-specific data stored in RAGE Analytics. [32] also envisions the EduGameConnector communicating with the P-CAP. The Domain Model Asset included in the P-CAP enables the creation of game-specific CFs. The Competency Assessment Asset provides the CQ States in which achieved CQs are stored. The design of game events in which CQs can be achieved is also done via the Competency Assessment Asset. The EduGameConnector should be able to access the CQ States and retrieve the CQ Frameworks from the P-CAP.

The P-CAP does not use RAGE Analytics components to date. The RAGE Ecosystem does not yet have access to the CQ Frameworks that QBL4Moodle enables. The GEF is equivalent in position to the P-CAP in the QSM. Both models allow the definition of game events in which CQs can be achieved. However, the P-CAP does not yet use RAGE Analytics components to track and store game data. The development status of the overall system consisting of Moodle and the RAGE Ecosystem does not allow for implementing the GEF in its envisioned form within this work scope. The lack of access to CQ Frameworks through RAGE Analytics prevents mapping game events with QBLM-based CQs. The yet-to-be-implemented EduGameConnector receives game data, without which no CQP is possible. Moreover, QBL4Moodle does not yet enable CQPs for individual users [32].

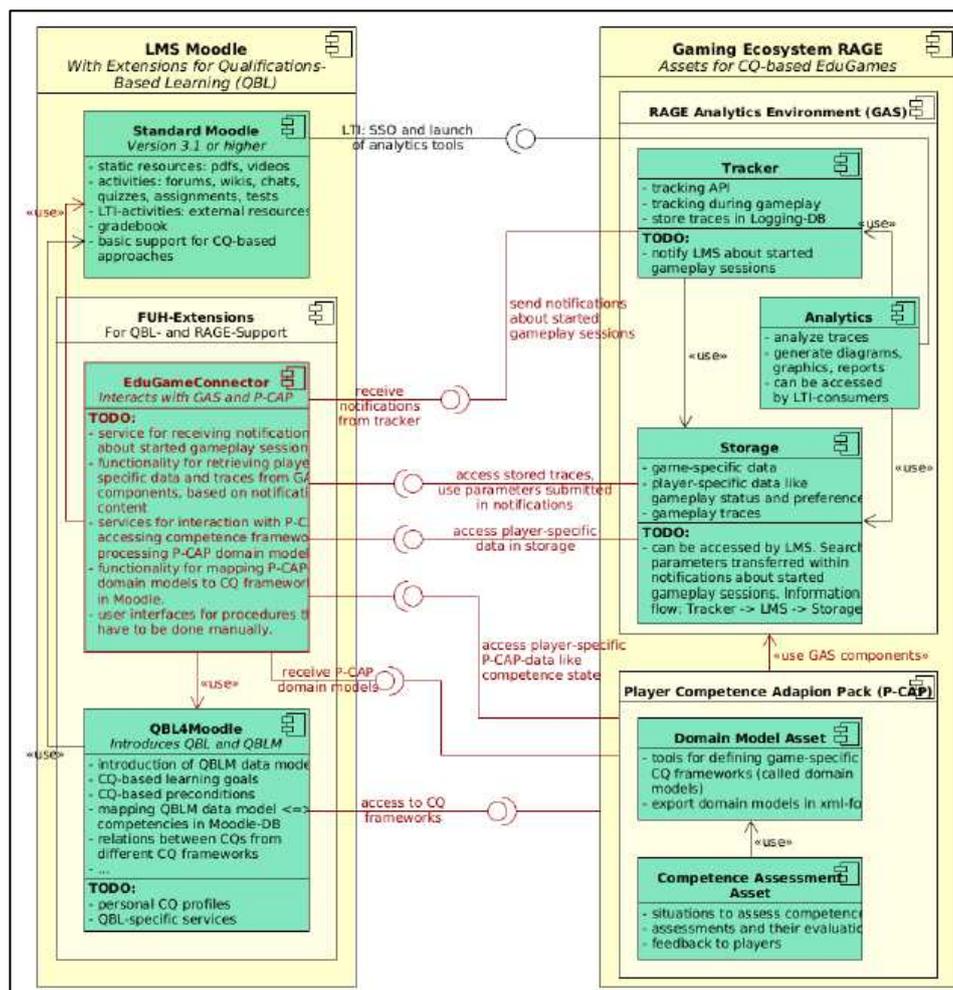


Figure 5: QBLM QSM [32]

3.3. Concept for an adapted System Design

The design will be adapted to map between CQs, and game events required for RQs to occur in the following subsection. The State of development of the overall system consisting of Moodle and RAGE Ecosystem does not allow to implementation of the GEF in its envisioned form within the scope of this work. The lack of access to the CQ Framework by RAGE Analytics prevents mapping game events with QBLM-based qualifications. The yet-to-be-implemented EduGameConnector receives game data, without which no CQP is possible. Besides, QBL4Moodle does not yet enable CQPs for individual users. Therefore, the design was adapted to map game results from RAGE Analytics to CQs defined independently from Moodle. In the context of RQ2, the player's interactions or results are to be accessed via the provided Representational State Transfer (REST) API [17]. The mapping and subsequent analysis will not take place within the RAGE Analytics environment as initially intended. Instead, the game events are transferred to the Analytics Engine via an in-game tracker (see Figure 6).

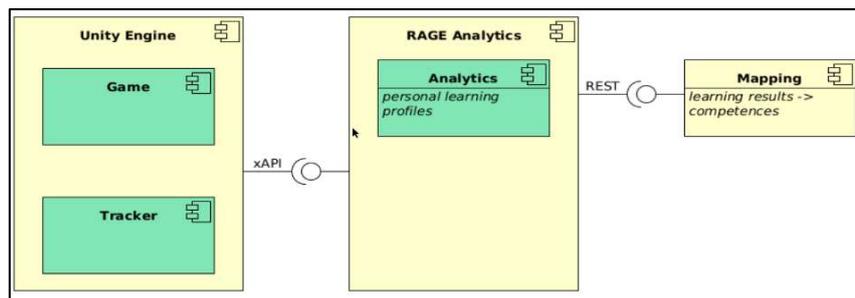


Figure 6: adapted system architecture

The mapping between the game results and the CQs will be made outside of RAGE Analytics. Therefore, the mapping is done in a simplified way. When answering the multiple-choice questions, not several game decisions lead to success. The game event can only be completed with the correctly selected answer. This means that no Proficiency Levels can be mapped either. Each of the multiple-choice questions will be mapped to one of the CQs before the game starts. This will be done via a Comma-separated values (CSV) file [23]. In a CSV file, the data is stored in text form separated by commas and not in column form [23]. The game is then started. The player's results are retrieved in real-time by a script to be created via RAGE Analytics' REST API. Immediately after the retrieval, the received JSON object is evaluated respecting the defined mapping. The success key expresses whether a question was answered correctly or incorrectly. The CQs achieved by answering the multiple-choice questions are written to another CSV file.

The identified game situations from the prototype of [28] can be mapped with standard interactions from the xAPI Serious Games (xAPI-SG) Profile [13]. The xAPI-SG Profile represents a standard tracking format for Serious Games. It was developed to collect general and game-independent information from Serious Games [13]. Only the identified game event Answer Questions are suitable for mapping to CQ. The other events, Open Game and Start Game do not depend on player decisions. The answer questions event can be transferred, e.g., via the xAPI Game Object Completable. Answering multiple-choice questions can be sent to RAGE Analytics via the tracker's Completed method. Correctly answered questions mean a call to the Completed method with the parameter *success = true*, while incorrectly answered questions result in a call with the parameter *success = false*.

4. Proof of Concept Implementation

In the following, the Proof of Concept (PoC) implementation and the results of the SDP will be described. First, the installation and configuration of RAGE Analytics will be defined. Afterward, the integration of a tracker for RAGE Analytics into the SGP will be addressed and followed by describing the retrieval of the game events and the automated mapping between the game events and the CQ model.

4.1. Installation and Configuration of RAGE Analytics

The user must have a Linux operating system, a Docker version, and Docker Compose installed to install RAGE Analytics. With the help of Docker, it is possible to separate the infrastructure from the application so that the software can be deployed faster. For this, Docker provides the ability to package and run applications in what is called a container. A container contains everything needed to run the application [8]. Multi-container Docker applications can be defined and executed using the Docker Compose tool [9]. In a Compose file, the application's services can be configured and then started via a single command. The A2 module [12] is used to manage permissions for users and connected services. Users create their account themselves and are assigned a role by "root". Next to the predefined roles, own roles can be added. The "root" password is created during the RAGE Analytics installation. In this prototype implementation, the users "developer," "teacher," and "student" were created and assigned to the appropriate roles. The Representational State Transfer (REST) API [17] is used to access and process the data received from the Unity game. The teachers' account is used for this purpose.

It must be ensured that the teachers' account has the appropriate access to the REST API endpoints. The REST API only provides access to aggregated match data in the form of a simplified JSON file. Elasticsearch [14] provides all the received data. Elasticsearch is an open-source distributed search engine and analytics engine for data. Raw data can be received from various sources, parsed, normalized, enriched, and then indexed in Elasticsearch. The data is stored as JSON documents, and Elasticsearch stores it in what is called an inverted index. In an inverted index, each word that occurs at least once in the documents is stored to indicate its occurrence [14]. This data structure enables a fast full-text search. This work is limited to the use of the REST API since the available data is elementary, and the feasibility should be shown. Elasticsearch is also freely accessible without further security protection and would have to be secured for productive use.

4.2. Tracker Integration of RAGE Analytics into the EduGame

To send game data from the SGP to RAGE Analytics, a tracker must first be integrated into the game. Therefore, the unity tracker from the RAGE project is used. An empty game object is created in the Unity editor in the first scene (here: MainMenu), renamed Tracker, and linked with the script Tracker.cs from the cloned directory. The game object Tracker can now be used to define parameters for the transfer of the game data. The tracker remains in the further scenes of the game.

For the extension of the Unity scripts, statements from xAPI-SG Profile [13] are used. The game logic is mapped via scripts in the Assets/Scripts folder. The Update() method has been added to the script to capture that the MainMenu is opened. When the player returns to the MainMenu after a game round, this action must not be captured again. As mentioned above, tracking is done using predefined xAPI statements. For tracking, a so-called Accessible object can be used. This is done in the code via the method Accessed(). The game logic for answering the multiple-choice questions is in the Multiple Choice Questions class. A Completable object describes whether the question was answered correctly or incorrectly. The script's implementation is done by the method Completed(). It is possible to create custom statements within the script if xAPI Profile statements are not enough, but for this PoC implementation, custom statements had not been required.

4.3. Automated Mapping of the Data

Game data submitted to RAGE Analytics is not accessed directly through Elasticsearch but through the REST API. Required for authentication and retrieval of learning activity results are the REST API endpoints /API/login and /API/activities/:activityId/results. Access to the REST API is tested using the Postman API [30] development environment. Postman is an API development environment and available for popular operating systems. Postman covers the entire API development process by providing test servers, API monitoring, automated documentation, and management of test scenarios and data through its own API [30]. A POST request generates a token that is needed for further

authentication. The account of the "teacher" is used since this person has access to the learning activity. The generated token is used for a GET request, which creates a JSON object containing the aggregated game data. This means that the number of entered screens (xAPI-Verb accessed) and the evaluation of answered questions according to true and false (xAPI-Verb completed) are displayed cumulatively.

The next step is to automate the retrieval of the data using a PHP script. The goal is to output the data in a CSV file, as this can be easily read in and further processed by Moodle. It should also be possible to map game events (here: multiple-choice questions) and CQs. A simple mapping of CQs to multiple-choice questions will occur in another CSV file shown in Figure 7. The CSV file with the contained mapping is provided to the PHP script before execution.

QuestionID	CQ	Question
1	A	Was liegt zwischen Berg und Tal?
2	B	Was wiegt schwerer: 1kg Blei oder 1000g Federn?
3	C	Ein Bauer hat 10 Schafe, alle sterben außer 9, wie viele Schafte hat er noch?
4	D	Wie viele Tiere nahm Moses mit auf die Arche?

Figure 7: Assignment of CQs to questions

To bundle results and achieved CQs for further processing, the mapping of CQs to questions should be included in the CSV file that stores the game results. The desired CSV file `file.csv` for outputting the game results with assigned CQs can be generated using a PHP script (see Figure 8).

Question	success	score	CQ
Was liegt zwischen Berg und Tal?	1	1	1 A
Was wiegt schwerer: 1kg Blei oder 1000g Federn?	1	1	1 B
Ein Bauer hat 10 Schafe, alle sterben außer 9, wie viele Schafte hat er noch?			1 C
Wie viele Tiere nahm Moses mit auf die Arche?			1 D

Figure 8: Output of game results with assigned CQs

The script requires the file `mapping.csv` for the assignment of CQs to multiple-choice questions. The results file `file.csv` is generated in real-time while the game is running. The functionality of the script is briefly explained below. The `login` function returns the token for authentication. Via `getRequest`-function, a GET request is made at the endpoint URL. Authentication is done with the passed parameter `token`. The `getResult`-function is the heart of the script. The script first executes the two explained functions and writes the JSON Object to the CSV file `file.csv`. It is essential to mention that a multiple-choice question is only in the JSON Object if the player has answered it (correctly or incorrectly). Afterward, the CSV file `mapping.csv` is read, the (answered) questions in `file.csv` are read via the `questionID`. The correct answers are identified, and the corresponding CQ is also written to the `file.csv` file. With the functions described so far, it is possible to query the data via the REST API by manually executing the PHP script. Below is a description of how the CSV file `file.csv` is updated only when there is new game activity. The `isChanged`-function reads the JSON object's timestamp and writes it to the `timestamp.txt` file. Then the function compares this timestamp with that of the most recent JSON object, sets the variable `"changed = true"`, and returns the string `"Changed"` if it is not equal. The function `start()` runs through a while-loop with a delay of 5 seconds and updates the file `file.csv` if the timestamp changes.

5. Initial Evaluation

In the form of a CW [38], an initial evaluation of the PoC implementation has been accomplished by domain experts in computer science education. The evaluation's main goal was to estimate the

productive capacity of the implementation and orientate the future development. During the CW, the following four tasks have been accomplished and evaluated:

1. Configuration of RAGE Analytics
2. Playing the EduGame
3. Evaluation of the EduGame outcomes visualization in RAGE Analytics
4. Mapping game events and CQ model

All four tasks are derived from the RQs mentioned in the first section of this paper. In the following four tasks, their goals and the evaluation results of the EP will be addressed.

The goal of the first task is to create the prerequisites in RAGE Analytics for the successful connection and evaluation of an educational game. This includes settings as well as the visualization of the data. First, the system components of RAGE and the related services have been configured and evaluated. Then, the prerequisites for the connection and evaluation of an EduGame could be created. Finally, a learning activity was successfully created, and a tracking code for the Unity game was generated, as shown in Figure 9, with it. Upon configuration, a default dashboard is displayed in RAGE Analytics to visualize the data received from the tracker. The dashboard can be customized with the "developer" role using Kibana [22]. Using "Edit" in Kibana, the dashboard is customized and saved to include only the AccessibleAccessedAllperTargetId, CompletablesCompletedperTarget, and xAPIVerbsActivity panels. This is done to be able to visualize the Accessibles (screens entered), Completables (questions answered) and xAPIVerbs (accessed, completed, testverb) used in the game according to the frequency of their use.

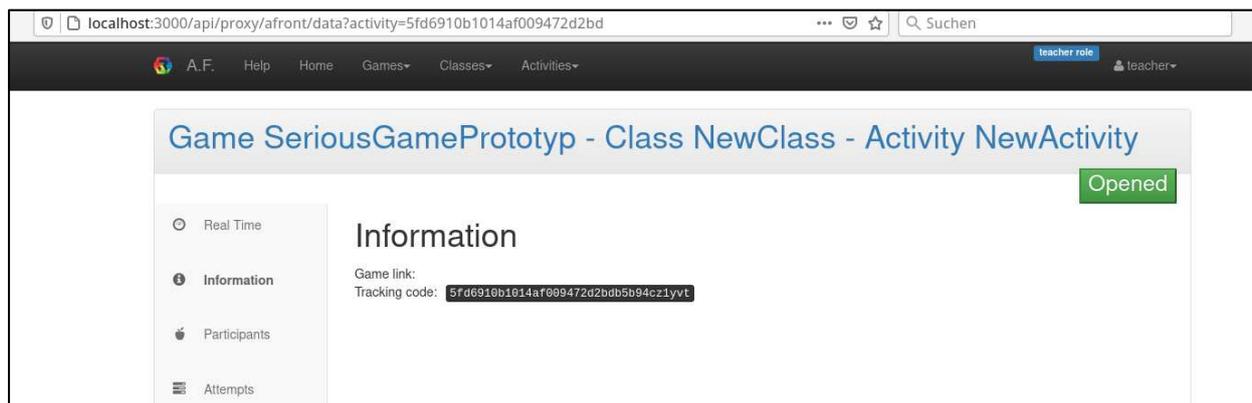


Figure 9: RAGE Analytics Frontend: newly created Learning Activity and corresponding Tracking Code

The second subtask aimed to run through the Unity game, which was extended by a tracker, without errors or aborts. Since the integration of the EduGame in Moodle as an External Tool was not implemented as part of the PoC implementation, the EduGame is started directly from the Unity environment. For this, the corresponding project is selected with the "Learner" role in Unity. The game is started by the user via "Play" and the players must enter their names manually (see Figure 10).

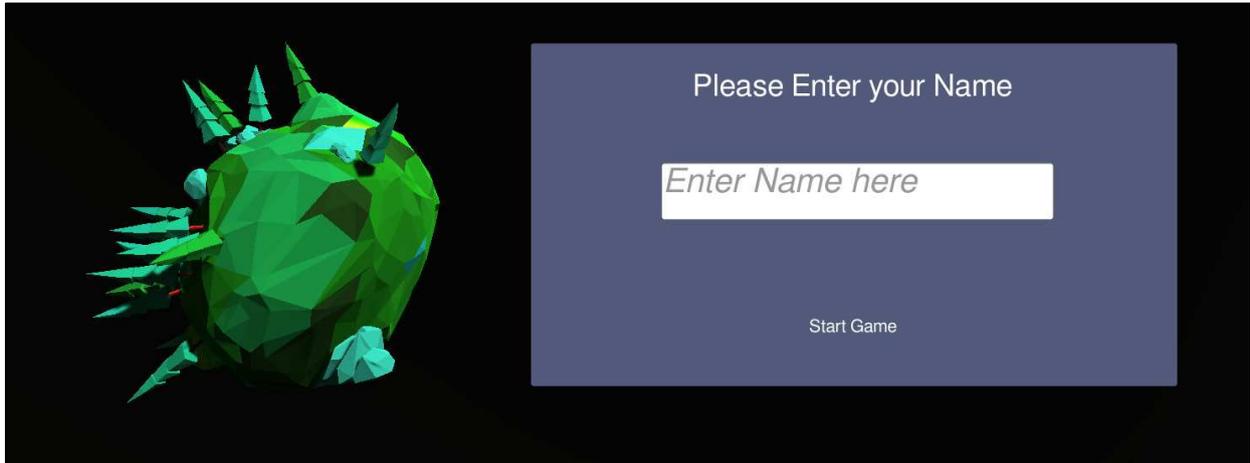


Figure 10: EduGame start screen

The game area includes a terrain, a display of raw materials (Wood, Stone) and three buttons for the construction of buildings (Woodfarm, Stonefarm, Tavern). The learner creates a first building with a left click on the Woodfarm button and another left click on the terrain (See Figure 11).

