NEMISA NPC
National Electronic Media Institute of South Africa

NEMISA 2023
DIGITAL SKILLS COLLOQUIUM
15-17 FEBRUARY 2023

Scaling data skills for multidisciplinary impact

PROCEEDINGS

Nkosikhona Theoren Msweli | Victoria Tau | Sibukele Gumbo | Hossana Twinomurinzi | Colin Surendra Thakur | Tendani Mawela

Coastlands Umhlanga Hotel & Convention Centre, Durban, South Africa

UNISA university of south africa

NEMISA Digital Skills Conference (Colloquium) 2023

Scaling data skills for multidisciplinary impact

NKOSIKHONA THEOREN MSWELI

VICTORIA TAU

SIBUKELE GUMBO

HOSSANA TWINOMURINZI

COLIN SURENDRA THAKUR

TENDANI MAWELA

ISBN Electronic: 978-0-6397-0845-4 (preprints)

ISBN Print: 978-0-6397-0844-7 (preprints)

15-17 February 2023

Coastlands Umhlanga Hotel & Convention Centre, Durban, South Africa

National Electronic Media Institute of South Africa

(NEMISA) and

University of South Africa (UNISA)

NEMISA Digital Skills Conference (Colloquium) 2023

Scaling data skills for multidisciplinary impact

ISBN Electronic: 978-0-6397-0845-4 (preprints)

ISBN Print: 978-0-6397-0844-7 (preprints)

Published by:

National Electronic Media Institute of South Africa (NEMISA)
and
University of South Africa

26 Canary Street

Auckland Park,

2006

South Africa,

Phone: +27 11 484 0583

Email: info@nemisa.co.za

Printed in the Republic of South Africa

Table of Contents

Acknowledgements

The proceedings have been made possible by funding from the National Electronic Media Institute of South Africa (NEMISA) in collaboration with the Knowledge for Innovation Unit (K4I) at the University of South Africa (UNISA).

We thank each of the NEMISA CoLabs, Innovation hubs and academic fraternity for the submissions and the efforts in making the NEMISA Digital Skills Conference (Colloquium) 2023, and the associated Postgraduate Symposium, a success.

List of Reviewers

We wish to thank the following reviewers for their contributions to the Proceedings. Each reviewer blind-peer reviewed at least two submissions. None of the authors or the editors were involved in reviewing their own submissions.

Billy Kalema	University of Mpumalanga
Chimwemwe Queen Mtegha	University of Cape Town
Colin Chibaya	Sol Plaatje University
Elias Tabane	University of South Africa
Lateef Amusa	University of Johannesburg
Lovemore Motsi	University of South Africa
Madzima Kudakwashe	Sol Plaatje University
Naomi Isabirye	University of the Witwatersrand
Nkosikhona Theoren Msweli	University of South Africa
Ntombizethu Lubisi	University of the Witwatersrand
Olawande Daramola	Cape Peninsula University of Technology
Priscilla Maliwichi	Malawi University of Science and Technology
Rodney Mushininga	University of KwaZulu-Natal
Sibukele Gumbo	Walter Sisulu University
Steven Mamorobela	Tshwane University of Technology
Tendani Mawela	University of Pretoria
Thembekile Mayayise	University of the Witwatersrand
Wallace Chigona	University of Cape Town
Yvonne Machaba	University of South Africa

List of Editors

Hossana Twinomurinzi	University of Johannesburg
Colin Surendra Thakur	Durban University of Technology
Tendani Mawela	University of Pretoria
Nkosikhona Theoren Msweli	University of South Africa
Sibukele Gumbo	Walter Sisulu University
Victoria Tau	National Electronic Media Institute of South Africa

Conference Organisers

Organising Committee

Ms Nkosikhona Msweli	University of South Africa
Ms Victoria Tau	National Electronic Media Institute of South Africa
Ms Sibukele Gumbo	Walter Sisulu University
Prof Colin Surendra Thakur	Durban University of Technology
Prof Hossana Twinomurinzi	University of Johannesburg

Conference Chairs

Prof Colin Surendra Thakur	Durban University of Technology
Prof Hossana Twinomurinzi	University of South Africa

Postgraduate Symposium Chairs

Ms Nkosikhona Msweli	University of South Africa
Ms Sibukele Gumbo	Walter Sisulu University
Ms Sarah Jembere	Durban University of Technology

Programme Committee

Ms Victoria Tau	National Electronic Media Institute of South Africa
Ms Nkosikhona Msweli	University of South Africa
Ms Sibukele Gumbo	Walter Sisulu University
Prof Hossana Twinomurinzi	University of Johannesburg

Datathon Committee

Ms Victoria Tau	National Electronic Media Institute of South Africa
Ms Nkosikhona Msweli	University of South Africa

Preface and Executive Summary

The NEMISA Digital Skills Conference (Colloquium) 2023, and the associated events (Postgraduate Symposium and Technology Innovation session) was held between 15-17 February 2023 at the Coastlands Umhlanga Hotel & Convention Centre in Durban, hosted by the Durban University of Technology (DUT) in collaboration with the Knowledge for Innovation unit (K4I) at the University of South Africa (UNISA) and the National Electronic Media Institute of South Africa (NEMISA).

A total of 496 delegates (184 in person and 312 virtually through Zoom) attended the three collocated events (Conference, Postgraduate Symposium and Technology Innovation) bringing together thought leaders, policy makers, software developers, technology enthusiasts, industry leaders, academics, students, social entrepreneurs, leaders of non-government organisations and innovation hubs to address how to scale data skills for multidisciplinary impact. The virtual delegates joined from 2 continents and 5 countries (excluding South Africa); Eswatini, Ethiopia, Germany, Kenya and Nigeria.

The purpose of the colloquium and events centred around the central role that data plays today as a desirable commodity that must become an important part of massifying digital skilling efforts. Governments amass even more critical data that, if leveraged, could change the way public services are delivered, and even change the social and economic fortunes of any country. Therefore, smart governments and organisations increasingly require data skills to gain insights and foresight, to secure themselves, and for improved decision making and efficiency. However, data skills are scarce, and even more challenging is the inconsistency of the associated training programs with most curated for the Science, Technology, Engineering, and Mathematics (STEM) disciplines. Nonetheless, the interdisciplinary yet agnostic nature of data means that there is opportunity to expand data skills into the non-STEM disciplines as well.

Technology Innovation

The Technology Innovation session was a culmination of the NEMISA Datathon held in November 2023 under the same theme of “Scaling data skills for multidisciplinary impact”. The datathon resulted in five talented winning teams being selected. It is noteworthy that each of these teams hailed from a different province of South Africa, including Mpumalanga, Gauteng, North West, Limpopo, and Free State. Over the course of two months following their victory, each team worked tirelessly with an innovation hub and mentors Koedr and Thenjiwe Kubheka to refine their solutions and compete for the ultimate winner prize at the conference. The panel of judges who evaluated their submissions was composed of esteemed professionals in the field, bringing a high level of expertise and credibility to the selection process. These judges were: Mr Tiyani

Nghonyama (Geekulcha), Mr Tumelo Baloyi (Monolith), Dr Naomi Isabirye (University of the Witwatersrand) and Dr Banele Mhlongo (Department of Health).

Mr Thato Semenya, Mr Letlhogonolo Korosi, Mr Nthabiseng Mathikge, Mr Itumeleng Pitje and Mr Katlego Mokgosane from Future DX were the winners in North West and worked with Mafikeng Digital Innovation Hub. Their solution was an online learning management system that offers future courses based on emerging technologies. The solution offers each student a portfolio of evidence on completion. The solution also AI to predict which jobs might be suitable to the learner and makes a recommendation.

The A-Team consisted of Ms Sanna Rasethaba, Ms Malebo Moleleki, Mr Kenan Kasongo and Mr Aaron Kibambe. They are from the Gauteng Province under the Technopreneurship Centre at the University of Johannesburg. Their solution focused on detecting misinformation mainly from social media using AI, and determining whether information being placed in the public domain is likely true or false. They saw an opportunity to offer reliable information to the public and to foster greater transparency and accountability.

Mr Lunga Feni and Mr Bahle Menziwa from the Cortex Hub were from the Limpopo Province. They presented a solution aimed to address the digitization gap in rural areas in South Africa, with a focus on network coverage, road access, electricity coverage, schools, and health care services, as well as census data. Their solution saw the opportunity from areas that do not have access to communications coverage and are therefore digitally excluded.

Techgro members were Mr Tshwarelo Madisha, Mr Bigboy Mogajane, Ms Maria Kanyane, Ms Patricia Kekana and Ms Tshegofatso. They were the datathon winners from Mpumalanga. They worked with the Vulingqondo Innovation Hub to refine their solution on Smart Agriculture. Being in a rural community that experiences high unemployment and poverty, their solution focused on rural smart agriculture initiatives that can unlock economic activities and create green job opportunities. They identified the opportunity as access to vast amounts of land but with limited water and long spells of heat.

Mr Jandre van Vuuren, Mr Ruan Britz, Mr Corne Combrink and Mr Schalk Hanekom of Elite Vine were from Free State. Their solution, Jobboard, is a cloud-based platform to connect job seekers and employers. It offers a convenient solution for individuals who are unemployed or facing resource limitations by providing access to job and education opportunities through both online and physical channels.

Postgraduate Symposium

The postgraduate symposium provided a facilitative platform for emerging researchers to showcase their work, to network, and to engage with seasoned academics and practitioners.

The day was opened by Prof Shawren Singh, the Head of the Department of Information Systems at the University of South Africa. He emphasised on the importance of continuous innovation across all sectors, driven by digital technologies, to improve South Africa's economy.

Prof Nixon Muganda Ochara from the University of the Witwatersrand urged students and researchers to incorporate AI and other digital technology tools into their postgraduate studies. He highlighted the significance of the current era, where access to information through digital assistants has become much easier. Prof Ochara emphasized the need for the practice of science to evolve with the available digital tools, rather than clinging to traditional methods. Additionally, Prof Ochara pointed out that a researcher's world view or philosophical stance has a significant impact on the culture of research writing, making it crucial for researchers to be mindful of this aspect. This perspective also influences the ethics surrounding research findings. Therefore, researchers should prioritize deep critical thinking and to develop the digital skills to use multiple digital assistants to enhance their work.

Prof Wallace Chigona, from the University of Cape Town, shared his insights on postgraduate supervision. He highlighted the increasing role of postgraduate students in their own supervision and drew from his wealth of experience in supervising postdoctoral fellows, PhD and Masters students. Prof Chigona shed light on a topic that is often overlooked - the supervisory experience of academics. He emphasized the need for more discourse on this subject, especially given the growing pressure on supervisors to produce more postgraduate students. The discussion also explored the intricate relationships between students and their supervisors, emphasizing the importance of emotional and social support in addition to technical guidance. The discussion also noted that some students may attempt the PhD for the wrong reasons and require additional support to successfully complete their studies. The discussion highlighted the need for more attention to be given to the supervision process and the importance of providing comprehensive support to postgraduate students.

Three PhD posters were presented: Chimwemwe Queen Mtegha from the University of Cape Town presented her research on institutional factors in government agencies that affect the realisation of national cybersecurity capacity building outcomes. She drew from Malawi. Thaneshni Moodley from the Durban University of Technology presented on framework to observe and analyse customer experience on the twitter platform using machine learning techniques. Teofelus Tonateni Tuyeni also from the University of Cape

Town discussed the research on identifying important factors that influence governments commitment to cybersecurity drawing from Namibia.

There were two workshops; one was on Natural Language Processing and Sentiment Analysis by Prof Colin Thakur and Mr Yassen Khan, and the other on Statistical Analysis using R by Dr Lateef Amusa and Ms Sibukele Gumbo. The Natural Language Processing and Sentiment Analysis workshop provided the audience with an in-depth understanding of how to extract meaning from textual data, identifying patterns and trends, and analysing the sentiment behind written content. Similarly, the Statistical Analysis using R offered an introduction to the R programming language and its use in statistical analysis. The audience learned how to visualize, explore, and analyse complex data sets using various R packages and tools, as well as how to perform hypothesis testing and regression analysis. The hands-on approaches of both workshops were particularly helpful in delivering an appreciation of data skills used on real-world data sets.

Colloquium

A total of 18 submissions were received for the colloquium of which 12 were accepted as full research papers after a double-blind peer review process, representing a 66.7% acceptance rate. The average peer review rate was 2.43, which means that each research paper was reviewed by at least 2 reviewers with some having more reviews.

Prof Shawren Singh, speaking on behalf of the Deputy Vice Chancellor of Teaching and Learning at the University of South Africa, gave a warm welcome to the esteemed attendees of the colloquium. In his opening remarks, Prof Singh emphasized the importance of creating an environment that fosters a thriving culture of research at the postgraduate level, which is inclusive of multiple sectors of society. Prof Singh went on to highlight the evolving demands placed on postgraduate research in today's world, which now requires not just theoretical contributions, but practical ones as well. He stressed the significance of globalisation in research and how it has intensified the competition at a global level, putting pressure on students to complete their postgraduate degrees in a shorter period of time. The discussion also delved into how research is driving organisations globally, but in South Africa, there seems to be a slow pace in focusing on solving critical societal problems. This puts a greater expectation on postgraduate students in South Africa today to deliver results that will have a real impact.

Prof Colin Thakur gave a welcome on behalf of the Deputy Vice Chancellor of Durban University of Technology. He stressed the significance of innovation in today's rapidly evolving business landscape and gave a reminder of the importance of asking questions, the right questions, and valuing all stakeholders in the creation of new knowledge. However, he cautioned that knowledge alone does not guarantee success. We must also possess the wisdom to use it appropriately and avoid potential destructive consequences. Innovation is an ever-changing target that requires constant improvement, for which we must remain vigilant and strive to continually develop innovative capabilities.

The Chair of the NEMISA Board, Ms Molebogeng Leshabane, extended a warm welcome to all guests and introduced the Acting Deputy Director General who officially opened the conference. During her remarks, Ms Leshabane noted the increasing complexity in the world that requires a multidisciplinary approach. She emphasized that data, in its agnostic nature, permeates all industries and is central to bringing together different sectors of society. Navigating these complexities to lead digital skills efforts in such a society is a great task for NEMISA. However, the organization is committed to reducing inequality and ensuring that no one is left behind. In addition, Ms Leshabane highlighted that youth unemployment is a pressing challenge that requires significant effort. The digital economy provides new opportunities, such as algorithmic auditing, which can help address this issue. Youth can also access markets that were not previously available, and government services can be improved through digital innovation.

The conference was officially opened by the Acting Deputy Director General of the Department of Communications and Digital Technologies (DCDT), Ms Nonkqubela Thathakahle Jordan-Dyani. On behalf of the Deputy Minister, My Philly Mapulane, she extended a warm welcome to all the partners of NEMISA who are committed to achieving the digital skills mandate, as well as the researchers who will be leading the way forward. Ms Jordan-Dyani emphasized the importance of addressing the curriculum and access to higher education to ensure that everyone has equal opportunities to develop the digital skills needed in today's world. She challenged NEMISA to fast-track the importance of digital skills in a changing world that is being driven by advanced digital technologies. With the skills South Africa already has, talent is not the challenge. What is needed is to share research findings and knowledge to enable everyone to benefit. She decried the silos in government and urged NEMISA to share the learnings from the conference widely through digital platforms. How can advanced digital technologies be filtered down to the every South African? This was a question she posed, as she highlighted the need to address the current challenges facing South Africa, such as unemployment, poverty, and socioeconomic equity. Skills development is a crucial part of South Africa's advancement, including responsive innovation to meet local demands. Achieving this requires strong partnerships from different sectors of society, including the rural areas. Addressing the issue of male dominance in the ICT field and making data and infrastructure more widely available are critical components of this effort. Ms Jordan-Dyani concluded by encouraging NEMISA to share the results and outcomes of the conference for others to draw upon. With everyone working together, South Africa can move forward with confidence and ensure that no one is left behind in the digital age.

The second keynote was offered by Dr Mark Nasila, the Chief Data Analytics Officer at the First Rand Group. Dr Nasila underscored the need to adopt new thinking in order to create a better future. He highlighted the importance of how advanced technologies such as AI are introduced into organisation, and emphasized the social aspects that must be considered in doing so. Dr Nasila compared how over $16 trillion is being invested in bringing AI into organizations worldwide, with China and the USA leading the way. He stressed the importance of skilling people to be a part of this wave of investment and to trust in AI. He also spoke about the ethical implications of AI, including the fear of job losses, and laboured the need for leaders to repurpose people into new roles that will not be taken over by machines. The keynote speech expressed confidence that NEMISA's investment in AI skilling will pay off and contribute to a better future for South Africa.

Ms Cheryl Benadie in her keynote emphasized the importance of focusing on the humans behind the data, science, and technology. Specifically, she brought attention to the area of wellness, which is often overlooked in the rapidly changing and advancing technology space. Ms. Benadie noted that the younger generations have grown up in a world where technology and reality are not distinguished, leading to a sense of disconnection from the real world. Shockingly, suicide has become the second leading cause of death among those aged 15-24, with the Gen Z facing a crisis of identity. Ms. Benadie called for purpose to be considered a critical digital skill for the younger generation to help combat this issue. It is crucial to connect the head and the heart to create a sense of wholeness among the Gen Z, who are more prone to burnout and stress. Ms. Benadie observed that technology can sometimes overshadow the humanity of people, leading to a loss of connection. Finally, she noted that jobs are becoming increasingly fluid, making it difficult to guarantee employment after completing studies – this is also a significant stressor today.

Panel on the Digital Skills Framework

Prof Leona Craffert, who chaired the panel session, offered an overview of the Digital Skills Framework (DSF) created in 2013, highlighting how the DSF provides an organizing framework on digital skills in South Africa, ranging from basic to advanced high-tech digital skills. Mr Trevor Rammitlwa, the CEO facilitated the panel, identifying how the transformation South Africa is going through is making digital skills essential across the board. However, there is a fragmentation of how digital skills are being approached, and the DSF serves as a means of bringing together different perspectives on digital skills. Prof Debbie Collier, an expert in Labour Law who works on digital work, the platform economy, and recently with Labour Unions and Nedlac on the regulatory (or policy) aspects of digital skills, noted that context plays a significant role in the world of work, and the strategy to roll out digital skills and the framework for digital skills need to interact in the holistic world of work. The framework is goal-oriented, offering a guide on what to look for, how to continually adapt to the changing world of work, and how to gain skills needed for new environments. M. Andy Searly, Director of Paladin Consulting and he original founder of the Digital Work Accelerator intiative, provided a perspective from industry on digital skills as part of collective action, noting the challenges of speed and scale demanded for skills for the market. He found that digital skills need to start at the primary level, extending to the workplace. The framework provides a common language and pathway to connect different paths a person may want to take, offering certainty to industry about what to expect from graduates of programs designed around the framework. Mr John April, Director in the Office of the CEO for QCTO, noted that the framework offers an opportunity to respond to the growing demand for digital skills, helping to set standards against which to develop content and qualifications, including verification and authentication. Mr Mlindi Mashologu, DDG for the Department of Communication and Digital Technologies (DCDT), emphasized that from a policy perspective, a digitally skilled society needs to have a reference point in the Future Skills plan of South Africa, which is supported by the ministry. Failing to have a digitally skilled population by 2030 risks alienation from a global society, a massive loss of jobs, and failure to be a competitive nation. Finally, it was mentioned that NEMISA has been working on an online platform from which digital skills are delivered. In conclusion, the panel discussed the importance of the DSF framework, which offers a common language between stakeholders, addresses the urgency of skills development, and emphasizes the need to prioritize areas to ensure a digitally skilled population in South Africa.

The CEO of NEMISA closed the conference and expressed his thanks for the rich engagement that had happened at the conference from the postgraduate symposium right to the final session, the panel on the Digital Skills Framework. Digital transformation when done well is a benefit for society, and that we cannot forget mental wellness and wholeness. The human element cannot be ignored and design thinking needs to be a part of efforts going forward. We need a post-colloquium review session so that we determine what next after the conference.

Universities represented at the Colloquium.

1. Durban University of Technology, South Africa
2. Ghent University, Belgium
3. Malawi University of Science and Technology, Malawi
4. Sol Plaatje University, South Africa
5. University of Cape Town, South Africa
6. University of Johannesburg, South Africa

7. University of Pretoria, South Africa
8. University of South Africa, South Africa
9. University of the Western Cape, South Africa
10. University of the Witwatersrand, South Africa
11. Walter Sisulu University, South Africa
12. Zimbabwe National Defense University, Zimbabwe

Summary of the accepted research papers

Thato Ditsele, Wallace Chigona, and Malebo Sephodi, investigated the sentiment of South African Twitter users towards COVID-19 vaccines. The study provides valuable insights for policymakers and healthcare organizations to shape effective strategies for promoting vaccine adoption.

Papama Mtambeka, Chimwemwe Queen Mtegha, Wallace Chigona, and Teofelus Tonateni Tuyeni, delves into the factors that affect students' compliance with universities' cybersecurity measures. The study provides insights into how universities can better protect their critical infrastructures and students against cyberattacks.

Prince Zaqueu and Tendani Mawela's identified the contributing factors for successful cybersecurity awareness, education, and training programs. The study offers several recommendations towards effective cybersecurity awareness, education, and training.

Vengai Musanga and Colin Chibaya proposed a machine learning model to forecast employee churn in organizations. Employee churn can be harmful to the quality of services, productivity, and customer loyalty. Therefore, retaining valuable employees is crucial for organizations. The authors used feature selection methods combined with strong classification models to predict employee churn. The results revealed that random forest is the most accurate in predicting employee churn.

Siphiwe Mndebele and Thembekile Mayayise presented a systematic literature review that investigated the challenges and impacts of implementing machine learning in the financial services sector. The authors uses three databases to search for relevant sources and conducted a thematic analysis. The review shows that more complex models are implemented in all the identified financial services sectors, followed by support vector machines. The paper concludes that data quality is crucial for predicting performance, efficiency, and accuracy of the model.