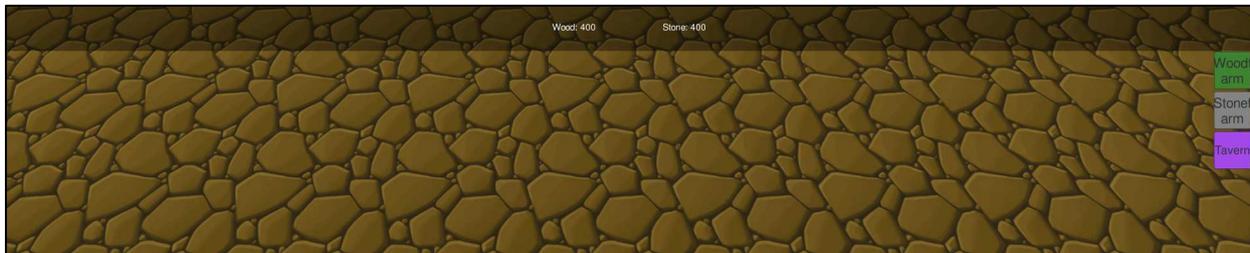


Figure 11: empty game interface of the SGP

The construction has cost raw materials, the new stock is displayed. After a left click on the Woodfarm a multiple-choice question appears. The learners answer the multiple-choice question (see Figure 12) by left clicking with the answer and receives a "Correct" as confirmation for the correct answer or "wrong" for an incorrect answer. If the answer is correct, the raw material Wood is continuously increased by two units. After a left click on the Woodfarm another multiple-choice question appears. The learners can end the game by clicking the pause button in the toolbar. The game has been played and evaluated if any errors or misbehaviors occurred. No errors or terminations were detected.



Figure 12: SGP game interface with a multiple-choice question displayed

The aim of the third subtask was to successfully evaluate the game data transmitted by the Unity game in the RAGE Analytics frontend. As a result, the dashboard in RAGE Analytics correctly maps student's game results (See Figure 13).

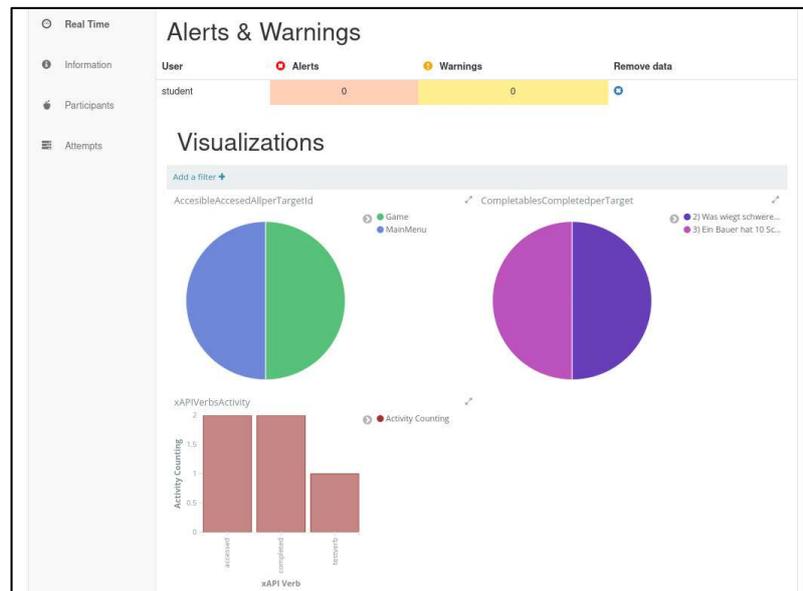


Figure 13: RAGE Analytics dashboard

The object of the fourth subtask is to obtain a CSV file with the achieved CQs after a game run. The mapping between game events (here: multiple choice questions) and CQs must be defined previously. The mapping between the multiple-choice questions and the CQs done via the CSV file is transferred to the results file. Together with the game results obtained via the REST API, the game's CQs can be read from the results file. For RQ1, the system components of RAGE and the related services have been configured and evaluated. No errors or misbehavior were detected. Based on the evaluation, improvements and renewals were identified. These will be implemented and considered in future work.

6. Conclusion and Future Work

In this paper, a mechanism for automated mapping for CQs gained in an EduGame and a player-specific evaluation of EduGame events have been introduced. In the concept section, the system's architecture and the mapping of CQs between the EduGame and the LMS had been presented. Afterward, the initial PoC implementation and the evaluation had been described. In the PoC implementation, a tracker was included in the Unity game, and the game scripts were adapted so that game events could be sent to RAGE Analytics. With the dashboard in RAGE Analytics, the game events could be successfully displayed. Therefore, RQ1 "Can player individual evaluations of EduGame events be achieved via Analytics Engines be achieve?" is addressed.

Regarding RO2, it was determined in the TBP that, based on current research, too many limitations exist to map game events in Moodle CQPs. Therefore, the conceptual system design was adapted. A mapping that is independent of CQ Frameworks mapping between game events and CQs could be implemented. Therefore, RQ2 "Can a model be designed to map learning game events and CQ in a CQ model?" is addressed.

Based on the evaluation, improvements and renewals were identified. The renewals and improvements found can be directly implemented as improvements from the CW. One of them is the easier usability of the RAGE Analytics dashboard. An improvement identified in the CW is that the dashboard in RAGE Analytics should be configurable to fit the game events used in the EduGame. Currently, the interface is only understandable for expert users and not for teaching staff. For this purpose, a configuration tool could be considered, which would allow configuring the dashboard in RAGE Analytics easily. A second direct follow-up effort could address the connection of the REST

API of RAGE Analytics to Moodle using the EduGameConnector component, which is yet to be developed, and CQPs. The more robust integration of EduGames or EduGame platforms in Moodle offers the opportunity to strengthen the LMS in its position as a central digital learning location and increase acceptance among learners and teachers. Moodle and the KM-EP's extensions, which are currently only conceptual, would have to be realized in future developments. Also, a follow-up effort could address user authentication between the Moodle, RAGE Analytics, and Unity Engine systems. It would be desirable if a user login to Moodle were sufficient to access RAGE Analytics and legitimize themselves in the Unity game.

Future investigation and evaluations need to be done in the case of the implementation of the EduGameConnector and P-CAP. Also, evaluations and concepts regarding the usage of CbKST based algorithms for the analysis of CQs and the certification of CQs. Since right now the prototype does not analyze the in-game behavior regarding CQs. Therefore, the projects Adaptive Learning with Knowledge Spaces (ALEKS) [4] and Enhanced Learning Experience and Knowledge Transfer (ELEKTRA) [15] are used as references and analyzed to what extent the algorithms and concepts used there can be used.

7. References

- [1] ADL Initiative U.S. Department of Defense: Project xAPI github, 2013. URL: <https://github.com/adlnet/xAPI-Spec/blob/master/xAPI-About.md#partone>.
- [2] ADL Initiative U.S. Department of Defense: Project xAPI, 2020. URL: <https://adlnet.gov/projects/xapi/#resources>.
- [3] D. Albert, J. Lukas. Knowledge Spaces, Lawrence Erlbaum Associates, Mahwah, 1999.
- [4] ALEKS: ALEKS 2021. URL: <https://www.aleks.com/>.
- [5] S. Axon: Unity at 10: For better- or worse- game development has never been easier, 2016. URL: <https://arstechnica.com/gaming/2016/09/unity-at-10-for-better-or-worse-game-development-has-never-been-easier>.
- [6] J. C. Beck: The videogame that teaches business strategy better than professors, 2015. URL: <https://qz.com/526534/the-video-game-that-teaches-business-strategy-better-than-professors/>.
- [7] C. Dalsgaard, Social software: E-learning beyond learning management systems in European Journal of Open, Distance and E-Learning, volume 9, 2016.
- [8] Docker: Docker 2021. URL: <https://docs.docker.com/get-started/overview/>.
- [9] Docker Compose: Docker Compose, 2021. URL: <https://docs.docker.com/compose/>.
- [10] R. Dillet: Unity CEO says half of all games are built on Unity, 2018. URL: https://techcrunch.com/2018/09/05/unity-ceo-says-half-of-all-games-are-built-on-unity/?guccounter=1&guce_referrer=aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnLw&guce_referrer_sig=AQAAAJzvctT-NWpjAyurVPYHBip0WcEuFQRo61nbUIsB7Yvr172go-9Chw4oIiaxmpNssMmFa_n-Y9hE2kbpPpje.
- [11] RAGE Analytics Overview: RAGE Analytics Overview, 2017. URL: <https://github.com/e-ucm/rage-analytics/wiki/Rage-analytics-Overview>.
- [12] RAGE A2: RAGE A2 overview, 2018. URL: <https://github.com/e-ucm/rage-analytics/wiki/A2-Overview>.
- [13] RAGE xAPI SG Profile: RAGE xAPI SG Profile, 2019. URL: , <https://github.com/e-ucm/rage-analytics/wiki/xAPI-SG-Profile>.
- [14] Elasticsearch: Elasticsearch – was ist Elasticsearch?, 2021. URL: <https://www.elastic.co/de/what-is/elasticsearch>.
- [15] ELEKTRA: ELEKTRA, 2021. URL: , <https://cordis.europa.eu/project/id/027986>.
- [16] FernUniversität in Hagen: FernUniversität in Hagen, Germany's State Distance-Learning University, 2021. URL: <https://www.fernuni-hagen.de>.
- [17] R.T. Fielding, Architectural Styles and the Design of Network-based Software in University of California, Irvine, 2000.
- [18] M. Freire, A. Serrano-Laguna, B. Manero, I. Martínez-Ortiz, P. Moreno-Ger, B. Fernández-Manjón, Game Learning Analytics: Learning Analytics for Serious Games in Learning, Design, and Technology, p. 1-29, 2016.

- [19] Game-Verband der deutschen Games-Branche e.V.: Jahresreport 2020, 2020. URL: <https://www.game.de/wp-content/uploads/2020/08/game-Jahresreport-2020.pdf>.
- [20] J. Heller, B. Mayer, D. Albert, Competence-based Knowledge Structures for Personalised Learning in 1st International ELeGI Conference on Advanced Technology for Enhanced Learning, Vico Equese, Italy, pp.8. fihal-00190482f, 2005.
- [21] IMS Global Learning Consortium: Learning Tools Interoperability, 2021. URL: <https://www.imsglobal.org/activity/learning-tools-interoperability>.
- [22] Kibana: Kibana, 2021. URL: <https://www.elastic.co/de/kibana>.
- [23] Microsoft: Outlook CSV, 2019. URL: [https://support.microsoft.com/de-de/office/erstellen-oder-bearbeiten-von-csv-dateien-zum-importieren-in-outlook-4518d70d-8fe9-46ad-94fa-1494247193c7#:~:text=Eine%20CSV%2DDatei%20\(durch%20Kommas,CSV%2DDateien%20durch%20Kommas%20getrennt,14.4.2021](https://support.microsoft.com/de-de/office/erstellen-oder-bearbeiten-von-csv-dateien-zum-importieren-in-outlook-4518d70d-8fe9-46ad-94fa-1494247193c7#:~:text=Eine%20CSV%2DDatei%20(durch%20Kommas,CSV%2DDateien%20durch%20Kommas%20getrennt,14.4.2021).
- [24] B. Medler, Generations of game analytics, achievements and high scores in Journal for Computer Game Culture, Eludamos, version 3(2), p.177-194, 2003.
- [25] Moodle.org: Moodle open-source learning platform, 2021. URL: <https://moodle.org>.
- [26] Moodle.org: External Tool, 2021. URL: https://docs.moodle.org/311/en/External_tool.
- [27] J.F. Nunamaker, M. Chen, T.D. Purdin, System development in Information System Research in Twenty-Third Annual Hawaii International Conference on System Sciences, pp. 631-640 vol.3, doi: 10.1109/HICSS.1990.205401, 1990.
- [28] T. Neuber, Konzeption und Realisierung einer interaktiven Spiele-Laufzeitumgebung in ein mandantenfähiges Weiterbildungsportal auf Basis einer Game-Engine und eines Learning Development Systems, Bachelor thesis, Hochschule RheinMain, 2018.
- [29] A. Nussbaumer, M. Maurer, S. Malicet, C. M. Steiner, D. Albert, A novel approach and software component for supporting competence-based learning with serious games in INTED2019 conference, 2018.
- [30] Postman: Postman, 2021. URL: <https://www.postman.com/>.
- [31] E. Soares de Lima, B. Fejió, A. Furtado, Player Behavior Modeling for Interactive Storytelling in Games in SBC Proceedings of SBGames, 2016.
- [32] M. Then, Supporting Qualifications-Based Learning (QBL) in a Higher Education Institution's IT-Infrastructure, PhD thesis, FernUniversität in Hagen, 2020. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001608, 18 June 2021. DOI: 10.18445/20200309-141118-0.
- [33] M. Then, B. Wallenborn, M. Fuchs, M. Hemmje, Towards a Domain Model for Integrating Competency Frameworks into Learning Platforms in Formamente – International Research Journal on Digital Future, issue 2016/3-4, p. 15-49, 2016.
- [34] Unity: Unity Analytics, 2021. URL: <https://docs.unity3d.com/Manual/UnityAnalytics.html>.
- [35] B. Vu, Taxonomy Management System Supporting Crowd-based Taxonomy Generation, PhD thesis, FernUniversität in Hagen, 2020. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001613, 18 Jun 2021. DOI: 10.18445/20200404-144028-0.
- [36] B. Wallenborn, Entwicklung einer innovativen Autorenumgebung für die universitäre Fernlehre, PhD thesis, FernUniversität in Hagen, 2018. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00001428, 18 Jun 2021. DOI: 10.18445/20180911-091907-0.
- [37] C. Wilson, User interface Inspection Methods: A User-Centered Design Method, Elsevier Science, Niederlande, 2013.

Examinations in the Context of Curriculum Content: Case Study of a 1926 Irish Mathematics Exam Paper

Hazel Murray^a, David Malone^b

^a*Munster Technological University (MTU), Bishopstown, Cork, Ireland*

^b*Maynooth University (MU), Maynooth, Co. Kildare, Ireland*

Abstract

Examinations are used worldwide as a method of assessing student ability and learning. The content and form of these examinations has developed over the years as university entry and higher education has become mainstream. In this work, we collected a selection of mathematics exam papers dating from 1926 to 2020. While we may be tempted to use this collection to determine the capabilities and complexity required by the mathematical content covered at the time, we demonstrate that examination papers alone cannot provide a clear insight into depth of learning. We select two questions from a 1926 exam paper and demonstrate that a targeted and potentially rote-learned curriculum can mean that seemingly difficult topics can be reduced to repetition exercises, potentially without analysis or comprehension.

Keywords

Education, Curriculum, Examination, Mathematics

1. Introduction

In our current technological world, we seek the skills of mathematicians and problem solvers in our economy and society. However, a reoccurring narrative is that the material covered in state-run examinations is becoming easier and tending away from in-depth pure mathematical explorations [1]. Developments in recent Irish examinations have sought to focus on students' understanding and comprehension of content, and therefore has been criticized for "simplifying the curriculum" [2].

In this paper, we investigate the content of a 1926 exam paper and uncover the context behind the seemingly complex mathematical problems by revealing the targeted and superficial curriculum content. We demonstrate that examination papers cannot be used as a means of assessing the depth of knowledge expected of students unless they are considered in the context of the curriculum content.

We begin in Section 1.1 with a sample of the related work in this field. Following this, Section 1.2 introduces the Irish examination structure and the taxonomy of Irish exam papers we collected. In Section 2, we introduce the exam paper questions we have chosen to focus on in this study. In Section 3, we present the modern solutions to these two exam questions. Then, in Section 4 we leverage contemporary textbooks from the time period to investigate how stu-

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ hazel.murray@mtu.ie (H. Murray); david.malone@mu.ie (D. Malone)

ORCID [0000-0002-5349-4011](https://orcid.org/0000-0002-5349-4011) (H. Murray); [0000-0002-6947-586X](https://orcid.org/0000-0002-6947-586X) (D. Malone)

© 2021 Copyright for this paper by its authors.



Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

dents might have been expected to answer the same questions. We observe that the textbooks of the time provide only segmented explanation, and include examples closely matched to the examination questions. Section 5 reviews our findings and discusses the impact they have on our understanding of the content and examinations of the past. Finally, we conclude in Section 6.

1.1. Related work

Many researchers have studied exam papers to determine the depth of understanding expected of students and to evaluate the complexity of the topic under examination. In 2017, Mac an Bhaird et al. [3] studied undergraduate calculus assessments including assignments and examinations to determine the amount of critical reasoning versus imitative reasoning (rote learning and repetition) opportunities in assessments for students. In 2020, Akçay et al. [4] studied Turkish exam papers for 5th, 6th, 7th and 8th classes in second level education. The exam papers were examined with respect to the number and type of questions, the language used in the questions, the visuals used, and the cognitive level (according to Bloom's taxonomy [5]) of the questions. They show that the questions are mostly at the comprehension level, and critical reasoning questions are rare. Cebesoy et al. [6] studied a collection of science examination papers with the goal of investigating the types of questions used, the consistency between questions and objectives in the unit, and whether there included questions related to mathematics. With knowledge of the unit content and the exam papers, all these questions could be answered and they found that most question were simple multiple choice but that they did attempt to test students' mathematical knowledge. Kang and Saeed [7] evaluated 2014-2015 mathematical examination papers from the Secondary School Certificate (SSC) Examinations and from the General Certificate of Education (GCE), which run in parallel in Pakistan. Their study showed that the SSC questions are highly focused on factual knowledge and routine procedures. They conclude that these items assess knowledge of facts and algorithms only, and do not measure essential mathematical skills.

Exam papers are a valuable source of information to assessors of current mathematical programs. They are also useful as a method to allow us to evaluate the historical development of educational priorities and gain insights into examinations from the past. However, as we will show in this paper, sometimes exam papers alone cannot reveal whether students are expected to employ critical or imitative reasoning unless we also look at the content these examinations are based on. We chose two Irish past examination questions, which on the surface seem to require critical thinking and complex algorithmic reasoning, but once taken in the context of the textbooks of the day, reveal that only superficial retention was expected.

1.2. Exam paper collection

The Irish secondary school system, broadly speaking, currently examines students after three years of secondary education at an age of approximately 15–16 (the *Junior Certificate*, formerly the *Intermediate Certificate*) and again, after a further 2–3 years at the *Leaving Certificate*. The system has run since 1925, with disruption in only exceptional circumstances, such as the 2020–2021 Covid pandemic. However, it is worth noting that there has been significant changes over

the years to the surrounding school system, such as the introduction of free secondary school education for all students in 1967.

In 2016, we began a collection of STEM (Science, Technology, Engineering and Mathematics) examination papers from the Irish second-level school system [8]. This collection was motivated by the sometimes imperfect memories of the syllabus and examinations that people had reported they had been through. As no collection of these examination papers was available to the public, or researchers, we aimed to construct an archive that would be of use to both the general public and researchers in STEM education.

A cross-section of the available Junior certificate and Leaving certificate subjects are included in this archive. At Junior Cert level this includes Maths, Science and Mechanical Drawing/Technical Graphics. At Leaving Cert level it includes Maths, Applied Maths, Physics, Chemistry, Biology, Technical Drawing/Graphics and Computer Science. Some of these subjects have not been available for the full period of the examinations (e.g. Biology and Computer Science). The archive is now largely complete, and has been compiled with the help of libraries, government bodies and the general public.

2. The questions

To examine a past examination paper, we chose the 1926 Leaving certificate mathematics paper [8]. This paper would be taken by students in their final examination in second level before potentially progressing to third level education. We select the 1926 examinations as these are some of the earliest examinations run by the newly established Irish state, and give us a good opportunity to view questions at a distance, where we have only limited preconceived knowledge of what students are being asked.

We have chosen two questions from the paper, Q3(a) and Q5, as these seem to differ the most from the mathematics exam questions that 21st century students would be expected to answer.

Q3: Convergence There are two parts to Q3, and you are asked to do either Q3(a) or Q3(b). We concentrate on the former, which is shown in Figure 2.

First, notice that the question includes some terminology that may not be familiar to modern readers. The phrase *decreases without limit* is used in a number of textbooks from the 1800s (including one of Augustus De Morgan's [9]), and appears to mean that the sequence decreases

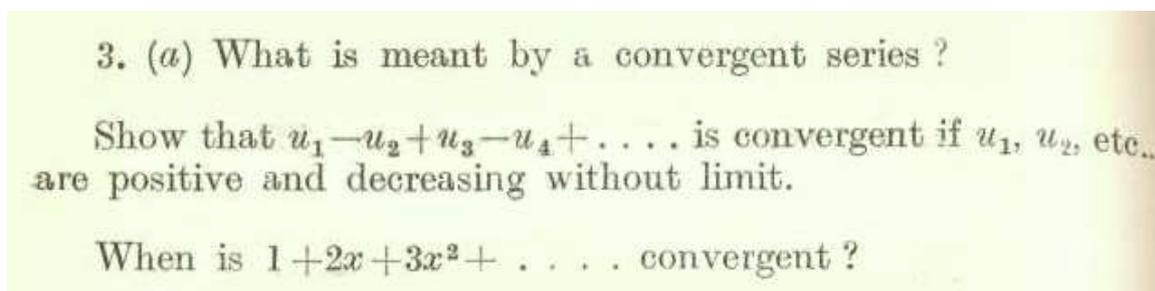
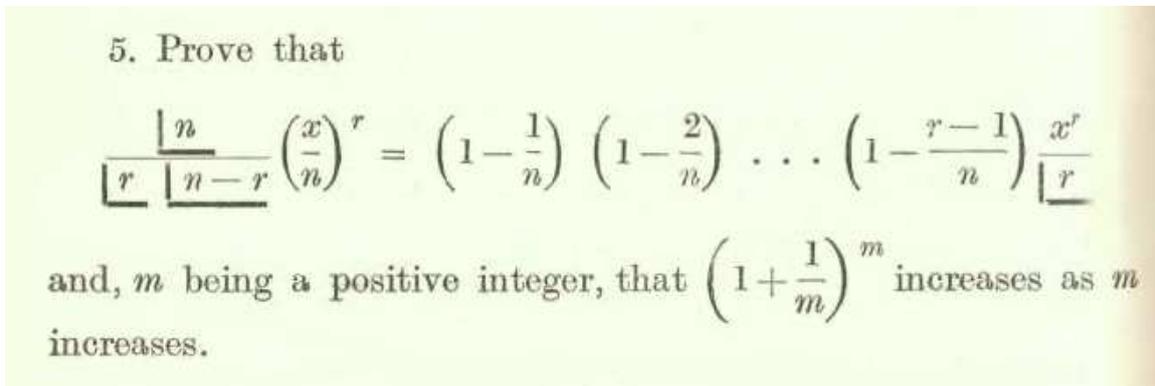


Figure 1: Leaving Certificate 1926 Higher Level Maths Q3(a).

to zero, without becoming constant. Thus, the question is asking about the Alternating Series Test [10].

Q5: Binomial Theorem Q5 is shown in Figure 2. It also contains some notation that could be unfamiliar to the modern reader [n is notation for $n!$]. It concerns binomial coefficients and a well-known limit for e .



5. Prove that

$$\frac{\binom{n}{r}}{\binom{n}{n-r}} \left(\frac{x}{n}\right)^r = \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{r-1}{n}\right) \frac{x^r}{n^r}$$

and, m being a positive integer, that $\left(1 + \frac{1}{m}\right)^m$ increases as m increases.

Figure 2: Leaving Certificate 1926 Higher Level Maths Q5.

3. Possible Modern Solutions

In this section, we will provide modern solutions for the two questions in the 1926 exam paper.

Q3(a): Convergence Q3 appears to require definitions and proofs from an initial third-level course in real analysis. For example, one might first define convergence as

A series with terms a_n converges if $\exists L \in \mathbb{R}$ so that $\forall \epsilon > 0$ we can find $N \in \mathbb{N}$ so that whenever $n \in \mathbb{N}$ and $n > N$ we have

$$\left| \sum_{k=1}^n a_k - L \right| < \epsilon.$$

The second part asks for a proof of the Alternating Series Test. We will not give the full details, but a typical proof of this might proceed as follows.

Let s_n be the partial sums, then note that s_{2n} is an increasing sequence and s_{2n-1} is a decreasing sequence. A little effort will show that $s_2 \leq s_{2n} < s_{2n-1} \leq s_1$. So, we see that the subsequence of even terms is increasing and bounded above, and so by the Monotone Convergence Theorem is convergent. Similarly, the subsequence of odd terms is decreasing and bounded below, and also convergent.

Now, armed with the convergence of these two subsequences, we observe that the difference of their limits is

$$\lim_{n \rightarrow \infty} s_{2n} - \lim_{n \rightarrow \infty} s_{2n-1} = \lim_{n \rightarrow \infty} s_{2n} - s_{2n-1} = \lim_{n \rightarrow \infty} -u_{2n} = 0,$$

so the subsequences have the same limit. A short $\epsilon - \delta$ argument shows that a sequence composed of odd and even terms converging to the same limit converges to that limit.

The final part of the question can be addressed using the Ratio Test.

The Ratio Test shows that the given series will converge when

$$\lim_{n \rightarrow \infty} \left| \frac{(n+1)x^n}{nx^{n-1}} \right| = |x| < 1.$$

The same test tells us that the series will diverge if $|x| > 1$. The case when $x = 1$ is obviously divergent and $x = -1$ gives an alternating non-convergent sequence.

Thus we have answered Q3(a).

Q5: Binomial Theorem Q5 has two parts. In modern notation, the first part asks to show that

$$\frac{n!}{r!(n-r)!} \left(\frac{x}{n}\right)^r \frac{x^r}{r!} = \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \left(1 - \frac{r-1}{n}\right) \frac{x^r}{r!}.$$

The solution can be given in a few lines of relatively straightforward algebra using factorials. One solution to the second part involves expanding $(1 + 1/m)^m$ using the Binomial Theorem and then observing that the terms of the binomial expansion can be rewritten using the first part of Q5 with $x = 1$ and $n = m$. The required result follows by noting that the terms on the right hand side decrease when $n = m$ increases, and that the binomial expansion for $n = m + 1$ has an extra positive term.

Thus these are what we would now consider to be correct solutions to the 1926 questions. Note, that the full derivation of these solutions might currently be considered outside the scope of a second-level mathematics course.

4. Solutions based on Contemporary Textbooks

We identified contemporary textbooks [11, 12] that use both the terminology of *decreasing without limit* and the older notation for factorials used in the exam paper. It seems reasonable that the students, or at least their instructors, must have been expected to have access to such a text. We aim to gain some insight into what might have been expected of a student in 1926 through these text books.

Q3(a): Convergence Beginning with Q3, for example, §123 of Hymers [11] provides the following definition for the convergence of a series:

A series $u_1 + u_2 + u_3 + \dots + u_n \dots$ (to ∞) is called convergent, if the sum s_n of any number n of its terms approaches continually to a finite quantity s as its limit, when n is indefinitely increased; and divergent in the contrary case.

The word *continually* does not appear to be formally defined in the text, and thus the student does not have to struggle with any arguments involving ϵ .

In fact, this text provides a proof of the Alternating Series Test in §132, using almost the same notation and the same terminology as the Leaving Cert question.

The alternating series $u_1 - u_2 + u_3 - u_4 + \&c.$, is convergent, if the numerical value of the terms decreases without limit.

For, by writing it in the forms $u_1 - (u_2 - u_3) - (u_4 - u_5) - \&c.$ and $u_1 - u_2, +(u_3 - u_4) + \&c.$; we see that it is $> u_1 - u_2$, and $< u_1$, and therefore is convergent.

This argument misses some of the subtleties of the full proof, and never uses the requisite fact that the sequence goes to zero. However, we suspect that a proof along this line is expected, possibly having been memorised by the student.

The Ratio Test is covered in §125 of Hymers' text, and §129 gives a version of the Ratio Test that can easily be applied to the last part of Q3(a) when $|x| \neq 1$. The same section seems to skirt around the situation when $|x| = 1$ without saying exactly what happens in this case.

Q5: Binomial Theorem Hymers does not spend much time on Binomial Theorem, so for Q5, we turn to Atkins' textbook [12], which has a chapter devoted to the topic. This chapter also introduces the \underline{n} notation for factorial and uses the phrase *decreases without limit* in an example in the chapter. In the example, the student is asked to

Find the limit of

$$\left(1 + \frac{1}{x}\right)^x,$$

when x increases without limit.

The solution shows that the terms can be expanded as

$$\left(1 + \frac{1}{x}\right)^x = 1 + \frac{1}{1} + \frac{1 - \frac{1}{x}}{1.2} + \frac{\left(1 - \frac{1}{x}\right)\left(1 - \frac{2}{x}\right)}{1.2.3} + \&c.$$

Putting $x = m$ gives the terms in exactly the form shown in the Leaving Cert Q5, and allows us to show the sequence is increasing in m by considering each term as m increases.

5. Discussion

The above analysis gives us an insight into the mathematics examination of its day. It seems to be more reliant on rote learning than one might have expected. The expected depth of student understanding, given that the questions seem so close to exercises from the textbook, seems limited. Indeed, the textbooks do not cover convergence at a level that would allow a student to fully understand the questions asked. Secondly, if one wishes to make the argument that examinations were more difficult in the past, then this suggests caution in the inspection of exam papers without the context of the curriculum content and textbooks. In fact, we would propose that investigation of historic learning and comparisons against such should be ideally undertaken with the examination papers, the textbooks of the day, the curriculum, and the expected marking schemes for such examinations.

6. Conclusion

The study of past exam papers is beneficial for informing our future education methodologies and curriculum, and can help us to improve and rectify mistakes from the past. Given context, they are an excellent source of information and we encourage anyone interested to delve deeper into our full STEM Irish exam paper archive [8].

References

- [1] NCCA, Project maths: Responding to current debate, <https://ncca.ie/media/2275/project-maths-responding-to-current-debate.pdf>, 2012.
- [2] NCCA, Post-primary overview project maths, https://ncca.ie/media/3153/project-maths-research_en.pdf, 2011.
- [3] C. Mac an Bhaird, B. C. Nolan, A. O'Shea, K. Pfeiffer, A study of creative reasoning opportunities in assessments in undergraduate calculus courses, *Research in Mathematics Education* 19 (2017) 147–162.
- [4] A. Akçay, M. Tunagür, A. Karabulut, Turkish teachers' assessment situations: A study on exam papers, *International Journal of Education and Literacy Studies* 8 (2020) 36–43.
- [5] B. S. Bloom, et al., *Taxonomy of educational objectives. vol. 1: Cognitive domain*, New York: McKay 20 (1956) 1.
- [6] Ü. B. Cebesoy, B. Yeniterzi, Investigation of science and technology exam questions in terms of mathematical knowledge, *Procedia-Social and Behavioral Sciences* 116 (2014) 2711–2716.
- [7] M. A. Kang, A. Saeed, Measurement of essential skills in mathematics: A comparative analysis of ssc (grade-x) and gce (o-level) exam papers, *Journal of Education and Educational Development* 7 (2020) 103–118.
- [8] D. Malone, H. Murray, Archive of mathsy state exam papers, <https://archive.maths.nuim.ie/staff/dmalone/StateExamPapers/>, 2016. Accessed June 2021.
- [9] A. De Morgan, *The differential and integral calculus*, Baldwin and Cradock, 1836.
- [10] K. Knopp, *Infinite sequences and series*, Courier Corporation, 1956, p. 68.
- [11] J. Hymers, *A Treatise on Differential Equations: And on the Calculus of Finite Differences*, Longman, Brown, 1858.
- [12] E. Atkins, *Pure Mathematics, volume 2*, William Collins and Son, 1858.

SplitFed Learning Without Client-Side Synchronization: Analyzing Client-Side Split Network Portion Size to Overall Performance

Praveen Joshi^a, Chandra Thapa^b, Seyit Camtepe^b, Mohammed Hasanuzzaman^a, Ted Scully^a and Haithem Afli^a

^a Munster Technological University, Ireland

^b CSIRO Data61, Australia

Abstract

Federated Learning (FL), Split Learning (SL), and SplitFed Learning (SFL) are three recent developments in distributed machine learning that are gaining attention due to their ability to preserve the privacy of raw data. Thus, they are widely applicable in various domains where data is sensitive, such as large-scale medical image classification, internet-of-medical-things, and cross-organization phishing email detection. SFL is developed on the confluence point of FL and SL. It brings the best of FL and SL by providing parallel client-side machine learning model updates from the FL paradigm and a higher level of model privacy (while training) by splitting the model between the clients and server coming from SL. However, SFL has communication and computation overhead at the client-side due to the requirement of client-side model synchronization. For the resource-constrained client-side, removal of such requirements is required to gain efficiency in the learning. In this regard, this paper studies SFL without client-side model synchronization. The resulting architecture is known as *Multi-head Split Learning*. Our empirical studies considering the ResNet18 model on MNIST data under IID data distribution among distributed clients find that Multi-head Split Learning is feasible. Its performance is comparable to the SFL. Moreover, SFL provides only 1%-2% better accuracy than Multi-head Split Learning on the MNIST test set. To further strengthen our results, we study the Multi-head Split Learning with various client-side model portions and its impact on the overall performance. To this end, our results find a minimal impact on the overall performance of the model.

Keywords

Distributed collaborative machine learning, Split learning, Multi-head split learning, Parameter transmission based distributed machine learning, Privacy preserving machine learning

1. Introduction

In the world of data, the security and privacy of individuals have now become one of the major concerns. To avoid data misuse, several restrictions such as the General Data Protection Regulation (GDPR) [1], Personal Data Protection Act (PDP) [2], and Cybersecurity Law of the People's Republic (CLPR) of China [3] have been introduced. These regulations are strictly practiced making data aggregation from distributed devices and regions almost impossible [4].

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ praveen.joshi@mycit.ie (P. Joshi); chandra.thapa@data61.csiro.au (C. Thapa); seyit.camtepe@data61.csiro.au (S. Camtepe); Mohammed.Hasanuzzaman@mtu.ie (M. Hasanuzzaman); Ted.Scully@mtu.ie (T. Scully); Haithem.afli@mtu.ie (H. Afli)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

To accommodate such restrictions along with the constraints placed by heterogeneous devices, improvised machine learning (ML) approaches were sought. Federated Learning [5] and Split Learning [6] are two such ML approaches that enable safeguarding the raw data and offload computations at the central server by pushing a part of the computation to the end devices.

Federated learning (FL) leverages the distributed resources to train an ML model collaboratively. More precisely, in FL, multiple devices collaboratively offer resources to train the ML model while keeping the raw data to themselves, as in no raw data leaves the place of its origin [4]. The main drawbacks of FL are two folds. Firstly, training a large ML model in resource-constrained end devices is difficult [7]. Secondly, all participating end devices and the server has the full trained model. This does not preserve the model privacy while training like in split learning [8].

To overcome these drawbacks, Split Learning (SL) enables model split and training the split model portions collaboratively at the client-side and the server-side separately [9]. The clients and the server never have access to the model updates (gradients) of each other's model portion once the training starts. This way, SL enables training large models in an environment with low-end devices such as internet-of-things and preserves the model's privacy while training. Also, it keeps the raw data to its origin (the analyst has no access to the raw data at all times). However, at a time, SL considers only one client and the server while training. This forces other clients to be idle and wait for their turn to train with the server [8].

To mitigate the drawback of FL having a lower level of model privacy while training and the inability of SL to train the ML model in parallel, specifically among the clients, the SplitFed learning (SFL) is recently proposed [8, 10]. SFL combines the best of the FL and SL. In this approach, an ML model is split between the client and the server (like in SL). In contrast to SL, multiple identical split of ML model, i.e., the client-side model portion, is shared across the clients. The server-side model portion is provided to the server. In each forward pass, all clients perform the forward propagation in parallel and independently. Then the activation vectors of the end layer (client-side model portion) are passed to the server. The server then processes the forward and backpropagation for its server-side model on the activation vectors. In backpropagation, the server returns the respective gradients of their activation vectors to the clients. Afterward, each client performs the backpropagation on the gradients they received from the server. After each forward and backward pass, all client-side models and server-side models aggregate their weights and form the one global model, specifically in SplitFedV1. The aggregation is done independently at the client-side (by using fed server) and server-side. In another version of the SFL called SplitFedV2, the authors changed the training setting for the server-side model. Instead of aggregating the server-side model at each epoch, the server keeps training one server-side model with the activation vectors from all the clients.

Despite the improvements in SFL, model synchronization is needed at the client-side that is obtained through model aggregation and sharing. This is done to make the global model (joint client-side model and server-side model) consistent at the end of each epoch. However, the model synchronization brings the computation and communication overhead at the client-side. This would be significant if the number of clients grows significantly. In this regard, this paper studies the SFL without client-side model synchronization. The resulting model architecture is called *Multi-head Split Learning* (MHSL). We summarize our contributions under two research questions stated in the following:

1.1. Our contributions

RQ1 Can we allow splitfed learning without client-side model synchronization?

We study the feasibility of MHSL. Our empirical studies on IID distributed MNIST and CIFAR-10 data among five clients find a similar result in MHSL and SFL. Moreover, SFL is slightly (1%-2%) better than MHSL on the MNIST. For CIFAR-10, SFL is better by around 10% than MHSL at the 20 global epoch. However, both SFL and MHSL performance is below 60% (low), thus requires further studies to make any conclusion.

RQ2 Is there any effect on the overall performance if we change the number of layers at the client-side model portions?

Performance of SFL and MHSL under different combinations of layers dispersed at the client-side, and the server-side behaved identically. No significant deviation in model convergence and their performance are observed for any of the client-side and the server-side model's combinations in our experiments.

2. Experiment setup

For the experiment purpose, we choose SplitFedV2 in this paper. This makes our analysis more focused on the split learning side. Moreover, we study if the federated learning part can be removed from the SFL, resulting in Multi-head Split Learning (MHSL). The overall architecture of MHSL is depicted in Figure 1. The model W is split into two portions; client-side model W_c portion and server-side model W_s portion. For the clients, their models are represented by W_c^i , where $i \in \{1, 2, \dots, N\}$ is the client's label. The global model W is formed by concatenating the W_c and W_s , i.e., $[W_c W_s]$ once the training completes.

How the final full model is formed in Multi-head Split Learning? Unlike SFL, MHSL removes the fed server and the synchronization of W_c^i at the end of each epoch. During the whole training, W_c^i are trained independently by their clients with the server. But, at the end of the whole training, the global full model W is constructed from any one W_c^i and concatenating it with W_s . To enable this way of constructing the final trained model, we keep the test data the same over all clients and only keep the training data localized. Thus, if the test results for all clients are similar, then it is reasonable to pick any W_c^i for the final full model.