Marungwane Leshego Mogale evaluated the impact of organizational resources and big data analytics on the business performance of South African e-commerce SMMEs. The author used a systematic approach to literature and identifies key organizational resources that enable the use of big data analytics. These resources include IT infrastructure, IT human resources, financial resources, risk-taking, innovativeness, and proactiveness. The paper develops a conceptual framework that may be tested in future research using empirical data.

Nixon Muganda Ochara analyzed how the emerging digitalization issues might be philosophically understood from a systems viewpoint. The paper identified five systemic digitalization challenges, including the circular economy, cyberphysical systems, sharing economy, digital transformation, and smart systems. The author used five systems metaphors to investigate the challenges.

Priscilla Maliwichi, Wallace Chigona, Address Malata, and Karen Sowon explored the challenge of low mobile phone ownership among women in poor-resource settings. The authors investigated how maternal healthcare clients who do not own mobile phones negotiate access to mobile phones for maternal healthcare in rural Malawi.

Nkosikhona Msweli focused on the competencies required to teach data science in a higher learning institution. She examined instructors' perceptions of their skills and competencies in teaching data science, which is a developing topic.

Elias Tabane investigated the use of dimensionality reduction techniques in machine learning to diagnose heart diseases in South Africa. The paper provides an overview of the different techniques and highlights the importance of considering the interpretability of the results and potential biases in the data and algorithms.

Sibukele Gumbo and Hossana Twinomurinzi explored the challenge of improving the completion rate of self-paced online learning courses, particularly in the context of Fourth Industrial Revolution (4IR) skills development in South Africa. The authors reported on the strategy implemented to improve the course completion rate of a self-selected sample of students who attended face-to-face data science introduction workshops.

Julia Keddie, Renette Blignaut, Fallo Kanye, Lieven De Marez, and Simon Perneel, investigated the use of the mobileDNA application to explore location information. The authors examined the behavior of mobileDNA users in terms of where, when, and how they utilized their smartphones daily.

Research in Progress, Short Papers and Abstracts

A framework for South African university students' online learning: Social presence, digital skills, and competencies

Ntombizethu Lubisi and Samwel Mwapwele

Over the years, institutions of higher learning across the world have embraced the use of digital technology to facilitate learning. University students require digital skills and digital competencies to take full advantage of online learning. Additionally, one of the most important factors in students' learning experience in an online environment is the sense of belonging. Students engaging in online learning, geographically separated, often feel isolated from their peers and instructors due to the lack of social presence. The purpose of the study was to explain the necessary digital skills, digital competencies, and social presence for effective online learning in a South African university. The study used the extended general technology competency and use (GTCU) framework which is combined with the social presence theory as a lens to study South African university online learning. A case study approach was used involving a highly ranked university in South Africa. We followed a mixed-method approach. Data was collected via an online questionnaire with 127 respondents and semi-structured interviews with seven participants. Quantitative data were analysed using descriptive statistics and the qualitative data was analysed using thematic analysis. Findings from the study indicate that online communication, social context, and interactivity were a challenge to students we collected data from when engaging in online learning. The interaction was a challenge, and participants felt isolated from their instructors which impacted their online learning experience. Students did not feel a sense of belonging to their courses. The study contributes to the theory by extending the GTCU framework. The study con-tributes to policies such as the South African National Development Plan (NDP) 2030 with a focus on lifelong learn-ing, and quality education from the United Nations (UN) Sustainable Development Goal 4. It also contributes to the ICT and education body of knowledge.

Exploring critical factors for effective digital skills delivery towards meaningful outcomes in an online-first South African context

Natasha Katunga, Carlynn Pokpas, Leona Craffert and Rorisang Molukanele

As South Africa embraces the Fourth Industrial Revolution (4IR) and pursues its digital transformation agenda (DCDT, 2020), there has been a significant focus on the digital skills development of citizens, ranging from basic digital literacy to more specialised skills in areas like robotics, data science and artificial intelligence. In this light, concerted efforts have been made to realise approaches and methods of skills delivery that are both feasible and effective in the given context – specifically in reaching the most vulnerable, under-resourced and digitally excluded communities.

Approaches towards digital skilling interventions across the country have typically been face-to-face, however, there has been considerable emphasis over the past years to shift gears. Online methods of teaching and learning have become central to the South African government's digital agenda. The COVID-19 pandemic unquestionably intensified and accelerated efforts in this regard. At a national level, research suggests that "during the lockdown, various 4IR tools were unleashed for primary education to higher and tertiary education where educational activities switched to remote learning (online learning) … [and] that South Africa generally has some pockets of excellence to drive the education sector into the 4IR, which has the potential to increase education access" (Mhlanga & Moloi, 2020, p. 1). Similarly, concerted shifts towards online learning approaches may potentially have significantly beneficial implications in enhancing the digital skills of South African citizens.

However, online learning certainly does not come without distinct challenges. The absence of social interaction and personal support from a teacher or peer has had negative consequences in cases where many require human intervention in learning (Haché, 2011). Factors like loneliness and lack of motivation (Precel, Eshet-Alkalai & Alberton, 2009), as well as poor language fluency and miscommunication, may negatively impact online learning attempts. More pressing concerns, however, are the individual's unfamiliarity with the technology required for online learning and the lack of cognitive skills which are required to engage effectively with (particularly online) technology (Precel et al., 2009). While technical difficulties and errors (for example, programs crashing) are to be expected in the online era, the far more significant dilemma is presented in the poor access to digital devices and the Internet – made glaring across the world during the current pandemic and particularly damaging and restricting to learning activities in developing regions.

Consequently, there is a need for education and training institutions across all sectors to investigate and devise the most effective teaching and learning methods (ITU et al., 2020), particularly considering the shift towards online learning. In this regard, a mixed-method case study guided by monitoring and evaluation techniques was conducted to explore critical factors which would enable digital skills training providers and key roleplayers to deliver digital skills most effectively for meaningful outcomes in an online-first South African context. The investigation focused on four digital skills development facilitators operating in under-resourced communities in the Western Cape province. These community-based organisations provide formal and informal digital skills learning opportunities to empower citizens and make them more employable, more entrepreneurial, and capable of generating income. The investigation included the perspectives of both the organisations and training beneficiaries.

The study results shed light on various key factors to be considered in the implementation of online digital skills training interventions in South Africa, including the digital inclusion profile of training beneficiaries and their preferences in terms of appropriate training approaches, and training facilitators' experience and perspectives on effective intervention delivery given the socioeconomic realities faced in their communities. Critical success factors in delivering digital skills training interventions effectively for meaningful outcomes in an online-first South African context, therefore, include adequate infrastructure and digital technologies, an ecosystem of broad and diverse networks and a nuanced people-centric teaching approach.

References

DCDT. (2020) National Digital and Future Skills Strategy (Issue 43730).

Haché, A. (2011) Under the radar: The contribution of civil society and third sector organisations to e-Inclusion. Seville, Spain: European Commission.

ITU, UNESCO, & UNICEF. (2020) The Digital Transformation of Education: Connecting Schools, empowering learners (Issue September).

Mhlanga, D., & Moloi, T. (2020). COVID-19 and the digital transformation of education: What are we learning on 4ir in South Africa? Education Sciences, 10(7), 1–11. doi.org/10.3390/educsci10070180.

Precel, K., Eshet-Alkalai, Y. and Alberton, Y. (2009) Pedagogical and design aspects of a blended learning course. The International Review of Research in Open and Distance Learning, 10(2), pp.1-16.

The Future of Work in the Broadcasting Sector: Preliminary Results

Hossana Twinomurinzi, Naomi Isabirye and Trevor Rammitlwa

The abstract shares preliminary findings from data about job susceptibility in the Broadcasting sector using a model measuring political, economic, social and technological factors. Five groups were measured: content creators (55), human resource practitioners (21), management (23), presenters and anchors (55), and technical personnel (52). It was found that overall, all the categories experience uncertainty about the availability of jobs in the broadcasting market, and that competition for jobs has increased. More specifically, content creators, presenters and anchors, and technical personnel feel a sense of job insecurity. HR practitioners are most cognisant though that jobs in the broadcasting sector are at risk, there is a new emergence of new forms of work in the broadcasting sector despite the sector providing fair wages for work done.

How Institutional Factors in Government Agencies Affect Developing Countries in the Realisation of National Cybersecurity Capacity Building Outcomes: A Case of Malawi

Chimwemwe Queen Mtegha

National cybersecurity capacity building is essential in expanding the countries' capabilities to fight against cybercrimes. Government agencies play a pivotal role in national cybersecurity capacity building as these are tasked with implementing the initiatives to build a country's cyber resilience. One of the key strategies countries use to develop national cybersecurity capacity is the National Cybersecurity Capacity Building Frameworks (NCCBF). These NCCBFs enable myriad opportunities for government agencies to realise national cybersecurity capacity building outcomes through the development of national cybersecurity policies, cybersecurity education and training, regulations, and incidence response mechanisms. In developing countries, the outcomes that have been realised from these NCCBFs have not been satisfactory. Studies have shown that the potential impact of these NCCBF to realise the cybersecurity capacity building outcomes are limited by the institutional context in which these interventions are being used. Research has seldom been done to explain how government agencies use these NCCBFs to generate the outcomes and how institutional factors affect the realisation of these national cybersecurity capacity building outcomes.

Evolving a framework to observe and analyse customer experience on the Twitter platform using Machine Learning Techniques

Thaneshni Moodley, Surendra Thakur and Alveen Singh

A customer's entire online experience with a brand is referred to as their "digital customer experience. The company website can serve as the initial touchpoint, but it may also include mobile apps, chatbots, social media, and other channels where the touchpoint is virtual (Pillarisetty and Mishra 2022; Roggeveen and Rosengren 2022). Some retailers have seen a decline in customer base as consumers, forced migrated to online shopping habits and modes of transacting (Stiegler and Bouchard 2020). Retail services comprise hardware, apparel, appliances, electronics, books, presents, and second-hand products, as well as incidental repairs, and involves the use of a building or portion of a building by persons engaged in the selling of goods to clients (Lømo and Ulsaker 2021).

It is clear that the 2020 Covid-19 pandemic – has forced a transformation retailers (Ogunlela and Tengeh 2020). Within months, a revolution has taken place, constituting major changes to how consumers view cash, how consumers shop online, what consumers expect from retailers, and which things consumers value as part of a positive buying experience (Rukuni and Maziriri 2020).

Consumers increasingly expected retailers to create seamless customer experiences (Mhlanga and Moloi 2020). This often meant leaning on digital capabilities to create a seamless, omni-channel experience by linking different aspects of the customer shopping experience. Some retailers were better at this than others(Haapio et al. 2021). Expectation is becoming a requirement for all retailers in order to operate, with many rapidly needing to rapidly include their operations to online(Lashgari and Shahab 2022). Retailers need a way to identify retail complaints and understand them. By identifying complaints, the retailers benefits by understanding the gap in their services and improve product quality, improve customer's loyalty and improve their business online reputation(Jiang and Stylos 2021; Nuninger 2022).

Among the many popular social media platforms available in RSA, Twitter is the fifth most popular. Twitter is a microblogging and social networking website based in the United States that allows users to send and receive messages known as "tweets." Unregistered users can only read tweets that are publicly visible, while registered users can write, like, and retweet them (Erskine and Hendricks 2021). With 100 million daily active users and 500 million tweets sent daily, Twitter, a social networking site started in 2006, is one of the most prominent social media platforms available today. At the 2007 South by Southwest Interactive conference, more than 60,000 tweets were exchanged, demonstrating Twitter's phenomenal growth (Okazaki et al. 2020b). The Twitter team used the conference to start expanding their user base. The 140-character limit was originally set by cell carriers, not Twitter, because Twitter started as an SMS-based site(Gunarathne, Rui and Seidmann 2018).

However, as Twitter evolved into a web platform, the limit was maintained because it was consistent with the Twitter brand: Twitter is a platform that tries to provide short, readable material (Ibrahim and Wang 2019; Singh et al. 2021).

This research works in the intersection of “establishing an understanding of how online retailers analyse customer experiences narrated on social media” and “customer experiences expressed on the Twitter social media platform”. Given recent and rapid advancements in data science, there are several possible tools that can be crafted to allow online retailers to intercept, interpret and respond to their customers. At this point, the question is “what are the best tools?” Therefore, the aim of this study is to determine which is the best ML algorithm work best in the data analytic framework to identify retail complaints posted in Twitter.

There are many recent developments in data science, and this research focuses on machine learning (ML). Classical ML includes supervised and unsupervised learning models while newer are ensemble ML models. These models have the potential to underpin these tools that online retailers can leverage upon to better interpret customer comments and in turn to better understand online customer needs. This study focuses on experimenting with ML techniques to determine which is the best for the purpose to understand and interpret online customer comments.

The research aims to answer the best ML technique found in supervised, unsupervised, and more recently, ensemble as well as deep learning. The process of studying raw data in order to draw conclusions about it is known as data analytics. Data analytics techniques and processes have been turned into mechanical processes and algorithms that operate on raw data for human consumption (Rahim et al. 2020). A company's performance can be improved by using data analytics.

Supervised ML is contained in a controlled way to oversee the outcome accordingly and learns what the user wants to know. Unsupervised ML explores the data that is given to it, the data is unlabelled and uncategorised. Deep learning is a subfield of ML concern with algorithms inspired by the structure and function of artificial neural network. The main objective of deep learning model is to reduce the optimisation functions which could be divided based on the classification and the regression problems.

This research sets out to determine which is the most suitable ML algorithm for the purposes of sentiment analysis of online customer complaints lodged on Twitter social media platform. This research is carried out using Pycharm, Microsoft Excel and selected ML techniques. Lexicon-based approach to be used to computer sentiment polarity. A performance metrics will be used to compare each algorithm. The performance metrics will contain the statistical measures reliability, accuracy, specificity, and F – measures. The way a data mining model performs on diverse data sets is measured by its Reliability(Merten et al. 2018). A data mining model is dependable if it provides predictions that find the same basic types of patterns independent of the test data provided. The Performance Specificity assesses how well a data mining technique performs on a certain dataset. A Classifier's accuracy is calculated as a percentage of total correct predictions divided by the total number of instances (Mezghani et al. 2019). The percentage of correctly positive data divided by the sum of correctly positive and false-negative data of the marker is referred to as Sensitivity and the usefulness and relevance of the data are measured by the data quality factor (Jeong, Cho and Lee 2018). In a highly competitive market, business decisions based on wild guesses have no place. Successful business owners collect and manage a specific type of information that aids in the development of future strategies. Only in this way will be able to tailor their products and services to precisely meet the wants of their customers.

The framework will start with data extraction and will extract group detection profiling, recommendations, hashtags, retweets, mentions and locations from Twitter. The extraction will identify customer complaints. The data will go through a series of NLP pre-processing techniques which are tokenisation, stemming and lemmatisation, part-of-speech (POS) tagger, name entity recognition (NER), and parser to extract emotions for the textual data from each tweet. The cleaned data will be modelled using the data mining algorithm which are Supervised Learning Algorithms, Unsupervised Learning Algorithms and Deep Learning Algorithms.

The research method to identify the appropriate machine learning algorithm to be use for twitter data analytics that produces the best result in terms of the clarity of customer complaint utilises Pycharm. Pycharm is an integrated development environment that combines python and c#. The research would then test each machine learning algorithm for reliability, accuracy, specificity, and F – measures. This aims to answer the research question on what appropriate machine learning algorithm can be proposed for Twitter data analytics that produces the best result in terms of the clarity of customer complaints based on the performance metrics attributes.

To answer the research question, a data analytic framework was developed in Pycharm. The appropriate data was extracted from Twitter within South Africa. The data was then cleaned before it can go through sentiment analysis. Data with a negative sentiment analysis was further used to best performing data mining technique. The data mining algorithms within the appropriate data mining technique was evaluated against the performance metrics. Supervised machine learning performed the best, followed by unsupervised machine learning and there after deep learning.

The aim of the study is to develop a data analytics framework to enhance the quality of data of customer complaints on Twitter. This research contributes to the current body of knowledge in using machine learning and data mining techniques to extract Twitter data to identify customer complaints more accurately. This is the first study that introduces data analytics to retail in South Africa.

References

Bao, Y., Quan, C., Wang, L. and Ren, F. 2017. The role of preprocessing in twitter sentiment analysis. In: Proceedings of International conference on intelligent computing . 2020. Springer, 615 624.

Caldwella, B., Allcockb , G., Waltona , S. and Kendrickb , N. 2020. Assessment of use of preprint articles in reporting news in the COVID 19 era. Innovation

Gunarathne, P., Rui, H. and Seidmann , A. 2017. Whose and what social media complaints have happier resolutions? Evidence from Twitter. Journal of Management Information Systems , 34 (2): 314 340.

Kundu, S. G., Ghosh, A., Kundu, A. and GP, G. 2022. A Ml Ai Enabled Ensemble Model For Predicting Agricultural Yield. Cogent Food & Agriculture , 8 (1): 2085717

Lohiniva, L., Nurzhynska , A., Hudi , h. and Anim , B. 2022. Infodemic Management Using Digital Information and Knowledge Cocreation to Address COVID 19 Vaccine Hesitancy: Case Study From Ghana. Jmir Infodemiology , 2 (2): e37134.

Okazaki, S., Díaz Martín, A. M., Rozano , M. and Menéndez Benito, H. D. 2020a. Using Twitter to engage with customers: a data mining approach. Internet Research

Zucco, C., Calabrese, B., Agapito , G., Guzzi , P. H. and Cannataro , M. 2020. Sentiment analysis for mining texts and social networks data: Methods and tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10 (1): e1333.

ICT SMMES in South African Communities: Current Impact and Future Advancement through 4IR

Annelie Jordaan and Rorisang Molukanele

The literature review focuses on ICT SMMEs in South Africa—especially in previously disadvantaged/unreached and marginalised communities—in the context of 4IR. The review covers the global emergence of 4IR, the worldwide impact of 4IR on business and people, 4IR in Africa, and then funnels down to ICT and the digital divide in South Africa. The role of SMMEs is explored within the framework of the country‘s ICT SMME Strategy, which forms the foundation for discussing ICT SMME training interventions in South

Africa. The role NEMISA plays in collaboration with the Northern Cape & Southern Gauteng e-Skills CoLab (amongst other CoLabs) to scale digital skills (e-skills) throughout South Africa, is explained. Specific SMME training interventions implement by the Northern Cape & Southern Gauteng e-Skills CoLab (or VUT CoLab in short) and its service providers are discussed extensively.

The Effects of Dark Patterns on Internet Addiction and Lifelong Learning of Adolescents

Naomi Isabirye

In South Africa, many studies on youth online behavior have emphasised raising awareness and educating youth, parents and educators about the general dangers of the Internet. However, there is a gap in studies that aim to understand the risks that specifically exist in the day-to-day use of the Internet amongst South African learners. While there is little research specifically on Internet addiction in South Africa, related studies suggest that there is a need to examine digital health and wellbeing amongst children and youth in South Africa. Despite the growing awareness of addictive digital experiences and dark patterns, there remains a need to promote access and usage of the Internet to youth to support learning. This paper explores the risks of addictive digital experiences and their effect on lifelong learning for adolescents. The paper takes the form of a narrative literature review and makes recommendations for policy makers in education.

Factors Influencing Governments' Commitment Towards Cybersecurity in Africa: A case of Namibia

Teofelus Tonateni Tuyeni

The increasing reliance on Information Communication and Technologies (ICTs) has seen exponential growth in cybersecurity incidents. This has caused substantial financial losses and disruptions to critical economic activities. Governments worldwide have a critical role in preventing and mitigating cybersecurity incidents. Prioritising cybersecurity can lead to a secure cyber environment and the attainment of the United Nation's

Sustainable Development Goals (SDGs). This study seeks to identify and explain the factors influencing the government's commitment towards cybersecurity in Africa by answering the following research question: What factors influence government's commitment towards cybersecurity?

The study will adopt an interpretivism philosophical stance. A conceptual model will be applied to provide a lens to assess factors influencing the government's commitment towards cybersecurity. This qualitative study uses the case of Namibia. Data will be collected through semi-structured interviews, online questionnaires, and document reviews. A purposive sampling method will be used to select participants. Data will be analysed using NVivo and SPSS. Thematic and descriptive analysis techniques will be applied to identify, analyse, and report patterns within data. The study is expected to contribute new knowledge by offering a novel theoretical explanation of factors influencing the government's commitment towards cybersecurity.

Awards

Best paper

Factors Affecting how University Students Comply with Cybersecurity Measures: A Case of South Africa

Papama Mtambeka	University of Cape Town
Chimwemwe Queen Mtegha	University of Cape Town
Wallace Chigona	University of Cape Town
Teofelus Tonateni Tuyeni	University of Cape Town

Runner up: AI & Data Science track

A Machine Learning Model to Forecast Employee Churn for HR Analytics

Vengai Musanga	Zimbabwe National Defense University
Colin Chibaya	Sol Plaatje University

Runner up: Information Systems track

The issues, challenges and impacts of implementing machine learning in the financial services sector: An outcome of a systematic literature review

Siphiwe Mndebele	University of the Witwatersrand
Thembekile Mayayise	University of the Witwatersrand

Best short paper

mobileDNA application used to explore location information

Julia Keddie	University of Western Cape
Renette Blignaut	University of Western Cape
Fallo Kanye	University of Western Cape
Lieven De Marez	Ghent University
Simon Perneel	Ghent University

Best reviewers

Justin Olawande Daramola	Cape Peninsula University of Technology
Tendani Mawela	University of Pretoria
Priscilla Maliwichi	Malawi University of Science and Technology
Lateef Babatunde Amusa	Illorin University

Preprints of Proceedings

Factors Affecting how University Students Comply with Cybersecurity Measures: A Case of South Africa

Papama Mtambeka[1], Chimwemwe Queen Mtegha[1], Wallace Chigona[1] and Teofelus Tonateni Tuyeni[1]

[1]University of Cape Town, Rondebosch, Cape Town, South Africa

mtmpap001@myuct.ac.za, mtgchi003@myuct.ac.za, wallace.chigona@uct.ac.za, tynteo002@myuct.ac.za

Abstract

Universities across the globe are experiencing a surge of cyberattacks due to the increased usage of information communication technologies (ICTs). To counteract cyberattacks, universities have implemented cybersecurity measures to ensure that students and the universities' critical infrastructures are protected. Unfortunately, universities in developing countries continue to face increased cyberattacks despite implementing cybersecurity measures. This study explores the factors that affect students' compliance with universities' cybersecurity measures.