Our program is written using python 3.7.6 and PyTorch 1.2.0 library. The experiments are conducted in a system having a Tesla P100-PCI-E-16GB GPU machine. We observe the training and testing loss and accuracy at each global epoch (once the server trains with all the activation vectors received from all clients). We consider the client-level performance. All the clients were selected to participate at least once at a global epoch without repetition for the current setup.

2.1. Dataset

For our experiments, two widely used image datasets, namely, MNIST and CIFAR-10, are selected. Moreover, this dataset maintains the closeness of our results with the reported results in the original paper SplitFedV2. MNIST [11] dataset consists of 60,000 images in the training

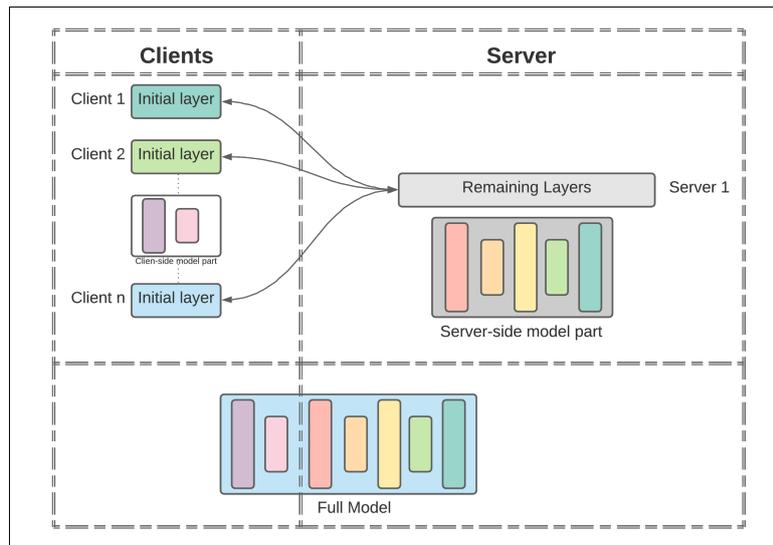


Figure 1: Multi-head split learning architecture.

Table 1

Datasets used in our experiment setup.

Dataset	Training samples	Testing samples	Image size
MNIST	60,000	10,000	28 × 28
CIFAR-10	50,000	10,000	32 × 32

dataset and 10,000 images in the test dataset. The dimension of each of the images in the MNIST dataset is 784 (28×28) in grayscale. Another dataset used for experimentation is CIFAR-10 [12], consisting of 50,000 images in the training set and 10,000 images in the test dataset. Each image corresponds to the dimension of 3072 (32×32). For the summary, refer to Table 1. Both of the datasets have ten classes for prediction. For the experimentation, color random horizontal flipping, random rotation, normalization, and cropping on MNIST and CIFAR-10 are conducted to avoid the problem of over-fitting. In addition, for all our experiments, data is assumed to be uniformly and identically distributed amongst five clients.

2.2. Models

ResNet-18 [13] network architecture is used for the primary experimentation on the MNIST and CIFAR-10 datasets. The ResNet-18 network was selected because of the discrete “blocks” structure in every layer of the architecture [13], and it is a standard model for image processing. Resnet-18 blocks were used to split the Resnet-18 between the clients and server to form the client-side and server-side models. Each block performs an operation; an operation in block refers to passing an image through a convolution, batch normalization, and a ReLU activation excluding the last operation in the block. Resnet-18 in the experiment is initialized with a learning rate of $1e-4$, and the mini-batch size of BN was set to 64 based on the initial experimentation 3.1. In addition, the first convolutional layer kernel size was set to 7×7 , remaining convolutional layers used 3×3 kernels as described in the model architecture Table 2.

Table 2

Model Architecture used in the experimental setup.

Architecture	No. of parameters	Layers	Kernel size
ResNet18 [13]	11.7 million	18	$(7 \times 7), (3 \times 3)$

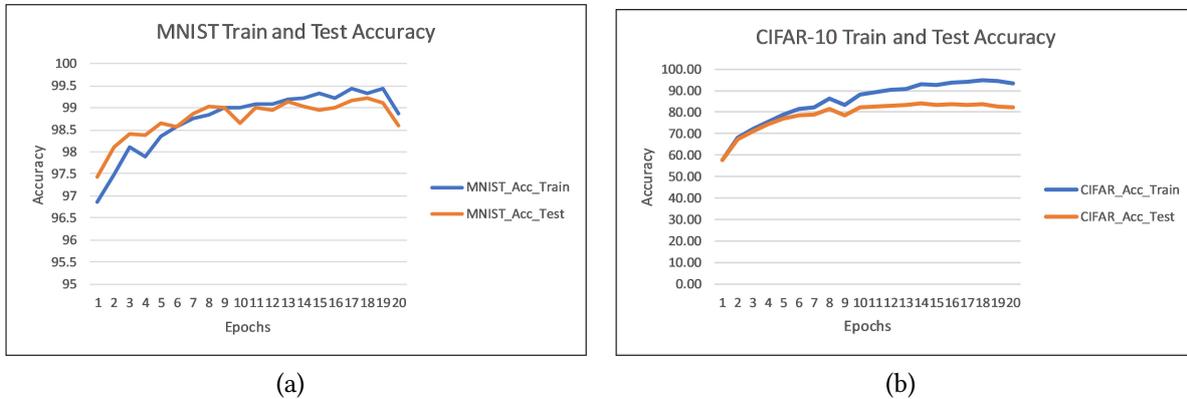


Figure 2: Train and test Accuracy of ResNet18 model on MNIST and CIFAR-10 in the centralized training.

3. Results

This section presents the empirical results on the MNIST and CIFAR-10 datasets. The results are divided into three parts. First, section 3.1 offers results obtained while training the centralized version of the Resnet-18 on the CIFAR-10 and MNIST datasets. In this section 3.2, we compare the results of SplitFedV2 and MHSL on MNIST and CIFAR-10 datasets. For both datasets, we consider five clients to have comparable results, as shown in SplitFedv2 research [8]. In both the architecture, we have kept the initial layer inside the clients (as a client-side model portion), and the rest of the layers reside in the server (as a server-side model portion). Finally, in section 3.3, we have presented our empirical results indicating the impact of the model split on the overall performance of the ResNet-18 model.

3.1. Baseline result

For the baseline, MNIST and CIFAR-10 are subjected to ResNet-18 model architecture. For both the datasets, data-augmentation techniques are the same as discussed in the section 2.1. Training of the ResNet-18 model is done in a centralized manner, i.e., the whole model resided in the server without any split, and all data are available to the server. The convergence curves of both the train and test accuracies for both datasets are shown in Figure 2.

3.2. Experiment1: Corresponding to RQ1

This section evaluated the impact of client-side aggregation by splitting the model on the first layer. The very first layer reside at the client-side (client-side model portion) and the remaining on the server-side (server-side model portion). Experimental results in terms of test accuracy on MNIST and CIFAR-10 dataset with and without client-side aggregation are shown in Figure 3.

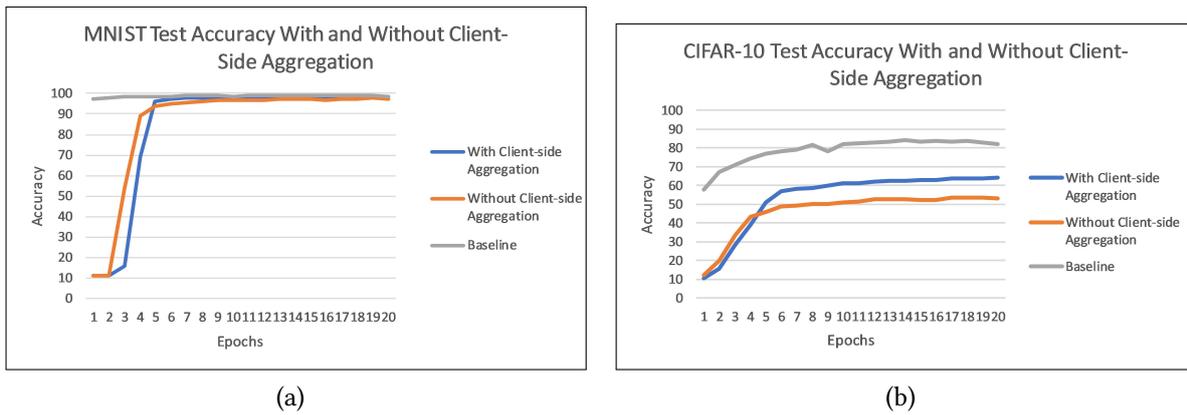


Figure 3: Test accuracy with client side aggregation (i.e., SFL) and without client-side aggregation (i.e., MHSL) on (a) MNIST and (b) CIFAR-10.

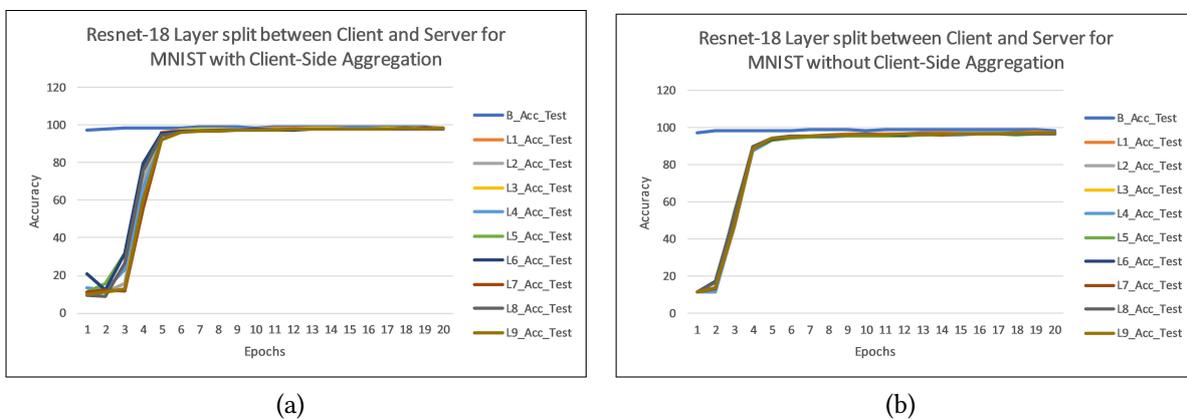


Figure 4: Test accuracy of ResNet-18 on MNIST (a) with client-side aggregation (i.e., SFL) and (b) without client-side aggregation (i.e., MHSL).

Table 3

Test Accuracy of ResNet-18 with the model split on different layers.

Split at layer	L1	L2	L3	L4	L5	L6	L7	L8	L9
Model with Client-Side Aggregation	98.54	98.46	98.56	98.54	98.37	98.21	97.84	98.13	98.25
Model without Client-Side Aggregation	97.23	97.36	96.98	96.71	96.79	96.92	96.93	96.95	97.19

From the results in Figure 3(a), it is evident that results are similar for SFL and MHSL. For CIFAR-10, the performance for both SFL and MHSL are quite lower than the baseline, but the result is better in the case of MNIST.

3.3. Experiment2: Corresponding to RQ2

This section evaluated the impact of the model split on the overall performance. Test accuracy on MNIST is shown in Figure 4.

From Table 3, it is evident that SFL and MHSL show a comparable test performance. Overall, our empirical results (both under RQ1 and RQ2 demonstrate that Multi-head Split Learning

(MHSL) is feasible, and there is no significant impact on the performance due to the model split at the various layers of the ResNet-18 model.

4. Conclusion and future works

This paper studied SplitFed Learning (SFL) without client-side model synchronization called Multi-head Split Learning (MHSL). Our experiments with ResNet-18 on the MNIST dataset demonstrated that MHSL is feasible. In other words, our studies suggested that the fed server and the client-side model synchronization can be removed from SFL to reduce the communication and computation overhead at the client side. In addition, our experiments with different combinations of model portion size at the client-side and the server-side found a negligible effect on the overall performance. This suggests the possibility of dynamic allocation of layers to the clients based on the computation power without any significant loss in the model performance.

This paper is the first step to find the feasibility of MHSL and the effect of the split network portion sizes to the overall performance. In the future, it will be interesting to see more exhaustive experiments and theoretical analysis on the convergence guarantee with the different models, various datasets, and under a larger number of clients in the experimental setup. Also, experimenting with the setup for non-IID data setup will be another research direction that can be explored.

Acknowledgments

This research was conducted with the financial support of the ADVANCE CRT PHD programme within the ADAPT SFI Research Centre at Munster Technological University. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. The project was partially supported by the Horizon 2020 projects STOP Obesity Platform under Grant Agreement No. 823978 and ITFLOWS under Grant Agreement No. 882986.

References

- [1] G. Chassang, The impact of the eu general data protection regulation on scientific research, *ecancermedicalscience* 11 (2017).
- [2] A. Azzi, The challenges faced by the extraterritorial scope of the general data protection regulation, *J. Intell. Prop. Info. Tech. Elec. Com. L.* 9 (2018) 126.
- [3] A. Qi, G. Shao, W. Zheng, Assessing china's cybersecurity law, *Computer Law & Security Review* 34 (2018) 1342–1354.
- [4] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (2019) 1–19.
- [5] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, *arXiv preprint arXiv:1610.05492* (2016).
- [6] M. G. Poirot, P. Vepakomma, K. Chang, J. Kalpathy-Cramer, R. Gupta, R. Raskar, Split learning for collaborative deep learning in healthcare, *arXiv preprint arXiv:1912.12115* (2019).

- [7] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, *IEEE Signal Processing Magazine* 37 (2020) 50–60.
- [8] C. Thapa, M. A. P. Chamikara, S. Camtepe, Splitfed: When federated learning meets split learning, *arXiv preprint arXiv:2004.12088* (2020).
- [9] A. Singh, P. Vepakomma, O. Gupta, R. Raskar, Detailed comparison of communication efficiency of split learning and federated learning, *arXiv preprint arXiv:1909.09145* (2019).
- [10] C. Thapa, M. A. P. Chamikara, S. Camtepe, Advancements of federated learning towards privacy preservation: from federated learning to split learning, *arXiv preprint arXiv:2011.14818* (2020). <https://arxiv.org/pdf/2011.14818.pdf>.
- [11] Y. LeCun, The mnist database of handwritten digits, <http://yann.lecun.com/exdb/mnist/> (1998).
- [12] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).
- [13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Chapter 5

Engineering and Society

Trust and Transparency in Data Protection in Online-Marketing – Differences between different Generations

Louis Kerker^a, Ingo Stengel^a, Stefanie Regier^a

^a *University of Applied Sciences Karlsruhe, Moltkestr. 30, Karlsruhe, Germany*

Abstract

This work aims to understand the importance of data protection for consumers of different generations. Especially under the aspects of trust and the already given options for transparency.

Trust in the services used has not emerged as a particularly important factor when it comes to using services online. The gap between the trust the consumers have in the processing of their personal data and the usage of them is too big to say that consumers really consider this point. It also became clear that consumers demand transparency and control options with regard to the processing of their data. The transparency that is currently provided is generally perceived as not being comprehensible enough and not helpful.

Keywords

Trust, Transparency, Online Marketing, Cookie Banner, Generation Z, Generation Y

1. Introduction

“Data is the oil of the 21st century”. This much-quoted phrase [1], [2] makes it clear: Data is one of the most important resources of the 21st century - Big Data is on everyone's lips. In today's digital times, more and more data are coming together. In 2018, there were 33 zettabytes worldwide, it is estimated that 27% is added every year and by 2025 it is expected to be 175 zettabytes [3]. Few people can really relate to this figure, as they have no real relation to the topic. Why should they care how much the amount of data is increasing or how much data is there in the world? Because there is a non-negligible part of the huge amount of data that needs to be important to everyone. This is growing with a similar speed. This is data that companies, marketing agencies and social networks store about each of us and compile into comprehensive user profiles. Through all this data, people are increasingly becoming transparent, revealing their interests and inclinations. These can be predicted and influenced if only the right data is compiled and used, e.g. personal data is used to influence elections through targeted advertising – see the Cambridge Analytica scandal - personal data is used to target consumers with advertising. In this way, personal data is used to make money.

Especially the younger generations, the "Millenials" and "Gen Z" of our society, of which far more than 90% use social media are online for many hours every day. As "digital natives", they don't know any different. Consequently they disclose vast amounts of information about themselves and do not seem to mind.

Of course, the outrage is great when another data leak or scandal is disclosed, or when you are suddenly shown an ad for a product that you have talked about or thought about, but never actively searched for.

Online advertising is a billion-dollar market that has been growing for years and already accounts for more than 50% of global advertising spending [4]. Personal data plays an immense role in online marketing. It allows consumers to be addressed in a more targeted manner. Scattering losses are to be

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ kerkerlouis@googlemail.com (L. Kerker); ingo.stengel@h-ka.de (I. Stengel); stefanie.regier@h-ka.de (S. Regier)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

avoided and advertising is to be played out more efficiently. Data, its collection and use are therefore extremely important for the advertising industry.

With this development data privacy and the protection of personal data becomes as relevant as maybe never before. But how do the consumers think about those topics? Is data privacy and the protection of their own data on their mind? Are they taking steps to protect themselves?

2. Basics

Digitalization and the increased shift of life into the online/digital space are progressing rapidly. Forms of communication and business models are changing or even being replaced, and new digital solutions are becoming established. These general digital developments do not stop at marketing. Countless new, data-driven processes have been developed in recent years [5]. Online marketing is a market worth billions and is becoming more relevant every year. According to the Zenith Forecast, spending on online advertising today already accounts for a share of more than 50% of total global advertising expenditure [4].

In its current form, online marketing is increasingly driven by the analysis of data and is thus becoming data-driven marketing [1]. Collecting data and analyzing it to make decisions is indispensable, especially in the digital field of advertising. This is because a change has taken place in online marketing in the recent past. Whereas the advertising environment used to be the focus, today the focus is on the users to be targeted. Advertising is no longer displayed on specific websites, but only to specific users who meet defined criteria [6]. This requires comprehensive knowledge about the target groups, which is obtained from user and interest profiles in which their activities and interests are collected. The information for user and interest profiles is collected primarily by tracking the activities of users, often across multiple websites. Tracking is the "collection, storage and analysis of user data and user behavior on websites"[7]. It is recorded which websites a user visits, how long he stays there and what he interacts with. From this, the interests, preferences, habits, and moods of the users are derived. Based on this, particularly targeted, personalized advertising messages are sent [8], [9]. Tracking takes place primarily through the embedding of cookies on websites or through so-called fingerprints[7].

In most of the current marketing activities, user data is of essential importance. Not only it is the basis of digital marketing, but its efficient collection and use is an essential factor for corporate success[1], [10]. Without personal data, modern business models in particular would not function successfully[11]. This is also the conclusion of the "Future Ready" study conducted by the digital agency Wundermann. In this study, 99% of the 250 respondents say that data will determine the success of marketing and sales [12]. But companies need to be aware of the risks that data processing entails. The tracking of users on their own website, but also across websites, represents a risk from a data protection perspective that cannot be neglected and for which companies should take comprehensive legal precautions[5]. Since as explained above, tracking as well as user and interest profiles are the basis for almost all significant online marketing methods and tools, this concerns the entire online marketing and its control.

General Data Protection Regulation (GDPR) was introduced to provide a regulatory framework for all these developments. GDPR is applicable throughout Europe and is partly substantiated, supplemented or modified by national laws of the member states[7], [11]. The intention of the GDPR is to replace the national data protection laws applicable and to create a uniform data protection law in the European Union[7], [11], [13]. It was created to "equip [Europe] for the digital age." According to Article 1 (1) of GDPR, the protection of individuals with regard to the processing of personal data is the object and purpose of the regulation. Through stricter regulations regarding the collection, processing, and storage of personal data, data subjects are to be strengthened in their rights vis-à-vis companies, no matter where they are located, and gain regulation over the use of their data [7], [13], [14]. In case of violations of the introduced rights and laws, the GDPR threatens with high fines. Depending on the severity of the violation, penalties of up to EUR 20 million or 4% of the total annual worldwide turnover, whichever is higher, are threatened under Article 83 (1) of the GDPR.