The study used a case of the University of Cape Town in South Africa, adopting qualitative research and an interpretive paradigm. We used a deductive approach to theory using Protection Motivation Theory (PMT) as the lens for inquiry. The sample for the study consisted of 40 participants, of which 35 were students and five were staff members of the University. The sample of the study was selected by convenience. We collected empirical data from the participants using semi-structured interviews. The data was then analysed using thematic analysis on NVivo software. The study found that students' compliance with cybersecurity measures is affected by their perceptions of the seriousness of the threats, the likelihood of the threats happening, their ability to protect themselves against threat, their belief in the effectiveness of the recommended solutions against cyber threats, and the costs associated with compliance to cybersecurity measures. When students perceive the risk as not severe enough to worry about, they do not find it necessary to comply with the University's cybersecurity measures. Similarly, when the students deem that the recommended compliance actions will not be practical or affordable, they do not adhere to the university cybersecurity measures.

Keywords: Protection Motivation Theory, University, Students, Cybersecurity Measures, Compliance

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

1 Introduction

With the rapid development and adoption of new technologies come new avenues for criminals to direct their cyberattacks as the number of potential victims increases (Gwebu et al., 2020). Cyberattack is defined as targeted attacks on computer systems with the aim of compromising the confidentiality, integrity and availability of data (Bendovschi, 2015). Therefore, organisations must implement and comply with cybersecurity measures to prevent cyberattacks (Khader et al., 2021). Cybersecurity is the protection of digital information assets from any attacks that may arise through internet usage (Von Solms & Von Solms, 2018). Universities are one of the main targets of cybersecurity threat, due to the substantial amounts of information they hold (Taha & Dahabiyeh, 2021). Research shows that between 2005 to 2021, higher education institutions experienced 1850 data breaches worldwide (Lukehart, 2022). In addition, universities use cyberspace platforms, such as student portals, and communication platforms, such as Microsoft teams, to manage various activities, such as admissions, examinations, administration, finances, and records, to facilitate the educational process (Li et al., 2019).

Despite the apparent advantages of automated processes that cyberspaces offer universities, they also pose cybersecurity threats and challenges to operations and information (Li et al., 2019). The increase in the recurrence of cyberattacks on universities' ICT infrastructures has led to the loss of sensitive information, finances, as well as social and intellectual property (Alharbi & Tassaddiq, 2021). The risk has increased in universities, especially in developing countries, for instance, South Africa, as the management of cyberspaces and resources is poor (Kabanda et al., 2018). In addition, South Africa has the third-highest cybercrimes in the world (Hubbard, 2019).

Despite being aware of the cybersecurity risk, some students do not comply with the universities' cybersecurity measures, such as policies on anti-virus, information and security passwords, internet and email use. These risks can be detrimental to a universities' information systems (Moallem, 2019). Given the rise in cyberattacks and the vulnerability of universities to cyberattacks, there is a need for research investigating the compliance behaviour of students with cybersecurity measures, as they are one of the primary users of university information systems. Unfortunately, there is still a dearth of studies on cybersecurity compliance in developing countries. Therefore, there is a need for more research on how students comply with cybersecurity. With this background, the study aims to answer the following question:

What factors affect students' compliance with Universities' cybersecurity measures?

We used the case of the University of Cape Town (UCT) to respond to the research question. We selected the University out of convenience. A deductive approach to theory was employed using protection motivation theory (PMT). We collected data through semi-structured interviews with students and staff from the department responsible for maintaining the ICT infrastructure of the University.

The study extends available knowledge in this regard, as literature is scarce on those factors that affect students' cybersecurity compliance. Further to this, the findings from the study will inform decision and policymakers in universities on how to implement cybersecurity measures to ensure compliance with the minimal cybersecurity measures.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

2 Literature Review

2.1 Cybersecurity

Cybersecurity has risen in significance in recent years due to the increased reliance on and adoption of information communication technologies (ICTs) (Alharbi & Tassaddiq, 2021). Cybersecurity threats are ubiquitous and may affect all organisations across industries, which may be costly (Al Moshaigeh et al., 2019). It is estimated that cybercrime will cost the global economy USD10.5 trillion from 2025 onward (Sausalito, 2020). The main reason for the worldwide trend of cybersecurity challenges is that most users do not follow their organisation's cybersecurity measures (Jeyaraj & Zadeh, 2020). Even though executive awareness of cybersecurity is expanding, most organisations remain inactive, whereas they would be more successful in dealing with cyberthreats if they were proactive. Personalising the risks for users would be beneficial, so that they know their susceptibility and the consequences of their non-compliance (Ergen et al., 2021).

While cybersecurity challenges are not new, the Covid-19 pandemic has significantly exacerbated these threats (Traxler et al., 2020). Covid-19 has shown a surge in cybercrime due to increased dependency on cyberspace, caused by the move towards virtual learning and working across the world (Traxler et al., 2020). In addition, the outbreak of the Covid-19 pandemic saw the emergence of more than 4,000 malicious websites within the first month of the pandemic in 2020, which is in correlation with the move towards virtual work that can be observed as people practised social distancing (Morgan, 2020).

2.2 Cybersecurity Challenges in Africa

Africa has one of the fastest internet penetration rates worldwide. This has resulted in an increasing trend of cyberattacks (Eboibi, 2020). Africa is perceived as the hub for cybercrime and cybercriminals due to the poor response to curb cybersecurity issues (Kabanda et al., 2018; Kshetri, 2019). The high rate of cybercrime on the continent has lowered Africa's GDP by more than 10%, costing the continent US$ 4.12 billion (Interpol, 2021). Online scams are the continent's most common and pressing cyber threats (Interpol, 2021). This entails the theft of personal information and banking information, which a threat actor subsequently uses to buy goods or services, siphon funds, or sell on the open market (Interpol, 2021).

Despite these developments, there is little research on cybersecurity issues in African countries (Kabanda et al., 2018). Along with these challenges, the continent faces a severe cybersecurity workforce shortage due to economic and institutional barriers (Kshetri, 2019). However, African countries have implemented cybercrime legal frameworks and policies in order to mitigate this issue, despite their obstacles, but have failed to enforce those policies (Eboibi, 2020).

In South Africa, there are some apparent signs of an effort to promote cybersecurity awareness and develop an effective cybersecurity culture (Kritzinger et al., 2017). Despite the global increase in research in cybersecurity globally, there is still a dearth of research on cybersecurity focusing on South Africa (Gwebu et al., 2020). The lack of understanding on the continent about the risks of accessing cyberspace contributes to a permissive climate for cybercrime (Serianu, 2017). Furthermore, the digital infrastructure development level in African countries directly impacts their security posture (Serianu, 2017). According to reports, cybercriminals rely on the general public's poor security practices; thus, policymakers ought to engage in public awareness campaigns due to the fact that there is substantial evidence that such programmes can effectively cut the success rate of cybercrime (Bada, Von Solms et al., 2019). White papers estimate that investing in cybersecurity awareness and training can affect user behaviour and minimise cyber-related risks by 45% to 70% (Bada, von Solms et al., 2019).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Unfortunately, current literature and reports show a bleak picture of the increase in cybercrime in Africa, attributed to the low ICT literacy levels that may hinder cybersecurity awareness efforts (Bada, Von Solms et al., 2019).

2.3 Cybersecurity Vulnerabilities in Universities

Universities are, by nature, open with a dense population and private data, which means that they attract a substantial number of cyberattacks due to a large amount of cyberspace usage (Yusif & Hafeez Baig, 2021). Some of the capabilities universities hold in this age of e-learning include online teaching and learning software, digital libraries, free Wi-Fi, and so on, which increases exposure to cybercrime susceptibility (Ajaero, 2020). Furthermore, universities position themselves at the forefront of technological innovation, opening them up to more vulnerabilities to increased security attacks (Yusif & Hafeez-Baig, 2021). Openness, which makes educational institutions susceptible to cyberattacks and data breaches, is a source of concern, with some scholars reporting that in excess of millions of data breaches are already being experienced by multinational organisations (Chapman, 2019). Still, the exposure of academic data is not as widely publicised (Chapman, 2019). Some scholars argue on the contrary that other industries do not report their breaches, due to a lack of investor confidence, and loss of competitiveness (Grama, 2014).

Vulnerabilities found in universities inferred from literature are classified into a few categories. The first category is administrative and cultural domains, which can clash with cybersecurity requirements. In this category, a lack of awareness and knowledge of cybersecurity best practices significantly affects the implementation of cybersecurity policies and leads to a violation of these policies (Ulven & Wangen, 2021). In addition, poor cybersecurity management can increase the vulnerability of universities to cyberattack, as they are unprepared to deal with an attack when it occurs (Nyblom et al., 2020). The second category is the technical domain, where vulnerabilities are caused due to shortfalls in the technology or systems in place (Ulven & Wangen, 2021). In this category, one aspect that increases the cybersecurity threat is the norm of students and staff bringing their own devices to work via which they connect to the network (Ulven & Wangen, 2021). The risk here is that these private computers may be used to penetrate the university networks, due to a lack of security protection systems on the devices, which increases the vulnerability to attacks on university data (Goni, 2022).

2.4 Cybersecurity Awareness and Behaviour of University Students

University students use technology and the internet for educational purposes and socialising, which became even more prevalent during the pandemic, when social distancing was encouraged (Alqahtani, 2022). As the future of the workforce, the impact of cybersecurity awareness behaviours of university students is particularly significant for society (Cheng & Wang, 2022). This makes students particularly vulnerable to cybercrime threats, as they make up most of the users of the information systems across universities (Taha & Dahabiyeh, 2021).

With the Covid-19 pandemic and online learning, university students always remain connected to the internet, and they do so by using various devices, which increase the danger of cyberattack if they do not remain vigilant about how they handle their online security (Matyokurehwa et al., 2021). Educational institutions are not taking proactive measures to raise awareness among college students about these issues and how to defend themselves from cyberattacks, such as identity theft or ransomware (Moallem, 2019). A significant risk is having a student body with an increased dependence on digital systems and is connected to the free Wi-Fi offered by the University, that is at once unaware of cybersecurity issues (Taha & Dahabiyeh, 2021). No matter the degree of technology and security the in which the institution invests, students remain the weakest link. Their lack of knowledge or ignorance makes them particularly vulnerable to targeted cyberattack (Taha & Dahabiyeh, 2021).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The trends in student awareness of cybersecurity show that students are unaware of the requisite knowledge and understanding of cybersecurity regulations and their practical application (Moallem, 2019). Another essential aspect to note is that a lack of cybersecurity awareness is not due to a lack of knowledge, but due instead to how students apply it in their daily lives (Moallem, 2019). Studies on the behaviour of students with cybersecurity reveal that engagement with cybersecurity issues was not satisfactory, where, if students were more aware of cybersecurity, some of these threats would be eliminated (Potgieter, 2019).

2.5 Human factor and compliance with cybersecurity measures

The human context is frequently regarded as the weakest link between cybersecurity and organisational information, as it is the main target for increasing cybercrimes (Chandarman & Van Niekerk, 2017). This determines the success or failure of the security chain. Scholars have identified individuals as the weakest link in the security chain as they fail to comply with cybersecurity best practices (Khader et al., 2021). According to a recent study, out of 874 cybersecurity incidents that were reported, 68% were caused internally by negligent individuals, 22% by external criminal individuals, and 10% by stolen credentials (Donalds & Osei-Bryson, 2020). The results of this study indicate that individuals' compliance with cybersecurity remains a challenge, and compliance behaviour is necessary to mitigate the risk of cyber incidents (Donalds & Osei-Bryson, 2020). Thus, cybersecurity depends not only on IT professionals, but also on educated users, who are highly aware of and employ cybersecurity best practices (Alsmadi & Zarour, 2018).

End-users of technology continue to break basic cybersecurity regulations, sustaining the cybercrime sector (Kabanda et al., 2018). User behaviour is crucial to mitigating and preventing cybersecurity issues (Taha & Dahabiyeh, 2021). When people are unaware that they are at risk, they often fail to recognise the attacks (Potgieter, 2019). Thus, students need to understand and be aware of cybersecurity threats and how to mitigate them (Potgieter, 2019).

Attacks against digital assets have not stopped and have become more varied and complicated, due to a lack of user cooperation and awareness, which causes many security approaches vulnerable to being misused or misread by users (Yusif & Hafeez-Baig, 2021). Compliance is based on the human component. This entails adhering to specified guidelines that aid in fulfilling predetermined objectives. Because cybersecurity regulations are viewed as guidelines, rather than as rules, the role of the human component in the bulk of cyberattacks or data breaches is emphasised (Yusif & Hafeez-Baig, 2021). Increased information availability has significant positive effects, but when it comes to changing human behaviour, merely presenting the information does not nearly have as much of an impact (Bada, Sasse, et al., 2019).

3 Theoretical Framework

Protection motivation theory (PMT) provides the theoretical framework for this study. PMT focuses on evaluating human behaviour regarding their motivation to respond to threats. The theory has been used in various studies to investigate individuals' protection behaviours.

3.1 Justification for selection of theory

The theory postulates that the past behaviours of an individual affect how they assess threats and their ability to handle them (Vance et al., 2012). Most cyberattacks are attributed to an inadequate level of user cooperation and knowledge. Thus, emphasis on the role of the human factor is being placed

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

(Van Niekerk & Von Solms, 2010). As a result, PMT was deemed appropriate to investigate the factors that affect the compliance behaviour of university students with cybersecurity measures.

3.2 Protection Motivation Theory (PMT)

PMT has been used in several studies as a tool to understand the motivations for individuals to comply with cybersecurity-related behaviours (Yusif & Hafeez-Baig, 2021). PMT predicts adopting or non-adoption compliance behaviours with cybersecurity (Ezati Rad et al., 2021). The theory has two primary constructs: threat appraisal, and coping appraisal. These two constructs are integrated to develop protection motivation (Ezati Rad et al., 2021). These constructs describe how individuals assess the level of risk they encounter in cyberspace and act as a protective measure (Yusif & Hafeez Baig, 2021). Figure 2. illustrates the conceptual model of the study.

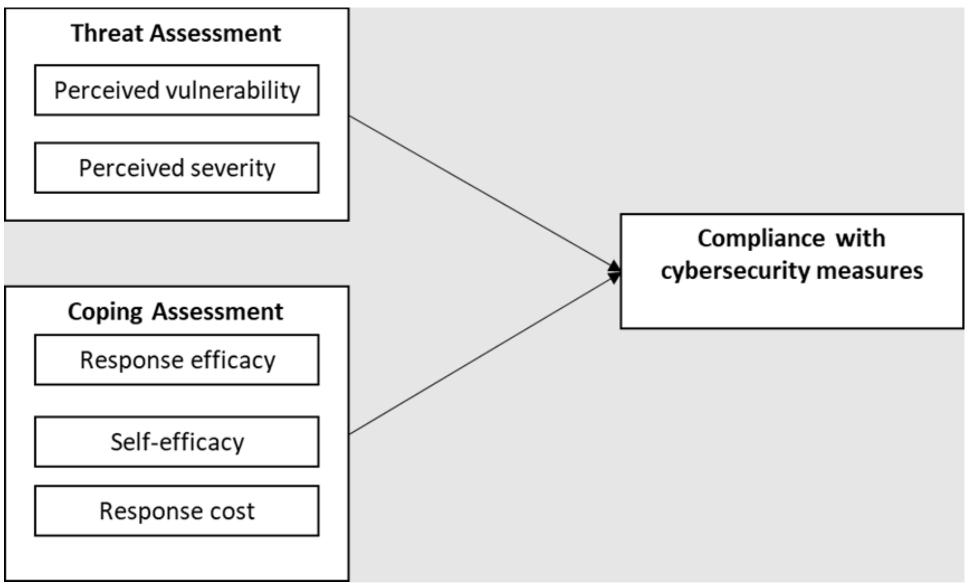


Figure 1: Conceptual model (Adopted from the Protection Motivation Theory (Rogers, 1975))

Threat Assessment

The threat assessment construct focuses on the severity and the evaluation of the threat. It consists of two main components: perceived vulnerability, and perceived severity. *Perceived vulnerability* refers to the individual's appraisal of the likelihood of being exposed to a threat (Rogers, 1975). The theory states that an action taken by an individual to overcome fear is influenced by the probability of its occurrence (Ezati Rad et al., 2021; Towbin, 2019). Therefore, if individuals perceive a low likelihood of the threat occurring, they are less likely to take action.

Perceived severity refers to the extent of the consequences from the threat if it occurs (Ezati Rad et al., 2021; Rogers, 1975; Towbin, 2019). The likelihood that the harm from the threat will have serious repercussions will compel an individual to act against the threat.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Coping Assessment

Coping assessment refers to a person's appraisal of their ability to respond to a threat (Crossler et al., 2019). This construct consists of three components: response efficacy, self-efficacy, and response cost. *Response efficacy* refers to the individual's belief that the recommended action is effective in mitigating the threat (Crossler et al., 2019; Tsai et al., 2016). Self-efficacy refers to an individual's belief in their ability to carry out the recommended action (Crossler et al., 2019). *Response cost* refers to the perceived costs related to conducting the recommended action against a threat or engaging in security behaviours (Crossler et al., 2019; Towbin, 2019), for instance, the costs of using Antispyware.

Overall, previous studies employing PMT show that self-efficacy and response efficacy have a significant relationship with the behavioural intentions of individuals in compliance with cybersecurity (Miraja et al., 2019). The constructs of the PMT used in this study are operationalised in Table 1.

Construct	Definition	Source
Perceived Vulnerability	The extent to which students think they are vulnerable to cyber threats.	Rogers, 1975; Yusif & Hafeez Baig, 2021
Perceived Severity	The degree to which the students perceive the seriousness of the threat/risk if it was to occur.	Rogers, 1975
Response efficacy	The degree to which students believe the recommended cybersecurity actions would help them avoid the threat.	Rogers, 1975; Yusif & Hafeez-Baig, 2021; Crossler et al., 2019
Self-efficacy	The extent to which students believe they can successfully perform the recommended tasks.	Tsai et al., 2016; Crossler et al., 2019
Response Cost	The costs of complying to cybersecurity measures set by the University, and the cost of non-compliance.	Towbin, 2019; Crossler et al., 2019

Table 1: PMT constructs

4 Context of the study

The University of Cape Town (UCT) is one of the top universities in Africa and the world. It is located in the Western Cape province of South Africa (UCT News, 2022). In 2021, it was estimated that the University enrolled over 30 000 students (UCT News, 2022). UCT boasts several advanced technologies to facilitate student learning. To ensure compliance with international ICT standards and guidelines, a secure cyber environment, and the protection of critical information, the University has developed various ICT policies such as an information security policy, account and password policy,

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

and anti-virus policy (UCT ICTS, 2022). The University also offers ICT services to support academic staff, students, and management on ICT-related matters.

Compliance with ICT policies has always been challenging for many organisations worldwide, including universities. Scholars have pointed to a lack of motivation, awareness, belief, and behaviour as the main contributing human factors in non-compliance (Alqahtani & Braun, 2021; Da Veiga et al., 2020).

5 Research Methodology

We followed an interpretive philosophy to conduct this research. The interpretive stance argues that truth and knowledge are subjective; thus, by adopting an interpretive philosophy, the researcher will be able to understand and interpret the experiences of respondents/subjects of the research (Kivunja & Kuyini, 2017). Adopting an interpretive approach is appropriate for this study, as it allows for a rich understanding of the phenomena and, in this case, an understanding of the factors affecting university students' compliance with cybersecurity measures (Orlikowski & Baroudi, 1991).

The study adopted a deductive approach. In this case, a conceptual model was developed using constructs from the PMT. We utilised a qualitative research approach by employing a case study of UCT. Data was collected using semi-structured interviews and observations of students' engagement with cybersecurity measures at the University. Semi-structured interviews allowed us to interact and fully engage with participants by asking follow-up questions.

The target populations were students and ICT staff at UCT. The students consisted of undergraduate students in different academic years across the University's six faculties. We chose this sample because undergraduate students make up most of the student population, constituting the University's largest users of information systems (UCT News, 2022). In 2021, UCT had 30 329 registered students, with 18 154 undergraduates, where undergraduate students make up more than 59% of the student population.
(UCT News, 2022). This makes undergraduate students the weakest link in the University's cybersecurity chain (Chandarman & Van Niekerk, 2017). We included staff members from ICTS to gain their perspectives as they are part of developing cybersecurity-related policies and interact with students on cybersecurity-related matters such as compliance with the University's ICT policies.

The study employed purposive sampling, a sampling technique where the researcher chooses participants based on personal judgement, and convenience sampling, which are selected participants at the researcher's convenience. The study also incorporates snowball sampling, where the initially chosen participants recruited or recommended more participants to participate. The purposive sampling technique was selected because the researchers were interested in university undergraduate students. Conversely, the convenience sampling method was applied because the researchers could conveniently access the sample. Therefore, the snowball sampling technique was used as purposive sampling could not yield the desired results.

A total of 40 responses were collected, of which 35 were students, while five respondents were staff members from ICTs. The students were coded as Respondent_student_X and the staff Respondent_staff_X. Table 2 summarises the demographic information of the respondents.

Respondents Positions:		
	Undergraduate students	35
	ICT staff	5
Age Range:		
	46 years above	2

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

	36-45 years	3
	26-35 years	11
	16-25 years	24
Faculty Departments:		
	Faculty of Commerce	13
	Faculty of Humanities	8
	Faculty of Science	6
	Graduate School of Business	3
	Faculty of Law	2
	Faculty of Engineering and Built Environment	3

Table 2: Demographic Information of Respondents

The interviews were recorded and later transcribed for analysis. The interviews lasted between eight and 20 minutes. Data was analysed using thematic analysis and Nvivo software. Collected data was organised and categorised according to themes as per the constructs. The research was conducted in line with the ethical standards of the University of Cape Town. We obtained ethics approval for the study from the University prior to the commencement of data collection.

6 Empirical analysis and Discussion

This section uses PMT to understand the underlying factors affecting students' compliance with cybersecurity measures. The findings and the discussions are further explained in the subsequent sections. Participant responses are cited verbatim.

6.1 Threat Assessment

We were set to understand how threatened the respondents felt about the possibility that they could encounter cyberattacks. In addition, it assisted in determining those factors would likely affect the respondents' decision to not comply with cybersecurity measures set by the University.

Perceived vulnerability

The majority of students did not think they were likely to be targeted by cybercriminals. The reason is that the students believed that they did not have any valuable information for cybercriminals to target.

"I am a student, and personally, I feel I do not have any valuable information that would be useful for the attackers. I think the attackers go for top government officials or high-profile people who may have confidential information. They easily target those because they can demand ransom." [Respondent_Student_4]

Furthermore, the students did not perceive that they were vulnerable, since they were protected by ICT protection measures put in place by the institution. The trust in the university's ICT service was consistent among the students, who believed that the likelihood of being a victim of cyberattacks was

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

low, because ICT service at their University offered them protection. As a consequence, they felt safe even if they did not take additional measures.

"I strongly believe that the University has built adequate infrastructure to protect the students and the staff. So, I believe our University is safe from cyberattacks." [Respondent_student_7]

"As a student, I am less likely to be a victim. Because we have a lot of software, and we have ICT services to help us with ensuring that there's protection from being attacked." [Respondent_student_3]

This finding reflects the broader literature, as many individuals rely on third-party protective measures. They do not believe that the responsibility lies with them (Interpol, 2021; Murphy et al., 2022).

In summary, students did not comply with the cybersecurity measures set out by University's ICT service department because they were not likely to be victims of cyberattacks, due to not owning valuable information and the protection given to them by ICT service at the University. However, when an individual believes that a threat of cybercrime is likely to occur, they are more motivated to comply with cybersecurity measures (Tsai et al., 2016).

Perceived severity of the threat

We asked the respondents to elaborate on the perceived severity if they were victims of cyberattacks. The respondents indicated that they did not perceive the true severity of the security predicament, because they had never encountered any cyberattacks.

"Since I started my studies at this University, I have never encountered any cyberthreats. So, when I browse on my device, I can download anything and do whatever pleases me." [Respondent_student_19]

The students' perception affected their compliance with the cybersecurity measures. These findings are supported by literature, which states that when individuals are unaware of the threats that they face, they are more likely to engage in unsafe behaviour, without taking the necessary protective measures (Potgieter, 2019).

The responses from the students corroborated those of UCT's staff members. The staff members indicated that even though the University's ICT service invited students to cybersecurity awareness events, most students did not attend them.

"Even if we have cybersecurity awareness sessions, students will rarely attend. Students will only really take it seriously once they are compromised. Otherwise, they don't really care until it happens to them; that's when they wake up and realise how important it is." [Respondent_staff_2]

The non-compliance of students with the basic cybersecurity measures set out by University's ICT service can be linked to them not feeling threatened by the cyberattack. As such, most of these students did not implement any cybersecurity measures, for instance, an anti-virus mechanism. The ICT anti-virus policy of the University stipulates that all computers in the university environment need to have an anti-virus mechanism installed to protect the computers from viruses and other malicious code. However, when we enquired from the staff members as to whether they had implemented incentives to promote cybersecurity compliance in the University, they did not have any mechanisms in place.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

"I honestly don't remember installing any anti-virus. The only cybersecurity measure I have installed is the authenticator, which the University's ICT service had put mandatory measures for me to continue using the UCT applications." [Respondent_student_30]

On the other hand, although some students did not perceive the severity of cybercrimes, they took personal protective measures against cyberattacks. The students indicated that they installed multi-step authentication and anti-virus applications on their devices. The students emphasised that they found the severity of the threat of becoming a victim of identity theft high, in response to which they took precautions. The students related their protective actions to the fear of their devices being compromised.

"I'm paranoid of identity theft. So, I make sure I install and update my anti-virus regularly. Everything associated with my personal information, I protect at any cost." [Respondent_student_14]

This finding is supported by the literature, which states that when individuals feel threatened by cybercrime, they are more motivated to comply with cybersecurity measures to remove the threat (Towbin, 2019).

6.2 Coping Assessment

Response Efficacy

Most of the students in the study knew what they had to do to protect themselves. When asked about which protective measures they thought students ought to take, their responses were similar as they based these on the measures set by University's ICT service. The respondents believed that taking these actions would help them avoid the threat of cyberattacks.

"make sure that your anti-virus software is always up to date? Make sure you go there and try CTS and have it installed if you don't understand. Make sure that you don't visit any of those illegal sites for movies for series. Rather install Netflix if you can't afford one, then I don't know man, but just don't visit those illegal movie sides and follow one to make a feature." [Respondent_student_23].

On the other hand, although some students knew about the various protective measures, they did not believe they were necessary to implement them.

"I don't think you need some anti-virus or ever. I think what's on your computer already is enough." [Respondent_student_5].

An individual's belief that the recommended protective action would help them avoid the threat motivates them to take action (Yusif & Hafeez-Baig, 2021). However, when an individual believes that installing anti-virus software on the device does not offer them any protection, they will not be motivated to install it, even if that is the recommended action to take. This matches what was found in the research, as students responded that they did not think installing anti-virus software would be necessary, so they did not install it, even though that does not comply with University's ICTs anti-virus policy.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Self-efficacy

The students had lower self-efficacy in terms of their ability to protect themselves. They believed that they had to have the technical skills to be able to protect themselves. This could have made them rely on ICT services in the institution. Even when they were aware of the risks, they did not initiate and were lenient in their protective measures.

> *"So students are normally very lenient, in the sense that they don't take note much of things. So there are plenty of security measures and tips to prevent cyber tech from happening, but students tend to ignore them. Students don't like reading notifications and things like that."* [Respondent_staff_24]

The response from the staff member showed that, even though students had the knowledge and awareness of cybersecurity, they were still lenient in complying with cybersecurity measures, as they did not believe they were responsible for their safety on the internet.

The respondent's lack of ability to protect themselves could have been attributed to the lack of engagement with cybersecurity issues, for instance, cybersecurity awareness. The study revealed that cybersecurity awareness is crucial to students' compliance with cybersecurity measures.

Many students had little to no engagement with cybersecurity as they had no knowledge of it. However, responses from students also showed that few amongst those students who were more interested in cybersecurity were in fact actively engaged with it. The students stated that they conducted research and read emails sent by University's ICT service. This made them more aware of the risks and motivated them to protect themselves against potential cyberattacks. This finding indicated that cybersecurity awareness was largely missing in the institution, which counteracted the responses from the staff members.

> *"The only cybersecurity awareness event I attended was during my first year. The UCT department gave a security talk to the students who were on financial aid. I believe they are more focused on those students because the students are given university laptops."* [Respondent_student_9]

Response Cost

The cost of complying was one of primary reasons why students did not comply with cybersecurity measures. Students cited that taking protective measures takes time, effort, and financial resources. The findings show that the higher the cost of compliance, the lower the motivation to comply with cybersecurity measures.

> *"Honestly, for you to install the software and also make sure they are updated, costs a lot of time and effort. In addition, I need to always have WIFI available to ensure that I update my devices."* [Respondent_student_8]

When the cost of compliance is higher than they are willing and able to pay, an individual is less motivated to comply (Alqahtani, 2022; Alqahtani & Braun, 2021). The finding is consistent with the existing literature.

When it comes to the cost of non-compliance, the students felt that in the case of an attack they would likely lose some information. However, they believed that they did not harbour the kind of information that thieves would likely be interested to steal. Consequently, the respondents believed that the cost of non-compliance was low.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

"Personally, nothing. When I think about it, I only have pictures, So I don't even mind in my opinion. Okay." [Respondent_student_11]

Furthermore, the study found that since the University did not have mechanisms to penalise students for non-compliance with cybersecurity measures, this reduced the perceived cost of non-compliance. A high cost for non-compliance may nudge individuals towards it (Towbin, 2019).

7 Conclusion and Recommendations

The study sought to understand the factors affecting university students in South Africa in complying with cybersecurity measures. The research adds to a body of literature to understand the underlying factors affecting compliance by students in developing countries. The study found that the perceptions of students regarding the severity of the threat, the likelihood of the threat to occur, their lack of belief in the cybersecurity measures in place and the cost associated with compliance affected cybersecurity compliance by students. In addition, a lack of cybersecurity awareness and knowledge may have contributed to compliance. Therefore, universities needed to create targeted cybersecurity awareness-raising initiatives. In addition, the University ought to implement metrics to measure the effectiveness of those cybersecurity awareness initiatives currently in place. Furthermore, universities ought to create incentives to promote cybersecurity compliance. This could be achieved through the hosting of competitions, along with sponsorships for cybersecurity seminars to motivate students' compliance and create awareness.

The sample was drawn from one University in South Africa, which may have limited the findings, and the generalisability of the study. Therefore, we recommend future research drawing from across a range of universities in South Africa.

References

Ajaero, C. C. (2020). *Behavioral characteristics computer users need to minimise ransomware exposure* (Publication No. 28149194) [Doctoral dissertation, Capella University]. ProQuest Dissertations & Theses Global.

Alharbi, T., & Tassaddiq, A. (2021). Assessment of cybersecurity awareness among students of Majmaah University. *Big Data and Cognitive Computing*, *5*(2). https://doi.org/10.3390/bdcc5020023

Al-Moshaigeh, A., Dickins, D., & Higgs, J. L. (2019). cybersecurity risks and controls. *The CPA Journal, 89*(6), 36–41.

Alqahtani, M.A. (2022). Factors Affecting Cybersecurity Awareness among University Students. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052589

Alqahtani, M.A., & Braun, R. (2021). *Examining the Impact of Technical Controls, Accountability and Monitoring towards Cyber Security Compliance in E-government Organisations*. https://doi.org/10.21203/rs.3.rs-196216/v1

Alsmadi, I., & Zarour, M. (2018, August 20). Cybersecurity Programs in Saudi Arabia: Issues and Recommendations. *1st International Conference on Computer Applications and Information Security, ICCAIS 2018*. https://doi.org/10.1109/CAIS.2018.8442013

Bada, M., Sasse, A. M., & Nurse, J. R. C. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *ArXiv Preprint ArXiv:1901.02672*.

Bada, M., von Solms, B., & Agrafiotis, I. (2019). Reviewing National Cybersecurity Awareness in Africa: An Empirical Study: an empirical study. *The Third International Conference on Cyber-Technologies and Cyber-Systems, CYBER 2018,* 78-83

Bendovschi, A. (2015). Cyber-attacks–trends, patterns and security countermeasures. *Procedia Economics and Finance*, *28,* 24-31.

Chandarman, R., & Van Niekerk, B. (2017). Students' cybersecurity awareness at a private tertiary educational institution. *The African Journal of Information and Communication*, *20*, 133–155.

Chapman, J. (2019). How safe is your data? Cyber-security in higher education (Vol. 12, pp. 1-6). Oxford, UK: Higher Education Policy Institute.

Cheng, E. C. K., & Wang, T. (2022). Institutional Strategies for Cybersecurity in Higher Education Institutions. *Information (Switzerland)*, *13*(4). https://doi.org/10.3390/info13040192

Crossler, R. E., Andoh-Baidoo, F. K., & Menard, P. (2019). Espoused cultural values as antecedents of individuals' threat and coping appraisal toward protective information technologies: Study of U.S. and Ghana. *Information and Management*, *56*(5), 754–766. https://doi.org/10.1016/j.im.2018.11.009

da Veiga, A., Astakhova, L. v., Botha, A., & Herselman, M. (2020). Defining organisational information security culture—Perspectives from academia and industry. *Computers and Security*, *92*. https://doi.org/10.1016/j.cose.2020.101713

Donalds, C., & Osei-Bryson, K. M. (2020). Cybersecurity compliance behavior: Exploring the influences of individual decision style and other antecedents. *International Journal of Information Management*, *51*. https://doi.org/10.1016/j.ijinfomgt.2019.102056

Eboibi, F. E. (2020). Concerns of cyber criminality in South Africa, Ghana, Ethiopia and Nigeria: rethinking cybercrime policy implementation and institutional accountability. *Commonwealth Law Bulletin*, *46*(1), 78–109.

Ergen, A., Ünal, A. N., & Saygili, M. S. (2021). Is it possible to change the cyber security behaviours of employees? Barriers and promoters. *Academic Journal of Interdisciplinary Studies*, *10*(4), 210–224. https://doi.org/10.36941/AJIS-2021-0111

Ezati Rad, R., Mohseni, S., Kamalzadeh Takhti, H., Hassani Azad, M., Shahabi, N., Aghamolaei, T., & Norozian, F. (2021). Application of the protection motivation theory for predicting COVID-19 preventive behaviors in Hormozgan, Iran: a cross-sectional study. *BMC Public Health*, *21*(1), 466. https://doi.org/10.1186/s12889-021-10500-w

Goni, O. (2022). Introduction to Cyber Crime. *International Journal of Engineering and Artificial Intelligence*, *3*(1), 9–23. https://doi.org/10.55923/jo.ijeal.3.1.701

Grama, J. L. (2014). *Legal Issues in Information Security: Print Bundle*. Jones & Bartlett Publishers.

Gwebu, K. L., Wang, J., & Hu, M. Y. (2020). Information security policy non-compliance: An integrative social influence model. *Information Systems Journal*, *30*(2), 220–269. https://doi.org/10.1111/isj.12257

Hathaway, O. A., Crootof, R., Levitz, P., Nix, H., Nowlan, A., Perdue, W., & Spiegel, J. (2012). The law of cyber-attack. *California Law Review*, 817-885.

Hubbard, J. (2019). SA business underplaying the danger of cybercrime. *Finweek*, *2019*(4), 37–38. www.thecyberacademy.co.za

Interpol. (2021). *African Cyberthreat Assessment Report*. Retrieved, December 19, 2022 from https://www.interpol.int/en/content/download/16759/file/AfricanCyberthreatAssessment_English.pdf

Jeyaraj, A., & Zadeh, A. (2020). Institutional Isomorphism in Organizational Cybersecurity: A Text Analytics Approach. *Journal of Organizational Computing and Electronic Commerce*, *30*(4), 361–380. https://doi.org/10.1080/10919392.2020.1776033

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Kabanda, S., Tanner, M., & Kent, C. (2018). Exploring SME cybersecurity practices in developing countries. *Journal of Organizational Computing and Electronic Commerce*, *28*(3), 269–282. https://doi.org/10.1080/10919392.2018.1484598

Khader, M., Karam, M., & Fares, H. (2021). Cybersecurity awareness framework for academia. *Information (Switzerland)*, *12*(10). https://doi.org/10.3390/info12100417

Kivunja, C., & Kuyini, A. B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, *6*(5), 26–41.

Kritzinger, E., Bada, M., & Nurse, J. R. C. (2017). A study into the cybersecurity awareness initiatives for school learners in South Africa and the UK. *IFIP World Conference on Information Security Education*, 110–120.

Kshetri, N. (2019). Cybercrime and Cybersecurity in Africa. *Journal of Global Information Technology Management*, *22*(2), 77–81. https://doi.org/10.1080/1097198X.2019.1603527

Li, L., He, W., Xu, L., Ash, I., Anwar, M., & Yuan, X. (2019). Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. *International Journal of Information Management*, *45*, 13–24. https://doi.org/10.1016/j.ijinfomgt.2018.10.017

Lukehart, A. (2022, July 5). *Top 5 Cybersecurity Threats Facing Higher Education*. Fierce Education. https://www.fierceeducation.com/technology/top-5-cybersecurity-threats-facing-highereducation

Matyokurehwa, K., Rudhumbu, N., Gombiro, C., & Chipfumbu-Kangara, C. (2022). Enhanced social engineering framework mitigating against social engineering attacks in higher education. *Security and Privacy*, *5*(5), e237.

Mlambo, C. (2021). Cybersecurity awareness in Zimbabwean universities: Perspectives from the students. *Security and Privacy*, *4*(2). https://doi.org/10.1002/spy2.141

Miraja, B. A., Persada, S. F., Prasetyo, Y. T., Belgiawan, P. F., & Redi, A. A. N. P. (2019). Applying protection motivation theory to understand Generation Z students intention to comply with educational software anti-piracy law. *International Journal of Emerging Technologies in Learning*, *14*(18), 39–52. https://doi.org/10.3991/ijet.v14i18.10973

Moallem, A. (2019). *Cybersecurity Awareness Among Students and Faculty* (1st ed.). CRC Press. https://doi.org/10.1201/97804

Morgan, H. (2020). Best Practices for Implementing Remote Learning during a Pandemic. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *93*(3), 135–141. https://doi.org/10.1080/00098655.2020.1751480

Murphy, C., Mtegha, C. Q., Chigona, W., & Tuyeni, T. T. (2022). Factors Affecting Compliance with the National Cybersecurity Policy by SMMEs in South Africa. *African Conference on Information Systems*. https://digitalcommons.kennesaw.edu/acist/2022/presentations/4

Nyblom, P., Wangen, G. B., Kianpour, M., Østby, G., & Wangen, G. (2020). The Root Causes of Compromised Accounts at the University. *ICISSP*, 540–551. https://www.researchgate.net/publication/334226938

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying Information Technology in Organisations: Research Approaches and Assumptions. *Information systems research*, *2*(1), 1-28. https://www.jstor.org/stable/23010611

Potgieter, P. (2019). The Awareness Behaviour of Students on Cyber Security Awareness by Using Social Media Platforms: A Case Study at Central University of Technology. *ICICIS*, 272–280.

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change1. *The Journal of Psychology*, *91*(1), 93–114.

Serianu. (2017). *Demystifying Africa's Cyber Security Poverty Line*. https://serianu.com/downloads/AfricaCyberSecurityReport2017.pdf

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Taha, N., & Dahabiyeh, L. (2021). College students information security awareness: a comparison between smartphones and computers. *Education and Information Technologies*, *26*(2), 1721–1736. https://doi.org/10.1007/s10639-020-10330-0

Towbin, R. S. (2019). *A protection motivation theory approach to healthcare cybersecurity: A multiple case study* [Doctoral dissertation]. Northcentral University.

Traxler, J., Smith, M., Scott, H., & Hayes, S. (2020). *Learning through the crisis: Helping decisionmakers around the world use digital technology to combat the educational challenges produced by the current COVID-19 pandemic*. https://wlv.openrepository.com/bitstream/handle/2436/623926/Traxler_et_al_Learning_through_the_crisis_2020.pdf?sequence=3

Tsai, H. Y. S., Jiang, M., Alhabash, S., Larose, R., Rifon, N. J., & Cotten, S. R. (2016). Understanding online safety behaviors: A protection motivation theory perspective. *Computers and Security*, *59*, 138–150. https://doi.org/10.1016/j.cose.2016.02.009

UCT ICTS. (2022). *Policies and guidelines*. https://www.icts.uct.ac.za/about-icts/policies-andguidelines

UCT News. (2022). *About UCT and Cape Town*. https://www.news.uct.ac.za/article/-2022-02-09about-uct-and-cape-town

Ulven, J. B., & Wangen, G. (2021). A systematic review of cybersecurity risks in higher education. In *Future Internet* (Vol. 13, Issue 2, pp. 1–40). MDPI AG. https://doi.org/10.3390/fi13020039

van Niekerk, J. F., & von Solms, R. (2010). Information security culture: A management perspective. *Computers and Security*, *29*(4), 476–486. https://doi.org/10.1016/j.cose.2009.10.005

Von Solms, B., & Von Solms, R. (2018). Cybersecurity and information security–what goes where?. *Information & Computer Security*, *26*(1), 2-9.

Vance, A., Siponen, M., & Pahnila, S. (2012). Motivating IS security compliance: Insights from Habit and Protection Motivation Theory. *Information and Management*, *49*(3–4), 190–198. https://doi.org/10.1016/j.im.2012.04.002

Yusif, S., & Hafeez-Baig, A. (2021). Cybersecurity Policy Compliance in Higher Education: A Theoretical Framework. *Journal of Applied Security Research*. https://doi.org/10.1080/19361610.2021.1989271

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

A Machine Learning Model to Forecast Employee Churn for HR Analytics

Vengai Musanga	Zimbabwe National Defense University	vengaimusanga@gmail.com
Colin Chibaya*	Sol Plaatje University, Kimberley, SA	colin.chibaya@spu.ac.za

Abstract

The challenge of employee churn is a major issue for many organizations. Unexpected employee departures can harm service delivery, the quality of services, productivity, and hurt customer loyalty. The ability to predict employee churn is crucial for retaining valuable employees. This study proposes a model that predicts employee churn by using feature selection through Pearson correlation methods, Information gain, and Recursive feature elimination, combined with strong classification methods such as random forest, logistic regression, decision trees, gradient boosting machines, and K-nearest neighbours. The IBM dataset was used for training and testing the proposed machine learning methods. The accuracy of the algorithms improved after applying particular feature selection methods. The results yielded showed that the random forest technique outperformed other models in terms of accuracy in predicting employee churn.

Introduction

The prediction of employee churn remains a major issue for most organizations globally. In this study, employee churn is defined as the loss of intellectual assets from a company [1]. Employees may decide to leave an organization for various reasons such as dissatisfaction with salary, bureaucracy in the organization, or limited career growth [2]. However, the departure of skilled employees is detrimental and can lead to decreased productivity. The need for intelligent systems that can predict employee churn is apparent [3].

Human resources are considered the most valuable assets within an organization [4] and employee motivation is crucial in determining their continued stay. Organizations should go an extra mile to retain their manpower. However, it is hard to predict an employee's plans to leave. To address the issue of employee retention, machine learning approaches can be used to analyse past employee data and identify patterns that predict the likelihood of employees leaving the organization [5]. In this study, we make a comparison to see which machine learning technique best predict employee churn.

Past research has explored the use of machine learning techniques for employee churn prediction, but the full potential of feature engineering and selection has not been fully exploited. This study aims to fill that gap by developing an integrated supervised machine learning model [6] that compares multiple feature selection and data balancing methods to improve employee churn predictive accuracy [7]. The IBM dataset was chosen for this study as it encompasses common factors found in most sectors.