The processing of personal data is permitted by three permissible circumstances, which are standardized in Art. 6 (1) of GDPR. First, personal data may be processed if the consent of the data

subject has been obtained. This represents the most significant possibility for advertisers to carry out data processing in accordance with the law [7]. According to GDPR recital 32, consent must be given voluntarily, which implies that there are real choices and that consent could also be refused. It must be actively and unambiguously declared, which means that the user must also implement his or her intention to give consent by an active act and tick a box. Boxes that have already been checked or consent through inactivity do not constitute an effective declaration of consent [7], [15]. This regulation is also known as the opt-in / opt-out regulation. Consent-less cookie banners are thus no longer permitted [5]. In addition, the consent must be revocable at any time in accordance with Art. 7 (3) DSGVO. The user must be informed of the possibility of revocation before consent is given. In addition, the entire consent request must be formulated in clear and simple language [7], [15].

The next important aspect of permission for online marketing is the balancing of interests according to legitimate interest, Art. 6 para. 1 p.1 lit. f) DSGVO. According to the Conference of the Independent Data Protection Authorities of the Federal Republic of Germany and its Countries (Data Protection Conference), if consent is not given, the permissibility of advertising must be based almost exclusively on this [16]. Accordingly, data processing is lawful if the "legitimate interests of the controller or a third party [...] override the interests or fundamental rights and freedoms of the data subject which require the protection of personal data" Art. 6 para. 1 p.1 lit. f) GDPR. According to GDPR recital 47ff, direct marketing is one of the legitimate legal, economic or ideal interests of the controller. In addition, the expectation of the data subject represents an important criterion for the weighing of interests [15]. The expectation is based on the predictability of the data processing and can also be influenced by information about the data processing. Furthermore, it can be argued nowadays that the use of, for example, cookies for the display of personalized advertising can no longer be described as unexpected [7], [15], [17].

3. Related Research

The importance of privacy, protection of personal data and the effects of different types of data procession on consumers, their behavior, but also on enterprises is the subject of numerous studies, researches and scientific articles. This topic has attracted the interest of various research communities reaching from economical to psychological areas. Research has deepened since online marketing has become the dominant form of advertising, since the importance of social networks has risen immensely and since the GDPR has become effective. The reoccurring key words of those studies are trust, vulnerability and transparency.

3.1. Trust and transparency

Trust and transparency are vital factors on the success of online advertising campaigns. Bleier and Eisenbeiss [18] were using field data to demonstrate, how much the efficiency of retargeting campaigns depends on the trust a consumer has in the retailer. They compared ad impressions and click through rates (CTR) of two German retailers. The main difference of the retailers was trust. The results show, that trust has a significant impact and more trustworthy retailers can benefit from deeper personalization. The laboratory study of Bleier and Eisenbeiss [18] emphasizes the importance of trust even more. Merchants that are more trustworthy achieve significantly higher CTR with higher personalized advertising, while the CTR of less trustworthy merchants drops sharply. Customers see more personalized banner ads as not useful, experience increased reactance and privacy concerns. Not to mention the negative effects on CTR.

In addition, Beier and Eisenbeiss [18] recommend that trusted merchants should take further steps to prevent privacy concerns from arising among customers, as these lead to reactance.

Aguirre et al. [19] confirm in their study also, that „a firm’s strategy for collecting information from social media websites is a crucial determinant of how customers react to online personalized advertising“. They show in their first step of the experiment how different levels of personalization

have an impact on the success of the advertisement. CTR will rise above previous levels, as consumers become accustomed to data collection and seem to appreciate transparency.

The authors then conducted a study that examined the interaction of higher personalization in overt and covert data collection and the impact on CTR as well as perceived vulnerability of participants. It showed that overt data collection with higher personalization leads to a significant increase in CTR as well as a decrease in perceived vulnerability.

In a direct comparison of more or less personalized ads with overt/covert data collection on a more trustworthy website (CNN) and a less trustworthy website (Facebook) the results showed: "When firms covertly collect data and present highly personalized advertisements, customers feel vulnerable [...] and express lower click-through intentions [...] if the advertisements appear on an untrustworthy website but not if they show up on a trustworthy website" [19].

Deloitte also attributes a decisive role to trustworthiness in its study "Dataland Germany – the Transparency Gap" [20]. The authors consider it particularly important that consumers are asked for consent and that there is transparency about the use of their data. Consumers expect companies to communicate clearly, what data is used for what purposes, as well as have clear guidelines on how to handle customer data. According to Deloitte, consumers' willingness to share data does not depend crucially on the benefits they see in it, "but rather on the transparency that companies show in dealing with data." [20].

Not only Deloitte [20] recognizes transparency and control as crucial factors, but also Martin et al. [21], who even get more specific and say that transparency without control options makes customers feel "more violation and lower trust."

Trust and transparency appear to be key factors for enterprises in regard to their marketing success. Whether the consumers value transparency and trust equally when they choose their service or provider or browse the Internet is matter of the later exploration.

4. Hypotheses

With the implementation of GDPR the users received various rights and possibilities to obtain more transparency about handling of their personal data and to be in control of "the collection, storage and processing of their own data" [7]. This includes the right to information, given by Art. 15 GDPR which gives the affected user the right to know whether and to what extent personal data about them are being processed. Article 17 of the GDPR, the right to erasure, obliges the responsible party to delete personal data immediately if the data subject revokes his or her consent to processing or requests deletion. Those transparency enhancing options include also that an explicit consent, defined in Art. 6 GDPR, is necessary for data processing. This explicit consent is commonly obtained through cookie banners. In this context, cookie banners without explicit consent and the invocation of legitimate interests are no longer sufficient [5]. Possekkel and Schiemann [5] state 2020 that "80 to 85 percent give their consent for cookies that are really necessary or bring convenience, but only about 20 to 40 percent of users give their consent for marketing purposes and advertising targeting and retargeting on the Internet, as well as for analytical purposes." However, Deloitte [20] recognizes among Generation Y that data protection notices are rarely read and tend to be perceived as a nuisance. This is also consistent with our own observations. Hypothesis H1 is therefore:

H1: Transparency enhancing options are appreciated but not particularly used.

Aguirre et al. [19], Bleier and Eisenbeiss [18] as well as Deloitte [20] identify trust as an important factor for the success of personalized advertisement. Facebook is used in their studies as an example of an untrustworthy website. Nevertheless, Facebook and its belonging social media and messengers (e.g. Instagram, WhatsApp) take the first two places of the top three used social medias – in each generation [22]. On top of that, scandals like the Cambridge Analytica case or data breaches are becoming much more frequent in the last years, but in her article for the Guardian Wong [23] points out that, despite of reoccurring data breaches and scandals a mass exodus from social media never happened. It is therefore

reasonable to assume that data privacy reasons and the (lack of) trust in the service / provider are not necessarily sufficient to switch. Hypothesis H2 reflects this assumption.

H2: Data privacy reasons and trust in the handling of personal data are no sufficient reasons to change the service / provider in use.

5. The Project

To answer the research question and hypotheses posed here, a quantitative research approach was chosen. This work aimed to understand the importance of data protection for consumers of the younger generations of our society as well as the feelings and preferences towards this topic. Therefore, all the results have been taken for the three different generations: Generation Z, Generation Y as well as the Generation “older” to summarize the generations born before 1980. The respondents have been addressed using a digital questionnaire that has been distributed through direct contacts as well as dedicated social media groups. This method has been chosen since the digital approach is the best way to reach respondents in younger generations. The scientific questionnaire was chosen as a survey technique because it allows capturing subjective experience in a past and private context. It can be conducted efficiently through the self-administration of the participants [24]. The respondents receive a questionnaire consisting of closed questions, which are to be answered by ticking or indicating numerical values. This type of questionnaire was chosen because the respondents can anonymously state their opinions and feelings in this format and the questionnaire is self-explanatory to fill out due to its format.

During the survey period, 243 persons participated in the data collection. Of these 243 questionnaires, however, only 198 met the predefined selection criteria for counting as valid cases. These were, among other things, the completion of the questionnaire up to the penultimate page, since no relevant questions for the study were asked on the last page. Furthermore, the data collected had to be cleaned of low-quality data. Study results suggest that lagging data can be most reliably identified by the completion time of the questionnaire [25]. Therefore, the recommendations of Leiner [25] were followed to filter the completion time as an indicator of meaningless data. This is done by using the "relative completion speed" or "relative speed index" (RSI), which allows a comparison between the different questionnaires. If the RSI exceeds 2.0, the data should be viewed critically. Furthermore, the proportion of missing answers was considered. After applying these additional quality criteria, 193 questionnaires could be identified as valid and qualitatively sufficient, resulting in a sample of $n=193$.

It is composed of 78 male participants and 111 female participants. 3 participants did not want to specify their gender. Broken down according to age groups, the sample consists of 70 participants from Generation Z, 75 from Generation Y and 47 participants from “older” generations. The division into generations was made according to the year of birth. Participants born before 1980 were assigned to the "Older" generation. Participants born between 1980 and 1996 were assigned to Generation Y and those born in 1997 and later were assigned to Generation Z.

To compare the different types of transparency enhancing options the respondents were presented with several different cookie information / cookie banners. These are not part of this paper. They were requested to describe their interaction with cookie banners as well as with other privacy options by describing how often they use this option (never, rarely, sometimes, mostly or always).

In order to answer the question whether trust is a sufficient reason to change the provider/service in use, the participants were asked to express how big their trust or mistrust regarding the handling of their personal data is. The websites to rate are a subset of the most used social networks with some additional messengers and news websites. In a second step the participants were asked to answer how often they use the website, social network or messenger. Finally, the respondents have been asked to choose which reasons are valid reasons for them to change the service in use.

6. Results

To answer the first hypothesis, cookie banners have been chosen to represent transparency enhancing options, as the vast majority might only know those as options to change their privacy settings while

browsing. The respondents were asked to give opinions about cookie banners in general. 16% of Generation Z and 11% each of Generation Y and Generation "Older" said they read through the cookie information before making a choice. However, only 8% of Generation Y, 3% of Generation Z and 0% of the "older" generation find cookie information understandable. Just as few state that the cookie information makes the handling of data really transparent (Generation Z 6%; Generation Y 5%; Generation "Older" 4%). In contrast, 80% of Generation Z, 70% of Generation Y and 68% of the "older" generation demand that the cookie information should be more comprehensible and easier to change. Cookie information is not considered useful or helpful by any generation. Only 11% of Generation Y would agree with this. Generation Y thus represents the maximum (Generation Z 7%, Generation "Older" 9%). This is also reflected in the fact that around 2/3 of each generation would describe cookie information as annoying. The behavior of the different generations with regard to cookie information was also recorded. This is shown in the diagrams in Figures 1 a-c.

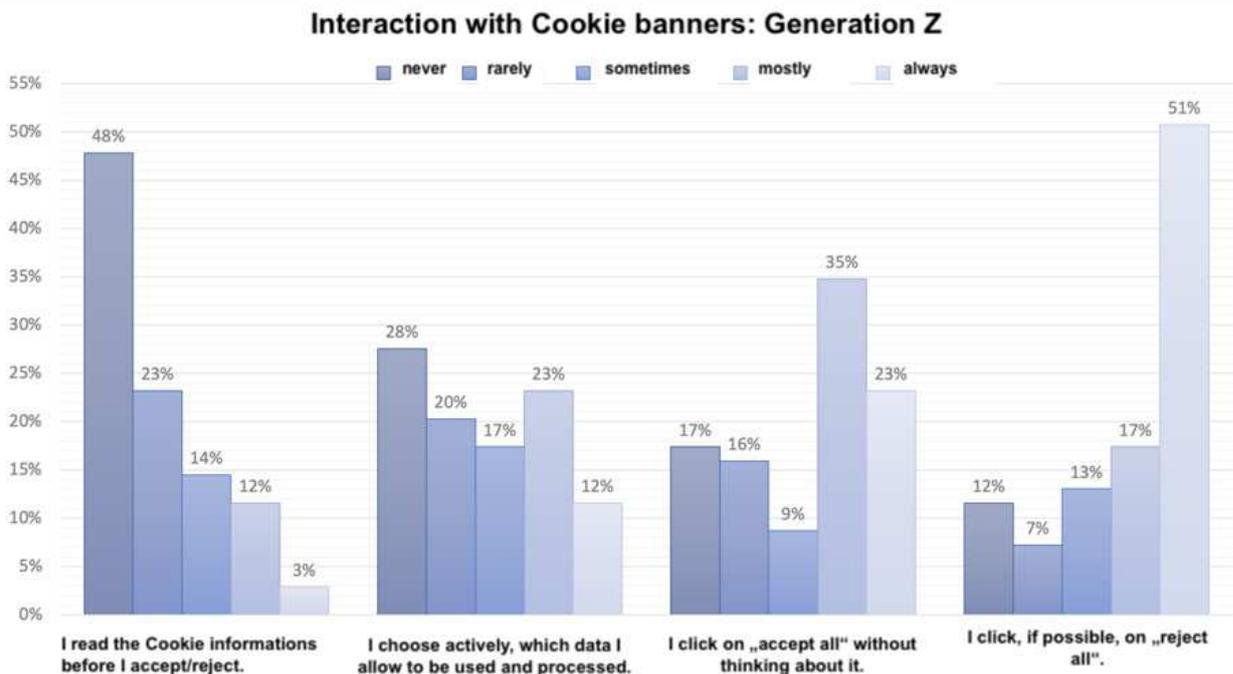


Figure 1a: Interaction with Cookie Banners: Generation Z

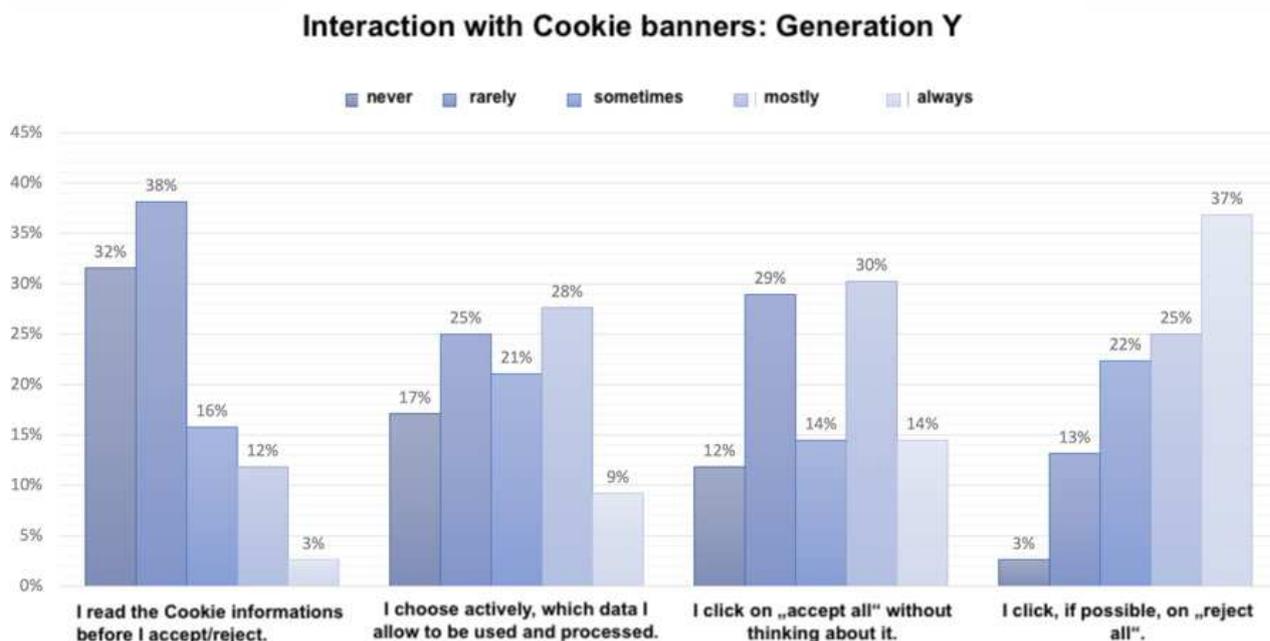


Figure 1b: Interaction with Cookie Banners: Generation Y

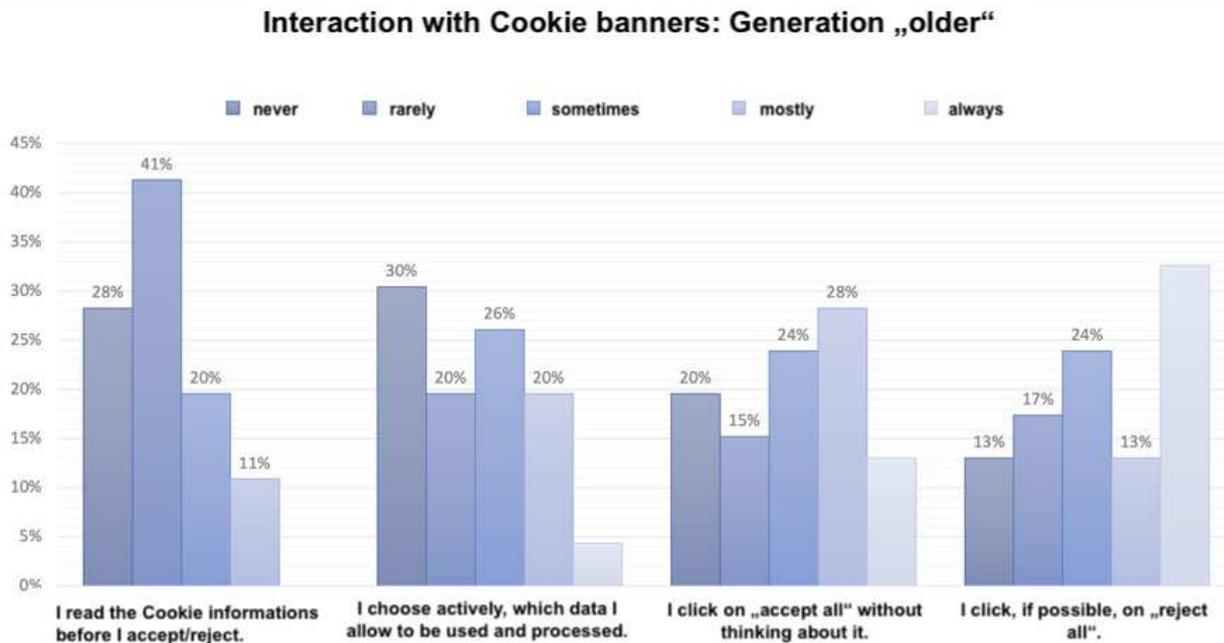


Figure 1c: Interaction with Cookie Banners: Generation “older”

It is striking that in all generations the cookie information is mostly "Never" or "Rarely" read. 69% of the "older", 70% of Generation Y and 71% of Generation Z stated this. The majority of Generations Z and Y click on "reject all" most of the time or always, if possible (Generation Z 68%; Generation Y 62%). Only 46% of the "older" generation indicated this. Approximately 1/3 of respondents from Generation Z and Y each indicated that they usually or always actively select which data may be collected. Again, fewer people from the "older" generation act in this way. Only one in four respondents stated that they usually or always actively select which data may be collected. When it comes to accepting all cookies, 58% of Generation Z usually or always click on this without thinking about it. Only 33% of Generation Z never or rarely do this. Among Generation Y, the rate of doing this most of the time or always is 44%, but 41% said they never or rarely do so. The generation "Older" achieves similar values, 35% say they never or rarely accept all without thinking about it, 41% say they do this most of the time or always. Across all generations, the number of people who read through the cookie information most of the time or always is only 11-15% (Generation Z 15%; Generation Y 15%; "Older" 11%). In contrast, around 70% of all generations stated that they never or rarely read the cookie information.