The proposed machine learning model is envisaged to be able to identify the critical factors in retaining valuable employees and assist human resources managers in their staff retention efforts. Subsequent sections of the rest of this study cover related literature, a description of the data used, the methods followed in completing this study, the results, and the discussions thereto. We close the work with a conclusion which mainly presents our recommendations, contributions of the work, and the likely direction for future research.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

RELATED WORK

Organizations predict employee churn to anticipate and understand the reasons for potential loss of employees, and to take proactive measures to retain valuable staff and minimize the impact of turnover on the business. By forecasting employee churn, organizations can improve their retention strategies, reduce hiring and training costs, and maintain a stable and productive workforce. Predicting employee churn offers several advantages for organizations. Firstly, improved retention strategies can be implemented by understanding the reasons for employee churn. This information allows organizations to develop targeted retention programs that aim to reduce turnover and improve employee satisfaction. Another advantage of predicting employee churn is cost savings. Minimizing employee turnover can help to reduce the costs associated with hiring and training new staff. This is important as these costs can add up quickly, especially for organizations with high levels of turnover. A stable workforce also leads to increased productivity. When employees are able to focus on their work without disruptions caused by high levels of turnover, they are better able to perform at their best. This, in turn, can lead to improved performance and increased productivity for the organization as a whole. Effective management of employee churn can also give companies a competitive advantage. Companies that are able to attract and retain top talent are better positioned to succeed in the marketplace, as they have a stable and productive workforce. This can give them a significant advantage over competitors that struggle to manage employee churn. Finally, high levels of employee churn can negatively impact customer satisfaction. Customers may need to constantly adapt to new staff, leading to frustration and a decrease in satisfaction levels. By reducing employee churn, organizations can improve customer satisfaction and strengthen customer relationships. This can lead to increased loyalty and repeat business, which is essential for long-term success.

Several studies have investigated the ability of machine learning algorithms to forecast employee turnover [7], [8], [9]. As a starting point in predicting employee churn, Hebbar et al. [10] utilized logistic regression on IBM's employee attrition dataset to determine the likelihood of an individual being part of the churn group. This allowed the researchers to get an initial understanding of the relationships between various employee characteristics and the likelihood of them leaving the company. By applying logistic regression to this dataset, Hebbar et al. [10] were able to identify the key factors that contribute to employee churn, such as job satisfaction, job involvement, and work-life balance. This information can then be used to develop targeted retention programs and reduce employee churn, leading to cost savings and increased productivity for the organization. The results of this study demonstrated the potential of logistic regression as a powerful tool in predicting employee churn and provided insights into the factors that organizations should focus on to reduce turnover.

Subsequently, a comparative study was conducted using random forest and support vector machine models to determine the key characteristics of the IBM employee attrition dataset. The study performed exploratory data analysis to understand the relationships between various employee attributes and the likelihood of churn. During this process, different data visualization techniques were used to represent the findings, such as bar charts, histograms, and scatter plots. The aim of this comparative study was to evaluate the performance of random forest and support vector machine models in predicting employee churn and compare their results with the findings of the previous logistic regression study. The results of this study helped to determine the strengths and weaknesses of each model, and to identify the most important factors that contribute to employee churn. This information can be used by organizations to develop more effective retention strategies, reduce employee turnover, and increase productivity. This comparative study demonstrated the potential of machine learning algorithms in predicting employee churn. The study provided a comprehensive evaluation of these algorithms and yielded results that showed that both random forest and support vector machine models had strong performance in predicting employee churn and were able to accurately identify the key factors that contribute to turnover. The study also highlighted the importance of performing exploratory data analysis and visualizing the findings, as this helps to understand the relationships between employee characteristics and the likelihood of churn. However, feature selection methods were not

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

explored, a gap we explore further in this study. Hopefully, accuracy may improve with the adoption of feature selection methods.

In a study conducted by Dam [11], the author investigated the use of various feature selection techniques for determining the most significant features in predicting employee churn. The author believed that by identifying the most informative features, a more accurate and efficient model for predicting employee churn could be developed. The author's findings provided insights into the importance of feature selection in the prediction of employee churn. In his study, Dam [11] compared the benefits of three feature selection methods: wrapper, filter, and embedded. He ultimately chose Recursive Feature Elimination, a wrapper method, as his method of choice due to previous research indicating that wrapper methods are effective in identifying the most important features in datasets of medium to large size. In a study conducted by Zhao and colleagues [12], the authors demonstrated the concept of feature importance in an XGBoost model that was trained on a dataset consisting of 1,000 items. The study showed how XGBoost models can be used to determine the relative importance of different features in a dataset, which can provide valuable insights for feature selection and model building. The results of the study demonstrated the practical applications of XGBoost models in evaluating feature importance and the potential benefits of using this approach in real-world data analysis. A lot of other research showed that Tenure was also a good predictor of employee churn [13].

The effect of employee satisfaction on employee churn was once investigated using regression methods [14]. The study found that employee satisfaction is a crucial factor in employee turnover, but the lack of advanced machine learning techniques for predicting employee attrition resulted in limited accuracy in predictions. Similarly, Yigit and Sourabizadeh [15] investigated employee turnover utilizing multiple techniques with differing levels of complexity (LR, NB, KNN, DT, SVM, and RF). They performed two experiments, one that included feature selection and one without. In both experiments, they found that SVM was the most effective method for forecasting employee churn. Falluchi et al. [16] studied a broad spectrum of algorithms and found that Gaussian NB had the highest recall rate at 0.541. However, using Gaussian NB on employee attrition data poses a challenge as it assumes all predictors are independent, which may not always be the case. This, combined with the algorithm's lower performance, results in a less precise prediction of employee turnover. Among the algorithms studied, decision tree performed second best, followed by logistic regression. Despite their limitations, conventional algorithms hold explanatory power and therefore merit consideration. Punnoose and Ajit [13], however, discovered that the XGBoost algorithm surpassed RF, LR, and Gaussian NB in terms of accuracy. Additionally, XGBoost's inherent regularization helps to avoid overfitting. The authors stated that the data used for predicting employee attrition contains noise, and XGBoost can effectively handle this by interpreting the noise as relevant information rather than ignoring it. This leads to the model learning the noise and resulting in a non-generalizable model [17]. In their research, the authors analyzed data from a global retailer with 73115 rows and 33 columns, and compared several machine learning techniques (LR, NB, RF, KNN, LDA, SVM, XGB). The results showed that XGBoost achieved an AUC score of 0.86 on the test set, while SVM and RF scored 0.52 and 0.51, respectively. The authors concluded that XGBoost is more effective for forecasting employee churn [13]. Also, Gabrani and Kwatra [18] determined that job satisfaction, length of employment, and evaluation are dependable indicators of employee turnover. The majority of research has shown that tree-based models, such as AdaBoost, Gradient Boosting Tree, Random Forest, and Extreme Gradient Boosting classifiers, consistently outperformed other models. The Multilayer Perceptron classification model was also used in some studies, with varying results.

Research on predicting employee churn was conducted using NB, LR, DT, and RF methods [19]. Another study by Khare et al. [20] developed an attrition risk equation using LR to forecast employee turnover. Nonetheless, both studies omitted feature engineering and feature selection techniques. Basha et al. [21] employed a gradient boosting classifier to construct the model. After evaluating the model's recall, precision, and accuracy, the gradient boosting tree outperformed other algorithms with an accuracy of 96%. The results also indicated that employees who are dissatisfied are more likely to

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

leave the organization, while those who have a longer tenure and are engaged in their work are less likely to depart. In their study, Sisodia et al. [22] assessed the performance of machine learning models for predicting employee turnover. The researchers aimed to compare the predictions of employee churn using various machine learning models such as SVM, NB, and DT. The results revealed that the RF model had a relatively high prediction accuracy. It was noted that a higher level of accuracy could have been achieved if more advanced machine learning techniques were used and the significance of features were validated through multi-criteria models [18]. A comparison of six machine learning models was performed in the IT sector [23]. The authors found that among the six evaluated models, Extreme Gradient Boosting was the top performer. This conclusion aligns with earlier research that reached similar results [12], [13]. Tharani and Raj [23] also regarded RF, Multilayer Perceptron, SVM, and KNN as good-performing models. These findings are in partial agreement with prior literature that evaluated RF and Multilayer Perceptron as good-performing models [13]. On the other hand, SVM and KNN were considered to be poor performers [12], [13].

DATA

The data used in the study was sampled from the IBM dataset which contains information about the employees such as personal details (name, employee ID, address, date of birth, and gender); contact details such as email and phone number; employment information such as hire date, job title, and department; compensation and benefits such as salary, bonuses, and insurance; performance evaluations such as reviews, ratings, and feedback; attendance and time off information such as sick leave, and vacation days; as well as education and training information, including degrees, certifications, and the courses taken. Initially the dataset had 35 features. However, 5 were removed due to redundancy, remaining with 30 features. Table 1 shows the 30 key features that were considered, along with the related data types.

Table 1. HR dataset features

No	Feature	Data Type
1	Age	Numeric Type
2	Business Travel	Categorical
3	Daily Rate	Numeric
4	Department	Categorical
5	Distance From Home	Numeric
6	Education	Categorical
7	Education Field	Categorical
8	Gender	Categorical
9	Environment Satisfaction	Categorical
10	Hourly Rate	Numeric
11	Job Involvement	Categorical
12	Job Level	Categorical
13	Job Role	Categorical
14	Job Satisfaction	Categorical
15	Marital Status	Categorical
16	Monthly Income	Numeric
17	Monthly Rate	Numeric
18	Number of Companies Worked	Numeric
19	Over Time	Categorical
20	Percent Salary Hike	Numeric
21	Performance Rating	Categorical
22	Relationship Satisfaction	Categorical
23	Stock Option Level	Categorical
24	Total Working Years	Numeric
25	Training Times Last Year	Numeric
26	Work Life Balance	Categorical
27	Years At Company	Numeric
28	Years In Current Role	Numeric
29	Years Since Last Promotion	Numeric
30	Years With Current Manager	Numeric

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

METHODS

The steps involved in predicting employee churn included importing the IBM dataset into a Jupyter Notebook as the first step. Subsequently, exploratory data analysis was conducted to gain a deeper insight into the data. Oversampling was done to address the problem of imbalanced data. To proceed, 70% of the data was sampled into training data, while the remaining 30% was designated as testing data. This ratio of 70/30 split was adopted based on research findings indicating optimal results [24].

Methods of feature selection, including Pearson correlation, Information gain, and Recursive feature elimination, were applied to identify the most crucial factors for prediction. Machine learning algorithms, including logistic regression, random forest, gradient boosting. decision trees, and K-nearest neighbour, were applied to each result of the feature selection methods. The accuracy, precision, recall, and F-Score of the algorithms were evaluated using test data for comparison.

The machine learning algorithms were rerun on the data without the use of feature selection methods. The classification results before and after feature selection were compared and the feature selection and classification approach with the highest accuracy and precision was selected. Finally, the key factors affecting employee turnover were analyzed and strategies for retaining employees were evaluated.

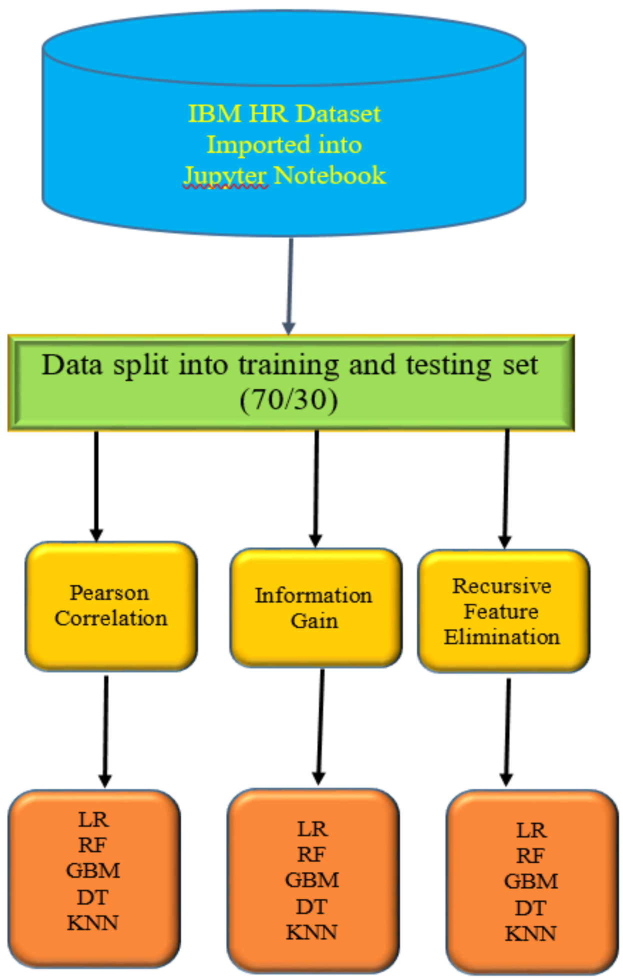


Fig 1

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

DATA IMBALANCES

The target variable data refers to the variable in a dataset that is being predicted or modeled. In this case, the target variable data is imbalanced, meaning that the distribution of the data between the different classes is not equal [11]. This can have negative effects on the performance of a machine learning model, as it can lead to decreased accuracy and bias towards the majority class. To mitigate these effects, two techniques were used to balance the data: Over-sampling and SMOTE. Over-sampling involves duplicating samples from the minority class to increase its representation in the dataset, while SMOTE (Synthetic Minority Over-sampling Technique) creates new synthetic samples for the minority class based on the existing samples that are closest to it in the feature domain. These techniques aim to balance the data, so that the machine learning model can be trained on more representative data and produce better results [11].

FEATURE SELECTION

In order to improve the prediction of employee turnover, the study used various feature selection methods to identify the most important features for the task [25]. This preprocessing step is beneficial for a number of reasons, including improvement in performance, reduction of overfitting, and reduced computational costs [11]. The study evaluated three different feature selection methods: Pearson Correlation, Information Gain, and Recursive Feature Elimination. Pearson Correlation (PC) selects features based on statistical measures and is known for its velocity and efficiency [11]. Features with a correlation greater than 0.8 were removed, leaving only columns with a correlation below 0.8. After this, classification algorithms were applied to the feature subset.

Information Gain (IG) evaluates the reduction of randomness in the data after transforming the dataset [25]. It analyzes each feature's contribution to the target variable by calculating its information gain. The features were ranked by descending information gain and those with a threshold of 0.005 or higher were included in the feature subset [26]. Again, classification algorithms were applied to this subset. Recursive Feature Elimination (RFE) is a feature selection method that iteratively measures feature importance and eliminates the least significant features [27]. The features were ranked by importance using the Random Forest (RF) method and the least important ones were removed until the desired number of features was reached. The RF-RFE algorithm was used to create the feature subset, followed by the implementation of classification algorithms.

CLASSIFICATION ALGORITHMS

The study utilized a diverse range of five classification algorithms to analyze the data. The first algorithm is Logistic Regression (LR), which is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The second algorithm used is Random Forest (RF), which is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The third algorithm is Gradient Boosting Machine (GBM), which is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The fourth algorithm used is Decision Tree (DT), which is a simple representation for classifying examples by recursively partitioning the feature space. The final algorithm applied is k-Nearest Neighbor (KNN), which is a non-parametric method used for classification and regression. The algorithm assumes that similar things exist in close proximity.

The Logistic Regression (LR) algorithm is a statistical technique used for binary classification, where the goal is to predict the probability of a categorical event happening (e.g., yes or no, true or false). The LR algorithm works by fitting a logistic function, also known as a sigmoid curve, to the data. The logistic function maps any real-valued number to a value between 0 and 1, which can then be interpreted as a probability value. In the context of LR, the logistic function represents the relationship

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

between the input features and the binary outcome. By estimating the parameters of the logistic function using maximum likelihood estimation, the LR algorithm can produce a model that predicts the probability of the event of interest for any new observation. Here, x represents the value of the output.

$$S(x)= \frac{1}{1+e^{-x}}$$

An analysis of the relationship between the dependent variable and independent variables is carried out using this algorithm. To accomplish this, a line must be fitted and the error between the line and the data points must be minimized. Weighting factors define the importance of the predictors [11]. The ease with which Logistic Regression can be implemented and trained made it a suitable baseline model for the study.

A Random Forest (RF) is a predictive modeling technique that makes use of Decision Trees (DT) as its foundation. The advantage of this approach lies in its ability to mitigate overfitting, a common problem in machine learning, by aggregating the predictions from multiple trees rather than relying on the result from a single DT. This aggregation helps to reduce the variance in the predictions, leading to improved performance compared to a single DT, while maintaining the same level of bias. In the model being described, all features were transformed into numerical values, after which the output variable was predicted using the RF technique. The use of DTs as the basis for prediction in RF offers several benefits, including its ability to handle complex non-linear relationships between the input variables and the output variable, as well as its ability to handle large amounts of data and to handle missing data. In summary, RF is a powerful technique for predictive modeling that can significantly improve the performance of a single DT by combining the results from multiple trees and reducing variance while maintaining bias. The conversion of all features to numeric values is an important step in this process, as it enables the model to handle a wide range of data and make accurate predictions.

Gradient Boosting Machines (GBM) is a powerful machine learning technique that combines multiple weak learners to form a strong ensemble model. In GBM, the weak learners are combined iteratively in a way that improves the overall accuracy of the model. The process of combining weak learners in GBM can be seen as an optimization problem, where the goal is to minimize the error rate of the ensemble model. This optimization is achieved through a gradient descent procedure, where the error rate is progressively and repeatedly reduced. In the study being described, all features were converted to numerical values, allowing the GBM algorithm to process the data and make predictions. The target variable in this study was "Attrition", which was predicted using the GBM technique. The ability of GBM to handle a wide range of data types and its ability to handle complex non-linear relationships between the input variables and the output variable make it a popular choice for predictive modeling. In summary, GBM is a powerful technique for predictive modeling that involves combining multiple weak learners to form a strong ensemble model. The optimization process used in GBM reduces the error rate of the model and improves its overall accuracy. The conversion of all features to numeric values is an important step in this process, as it enables the GBM algorithm to process the data and make accurate predictions.

Decision Trees (DT) are a widely used machine learning technique for data prediction and classification. They have a tree-like structure that can be easily visualized and understood, making them a popular choice for both researchers and practitioners.. In a DT, every internal node represents a test on an attribute of the data, and every branch represents the outcome of that test. The leaf nodes of the tree represent the class labels of the data. This flowchart-like layout makes it easy to understand how the model arrived at its predictions and can help in interpreting the results. DTs are effective for both regression and classification problems, and they can handle both continuous and categorical data. They are particularly useful for dealing with complex non-linear relationships between the input variables and the output variable. DTs can also handle large amounts of data and can handle missing data effectively. In summary, DTs are a powerful technique for data prediction and classification that can handle complex relationships between the input variables and the output variable. The tree-like structure of

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

DTs makes them easy to understand and interpret, and they can handle a wide range of data types and missing data.

K-Nearest Neighbors (KNN) is a popular machine learning algorithm used for binary classification tasks. In KNN, data points are classified based on the class of their nearest neighbors. The number of neighbors considered is specified by the user and is represented by the parameter "k". To determine the nearest neighbors, a distance measure is used. The most common distance measure used in KNN is the Euclidean distance, which calculates the straight-line distance between two points in a multi-dimensional space. This distance is calculated for each instance in the dataset, and the k instances with the smallest distances are identified as the nearest neighbors. Once the nearest neighbors have been identified, the unlabeled instances are classified based on the majority class of their k-nearest neighbors. If k is set to 3 and 2 out of the 3 nearest neighbors belong to class A and 1 belongs to class B, the unlabeled instance will be classified as belonging to class A. This process is repeated for all instances in the dataset, and the resulting classifications form the output of the KNN algorithm. In conclusion, KNN is a simple and effective method for binary classification tasks, as it makes use of the class information of nearby instances to make predictions. By using the Euclidean distance and a specified value of k, KNN is able to effectively classify new, unlabeled instances.

RESULTS

To evaluate the various strategies, the feature selection techniques and corresponding algorithms were executed and the results of their accuracy were summarized in Table 2. The purpose of comparing the different approaches was to determine which method performed the best in terms of accuracy. The feature selection methods used were applied to identify the most important features in the data set, which then served as inputs for the relevant algorithms. The accuracy scores of the algorithms were recorded and presented in Table 2 for easy comparison and analysis. This helped to determine the most effective combination of feature selection and algorithm for the specific problem at hand.

Table 2. Classification accuracy with feature selection

Feature Selection Method	Accuracy %				
	LR	RF	GBM	DT	KNN
PC	91.62	91.76	91.49	82.16	88.51
IG	88.78	92.57	90.68	82.57	87.43
RFE	87.43	92.30	90.14	79.59	87.70

In addition to evaluating the performance of the algorithms with the use of feature selection methods, the algorithms were also tested without these methods to compare the results. The accuracy scores of the algorithms in this scenario were recorded and presented in Table 3. The purpose of this comparison was to determine if feature selection improved the accuracy of the algorithms, or if the algorithms could produce acceptable results without it. By presenting the accuracy scores without feature selection in Table 3, a clearer understanding of the impact of feature selection on the performance of the algorithms could be obtained. This information would then inform the decision of whether or not to use feature selection in a given scenario.