Other options which promote transparency from the consumer's point of view are also used by only a small proportion of respondents. Only 22% of Generation Z and 25% of Generation Y have already used the option to view their assigned advertising profile. Among the "older" generation, only 9% have done so. 32% of Generation Z and 33% of Generation Y said they had not done this before, but 46% / 42% had not done it because they did not know it was possible. Among the "older" 45% have not done it and 47% did not know it was possible. Requesting and downloading the collected data is also taken up by few of the respondents. Here it is only 14% of Generation Z, still 21% of Generation Y but only 6% of the "Older" generation. Here, the "not taking advantage of the possibility" among Generation Z is divided into 30% who did not and 55% who did not know it was possible. 43% of Generation Y answered "no" and 36% "no, I didn't know it was possible." Among "older" people, more than half did not know it was possible and 40% did not. The right to request deletion of collected data is also used by only a small proportion of respondents (14% Generation Z; 21% Generation Y and 4% "Older"). While only 1/3 of Generation Y and Generation "Older" respondents were unaware of the option and 43% (Generation Y) and 60% ("Older") did not use it, 46% of Generation Z did not know that it was possible and 39% did not do so. The option to deactivate personalized advertising is used by 1/3 of Generation Z respondents, not used by 1/3 and not known by 1/3. Among Generation Y, 36% deactivate personalized advertising, 28% do not do so and 7% are not aware of the option. Among the "older" respondents, 36% do not know or do not use this option and 28% deactivate personalized advertising.

To answer the second hypothesis, the first step was to learn which services and social networks are the most used by the participants.

The services and social networks most used by Generation Z are those belonging to the Facebook Group or Alphabet/Google. WhatsApp and Google (search engine, maps, etc.) share first place. They are used by 93% of the respondents belonging to Generation Z more often than four days a week. Instagram comes in second place with 80% of respondents and then YouTube with 79%. The last place of the top 5 is occupied by Snapchat with 54%.

Comparing this to the trust that users have in the service's handling of personal data shows that the most used services have the lowest percentage scores for trust. In this negative top ranking, the services belonging to Facebook again lead the way. Only 6% of the users surveyed have slight or full trust in Facebook's handling of personal data. For Facebook Messenger, the value is only 5% and Instagram also does only marginally better with 8%. The 16% of Generation Z who trust WhatsApp when it comes to handling data represent a significant improvement compared to the previous figures. However, they are also contrasted by 57% who distrust or even strongly distrust WhatsApp.

Figures 2 a-c depict the use of selected services and the trust that users have in the data processing of the respective service.

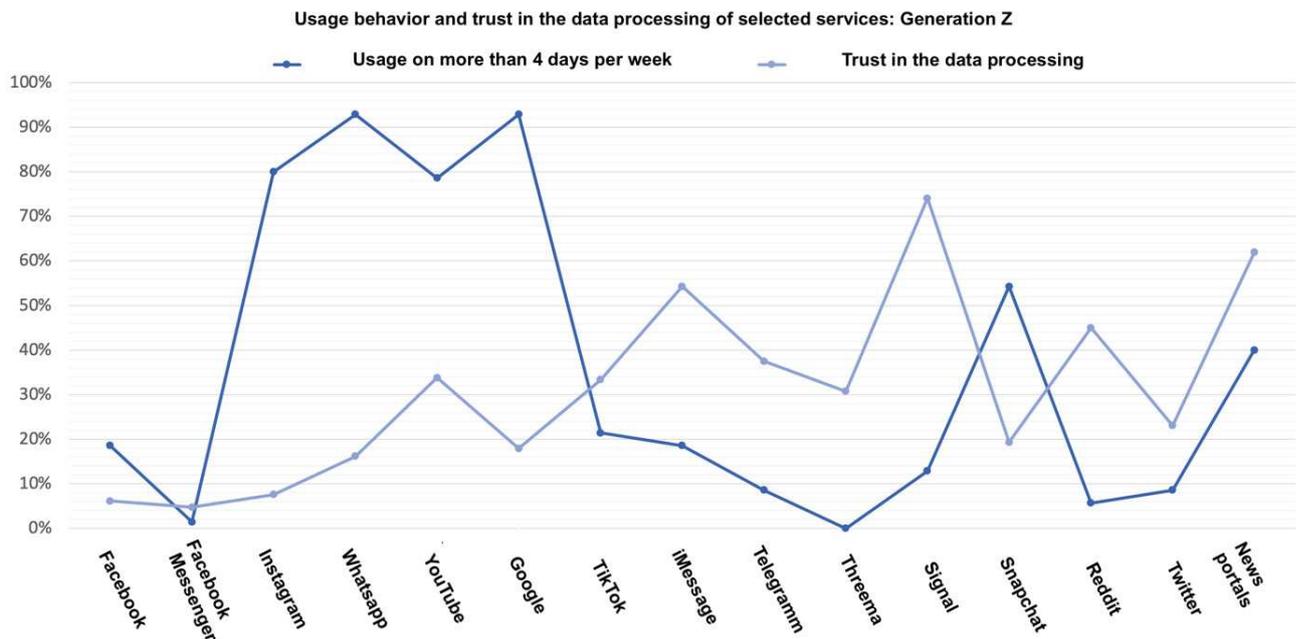


Figure 2a: Usage behavior and trust in the data processing of selected services: Generation Z

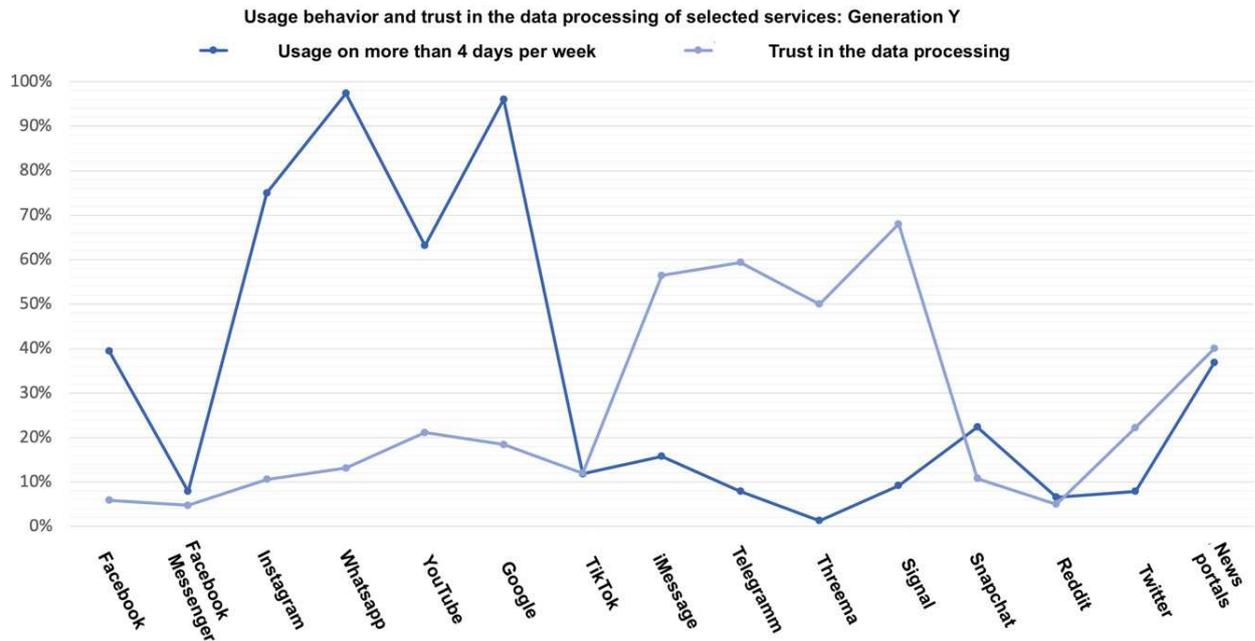


Figure 2b: Usage behavior and trust in the data processing of selected services: Generation Y

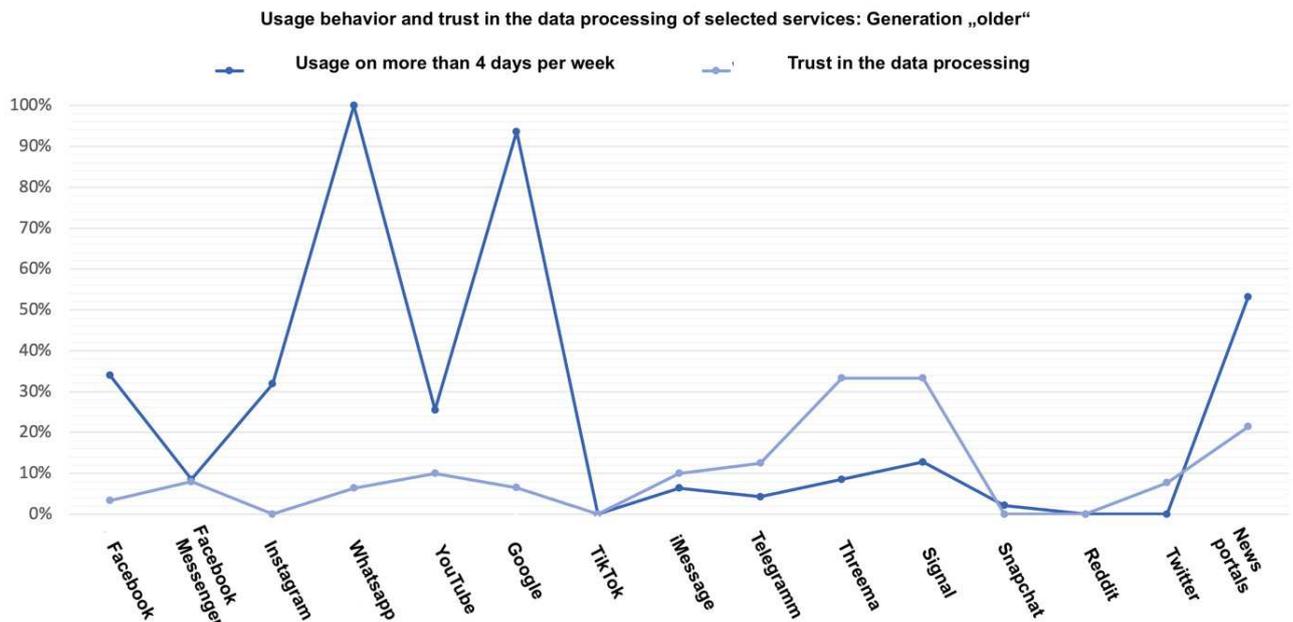


Figure 2c: Usage behavior and trust in the data processing of selected services: Generation “older”

The largest percentage of users who trust the data processing of the service is given to Signal with 74%, news portals with 62% and iMessage with 54%. While news portals are at least used by 40% of respondents more often than 4 days a week, iMessage is already used significantly less (19%) and Signal by only 13% of respondents more often than 4 times a week. Although Signal achieves the highest level of trust among respondents, it is not used at all by 80% of respondents.

Threema, which of the services available for selection can be named as one of the services with the comparatively high data protection [26], is used the least by Generation Z. Of the 70 respondents, 64 are not logged in, 4 are logged in but do not use it, and only one person uses it 1-3 days per week.

The results for Generation Y are similar. WhatsApp and Google (search engine, maps, etc.) also occupy the first two places among the respondents of this generation and are used more than 4 times a

week by 97% and 96%, respectively. This is followed by Instagram (75%), YouTube (63%) and Facebook (39%). Generation Y distrusts Facebook and Facebook Messenger the most, analogous to Generation Z (76% / 78%). Unlike Generation Z, however, TikTok (64%), Instagram (58%) and Snapchat (57%) follow.

Generation Y, like Generation Z, has the least trust in the services belonging to Alphabet/Google and Facebook when it comes to handling personal data. YouTube still receives the highest value here, with 21% of users stating slight to full trust (FB 6%, FB Messenger 5%, Instagram 11%, WhatsApp 13%, Google 18%).

The messengers Signal, Threema, Telegram and iMessage receive the most trust from users in handling data. Signal receives 68%, iMessage, Telegram and Threema between 50% and 59%. It is surprising that, as with Generation Z, the services with the highest trust are not among the most used services.

Among the "older" generation, news portals are used most frequently after WhatsApp and Google. Only then, Facebook, Instagram and YouTube follow. It is unusual that among the "older" generation, only Signal and Threema are trusted by more than 30% of users. Even with these two messengers, distrust about the handling of personal data is over 10 percentage points higher than trust. Twitter, Reddit and Snapchat in particular score exceptionally poorly in terms of trust scores. However, these services are also the least used by the respondents of this generation. The much-used news portals achieve a medium level of trust. The relatively heavily used services Facebook, Instagram and WhatsApp score high values for distrust, analogous to the two younger generations. A maximum of 10% of users said they had confidence in the data processing of these services. On average, 31% of Generation Z, 27% of Generation Y and 10% of the "older" generations say they have at least slight trust in the services' handling of personal data.

When asked directly whether users use services even though they have concerns about data privacy, a consistent picture emerges across all generations. 45% of respondents from Generation Z fully agreed to use services even though they had concerns about data privacy. The figure for Generation Y was 50% and for the older generations 51%. If the proportion of people who tend to agree with this statement is also added, the result is 87% for Generation Z, 87% for Generation Y and 89% for the older generation. After a Chi Square Test for independence, it can be seen that there is no stochastically significant correlation between "generation" and "use of services despite data protection concerns".

7. Discussion

Hypothesis H1 assumes that opportunities to promote transparency are valued, but not actually used. To prove this hypothesis, the right to information under Article 15 and the right to delete one's own data were used as examples of such transparency-promoting measures. The use of cookie banners was also queried, as this is probably the most universally known means by which consumers can track their data processing or restrict data processing. Aguirre et al. [19], Bleier and Eisenbeiss [18], and Deloitte [20] identified both transparency about data processing and trust as key factors. As Table 1 shows, there are differences between the generations. Generations Z and Y generally rated the cookie banners available for selection slightly better than the "older" generation. It is noticeable that the cookie banners with a high level of information or transparency about the use of data did not necessarily receive the best rating. The best ratings in each case in all generations were given to the cookie banners that have only a medium level of information content, but offer several different options for selecting which data may be processed directly, offer better selection options than "accept all" or have additional settings set up that are not, however, directly mapped. It is also noticeable that a cookie banner with little information receives a relatively good rating if it has "Accept all" and "Reject non-essentials" as default selections. The cookie banners with the most detailed information and thus the highest transparency about data processing did not achieve exceptionally good ratings. It seems as if the options to limit the data processing are more important than a detailed explanation.

The reason for this can be inferred from other survey results. With a share of only 16%, Generation Z was the generation whose most users read through the cookie information before selecting it at all. Conversely, this means that the vast majority do not read the information. Even if the information were

actually read, very few would find the information clear and understandable or would find that it makes transparent how the data is handled. Most respondents found it annoying and unhelpful. These findings were also similar to those drawn from the survey by Deloitte [20] and can therefore be seen as confirmed.

Most respondents of the younger generations also click on "Reject all" most of the time to all of the time, if possible. Across all generations, the selection options are not considered sufficient. More than 2/3 of the "older" generation, 70% of Generation Y and even 8 out of 10 Generation Z respondents also demand that the cookie information should be easier to understand and change. Since the majority of all generations also feel that the cookie information does not make transparent what happens to the data, the following conclusion can be drawn:

Transparency is valued. In its current form, however, cookie information does not provide the transparency that consumers want. Even detailed information does not provide the transparency that would be necessary, because consumers find cookie banners annoying, rarely read them at all, and when they do read them, do not understand them sufficiently. Therefore, the banners that are rated better are those that provide just enough to understand what is most important. It can also be concluded from the present results that control options are more important to the respondents than transparency.

The other transparency-promoting options, such as the rights stemming, from Articles 15 and 17 of GDPR to obtain information about the data collected or to have it deleted, are used by few users. They tend not to be used because they are not viewed positively, but because the majority of users of all generations do not know that these options exist.

Hypothesis H1 and the research findings of Deloitte [20], Acquisti et al. [19] and Bleier and Eisenbeiss [18] as well as Martin et al. [21], who consider transparency with additional control options to be extremely important, can therefore be confirmed.

Hypothesis H2 states: "Data protection reasons and trust in the handling of personal data alone are not sufficient to switch services/providers. As explained in section Related Research Aguirre et al. [19], Bleier and Eisenbeiss [18] and Deloitte [20] in particular identified trust as a key factor. Aguirre et al. [19] and Bleier and Eisenbeiss [18] relate this more to the context of the effectiveness of advertisements and Deloitte [20] to the willingness to share data. The present research extends this approach by asking whether trust is also such an important factor that can cause users to turn their backs on services and providers. The results presented here paint a clear picture across all generations: for all generations, social networks of the Facebook Group are among the most used services. The fact that Facebook has a questionable reputation when it comes to the use of data is shown not only by the research of Cabañas et al. [27], but also by Aguirre et al. [19], which lists Facebook as an untrustworthy website. The users surveyed feel the same way. Only a vanishingly small proportion of all three generations express their trust in these social networks or services. However, the astonishing findings of the survey are only revealed in the next step, when the trust that users place in these services is compared with the extent to which these services are used. It becomes apparent that none of the most frequently used services receives a high rating in this regard. Rather, it becomes clear that the social networks or messengers that are given a high trust rating have very low usage rates across all generations. In addition, it could be shown that the overwhelming majority of respondents, stochastically independent of generation, state that they use services despite data privacy concerns. This is despite the fact that an even larger proportion of respondents in each generation stated that they saw data protection reasons as a reason to switch providers. If the environment was seen as a primary reason for switching services for all generations, this could explain why the highest usage rate does not correlate with the level of trust that exists in providers. Since the majority of the environment uses the service and does not switch, the individual consumers also use this service despite the lack of trust.

Although the proportion of respondents who see data privacy as a reason for switching is high in all generations, it continues to rise with increasing age. A stochastically significant correlation between age and data protection as a reason for switching was demonstrated. The present results can neither confirm nor disprove the research findings of Aguirre et al. [19] or Bleier and Eisenbeiss [18] regarding the influence of trust on the effectiveness of advertising measures. However, they can extend them to the effect that privacy reasons and trust can be seen as important influencing factors on consumer behavior, but do not influence the actual behavior of consumers regarding their choice of service or provider. Thus, Wong [23] can be confirmed to the effect that, despite recurring data leaks, there has not yet been a mass exodus from social networks, and this despite the fact that in the present survey

about 2/3 of each generation stated that data leaks were the reason for switching. It is interesting that the environment is the more decisive reason for a change. This was indicated by most of the respondents across all generations. It was found that the respondents' environment has a major influence on behavior. This is consistent, especially with respect to Generation Z, with the findings of current research [28]. It showed the significant influence of friends and family or the environment on actions and use of services.

Hypothesis H2, that data protection reasons and trust in the handling of personal data are not sufficient reasons to switch providers, can thus be confirmed.