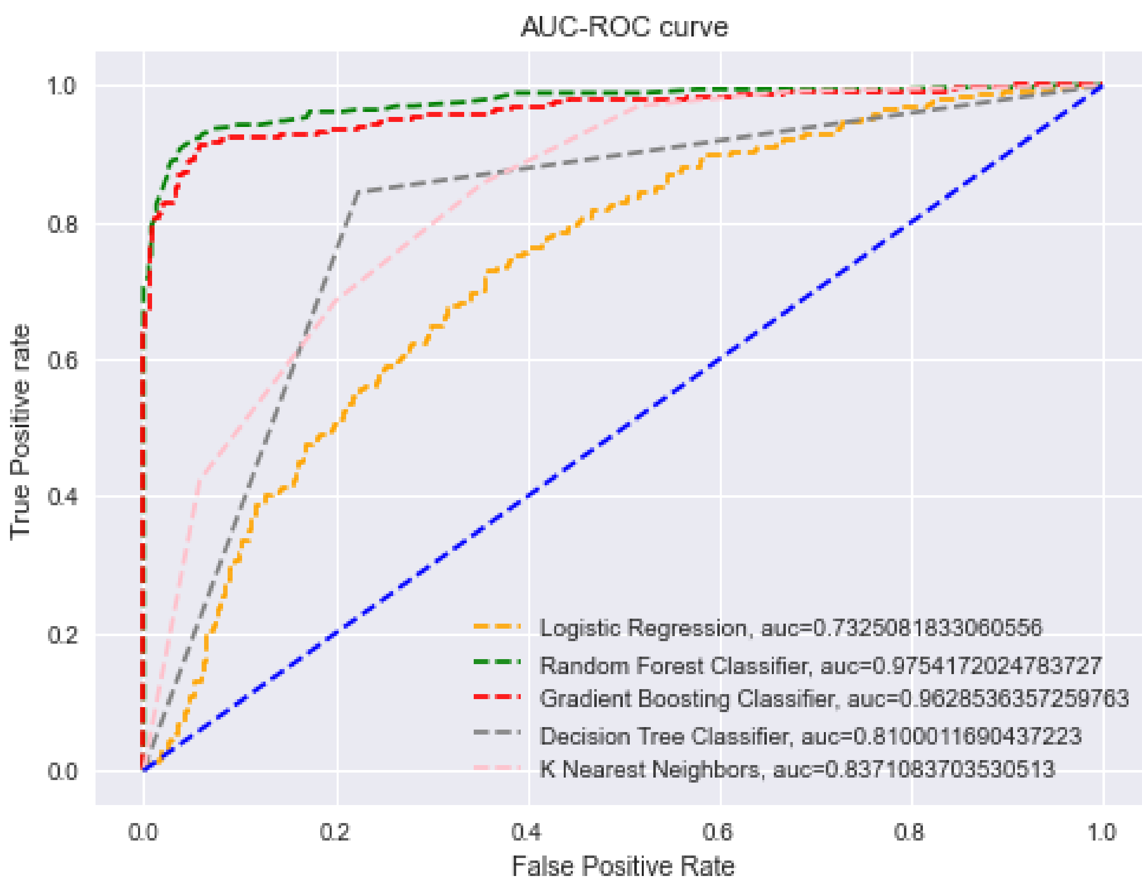
Table 3. Classification accuracy without feature selection

	Accuracy %				
	LR	RF	GBM	DT	KNN
Original, untreated data	84.35	86.17	87.53	76.64	84.13

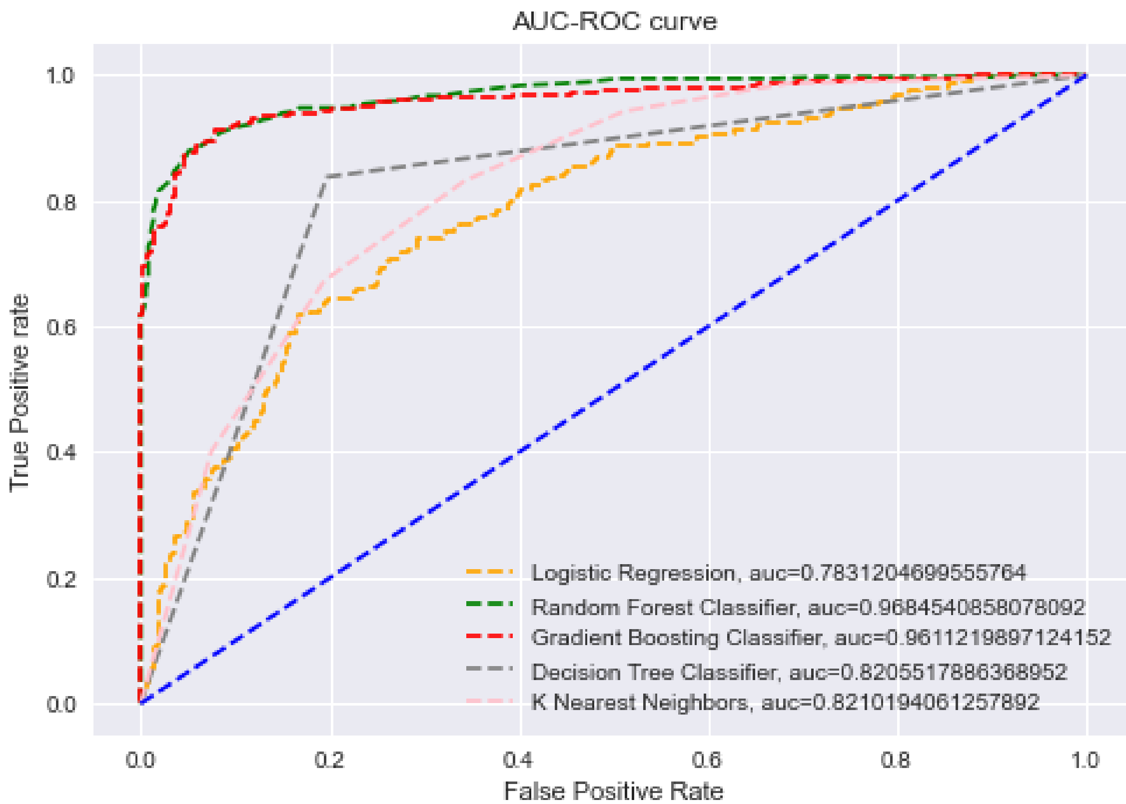
ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

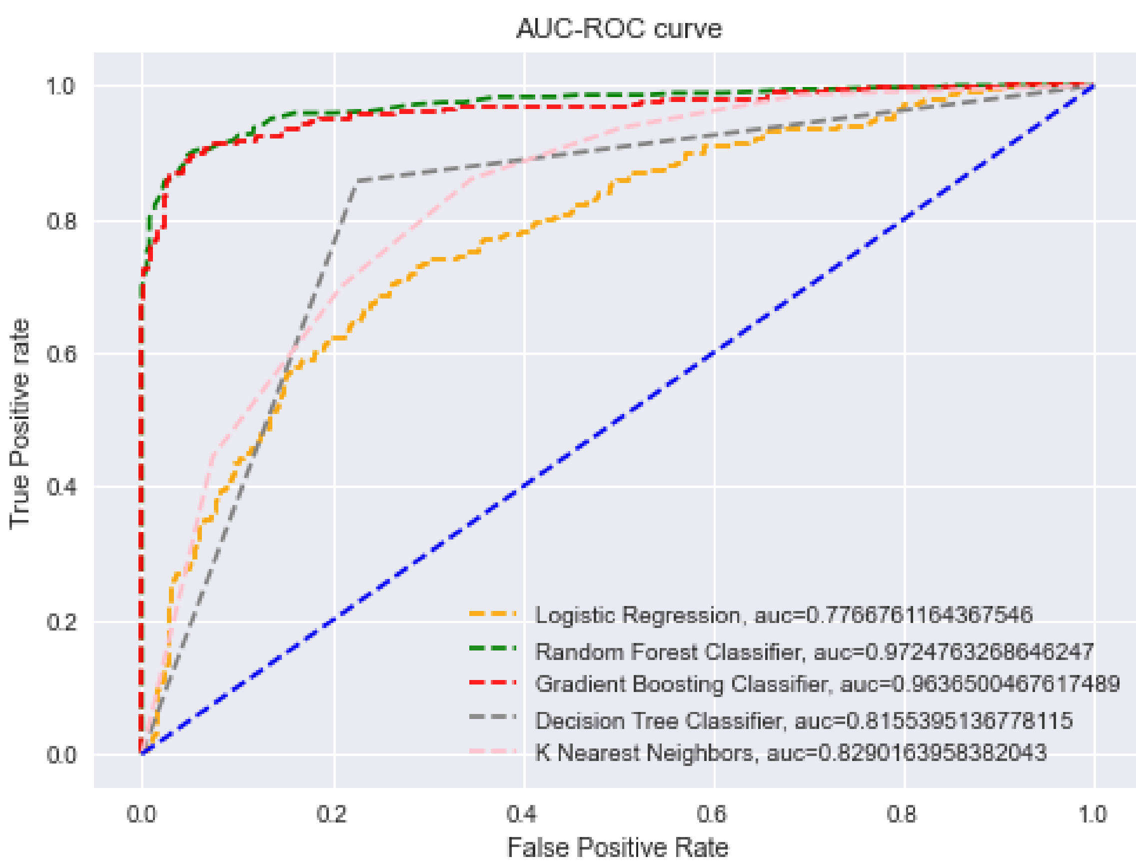
When analyzing Table 2 and Table 3, it becomes evident that the accuracy of the results continually improves with the use of feature selection methods. One method in particular, Pearson Correlation, stood out as being particularly effective, resulting in a significant increase in accuracy when compared to other methods. Out of all the classification techniques used, the Random Forest (RF) Classifier demonstrated the highest accuracy, showing a 6.84% improvement when Pearson Correlation was employed. The accuracy of the Gradient Boosting Machine (GBM) also saw an improvement of 4.52%, while the decision tree (DT) method showed a 7.2% increase. The k-Nearest Neighbors (KNN) method saw a 5.21% increase in accuracy, and the Logistic Regression (LR) method had the largest improvement of all, with an 8.61% increase in accuracy.

To evaluate the performance of different classification thresholds, Receiver Operating Characteristic (ROC) curves are utilized, and the Area Under the ROC Curve (AUC-ROC) is calculated both with and without feature selection. The AUC-ROC summarizes the results of each threshold's confusion matrix and a high AUC (close to 1) indicates that the model is highly separable. The closer the AUC is to 1, the better the model's performance. The results showed that the AUC-ROC of the Gradient Boosting Machine (GBM) and Random Forest (RF) algorithms was higher than the other algorithms in all cases, implying that the two ensemble algorithms are superior models. The AUC-ROC graph clearly indicates that GBM and RF are strong models, with AUC values close to 1, particularly when Pearson Correlation is used for feature selection. The ROC curves for the different algorithms are presented below.

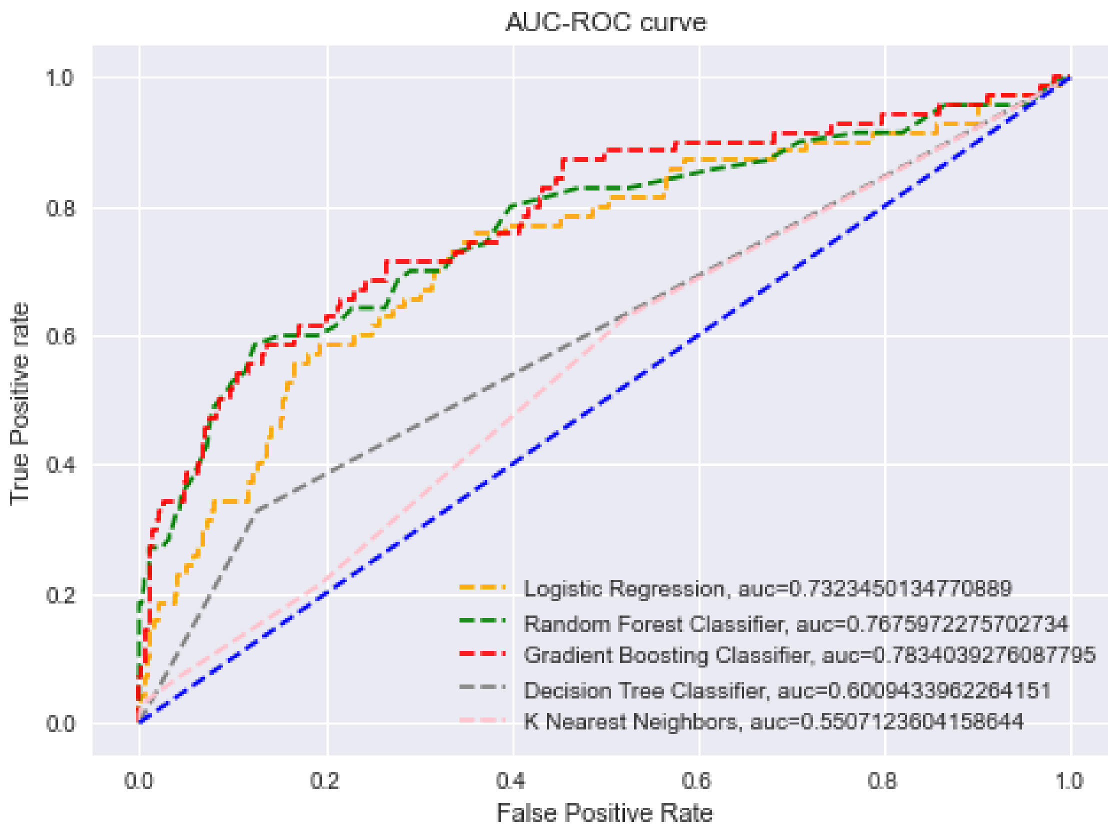


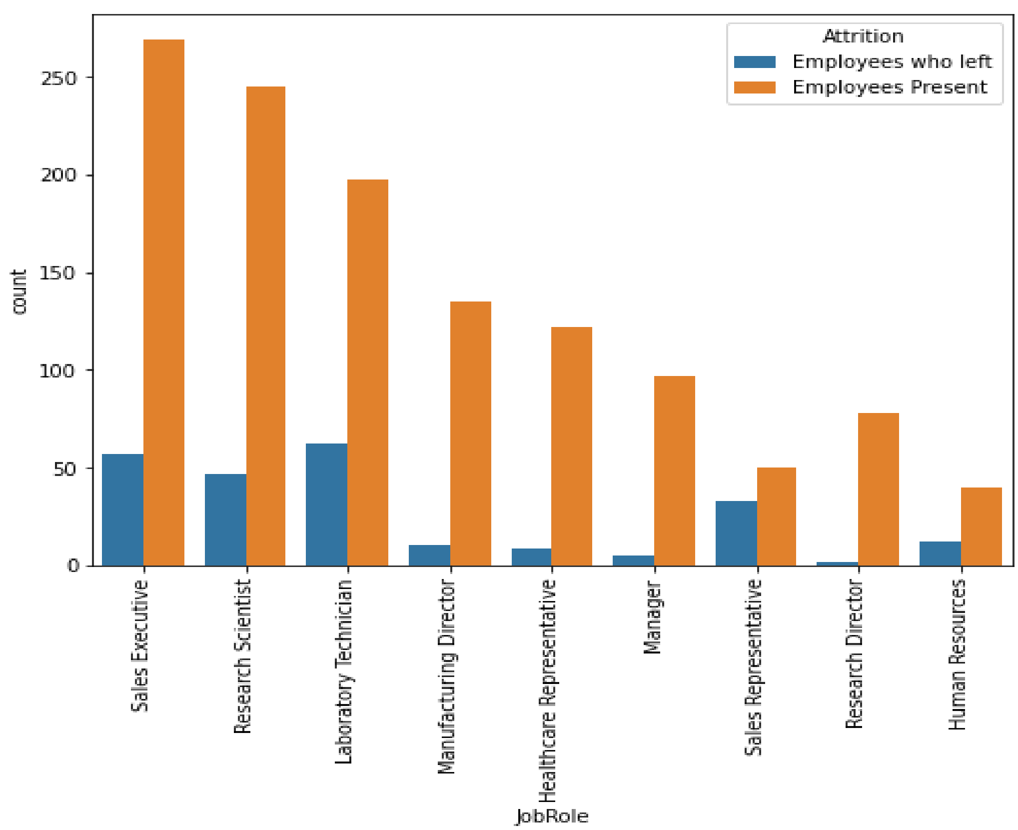
ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)





ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)





ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

CONCLUSION

The purpose of the study was to build a supervised machine learning model for employee attrition prediction. To do this, the study compared the performance of five different algorithms: Logistic Regression (LR), Random Forest (RF), Gradient Boosting Machine (GBM), Decision Tree (DT), and K-Nearest Neighbours (KNN). These algorithms were evaluated both with and without feature selection to determine the impact of feature selection on each algorithm performance. The results of the study showed that the Random Forest (RF) model had the highest accuracy and AUC (Area Under the Curve) when compared to the other algorithms. On the other hand, the Gradient Boosting Machine (GBM) model performed the best on untreated data. The study also found that ensemble algorithms (algorithms that make predictions by combining the outputs of multiple models) showed greater predictive power. In terms of feature selection, the results showed that it improved the performance of the algorithms. The study found that the Pearson Correlation method was the most effective feature selection technique. The study concluded that the employee attrition prediction model built using machine learning can assist management in developing effective retention strategies. The results of this study suggest that future studies should explore the use of unsupervised machine learning and deep learning techniques for employee attrition prediction.

ACKNOWLEDGEMENTS

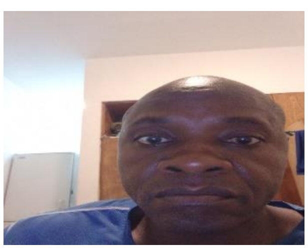
We acknowledge both the moral and technical support given by colleagues at the Sol Plaatje University, Department of Computer Science, Data Science, and Information Technology. This research was also partly funded through an ABSA research grant and the CAIR (Centre for Artificial Intelligence Research) fund, grant agreement number: CSIR/BEI/HNP/CAIR/2020/10, supported by the Government of the Republic of South Africa, through its Department of Science and Innovation (DSI).

REFERENCES

1. R. K. Pradhan, L. K. Jena and R. Pattnaik, "Employee Retention Strategies in Service Industries: Opportunities and Challenges," Employees and Employers in Service Organizations, Apple Academic Press, 2017, pp 53-70.
2. A. Khaled, Z. Safeya, A. Abdullah and T. M. Usmani, "Employee Retention Prediction in Corporate Organizations using Machine Learning Methods," Academy of Entrepreneurship Journal, vol.27, 2021, pp 3-4.
3. Z. Xiaojuan, "Forecasting Employee Turnover in Large Organizations. " PhD diss., University of Tennessee, 2016.
4. I. Fulmer, R. Ployhart, "Our Most Important Asset", Journal of Management, vol.40, 2013, pp 2-3
5. A. Reyes, C. Aquino, D. C. Bueno, "Why Employees Leave: Factors that Stimulate Resignation Resulting in Creative Retention Ideas",Researchgate Publication, 2019, pp7-8.
6. A. Raza, "Predicting Employee Attrition Using Machine Learning Approaches", vol.12, Applied Sci. Journal, 2022.
7. A. Omar, F Hossam, J Khalid, H Osama and N. Ghatasheh, "Predicting customer churn in telecom industry using multilayer preceptron neural networks: modeling and analysis", vol.11(3), Life Science Journal, 2014, pp 23-24.
8. V.Umayaparvathi and K Iyakutti, "A survey on customer churn prediction in telecom industry: datasets, methods and metric",vol.3, IRJET, 2016, pp 12-16.
9. A. Amin , S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, A. Hussain, " Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study", vol. 4, IEEE Access, 2016, pp.7940-7957.
10. A. Hebbar, P. Sanath, S. Rajeshwari, S Saqquaf, "Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees", 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, May 2018, pp 934-938.
11. R. V. Dam, "Predicting Employee Attrition", Masters Thesis, University of Tilburg, 2021.
12. Y. Zhao, M. Hryniewicki, F. Cheng, B.Fu and X. Zhu, "Employee Turnover Prediction with Machine Learning: A Reliable Approach", PhD diss, University of Toronto, 2019.
13. R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms", International Journal of Advanced Research in Artificial Intelligence, vol. 5 Issue 9, October 2016, pp 32-33.
14. N. El-Rayes, M. Fang, M. Smith and S. Taylor, "Predicting employee attrition using tree based models", International Journal of Organizational Analysis, vol 28, pp 33-36, 2020.
15. I. Yigit and H. Sourabizadeh, "An Approach for Predicting Employee Churn by Using Data Mining", IEEE, 2017.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

16. F. Falluchi, M. Codalangelo, R. Giuliano and L. William, “Predicitng Employee Attrition Using Machine Learning Techniques”, Computers Journal, vol.9, 2020.
17. X. Ying, “An Overview of Overfitting and its Solutions”, IOP Publishing, vol.1168, No.2, 2022.
18. G. Gabrani and A. Kwatra, “Machine Learning Based Predictive Model for Risk Assessment of Employee Attrition”, Computational Science and Its Applications,Springer International Publishing, 2018.
19. V. Saradhi and G. K. Palshikar, “Employee churn prediction,” Expert Sys. with Applications, vol. 38, no. 3. 2011
20. Khare, R. Kaloya ,D. Choudhary, CK and Gupta, G. “Employee attrition risk assessment using logistic regression analysis,” Computational Science and Its Applications, vol. 34, 2013.
21. M. Basha, D. S. Rajput, V. Vandhan, “Impact of Gradient Ascent and Boosting Algorithm in Classification”, International Journal of Intelligent Engineering Systems, vol.11, February 2018.
22. S. Sisodia, S. Vishwakarma and A. Pujahari, “Evaluation of machine learning models for employee churn prediction”, International Conference on Inventive Computing and Informatics (ICICI) (pp. 1016-1020), IEEE, 2017
23. M. Tharani and V. Raj, “Predicting employee turnover intention in itites industry using machine learning algorithms”, Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) , 2020.
24. A. Gholamy, V. Kreinovich and O. Kesheleva, “Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation”, Technical Report: UTEP-CS-18-09, University of Texas, February, 2018.
25. B. Pranto, “Entroppy Calculation, Information Gain & Decision Tree Learning”, https://medium.com/analytics-vidhya/entropy-calculation-information-gain-decision-tree-learning-771325d16f. Accessed 20 August 2022.
26. A. Gupta, “Feature Selection Techniques in Machine Learning”, https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/. Accessed 20 August 2022.
27. N. V. Sharma and N. S. Yadav, “An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers”, Microprocessors and Microsystems Journal, vol. 85, 2021
28. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene selection for cancer classification using support vector machines”, Machine Learning, vol 46, issue1/3, pp 389-422, 2018.
29. T.W. Edgar and D.O. Manz, “Research Methods for Cyber Security”, Science Direct Publishers, 2017, pp 36-38.
30. G. James, D. Witten, T. Hastie and R. Tibshirani, “ An introduction to statistical learning”, Springer 2018,
31. S. Han, H. Kim and Y.S. Lee, “Double random forest”, Mach Learn vol 109, 2020, pp12-13.
32. [35] E. Scornet, “Trees, Forests and Impurity-based Variable Importance”, HAL Open Science, 2020, pp 10-11.
33. M. Chen, Q. Liu, S. Chen, Y. Liu, C. H. Zhang and R. Liu, “ XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System”, IEEE Access, 7, 13149–13158. https://doi.org/10.1109/access.2019.2893448.
34. J. Han, J Pei, M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2011.
35. S. Zhang, X. Li, M. Zong, Z. Ming and D. Cheng, “ Learning k for kNN Classification”, ACM Transactions on Intelligent Systems and Technology, ResearchGate, 2017, pp 7-8.
36. D. Subramanian, “A Simple Introduction to K-Nearest Neighbors Algorithm”, ScienceDirect Publishers, 2019.
37. L. Y. Hu, M.W. Huang, S.W. Ke, “The distance function effect on k-nearest neighbor classification for medical datasets”, SpringerPlus, vol 5, 2016.