8. Conclusions and Outlook

This worked aimed to understand the importance of data protection for consumers of different generations. Especially under the aspects of trust and the already given options for transparency.

The results have shown that trust in the services used has not emerged as a particularly important factor when it comes to using services online. The gap between the trust the consumers have in the processing of their personal data and the usage of them is too big to say that consumers really consider this point. In general, data protection reasons are one reason for switching providers. This applies to all consumers, but a correlation between age and this reason could be demonstrated. For the "older" consumers, this was a stronger argument than for the younger ones. Nevertheless, all generations use services despite their privacy concerns. It was shown that it is primarily the consumer's environment that influences which services are used. It also became clear that consumers demand transparency and control options with regard to the processing of their data. The transparency that is currently provided is generally perceived as not being comprehensible enough and not helpful.

9. References

- [1] S. Boßow-Thies, C. Hofmann-Stölting, and H. Jochims, *Data-driven Marketing: Insights from Wissenschaft und Praxis*. 2020.
- [2] C. Höinghaus, 'Daten: Das Öl des 21. Jahrhunderts: Big Data wirtschaftlich sinnvoll einsetzen', Aug. 28, 2015. <https://www.cio.de/a/big-data-wirtschaftlich-sinnvoll-einsetzen,3246278> (accessed Jul. 30, 2021).
- [3] M. Kroker, 'Weltweite Datenmengen sollen bis 2025 auf 175 Zetabytes wachsen – 8 mal so viel wie 2017', Nov. 27, 2018. <https://blog.wiwo.de/look-at-it/2018/11/27/weltweite-datenmengen-sollen-bis-2025-auf-175-zetabyte-wachsen-8-mal-so-viel-wie-2017/> (accessed Jul. 26, 2021).
- [4] H. Rampe, 'Zenith Forecast: 2021 fließt die Hälfte aller Werbespendings in Internet-Werbung', <https://www.horizont.net>, Jul. 09, 2020. <https://www.horizont.net/medien/nachrichten/zenith-forecast-2021-fliesst-die-haelfte-aller-werbespendings-in-internet-werbung-176029> (accessed Jul. 26, 2021).
- [5] M. Possekel and S. Schiemann, 'Data-driven Marketing als Risiko', *Control Manag Rev*, vol. 64, no. 2, pp. 52–57, Feb. 2020, doi: 10.1007/s12176-019-0079-5.
- [6] I. Kamps and D. Schetter, *Performance Marketing Der Wegweiser zu einem mess- und steuerbaren Online-Marketing - Einführung in Instrumente, Methoden und Technik*. 2020. Accessed: Jun. 25, 2021. [Online]. Available: <https://doi.org/10.1007/978-3-658-30912-1>
- [7] M. Lutz, 'Datenschutz im Onlinemarketing', in *Datenrecht in der Digitalisierung*, L. Specht-Riemenschneider, N. Werry, and S. Werry, Eds. Erich Schmidt Verlag, 2020, pp. 203–250.
- [8] N. Fabisch, 'Ethische Grenzen der Datennutzung im Marketing', in *Data-driven Marketing*, S. Boßow-Thies, C. Hofmann-Stölting, and H. Jochims, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 283–312. doi: 10.1007/978-3-658-29995-8_13.
- [9] M. Nebel, 'Privatsphäre und Privatheit', in *Persönlichkeitsschutz in Social Networks: Technische Unterstützung eines grundrechtskonformen Angebots von Social Networks*, M. Nebel, Ed. Wiesbaden: Springer Fachmedien, 2020, pp. 39–51. doi: 10.1007/978-3-658-31786-7_4.
- [10] H. M. Wolters, 'Qualität von Kundendaten – Ansätze zur Analyse und Verbesserung als Basis für effiziente Marketingentscheidungen', in *Data-driven Marketing*, S. Boßow-Thies, C. Hofmann-Stölting, and H. Jochims, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 29–41. doi: 10.1007/978-3-658-29995-8_2.

- [11] C. Westerkamp, ‘Datenschutz gemäß DSGVO im datengetriebenen Marketing – ein Überblick’, in *Data-driven Marketing*, S. Boßow-Thies, C. Hofmann-Stölting, and H. Jochims, Eds. Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 237–256. doi: 10.1007/978-3-658-29995-8_11.
- [12] D. Hein, ‘Wunderman-Studie: Zwei Drittel aller Unternehmen können aus Daten keine Marketing-Maßnahmen ableiten’, <https://www.horizont.net>, Mar. 13, 2018. <https://www.horizont.net/marketing/nachrichten/Wunderman-Studie-Zwei-Drittel-aller-Unternehmen-koennen-aus-Daten-keine-Marketing-Massnahmen-ableiten-165526> (accessed Jun. 25, 2021).
- [13] T. Becker, A. Freiherr von Bussche, J.-M. Grages, and A.-M. Frey, *DSGVO/BDSG: Kommentar zu DSGVO, BDSG und den Datenschutzbestimmungen von TMG und TKG*, 3. Köln: Verlag Dr. Otto Schmidt KG, 2018.
- [14] J. Behrens, ‘DSGVO im Digitalen Marketing – heutige und künftige Herausforderungen für den CMO’, in *Digitales Marketing – Erfolgsmodelle aus der Praxis*, M. Terstiege, Ed. Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 17–42. doi: 10.1007/978-3-658-26195-5_2.
- [15] U. Schläger, J.-C. Thode, C. M. Borchers, C. S. Conrad, and M. Cyl, Eds., *Handbuch Datenschutz und IT-Sicherheit*. Berlin: Erich Schmidt Verlag, 2018.
- [16] DSK, ‘Kurzpapier Nr. 3’, presented at the Konferenz der unabhängigen Datenschutzbehörden des Bundes und der Länder (Datenschutzkonferenz), 2017.
- [17] R. Schwartmann, A. Jaspers, G. Thüsing, D. Kugelmann, M. Atzert, and A. Buchmann, *DSGVO/BDSG: Datenschutz-Grundverordnung, Bundesdatenschutzgesetz*. Heidelberg: C.F. Müller, 2018.
- [18] A. Bleier and M. Eisenbeiss, ‘The Importance of Trust for Personalized Online Advertising’, *Journal of Retailing*, vol. 91, no. 3, pp. 390–409, Sep. 2015, doi: 10.1016/j.jretai.2015.04.001.
- [19] E. Aguirre, D. Mahr, D. Grewal, K. de Ruyter, and M. Wetzels, ‘Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness’, *Journal of Retailing*, vol. 91, no. 1, pp. 34–49, Mar. 2015, doi: 10.1016/j.jretai.2014.09.005.
- [20] Deloitte, ‘Datenland Deutschland Studie - Die Transparenzlücke’, *Deloitte Deutschland*, 2014. <https://www2.deloitte.com/de/de/pages/trends/studie-datenland-deutschland.html> (accessed Jul. 26, 2021).
- [21] K. D. Martin, A. Borah, and R. W. Palmatier, ‘Data privacy: effects on customer and firm performance’, *Journal of marketing*, vol. 81, no. 1, Jan. 2017.
- [22] Adobe, Ed., ‘Adobe Across the Ages Study - Vertrauen in Marken und Medien: Worauf es den Konsumenten unterschiedlicher Generationen wirklich ankommt.’ Aug. 2019.
- [23] J. C. Wong, ‘The Cambridge Analytica scandal changed the world – but it didn’t change Facebook’, *the Guardian*, Mar. 18, 2019. <http://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook> (accessed Jun. 25, 2021).
- [24] N. Döring and J. Bortz, *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, 5. vollständig überarbeitete, Aktualisierte und erweiterte Auflage. Berlin Heidelberg: Springer, 2016.
- [25] D. J. Leiner, ‘Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys’, *Survey Research Methods*, pp. 229–248 Pages, Dec. 2019, doi: 10.18148/SRM/2019.V13I3.7403.
- [26] I. Bauer, ‘Wie sicher ist Threema?’, *heise online*, Apr. 21, 2021. <https://www.heise.de/tipps-tricks/Wie-sicher-ist-Threema-5043167.html> (accessed Jul. 12, 2021).
- [27] J. G. Cabañas, Á. Cuevas, A. Arrate, and R. Cuevas, ‘Does Facebook use sensitive data for advertising purposes?’, *Commun. ACM*, vol. 64, no. 1, pp. 62–69, Jan. 2021, doi: 10.1145/3426361.

Three Source Electron-Beam Co-deposited Thermoelectric BiSbTe Thin Films.

Philipp Lorenz^a, Gabriel Zieger^a and Heidemarie Schmidt^{a,b}

^aLeibniz-IPHT, Albert-Einstein-Str. 9, 07745 Jena, Germany

^bInstitut für Festkörperphysik, Physikalisch Astronomische Fakultät, Friedrich-Schiller-Universität Jena, 07743 Jena

Abstract

We present our ongoing research on three source electron-beam co-deposition of thin Bi_xSb_yTe_z-films. We show first promising results of the thin films while taking advantage of the high speed, cleanliness, reproducibility and low operating costs of e-beam co-evaporation.

A variation of the film thickness of the Bi_xSb_yTe_z thin films was performed in the range of $d = 240$ nm to 620 nm as well as two different stoichiometries to compare the resulting layer properties. Additionally, one annealing step was applied to analyse its impact on said properties. For all samples Seebeck coefficient and electrical conductivity have been measured, and thermoelectric power factor has been derived before and after annealing.

Keywords

electron beam deposition, co-evaporation, BiSbTe, bismuth, antimony, tellurium, thin films, thermoelectric materials, Seebeck effect

1. Introduction

Thermoelectric materials are used in a wide field of applications, for example IR-sensing, thermometry, or electrical power generation using thermoelectric generators [1, 2, 3, 4]. A temperature gradient in a thermoelectric material leads to the forming of internal diffusion currents and generates a thermoelectric voltage (Seebeck effect).

The performance of a thermoelectric material with Seebeck coefficient α and specific electrical conductivity σ is described by the power factor:

$$PF = \alpha^2 \cdot \sigma \quad (\text{in W/m/K}^2)$$

where α is given in units V/K and σ is given in units of $1/\Omega\text{m}$. For thermoelectric microdevices it is necessary to develop fast, clean, and reproducible thin film deposition techniques. The deposition of Bi_xSb_yTe_z-thin films has been shown in the past by several methods like flash evaporation [5, 6, 7], sputtering [8, 9, 10], electrodeposition [11, 12] or pulsed laser deposition [13]. However, as a fast, clean and reproducible method with low operating costs [14] electron beam (e-beam) co-evaporation should be taken into account.

Compared to other common methods for the deposition of thin films on silicon wafers at production level, it also reduces the preparation time and cost. The produced thin films are then investigated in terms of σ , α and the resulting PF .

2. Experimental details

All investigated films were deposited on 10 cm diameter borofloat glass discs. The samples were co-deposited out of separate crucibles filled with 5n purity bismuth, 5n purity antimony, and 5n purity tellurium. The distance between the three separately controlled electron beam evaporator systems and the substrate amounts to 50 cm.

The Bi_xSb_yTe_z thin films were deposited with a variation in stoichiometry and thickness in the range of $d=240$ nm to 620 nm. The film deposition was performed in high vacuum at a working pressure of 5×10^{-6} mbar or lower with a rotating substrate holder. For the whole e-beam co-evaporation process a total deposition rate of 7 \AA s^{-1} to 8 \AA s^{-1} was targeted and monitored by quartz crystal oscillators. For the e-beam co-evaporation the evaporation rates from the three separate targets were measured separately. After the deposition process, the glass wafers are coated with a protective layer of photoresist for further processing, and in this step they are subjected to a short annealing step at 80 °C for 20 min. After separation the resist is removed with acetone and afterwards cleaned with isopropyl alcohol and deionized water. Then the layers were characterized to determine their Seebeck coefficient and

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ philipp.lorenz@leibniz-ipht.de (P. Lorenz);

gabriel.zieger@leibniz-ipht.de (G. Zieger);

heidemarie.schmidt@leibniz-ipht.de (H. Schmidt)

ORCID 0000-0003-0085-5386 (P. Lorenz); 0000-0002-6194-6559 (G.

Zieger); 0000-0001-7100-6926 (H. Schmidt)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the sheet resistance was measured with a four point probe. The composition of the films was measured via wavelength-dispersive X-ray spectroscopy (microprobe analyzer Jeol JXA8800). Film thicknesses were measured with quartz crystal oscillators during the deposition and afterwards checked using a Dektak profilometer (Veeco DekTak 8), to evaluate the electrical conductivity of the films by using the before measured sheet resistance. As in several practical fabrication processes thermal annealing steps are necessary, for the films annealing steps of 14 h at 200 °C in nitrogen atmosphere were performed with a repetition of the above mentioned measurements, respectively.

3. Results & discussion

Bi₁₀Sb₃₀Te₆₀

For different film thicknesses the Seebeck coefficient of Bi₁₀Sb₃₀Te₆₀-thin films is shown in Figure 1. We deposited three layers of that composition on unheated glass substrates and one layer on a glass substrate that was heated to about 185 °C. The as deposited films without heating show Seebeck coefficients close to 100 μV/K while for an increase in film thickness, the Seebeck coefficient slightly decreases. After annealing for 14 h at 200 °C, the Seebeck coefficient increases by more than 100 % for each film deposited on unheated glass substrates with the highest value at 233 μV/K for 402 nm thickness. While the Seebeck coefficient for the thin films with 283 nm and 402 nm are nearly identical, it slightly decreased for the 577 nm-film with 197 μV/K. The Bi₁₀Sb₃₀Te₆₀-thin film deposited onto a heated substrate has a higher Seebeck coefficient of 159 μV/K as deposited compared to the films deposited on the non heated substrates. After annealing the Seebeck coefficient has increased to 209 μV/K, but is still even lower than of the before mentioned annealed even thinner films deposited without active substrate heating. What catches the eye is that the Seebeck coefficient for all Bi₁₀Sb₃₀Te₆₀-thin films are on a comparable range after annealing, but decreases slightly for higher film thickness. For the electrical conductivity of the Bi₁₀Sb₃₀Te₆₀-thin films, shown in Figure 2, there is a rather big difference between the deposition on the heated and non heated substrates. While the as deposited thin films on unheated substrates have electrical conductivities of about $6.9 \times 10^3 \Omega/\text{m}$ to $8.4 \times 10^3 \Omega/\text{m}$, the as deposited film on the heated substrate has an electrical conductivity of $37.2 \times 10^3 \Omega/\text{m}$. After annealing, the films on unheated substrates show a slight increase in electrical conductivity, while the annealing of the film deposited

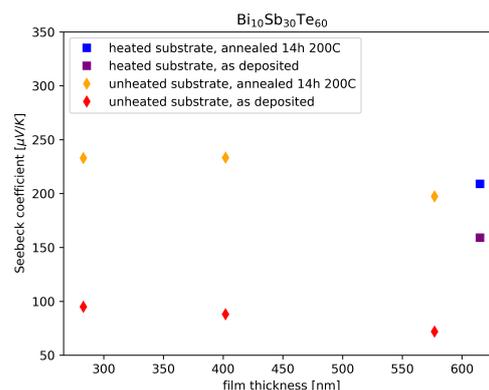


Figure 1: Seebeck coefficients of Bi₁₀Sb₃₀Te₆₀ thin films for different thicknesses. After annealing at 200 °C for 14 h there is an increase in value for all measured films.

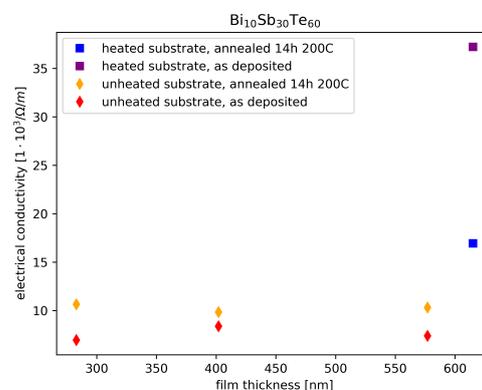


Figure 2: Electrical conductivity of Bi₁₀Sb₃₀Te₆₀ thin films for different thicknesses. The annealing at 200 °C for 14 h caused only small a increase for the depositions on unheated substrates and strong decrease for the film deposited on the heated substrate.

on the heated substrate led to a significant reduction to about $16.9 \times 10^3 \Omega/\text{m}$.

The power factor of the Bi₁₀Sb₃₀Te₆₀-thin films is shown in Figure 3. For the thin films deposited on unheated substrates a significant increase in the thermoelectric power factor is achieved through the annealing for 14 h at 200 °C, with a decrease towards higher film thickness. The power factor of the film deposited on the heated substrate is also significantly higher with 0.94 mW/m/K, while through annealing it decreases to 0.74 mW/m/K, but is still higher than the films deposited on unheated substrates.

Bi₁₁Sb₃₅Te₅₄

The results for the Seebeck coefficient of the second stoichiometry studied can be seen in Figure 4. All films

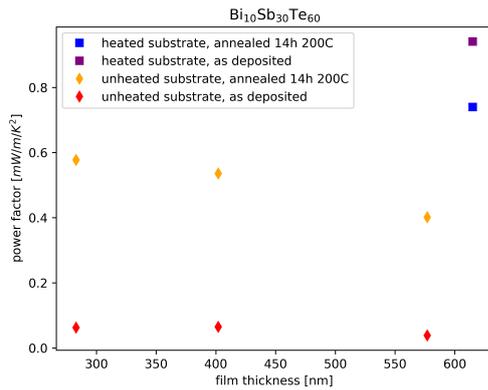


Figure 3: Power factor of $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ thin films for different thicknesses. After annealing at 200°C for 14 h the power factor of all films deposited on unheated substrates have a strong increase. For the film deposited on the heated substrate the power factor decreased with annealing but is still higher compared to the before mentioned films.

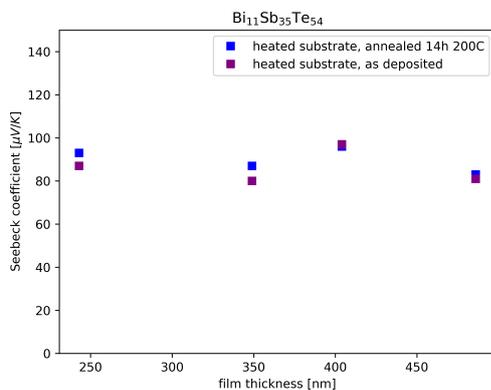


Figure 4: Seebeck coefficient of $\text{Bi}_{11}\text{Sb}_{35}\text{Te}_{54}$ thin films for different thicknesses. The values are similar for all investigated films, even after annealing at 200°C for 14 h, but are lower compared to the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin films.

were deposited on heated the substrates at 185°C and show a comparable Seebeck coefficient, despite the variation of thickness. Furthermore after annealing for 14 h at 200°C the Seebeck coefficient just slightly changes which can be expected due to the small change in temperature. Compared to the previous shown stoichiometry the Seebeck coefficient of this films is significantly lower, which shows that stoichiometry is probably more important than layer thickness for good thermoelectric performance of $\text{Bi}_x\text{Sb}_y\text{Te}_z$ -thin films. The electrical conductivity of the thin films is shown in Figure 5 which again does not seem to show any dependence on the layer thickness, but the annealing shows more of an impact then for the Seebeck coefficient,

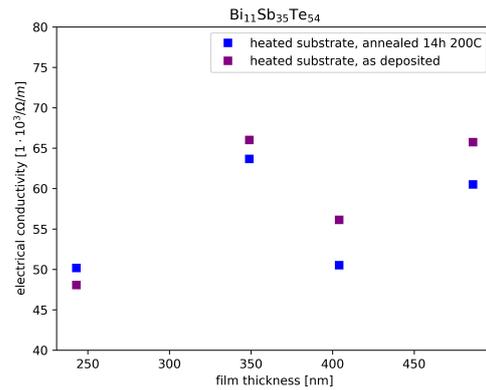


Figure 5: Electrical conductivity of $\text{Bi}_{11}\text{Sb}_{35}\text{Te}_{54}$ thin films for different thicknesses. Before and after annealing the values are significantly higher then that of the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin films .