Mr Vengai Musanga holds an MSc in Data Science from the University of Zimbabwe, Zimbabwe and a BTech (Honours) Degree in Computer Science from the Harare Institute of Technology, Zimbabwe. His main research area is Artificial Intelligence with a particular focus on Machine Learning Engineering. Mr. Musanga is currently the Chairperson of the Computer Science and Software Engineering Department at the Zimbabwe National Defence University, Zimbabwe.

Dr. Colin Chibaya holds a PhD (Computer Science), Rhodes University, South Africa; MSc (Computer Science), National University of Science and Technology, Zimbabwe; and a BSc (Mathematics and Computer Science), Enrique Jose Varona, Cuba. His two main research focuses are on the fields of Artificial Intelligence, particularly, Swarm Intelligence Systems, as well as Cryptography. Dr. Chibaya is currently a senior Lecturer in the Department of Computer Science, Data Science, and Information Technology in the School of Natural and Applied Sciences at the Sol Plaatje University

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The Issues, Challenges and Impacts of Implementing Machine Learning in the Financial Services Sector: An Outcome of a Systematic Literature Review

Siphiwe Mndebele[1,2*] and Thembekile Mayayise[1†]
[1] EasyChair
[2] University of the Witwatersrand, Johannesburg, South Africa.
784297@students.wits.ac.za, Thembekile.Mayayise@wits.ac.za

Abstract

Machine learning (ML) is another branch of technology deemed valuable in the financial sector because of its ability to assist organisations in identifying fraudulent transactions and predicting the ability of customers to repay their bank-issued loans. However, like any type of technology, the adoption of ML introduces changes that impact the processes and operations of the financial service sector. Research on the merits of implementing ML is well captured; however, research on such developments' challenges, issues, and impact is scant. To address this gap, a systematic literature review was undertaken to contribute to the research discourse by investigating the issues, challenges and impacts of implementing ML in the financial business sector. The ScienceDirect, EBSCOhost and ProQuest databases were used to search for the relevant scholarly sources published from 2013-2022. The literature was reviewed based on the PRISMA flow diagram and a thematic analysis of the 35 articles that met the inclusion criteria. The outcome of the review revealed that more complex models, such as artificial neural networks, were implemented in all the identified financial services sectors, followed by support vector machines. This review concludes that the larger the quantity and complexity of financial data, the less the data quality, which significantly reduces the prediction performance, efficiency, and accuracy of the model, which can significantly impact the operations, financial aspects, and the overall reputation of the firms. Future research must explore the impact of ML on the operational, adoption and skills shortages in the financial sector.

[*] Masterminded EasyChair and created the first stable version of this document
[†] Created the first draft of this document

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

1 Introduction

Machine learning (ML) involves the capability of systems to use curated algorithms to iteratively learn by training available data sets to establish patterns in existing data and make predictions about new data (Hall et al., 2016). ML draws from various fields such as data mining, statistics, artificial intelligence, analytics, and optimisation studies. Artificial intelligence(AI) is often regarded as the predecessor of machine learning (Marr,2016). More recently, the complexity of data generation and the quest to make machines assimilate human intelligence have pushed the boundaries of AI to construct intelligent machines that can learn independently of human input (Agavanakis et al., 2019).

Since the inception of machine learning in the early 1950s, it has gained so much prominence. Its demand is expected to grow substantially due to increased data volume and variety. ML is now used in scientific, engineering, manufacturing, agricultural, commercial and social settings (Jordan & Mitchell,2015). The adaptability of machine learning algorithms facilitates the generation of insights from a variety of complex and very large data sets. This has promoted the adoption of machine learning across a variety of fields to generate necessary insights to support decision-making.

In totality, Digital transformation is the driver of enormous changes which embeds integration of digital technology into business operations, processes, and management to deliver better, faster, and more comprehensive insights about a business, products, services, and customers with the aim of creating efficiencies and value for its customers and profit for the business (Müller, Junglas, Brocke & Debortoli, 2016; Grover, Chiang, Liang, & Zhang, 2018). Hence the investments in digital transformation initiatives are often driven by companies who want to retain or improve their competitive advantage. Hence it is no surprise that an increase in investments is predominantly driven by companies that want to keep up with the demand from their customers through insights such as those gained from Business Intelligence(BI) capabilities (Grover et al., 2018; & Müller, Kiel, & Voigt, 2018). However, as more corporate and academic research points to the benefits associated with the adoption and implementation of machine learning technologies by private businesses, there appears to be limited literature highlighting the issues, challenges, and impacts of implementing machine learning technologies within the business context, specifically in the financial services sector(Hall et al., 2016; Arunachalam, Kumar, & Kawalek, 2018). The aim of this paper is to address the research gap from an academic context by conducting a systematic literature review of the scholarly literature on the issues, challenges, and impacts of implementing machine learning in the financial services sector. The paper is arranged in the following manner: section 1 is the introduction, section 2 covers the literature review, section 3 delves into the research methodology, section 4 covers the research findings, and section 5 concludes the study.

1.1 Objective

The main aim of this paper is to uncover the extent to which machine learning has been researched in the context of the issues, challenges, and impacts of implementing ML in the financial services sector and to identify potential areas for future research. This will be achieved by exploring the different implementations of ML techniques in the financial services sector to identify, analyse and evaluate the issues, challenges, and impacts of implementing those machine learning models/techniques (Jordan & Mitchell, 2015).

1.2 Research questions

The primary research question that this study seeks to address is:

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

What are the issues, challenges and impacts of implementing Machine Learning in the financial services sector?

Sub-Questions:

- In what financial business contexts are ML technologies implemented?
- What ML methods/techniques are implemented in each financial business context?

2. Literature review

2.1 Applications of machine learning in the financial services sector

The financial services sector has long been using statistical techniques to analyse and mine various financial data from different sources to gain new insights such as valuing investments, assessing customer creditworthiness and assessing risk. In recent times, the advancements in computing power, cheaper cloud storage, and demand to make more informed business decisions have invested in machine learning-driven technologies more common and influential within the financial services sector (Manlangit, Azam, Shanmugam, & Karim, 2019). According to Leo et al. (2019), implementing ML is important, especially when financial services institutions require the necessary analytical capability, as the implementation is likely to impact every aspect of their business model. In certain financial services environments, unsupervised ML is often used to explore the complex input data from their clients' loan records, where regression and classification methods are used to predict key credit risk variables to determine the risk of clients defaulting on loan payments (Khandani et al., 2010; Van Liebergen, 2017). However, one of the major challenges in financial institutions is the detection and prevention of credit card fraud. Credit card transactions, by design, produce very large datasets that require advanced algorithm training to detect complex patterns for the classification of fraudulent transactions from non-fraudulent ones and to detect potential money laundering transactions from normal transactions. Manlangit et al. (2019) applied ML techniques to transform the raw credit card transaction data into meaningful patterns, followed by K- Nearest Neighbours (k-NN) analysis to classify fraudulent transactions from non-fraudulent transactions. It was then concluded that their proposed technique improves the accuracy of detecting fraudulent transactions faster and more accurately than the conventional rule-based systems used by banks. This proves that innovative ways of combating credit card fraud are being investigated and are much needed due to the magnitude of the problem. In many instances, machine learning techniques are used in surveillance and monitoring for investment banks and securities exchange regulatory bodies to identify misconduct and trading breaches that result in high financial and reputational costs. Evidence of studies is growing in understanding the application of various machine learning techniques to predict things such as bankruptcy of clients and businesses, as well as research into the use of ML methods to evaluate and assess risk in the banking sector(Leo et al., 2019; Qu, Quan, Lei, & Shi, 2019).

2.2 Related works and knowledge gaps

Due to the widespread use of machine learning, questions about the issues, challenges, and impacts of implementing machine learning in a business context are becoming more relevant and necessary to be answered (Müller et al., 2016; Grover et al., 2018). According to a comprehensive literature review

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

conducted by Trieu (2017) to evaluate the evidence published between 2000 to 2005 on the value derived from BI systems by organisations, it is argued that although the impact of BI is significant, the conditions need to be favourable to an organisation to reap the benefits of implementing a BI system as some of the internal and external factors that may hinder the value realisation by an organisation (Trieu, 2017). Since the review focused on literature between 2000 and 2005, it was highlighted that there was a lack of studies looking into how BI impacted the internal and external factors to create business value. It was then recommended that such studies should be conducted. In another systematic literature review between 2000 and 2017 exploring BI in Small and Medium-sized Enterprises (SMEs) the implementation of BI presented a new set of challenges that could potentially hinder the realisation of the benefits of implementing BI by SMEs (Llave, 2017). Challenges such as the lack of established frameworks and standards guiding the governance, security, and guarantee of privacy when utilising BI technologies were identified.

Both systematic literature reviews highlight an adequate body of evidence on implementing BI technologies in some business sectors and the resultant challenges associated with these technologies. However, neither of the reviews was primarily focused on ML, nor did they highlight challenges, issues, or impacts that are specifically associated with or as a result of machine learning in particular. In another study, Devi and Radhika (2018) compared various statistical and ML techniques by evaluating their performance ratios, including accuracy, specificity, sensitivity, and precision in detecting and predicting bankruptcy. Some notable benefits identified in the study include the finding that ML techniques have better performance accuracy. However, their discussion of limitations is limited to the inherent technical applicability of various methods and their performance and not so much to the impact or challenges presented to the broader financial services sector due to the implementation of these ML techniques. Similar reviews evaluating the applicability and performance of various statistical and ML techniques in bankruptcy prediction have been conducted (Lin et al., 2011; Qu et al., 2019). Similarly, they both fall short of discussing the overall impact or challenges of implementing ML techniques in the business models of the broader financial services sector.

Leo et al. (2019) reviewed literature that evaluates and discusses ML techniques that have been researched in the context of risk management in the banking sector but have yet to delve into broader ML issues in the financial sector.

Van Liebergen (2017) discusses the application of ML in the area of credit risk modelling and comes to conclude that non-parametric and non-linear ML methods are too complex to understand and audit, which is to the displeasure of most financial supervisors who typically require models to be clear and simple to understand, verify and validate. Regulations around data sharing and data usage were also identified as key impediments to achieving the data capacity to make meaningful interpretations and conclusions about the data analysed. As such, there remains a knowledge gap about the comprehensive documentation of the issues, challenges, and impacts of implementing ML in business, particularly in the financial services sector.

3. Methodology

This systematic literature was conducted by following the guidelines by Hagen-Zanker & Mallett (2013) on conducting a qualitative systematic literature review in the field of information systems. A description of search criteria and terms used for identified databases, the inclusion and exclusion criteria, the bias identified, and the data extraction and thematic analysis of the evidence.

To answer the research questions outlined in section 1.2, a qualitative research approach in the form of a systematic literature review, as depicted in the PRISMA flowchart in Figure 1, was adopted. Scholarly literature published from 2013 to 2022 from the following databases: Science Direct,

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

ProQuest and Ebsco Host were consulted. Instead of consulting one source, three sources were chosen to minimise bias by choosing databases containing various relevant and current literature on ML studies. The systematic literature review method has been used in different disciplines and relied upon for issues identification for future research. The PRISMA diagram in Figure 1 summarises the steps taken to conduct the systematic literature review. The inclusion and exclusion criteria are detailed in Table 1.

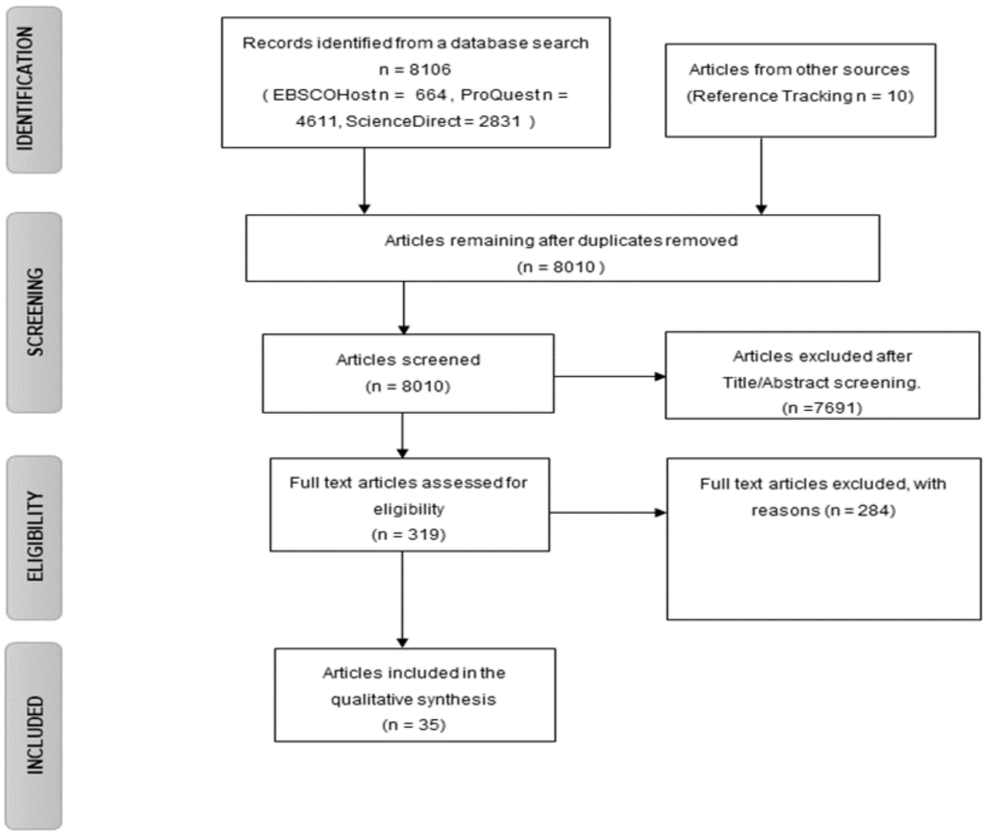


Figure 1: PRISMA

The following section explains the activities in the stages contained in Figure. 1: These stages were guided by the inclusion and exclusion criteria for the literature survey, as outlined in Table 1.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Table 1. Inclusion and exclusion criteria

Included	Excluded
Scope of coverage: 2013-2022	Literature published before 2013 and after 2022 is excluded
Exposure of interest: machine learning, business analytics, predictive analytics, artificial or business intelligence with a specific focus on machine learning	Non-Machine learning-related content
Language: English	Non-English written
Source of literature: scholarly books, journal articles, conference proceedings, and grey literature, including dissertations and theses	Any literature not in the form of scholarly books, journal articles, conference proceedings, and grey literature, including dissertations and theses.
Study Design: include quantitative, qualitative, and mixed-methods studies	Unspecified study design

The literature review was conducted using the academic institution's library resources where the researchers are affiliated. The literature search was conducted in September 2020 and December 2022. Only the literature that was available and accessible from the chosen databases on the date of the search was used.

Following the adapted PRISMA flow chart outlined in Figure.1, the following steps were carried out during the systematic literature review.

Identification phase

After selecting the following database sources, i.e. EBSCO Host, Science Direct and ProQuest, the relevant literature was selected using the keywords/phrases in Table 2.

Table 2: Keywords and Database Search Terms

Database	Query
ProQuest, EBSCO host & Science Direct	Banking, Investment, Insurance, Auditing, financial, Accounting, Credit Machine Learning, ML, predictive analytics, supervised learning, unsupervised learning, business analytics, artificial intelligence, business intelligence, statistical learning Issues, challenges, impacts, barriers, limitations, Inhibitors, disadvantages, pitfalls, threats, implications, impediments, shortcomings

The initial search results yielded 8106 records in total from the three database sources. After duplicate removal,8010 articles remained for further review.

Screening Phase

The 8010 articles were reviewed based on the content of their abstracts and titles. From that review, 7691 articles were excluded after the title and abstract review, and 319 articles remained for further analysis.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Eligibility

A detailed analysis of the 319 articles was further done, and 284 articles were excluded as they were found irrelevant. Thirty-five articles were found suitable for further review and analysis to solicit the answers to address the research questions posed for this study.

Synthesis

In total, 35 articles were synthesised to obtain answers to the research questions, as outlined in section 1.2.

3.1 Bias identified and addressed

To minimise bias in this study, measures were taken to address bias in the following manner:

In addressing location bias which comes as a result of restricting search limiters or sources which can potentially ignore a relevant body of evidence (Keenan, 2018), sources from published sources and grey literature were consulted to have a representative balanced collection of evidence (Drucker, Fleming, & Chan, 2016; Keenan, 2018). Evidence Selection Bias: is often introduced when a study fails to equitably identify all available unclear sources, which can be due to unclear definitions of key topic concepts (Drucker et al., 2016; Keenan, 2018;). To address this, the steps outlined in Figure 1 were followed within the inclusion and exclusion criteria confines.

Synthesis / Reporting Bias-This form of bias can arise when researchers selectively report some findings and exclude others (Drucker et al., 2016). This was addressed by making use of the inclusion criteria designed for this study.

Competing Bias usually arises when funding was obtained to support the study, and there is some connection with the funded entity (Drucker et al., 2016). The researchers have no conflict of interest in relation to this study.

3.2 Data analysis

Data were extracted from the 35 remaining articles for qualitative synthesis and put into an analysis matrix, and thereafter, thematic analysis was applied as this is commensurate with qualitative studies of this nature. Thematic analysis is a technique generally used to identify, analyse, organize, and report data in themes in a specific data set (Roberts, Dowell, & Nie, 2019).

As this systematic literature review aimed to qualitatively review the issues, impacts, and challenges of implementing ML in the financial service sector, a six-phase step practical guide on conducting a reliable thematic analysis was adopted(Nowell et al., 2017). The first step was concerned with getting the researchers familiar with their research area to stimulate their thoughts and get them to start asking the right and relevant questions. The codes for the financial sector identified, models, implemented, and identified challenges were generated in the second phase to simplify and ensure focus on the research questions (Nowell et al., 2017; Roberts et al., 2019).

The third phase was concerned with searching for themes which only starts once all coded data extracts are collated into meaningful sub-themes (Nowell et al., 2017). In the third phase, the review of the set of themes devised is conducted to check for the validity and meaning of patterns. It is in this phase that errors in the initial coding start to emerge (Nowell et al., 2017), and where errors may start to converge or diverge. In this study, based on the outcome of the literature review, the *identified*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

challenges code was broken into two codes, the *intrinsic challenges and extrinsic challenges* to make sense of challenges that are intrinsic to the model and those that are as a result of the sector(extrinsic).

The second last phase was concerned with the labelling of the themes based on the aspect of the data each theme captures and the overarching link it makes with the research focus and research questions. This was done to ensure alignment that all the developed themes are linked to the global organizing theme of the issues, impacts, and challenges of implementing ML techniques in the different financial services activities.

The last and final step was for the researchers to verify that they had captured all the themes and thus could start with the final analysis and proceed with the write-up. The codes identified and the themes are depicted in Figure 4.

4. Findings

Following the literature search conducted by ProQuest,Ebsco Host and Science Direct, using the chosen keyword strings, only 35 articles remained for synthesis as those articles fully satisfied the inclusion criteria in Table 1.

Figure 2 depicts the number of articles in the financial services sector. According to the information in Figure 2, most of the articles were on challenges regarding using ML in the financial services sector, focusing on Bankruptcy prediction, followed by fraud detection. This was followed by credit default prediction, forecasting of the stock market performance and general use, penultimate, and the least of the sources focused on risk management.

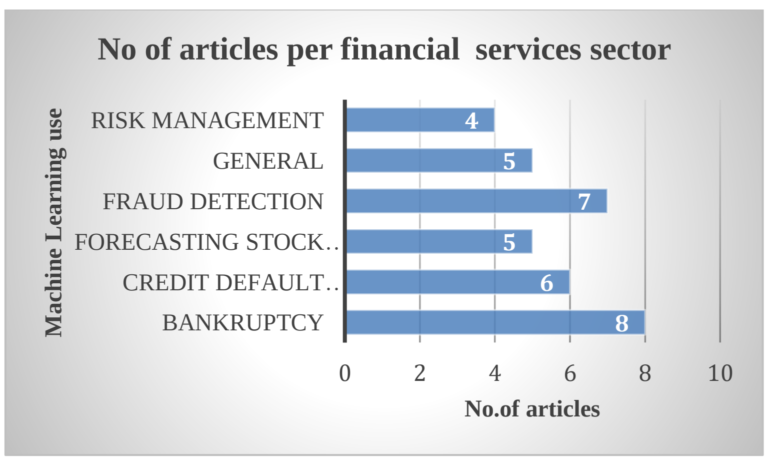


Figure 2: Number of articles included per financial services sector.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Figure 3 shows the common machine learning models per financial services sector, with the hybrid or ensemble model commonly used in Bankruptcy prediction and credit default prediction. Deep learning and neural networks are commonly used in risk management.