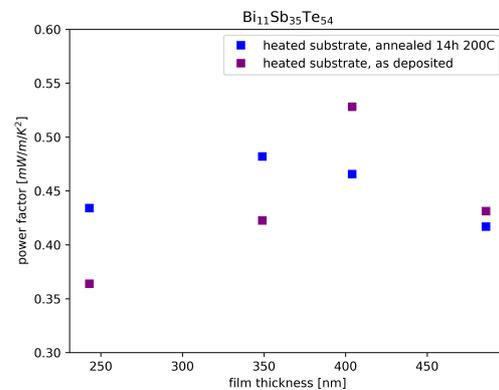


Figure 6: Power factor of $\text{Bi}_{11}\text{Sb}_{35}\text{Te}_{54}$ thin films for different thicknesses. The values are comparable to the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin films deposited on unheated substrates even after annealing.

which is decreasing for the majority of the films, but not as much as for the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin film showed earlier. In general, the electrical conductivity in these thin films is higher than that of the previous shown films of other stoichiometry. The power factor of the $\text{Bi}_{11}\text{Sb}_{35}\text{Te}_{54}$ -thin films, as presented in Figure 6, is in general slightly lower compared to those of the the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin films, especially when compared to the film deposited on the heated substrate. This leads to the conclusion, that in terms of thermoelectric power factor, the $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ stoichiometrie is the better composition.

4. Conclusion & Outlook

We presented our ongoing research on three source electron-beam co-deposition of thin $\text{Bi}_x\text{Sb}_y\text{Te}_z$ -films. We showed first promising results of the thin films.

A variation of the film thickness of the $\text{Bi}_x\text{Sb}_y\text{Te}_z$ thin films was performed in the range of $d = 240$ nm to 620 nm to compare the resulting layer properties. Additionally, one annealing step was applied to analyse their impact on said properties. For all samples Seebeck coefficient and electrical conductivity have been measured, and thermoelectric power factor has been derived before and after annealing.

Promising thermoelectric properties were obtained for these preliminary samples. The promising results of $\text{Bi}_{10}\text{Sb}_{30}\text{Te}_{60}$ -thin films will be verified by additional samples and further investigation by structure analysis to find the reason for the decrease in electrical conductivity.

References

- [1] U. Dillner, V. Baier, E. Kessler, J. Müller, A. Berger, D. Behrendt, H.-A. Preller, A high sensitivity single-chip 4-element thermoelectric infrared sensor, in: Proceedings of the 8th International Conference for Infrared Sensors and Systems, Nuremberg, Germany, 2004, pp. 25–27.
- [2] B. Habbe, J. Nurnus, Thin film thermoelectrics today and tomorrow, *Electron. Cooling* 17 (2011) 24–31.
- [3] D. Kraemer, B. Poudel, H.-P. Feng, J. C. Caylor, B. Yu, X. Yan, Y. Ma, X. Wang, D. Wang, A. Muto, et al., High-performance flat-panel solar thermoelectric generators with high thermal concentration, *Nature materials* 10 (2011) 532.
- [4] S. H. Lee, H. Shen, S. Han, Flexible thermoelectric module using bi-te and sb-te thin films for temperature sensors, *Journal of Electronic Materials* (2019) 1–7.
- [5] M. Kashiwagi, S. Hirata, K. Harada, Y. Zheng, K. Miyazaki, M. Yahiro, C. Adachi, Enhanced figure of merit of a porous thin film of bismuth antimony telluride, *Applied physics letters* 98 (2011) 023114.
- [6] M. Takashiri, S. Tanaka, K. Miyazaki, Improved thermoelectric performance of highly-oriented nanocrystalline bismuth antimony telluride thin films, *Thin Solid Films* 519 (2010) 619–624.
- [7] F. Völklein, V. Baier, U. Dillner, E. Kessler, Transport properties of flash-evaporated $(\text{bi}_{1-x}\text{sb}_x)_3\text{te}_3$ films i: Optimization of film properties, *Thin solid films* 187 (1990) 253–262.
- [8] A. A. Marinho, N. P. Costa, L. F. C. Pereira, F. A. Brito, C. Chesman, Thermoelectric properties of bisbte alloy nanofilms produced by dc sputtering: experiments and modeling, *Journal of Materials Science* 55 (2020) 2429–2438.
- [9] M. Bala, A. Masarrat, V. Kumar, S. Ojha, K. Asokan, S. Annapoorni, Effect of thermal annealing on thermoelectric properties of $\text{bixsb}_2\text{-xte}_3$ thin films grown by sputtering, *Journal of Applied Physics* 127 (2020) 245108.
- [10] S.-j. Jeon, H. Jeon, S. Na, S. D. Kang, H.-K. Lyeo, S. Hyun, H.-J. Lee, Microstructure evolution of sputtered bisb–te thermoelectric films during post-annealing and its effects on the thermoelectric properties, *Journal of Alloys and Compounds* 553 (2013) 343–349.
- [11] S. Lal, D. Gautam, K. M. Razeeb, Optimization of annealing conditions to enhance thermoelectric performance of electrodeposited p-type bisbte thin films, *APL Materials* 7 (2019) 031102.
- [12] C.-K. Yang, T.-C. Cheng, T.-H. Chen, S.-H. Chu, The thermoelectric properties of electrochemically deposited te-sb-bi films on ito glass substrate, *INTERNATIONAL JOURNAL OF ELECTROCHEMICAL SCIENCE* 11 (2016) 3767–3775.
- [13] T.-H. Chen, P.-H. Chen, C.-H. Chen, Laser co-ablation of bismuth antimony telluride and diamond-like carbon nanocomposites for enhanced thermoelectric performance, *Journal of Materials Chemistry A* 6 (2018) 982–990.
- [14] S. Schiller, U. Heisig, S. Panzer, *Elektronenstrahltechnologie*, VEB VerlagTechnik, 1976.

Temperature Determination During Flash Lamp Annealing

Viktor Begeza^a, Lars Rebohle^{a,b} and Thomas Schumann^{a,b}

^a *Helmholtz-Zentrum Dresden - Rossendorf, Institute of Ion Beam Physics and Materials Research, Bautzner Landstraße 400, 01328 Dresden, Germany*

^b *Helmholtz Innovation Blitzlab, Bautzner Landstraße 400, 01328 Dresden, Germany*

Abstract

Flash lamp annealing (FLA) is a modern technology for the thermal treatment of materials which currently opens up new application areas. During FLA, an intense pulse of light with a pulse duration of milliseconds and below is applied to the surface of a material. In contrast to traditional methods like furnace annealing, temperature now strongly depends on the material properties and the thickness of the sample. In addition, the short time scale leads to a temperature distribution over depth and makes direct temperature measurements very challenging.

In this work we first review in brief the existing possibilities for a direct temperature measurement during FLA. The main part presents our own concept which is a combination of direct measurements, calibration and thermodynamic simulation. The latter point is of special interest as it allows to get information about the temperature distribution within the material, provided that the relevant material parameters are known. Finally, the impact of such temperature distributions on physical processes like diffusion, crystallization and phase formation is discussed.

Keywords ¹

Flash Lamp Annealing, Temperature Distribution, Millisecond Thermal Treatment, Thermodynamic Simulation, Diffusion, Crystallization, Phase Formation

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ l.rebohle@hzdr.dw (L. Rebohle);

🌐 www.hzdr.de



© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Workshop »Future of Education«

The Future of Education – Lessons learned from CERC Workshop “Future of Education” and an Online Course on “Digital Basics”

Tanja Kranawetleitner^a, Heike Krebs^a, Nina Kuhn^a and Robert Loew^b

^aApplication Center for Materials and Environmental Research (AMU), Augsburg University, Germany

^bDIPF | Leibniz Institute for Research and Information in Education, Frankfurt, Germany

Abstract

Based on the findings in the development of an interactive self-learning course as an additional offer in vocational education, the participants discussed future educational concepts at a workshop during CERC 2021.

Keywords

Education, Vocational Training

1. Introduction

In the project »Education 4.0 for SMEs« we asked teachers and students from vocational schools which digital competences they want to foster which are not or only superficially included in the normal curriculum. Based on this survey we developed an interactive online course »Digital Basics« with the following topics:

1. Office Basics,
2. Data protection on the internet,
3. Motivation – especially for education,
4. Tinkercad – creative ideas for future education,
5. Job application coaching.

Didactically, »classic eLearning« is implemented here, but the materials provided are just as suitable for use in normal school operations as they are for the »inverted classroom method«. Each topic is offered in a module independent of the other topics.

In the workshop, we gave a brief insight into this course, and then discussed how education could basically be designed in the future.

This article aims to present the insights gained in the workshop based on the principles developed for the online course. These insights were incorporated both in the revised version of the course before its actual implementation with learners and in various other events based on the course content (see chapter 5). The conclusions (see chapter 7) are based on the findings of the course implementation as well as the further events.

The course is offered in German via the moodle eLearning platform. An insight into the course content can be obtained via <https://rloew.eu/digitalbasics>.

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland

✉ mail@robertloew.de (R. Loew)

🌐 <https://info.robertloew.de> (R. Loew)

🆔 0000-0002-3696-4261 (T. Kranawetleitner); 0000-0003-2649-657X (H. Krebs); 0000-0002-3566-8890 (R. Loew)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Basic Preliminary Considerations

The aim of the course is not to impart complete technical knowledge, but rather to provide an introduction to the individual topics. The curiosity of the participants is to be awakened so that after working through the course they can continue to explore the topics of interest to them independently.

To this end, the course attempts to create an awareness of problems in order to recognize pitfalls in professional practice and to be able to react correctly.

These principles are now presented on the basis of the topic modules, which differ considerably from one another in their implementation. From this difference we would like to gain experience on how these methods are accepted by the learners in order to shape education of the future.

3. Modules Implementations

3.1. Office Basics

Using numerous materials and learning videos on word processing, presentation and spreadsheet, an introduction to Office is given, which is supplemented by exercises. The course is essentially limited to technical knowledge for implementing commonly needed functions. For example, design basics are deliberately excluded, although how text formatting works is shown.

Learning content already available on the web for self-study was deliberately used in a slightly modified form to give learners the opportunity to deepen their knowledge with content from the original course after completing the course.

3.2. Data Protection

After a brief introduction to the basics of data protection, the course looks at opportunities and risks in social media. The aim is to raise awareness that it is problematic to simply post what interests you. For example, you should consider whether a photo might violate a person's personal rights or contradict the confidentiality of company information. You should also recognize whether it is better not to comment on a post or answer a question in social media under any circumstances, since the intention behind this is to spy on a person in a personal or professional environment.

Digressions on creating »good passwords«, for example, show other areas of data protection that can be directly implemented in practical everyday life.

This topic usually involves working with examples in which the participants think of the answers on their own, after which the solution is displayed together with an explanation (see fig. 1 and fig. 2).

3.3. Motivation

How can I motivate myself? Based on an existing podcast series on the subject, participants are introduced to the topic. Here, too, the entire podcast series is then available for further self-study. Interactive quizzes are used to playfully check learning progress.

3.4. Tinkercad

The task is to design your own vision of the »education of the future« as a CAD model. For this purpose the CAD software Tinkercad is used, which is offered free of charge web-based by its producer. This way we want to show that nowadays you should be able to learn and use new tools quickly. If

I have a private social media profile where I regularly post texts and photos from my personal and professional life.

What can I post ?

Class photo of my vocational school class

Photo of me in front of the company headquarters

Photo of me at a machine / in my office

List of grades of an exam

Figure 1: Example of an interactive question of module data protection.

Post :

»Name a movie you've watched over 5 times and still enjoy !«

This post has been commented more than 500,000 times !

What is the intention behind the post ?

Answer :
It seems to be so-called **»Social Engineering«** !
(Fishing passwords, passphrases and answers to security questions.)

Figure 2: Example of a post shown in module data protection.

you have ever played with Lego™ building blocks, you should be able to implement your own creative ideas quite quickly with Tinkercad (see fig. 3).

3.5. Job Application Coaching

This section provides tips on how to create your own application documents. The result of this part of the course are documents that can be used directly in an application. Part of the offer is to submit the created documents and receive comments from experts.

Knowledge is provided on how to behave in job interviews by means of a number of quizzes, in which a question is first answered by the participant and then the correct answer is displayed (see fig. 4).

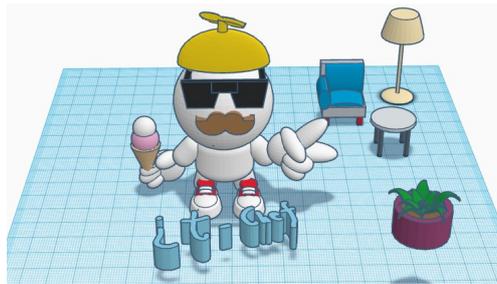


Figure 3: Example of a model on »future workplace«.

What should you definitely do at an interview ?

Appear well groomed for the interview

Better to arrive a little later to appear »cool«

First ask how much vacation I have

Prepare myself before the interview

Figure 4: Example of an interactive question of module job application coaching.

4. Workshop

In the context of an impulse lecture of the workshop, the modules of the course and the different learning methods used were shown as examples in order to form a basis for discussion.

Likewise, other »modern« learning methods were briefly discussed, such as Inverted Classroom [1].

An important result of the workshop discussion was that the evaluation of learner behavior can provide detailed conclusions about the acceptance and motivation of the learners. Therefore, this should already be planned during the design of the course, as one goal of the course was to gain insights into how learning in vocational education will be optimally designed in the future.

Here we would like to thank the workshop participants for their active contribution and many creative suggestions, which we were able to incorporate before the start of the course. For example, the logging of participant behavior in the »Data Protection« module was optimized so that a summary can be viewed online at any time (see fig. 6).

5. Using the Course Content in other Project Offerings

Based on the course content, interactive lectures were offered as part of several events – here is a selection:

5.1. AzubiCamp 2022 : Digital Learning Camp for Apprentices

Interactive presentations were offered to the participating trainees on the following topics

- Trainees recruit trainees – apply for a job and your company (based on the module »Job Application Coaching«),
- Pimp my data protection (based on the module »Data Protection«),
- Where did it go? – Finding and retaining your own motivation (based on the module »Motivation«).

5.2. alles digital – Skills and Tips for Trainers

- Digital basics for trainees – an e-learning course (presentation of course content)

6. Analysis of the eLearning Course

To find out which types of learning were popular for participants and how, we used the following methods of analysis:

- Survey of learners immediately after their completion of the course,
- Discussion during the final event,
- Log evaluation (learning analytics),
- Feedback during the events mentioned in section 5.

6.1. Survey Result

The various modules were all rated as at least good, both in terms of content and the learning method selected in each case. In the Office module, it was suggested that there should optionally be more in-depth courses or courses that go beyond the basics. The usability in the professional environment was also rated as good for all modules. The data protection module, which was intended to raise interest in the topic rather than to convey technical content, was very well received by the participants thanks to its playful and very example-oriented approach.

6.2. Final Discussion

It became apparent that the participants liked the various methods. The interactive learning content was particularly well received.

Fig. 5 shows that the participants enjoyed it and learned something new. They also saw the applicability of what they had learned in practice as good.

In a quiz at the final event, it showed that at least the most important basics stuck with many participants and that they had a very good understanding of the principles.

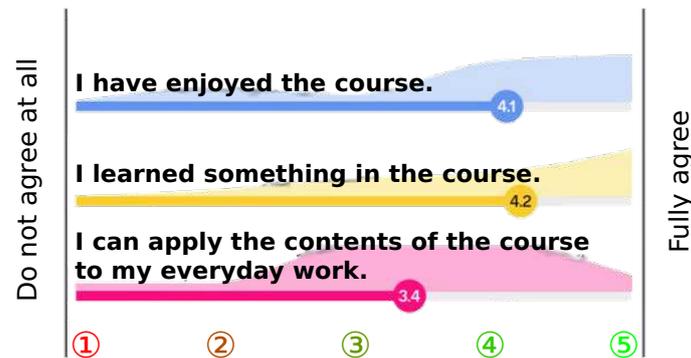


Figure 5: Survey during the closing event of the eLearning course.

Log Part 1 of »Social Engineering« from module »Data Protection«



Figure 6: Evaluation example Learning Analytics.

6.3. Learning Analytics

The evaluation of the logs of the first run of the course is not yet completely finished, since the analysis is to be done together with the second run; however, the following statements can be made:

1. If a topic was processed by a participant, most of the learning content from it was usually also reviewed.
2. The order of working on the topic mostly followed the order suggested in the course, although this was not mandatory.
3. Exception with 1. and 2. was here the topic Office: Here the focus was often on one of the topics of the course. Exception with 1. and 2. was here the topic Office: Here the focus was often put on one of the subtopics word processing, presentation or spreadsheet. However, this was to be expected, since the previous knowledge of the participants was extremely different according to the pre-survey.
4. Further offers – e.g. links to websites with more in-depth information – were rarely used.

The evaluation example in Fig. 6 shows how many participants viewed the detailed information on the posts in the »Social Engineering« section (period 2 days, 21 participants).

7. Conclusions

As we suspected, a mix of learning methods turned out to be positively valued by learners. It seems worthwhile to carefully consider how each learning content is presented, reinforced, and motivated.

In an online-only learning experience, interactivity is critical to learner motivation due to the elimination of all interaction between participants that occurs in almost all other forms of learning.

The ability to contact the provider team with questions was also rated well; although this was rarely used by participants during the course. The participants were mainly trainees who were supported by their training company during the course. Therefore, there was also a further possibility of support by the trainers.

Hint

An overview of the project results can be found in the recommendations for action [2].

Acknowledgements

The Project “Bildung 4.0 für KMU” (“Education 4.0 for small and medium-sized manufacturing enterprises”, Grant number 01PA17014) is funded by the German Federal Ministry of Education and Research and the European Social Fund for Germany within the “Digital Media in Vocational Training” program.

We would like to thank the participating companies and vocational schools for their support, as well as the trainees who made use of the course (some of them in their free time).

References

- [1] Bergmann, J., Sams, A.: Flip Your Classroom: Reach Every Student in Every Class Every Day. Flipped Learning Series, International Society for Technology in Education (2012), <https://books.google.de/books?id=nBi2pwAACAAJ>
- [2] Filipenko, M., Kranawetleitner, T., Krebs, H., Lechler, K., Loew, R., Pistoll, D., Priesmeier, F.: Handlungsempfehlungen. Projekt Bildung 4.0 Self Publishing, Augsburg, Germany (2022), <https://www.b4kmu.de/Handlungsempfehlungen>

Award Winners

Best Paper Award

A Robust Martingale Approach for Detecting
Abnormalities in Human Heartbeat Rhythm

Jonathan Etumusei, Jorge#Martinez Carracedo, Sally McClean

Best Presentation Award

Trust and Transparency in Data Protection in
Online-Marketing — Differences Between Different
Generations

Louis Kerker, Ingo Stengel, Stefanie Regier

CERC 2021
Collaborative European
Research Conference
Cork, Ireland
9 - 10 September 2021
www.cerc-conf.eu