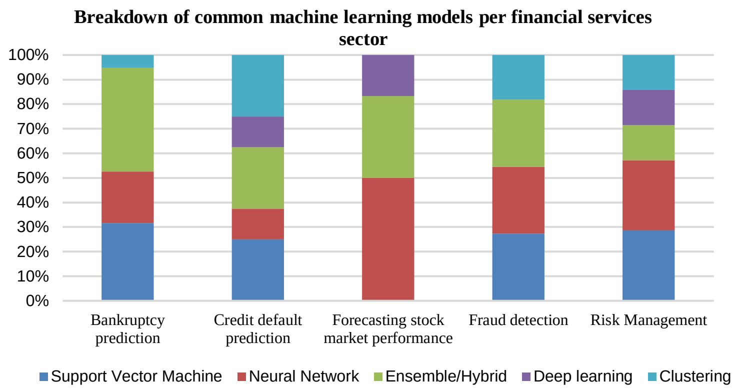


Figure 3: Common machine learning models per financial institution

4.1 Machine language uses in the financial services sector

4.1.1 Bankruptcy

Financial institutions, especially banks and credit bureau agencies, use machine learning to predict bankruptcy for their clients. Predicting bankruptcy assists financial institutions in preventing financial losses through the granting of loans to risky entities. Support Vector Machine (SVM) is generally used as a starting point for bankruptcy prediction; however, due to their limitation in handling large complex data, hybrid and ensemble models are often used because of their superiority when it comes to predictive accuracies (Xie et al., 2013; Lu et al., 2015).

4.1.2 Fraud prediction

Different crimes are committed for financial gain, such as tax evasion and money laundering. Money laundering is one of the main challenges banks seek to address. The ability to accurately identify fraudulent activities or predict fraudulent activities is what ML offers as a solution to reduce the occurrence of fraudulent activities which affect the banks. According to Chen et al. (2018), the identification of fraudulent activities is not a simple task since, in reality, there are fewer fraudulent transactions than fraudulent ones, and the identification of fraudulent activities requires going through large volumes of transactions, particularly for SVM. It is argued that the complexity of transactions data makes SVM to poorly perform, and it makes ANN very difficult to implement due to poor data quality issues (Chen et al. 2018;Aslam et al., 2022; Roseline et al., 2022; Lokanan & Sharma, 2022;Aslam et al., 2022;Canhoto, 2021) . Moreover, the unavailability and poor data quality due to compounding changes in transaction behaviour make ensembles and deep learning perform poorly. The risk of false positives is also high, which threatens banks as they stand to annoy customers(Khang et al., 2021) by constantly erroneously flagging them.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

4.1.3 Credit Default Prediction

Knowing the credit score and default prediction is paramount for organisations in the financial services sector as they need to accurately determine which clients will be able to repay their loans so that they only loan to clients with satisfactory credit scores or records. Financial service institutions rely on accurate and efficient credit default prediction for the profitability and operations of their institutions. According to Shi & Xu, 2016; Moula, Guotai, & Abedin;2017 & Liu et.al,2022, the ability to accurately predict whether credit applicants will be able to pay back the money is a key to determining how much to loan them. Customer data quality issues (Twala,2013) make it difficult to rely on some of the ML models because of the risk of inaccurate decisions.

4.1.4 Forecasting stock market performance

Machine learning is also used to predict stock market movements and assist potential buyers in deciding on when to buy stock or commodity and when to sell. This decision is also essential for fund managers, stockbrokers and investment analysts at large (Kraus & Feuerriegel, 2017; Sigo, Selvam, Venkateswar, & Kathiravan, 2019). The data in stock market conditions are highly volatile; as a result accurate prediction is required to accommodate the volatility. The ML models often implemented for such purposes are often **A**rtificial **N**eural **N**etwork (ANN), Ensemble, and deep learning (Kraus & Feuerriegel, 2017; Sigo et al., 2019). However, the accuracy of these models is negatively affected by unpredictable situations such as stock price volatility. In addition, the technical analysis models exclude factors such as regional politics and transaction costs.

4.1.5 Risk management

The financial services sector faces many challenges which include the implementation of new technological solutions such as ML and how it affects its processes and the way to comply with regulations. When it comes to fraud prediction, accuracy of flagging the exceptions is pivotal as resources are dedicated to investigating the suspected fraudulent transactions. Incorrect identification of fraudulent transactions could result in wasteful expenditure and resources(Zhang et al., 2021). Hence financial institutions need to conduct risk assessments to determine if ML is successfully implemented in an organisation or not before they can extend its scope. The ML models especially the predictive ones (Horak et al. 2020) are not perceived to be perfect as they may omit certain indicators such as risky decisions and unfavourable business markets.

ML techniques are also used for other uses which can be regarded as general use such as house prices prediction((Forys,2022) and airline prices prediction (Vadlamani et.al,2022). In such instances a hybrid of models may be applied to reach desirable outcomes.
According to Mahalakshmi et al.,(2022) , ML and AI are commonly used in the financial sector in document analysis (Mahalakshmi et al., 2022); of which in our study falls under the general use as there are various documents which can be analysed for different purposes.

Based on the outcome of the analysis, it was found that most of the evidence focuses more on the internal ML implementation but is not necessarily specific on the impact it has on the specific sector where ML is applied. The remainder of the literature found is in the form of evidence which focuses a lot more on the challenges of applying ML, which is not business model specific. The challenges were grouped into six thematic themes, as depicted in Figure 4, which were aligned to the ML models. These themes were merged into one all-encompassing general theme: The research into the implementation of machine learning models is still very academic and focused on the intrinsic challenges of individual models than their impact on the sector, such as regulation, transparency, and skills shortage. The general theme was to address the research question, which asks: What are the issues, challenges and impacts of implementing Machine learning in each financial business context?

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

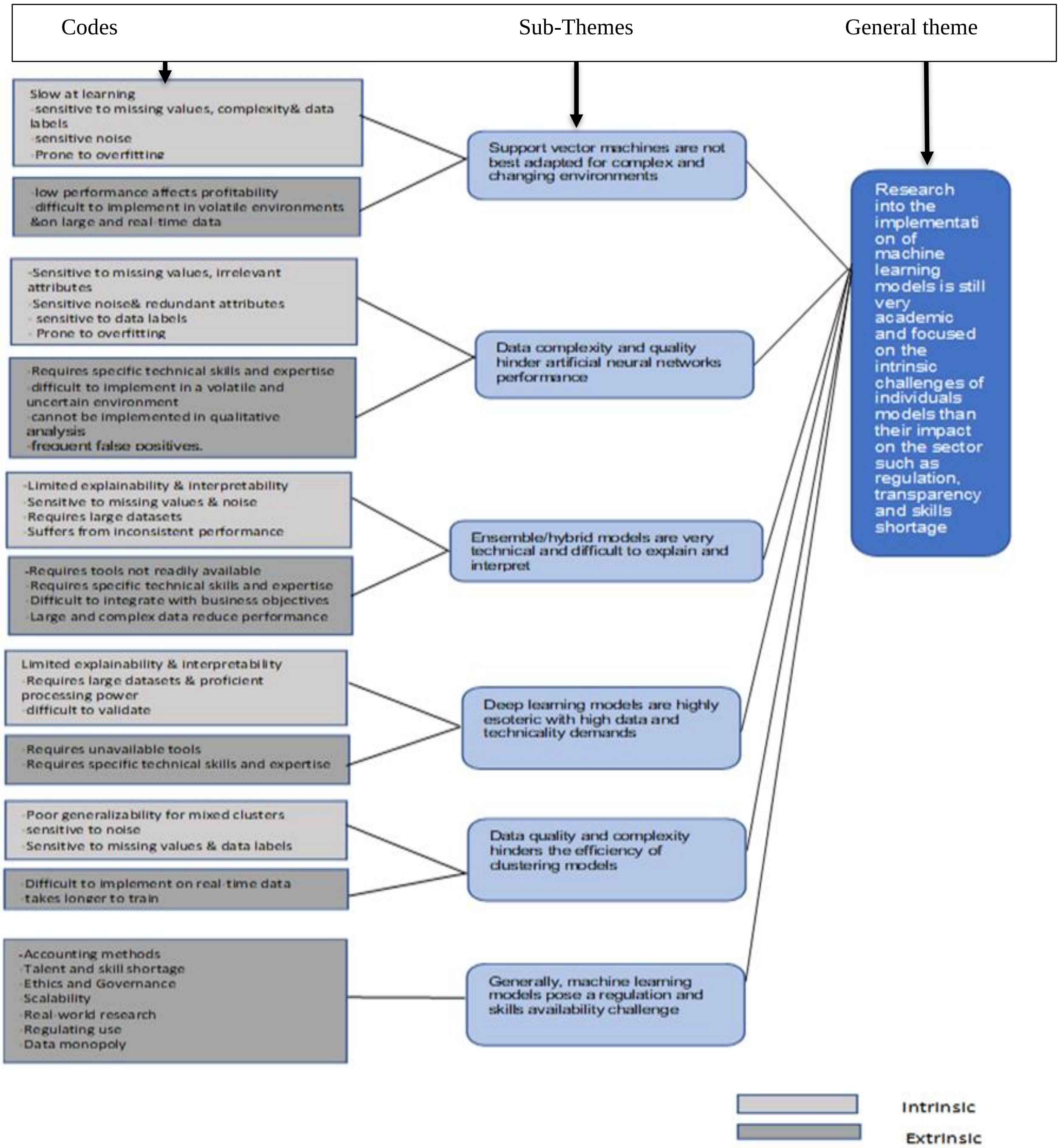


Figure 4: Machine learning themes: Issues, Challenges, and impacts

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Figure 4 outlines the six themes identified, with five clusters of codes associated with the intrinsic challenges of each model and six extrinsic or systemic challenges due to ML implementation. For each ML model, we combined the codes into sub-themes.

The six themes that emerged are as follows:

The six themes that emerged are as follows:

- *Support vector machines are not very suitable to implement for complex and changing environments.*

The issues around the implementation of SVM are of handling the missing data values and the noise and sensitivity to data labels. The SVMs seem ideal for performance in more static environments.

- *The complexity of data and quality issues hinder artificial neural networks performance*

Data is fundamental in generating relevant insights in the financial context, however, the sensitivity to noise, data labels and redundant attributes often temper the optimum performance of artificial neural networks.

- *Ensemble/hybrid models are often technical and cumbersome to explain and interpret.*

In performing certain types of calculations, hybrid models are often used to strengthen the functioning of the models to achieve the desired outcomes. However, when it comes to ensemble/ hybrid models, the primary concern is that these models require specific technical skills to implement. A financial institution will thus need to invest on the acquisition or training of resources who will successfully implement or operationalise these models. From a productivity perspective, ML and AI have been seen to enhance employee productivity or work outcomes in organisations (Ramachandran et al., 2022)

- *Deep learning models are highly esoteric with high data and technicality demands*

The implementation of deep learning models also required the use of resources with the relevant technical skills. From a functionality perspective, the implementation may also require the use of tools which may not be readily available.

- *Data quality and complexity hinders the efficiency of clustering models*

Clustering models are also sensitive to noise, missing values and data labels. They may also be difficult to implement in a changing environment.

- *Generally, machine learning models pose a regulation and skills availability challenge.*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Machine learning methods pose a challenge with respect to the skills that are required to implement these models and the ability to regulate the use of ML as the scope of use often require the use of personal information which pose a challenge from an ethics and governance point of view. It remains a challenge on how the evolution and implementation of ML can be implemented.

Lastly, the six sub-themes were merged into a single general and all-encompassing theme which states that "research into the implementation of machine learning models is still very academic and focused on the intrinsic challenges of individuals models than their impact on the sector such as regulation, transparency, and skills shortage. The general theme was to address the question by encapsulating the answer to the main research question.

5.Conclusion

The study aimed to understand the different business contexts within the financial services sector where ML techniques are implemented. It was found machine learning is predominantly implemented in bankruptcy prediction, followed by credit default prediction, forecasting of stock market performance, fraud detection, and risk management. The ML methods/techniques that were generally implemented in financial sectors include highly complex artificial neural networks and ensemble models, which are the most implemented models across all sectors, followed by support vector machines, clustering, and deep learning techniques. Numerous issues, challenges and impacts were identified in Figure 4. They are mainly concerned with the performance of ML, which is linked to the quality of the data that is used and the complexity of the calculations which are performed. The more complex the ML model is, the more likely the complexity of the analysis will be. Prediction inefficiency and inaccuracy can negatively impact the financial institution's operations, finances, and overall reputation; hence it is essential that risk management is done when implementing ML. Overall, Machine learning research is still very academic and focused on the intrinsic challenges of individual models than their impact on the sector, such as regulation, transparency, and skills shortage. The study was limited in the sense that only specific keywords strings were used, three database sources were consulted, and if the scope was widened, it could have resulted in deeper insights. In addition, future research can look at specific external challenges of implementing ML, such as operational, adoption & skill shortage.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

References

Agavanakis, K. N., Karpetas, George. E., Taylor, M., Pappa, E., Michail, C. M., Filos, J., Trachana, V., & Kontopoulou, L. (2019). Practical machine learning based on cloud computing resources. *AIP Conference Proceedings*, 2123(1), 020096–1. asn. Software Engineering (pp. 3-4). ACM.

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: A systematic review and agenda for future research. *Electronic Markets*, *26*(2), 173–194.Retrieved from https://doi.org/10.1007/s12525-016-0219-0

Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from artificial intelligence and machine learning. *Research in International Business and Finance*, 62.Retrieved from https://doi.org/10.1016/j.ribaf.2022.101744

Arunachalam, D., Kumar, N., & Kawalek, J. P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review*, *114*, 416–436. Retrieved from https://doi.org/10.1016/j.tre.2017.04.001

Bettany-Saltikov, J. (2010). Learning zone. Learning how to undertake a systematic review: Part 2. *Nursing Standard*, *24*(51).

Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, 131, 441–452. Retrieved from https://doi.org/10.1016/j.jbusres.2020.10.012

Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: *A review- Knowledge and Information Systems*, *57*(2), 245–285.

Devi, S. S., & Radhika, Y. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing*, *8*(2), 133–139.

Drucker, A. M., Fleming, P., & Chan, A.-W. (2016). Research Techniques Made Simple: Assessing Risk of Bias in Systematic Reviews. *Journal of Investigative Dermatology*, *136*(11), e109–e114.Retrieved from https://doi.org/10.1016/j.jid.2016.08.021

Du Jardin, P. (2019). Forecasting bankruptcy using biclustering and neural network-based ensembles. *Annals of Operations Research*, 1–36.

Foryś, I. (2022).Machine learning in house price analysis: regression models versus neural networks. Procedia Computer Science, 207, 435–445.Retrieved from https://doi.org/10.1016/j.procs.2022.09.078

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems, 35*(2), 388–423.Retrieved from https://doi.org/10.1080/07421222.2018.1451951

Hagen-Zanker, J., & Mallett, R. (2013). How to do a rigorous, evidence-focused literature review in international development: A guidance note. *London: Overseas Development Institute*.

Hall, P., Phan, W., & Whitson, K. (2016). *The Evolution of Analytics: Opportunities and Challenges for Machine Learning in Business*. O'Reilly.Retrieved from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/evolution-of-analytics-108240.pdf

Horak, J., Vrbka, J., & Suler, P. (2020). Support Vector Machine Methods and Artificial Neural Networks Used for the Development of Bankruptcy Prediction Models and their Comparison. *Journal of Risk and Financial Management*, *13*(3), 60.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. Retrieved from https://doi.org/10.1126/science.aaa8415

Keenan, C. (2018, April 18). Assessing and addressing bias in systematic reviews. *Meta-Evidence*.Retrieved from http://meta evidence.co.uk/assessing-and-addressing-bias-in-systematic-reviews/

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26.

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38–48. bsu.

Khang, P. Q., Kaczmarczyk, K., Tutak, P., Golec, P., Kuziak, K., Depczynski, R., Hernes, M., & Rot, A. (2021). Machine learning for liquidity prediction on Vietnamese stock market. *Procedia Computer Science*, *192*, 3590–3597. Retrieved from https://doi.org/10.1016/j.procs.2021.09.132

Lagasio, V., Pampurini, F., Pezzola, A., & Quaranta, A. G. (2022). Assessing bank default determinants via machine learning. *Information Sciences*, *618*, 87–97. Retrieved from https://doi.org/10.1016/j.ins.2022.10.128

Liu, Y., Yang, M., Wang, Y., Li, Y., & Xiong, T. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis*, *79*. Retrieved from https://doi.org/10.1016/j.irfa.2021.101971

Lokanan, M. E., & Sharma, K. (2022). Fraud prediction using machine learning: The case of investment advisors in Canada. *Machine Learning with Applications*, *8*, 100269.Retrieved from https://doi.org/10.1016/j.mlwa.2022.100269

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 1–22.

Lin, W.-Y., Hu, Y.-H., & Tsai, C.-F. (2011). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 421–436.

Llave, M. R. (2017). Business Intelligence and Analytics in Small and Medium-sized Enterprises: A Systematic Literature Review. *CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017* . Retrieved from https://doi.org/10.1016/j.procs.2017.11.027

Lu, Y., Zeng, N., Liu, X., & Yi, S. (2015). A New Hybrid Algorithm for Bankruptcy Prediction Using Switching Particle Swarm Optimization and Support Vector Machines. *Discrete Dynamics in Nature and Society*, *2015*. Computer Science Database; Middle East & Africa Database; Publicly Available Content Database. Retrieved from https://doi.org/10.1155/2015/294930

Mahalakshmi, V., Kulkarni, N., Pradeep Kumar, K. v., Suresh Kumar, K., Nidhi Sree, D., & Durga, S. (2022). The Role of implementing Artificial Intelligence and Machine Learning Technologies in the financial services Industry for creating Competitive Intelligence. *Materials Today: Proceedings*, *56*, 2252–2255. Retrieved from https://doi.org/10.1016/j.matpr.2021.11.577

Manlangit, S., Azam, S., Shanmugam, B., & Karim, A. (2019). Novel Machine Learning Approach for Analyzing Anonymous Credit Card Fraud Patterns. *International Journal of Electronic Commerce Studies*, *10*(2), 175–201. bsu.

Marr, B. (2016). *What Is The Difference Between Artificial Intelligence And Machine Learning?* Forbes.Retrieved from https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/

Moula, F. E., Chi Guotai, & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management*, *19*(2), 158–187.

Müller, J. M., Kiel, D., & Voigt, K.-I. (2018). What drives the implementation of Industry 4.0? The role of opportunities and challenges in the context of sustainability. *Sustainability*, *10*(1), 247.

Müller, O., Junglas, I., Brocke, J. vom, & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*, *25*(4), 289–302.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*, 1–13.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, *37*(1), 43.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science*, *162*, 895–899.

Ramachandran, K. K., Apsara Saleth Mary, A., Hawladar, S., Asokk, D., Bhaskar, B., & Pitroda, J. R. (2022). Machine learning and role of artificial intelligence in optimizing work performance and employee behavior. *Materials Today: Proceedings*, *51*, 2327–2331. Retrieved from https://doi.org/10.1016/j.matpr.2021.11.544

Roberts, K., Dowell, A., & Nie, J.-B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Medical Research Methodology*, *19*(1), 1–8.

Shi, J., & Xu, B. (2016). Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function. *Journal of Risk and Financial Management*, *9*(4).

Sigo, M. O., Selvam, M., Venkateswar, S., & Kathiravan, C. (2019). Application of Ensemble Machine Learning in the Predictive Data Analytics of Indian Stock Market. *Webology*, *16*(2), 128–150. Middle East & Africa Database.

Trieu, V.-H. (2017). Getting value from Business Intelligence systems: A review and research agenda. *Decision Support Systems*, *93*, 111–124. bsu.

Twala, B. (2013). Impact of noise on credit risk prediction: Does data quality really matter? *Intelligent Data Analysis*, *17*(6), 1115–1134. asn.

Van Liebergen, B. (2017). Machine learning: A revolution in risk management and compliance? *Journal of Financial Transformation*, *45*, 60–67.

Xie, G., Zhao, Y., Jiang, M., & Zhang, N. (2013). A Novel Ensemble Learning Approach for Corporate Financial Distress Forecasting in Fashion and Textiles Supply Chains. *Mathematical Problems in Engineering*, *2013*. Computer Science Database; Middle East & Africa Database; Publicly Available Content Database. Retrieved from https://doi.org/10.1155/2013/493931.

Yao, J., Wang, Z., Wang, L., Zhang, Z., Jiang, H., & Yan, S. (2022). A hybrid model with novel feature selection method and enhanced voting method for credit scoring. *Journal of Intelligent and Fuzzy Systems*, *42*(3), 2565–2579.Retrieved from https://doi.org/10.3233/JIFS-211828

Zhang, C., Li, M., & Li, Y. (2021). Financial risk analysis of real estate bubble based on machine learning and factor analysis model. *Journal of Intelligent and Fuzzy Systems*, *40*(4), 6493–6504. Retrieved from https://doi.org/10.3233/JIFS-189488.

Vadlamani, S. L., Shafiq, M. O., & Baysal, O. (2022). Using machine learning to analyze and predict entry patterns of low-cost airlines:*A study of Southwest Airlines. Machine Learning with Applications*, 10, 100410.Retrieved from https://doi.org/10.1016/j.mlwa.2022.100410

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

mobileDNA application used to explore location information

Fallo Khanye[1], Dr Julia Keddie[1], Prof Renette Blignaut[1], Prof dr Lieven De Marez[2] and Simon Perneel[2]
[1] University of Western Cape, Cape Town, S.A.
[2] imec-mict-ugent, Gent University, Belgium
3653309@myuwc.ac.za, jkeddie@uwc.ac.za, rblignaut@uwc.ac.za, Lieven.DeMarez@ugent.be, Simon.Perneel@ugent.be

Abstract

Global positioning systems (GPS) are more trustworthy than other techniques for gathering location data and have opened new research opportunities. Researchers and partners of Media, Innovation and Communication Technology at Gent University are interested to know how Belgians utilise specific applications on their smartphones. The purpose of this study is to investigate mobileDNA users' behaviour in terms of where, when, and how they utilised their smartphone daily. MobileDNA is a free smartphone logging application that was used as the data collection tool for this research. The GPS metadata is collected every 15 minutes by mobileDNA. We discovered that the application usage sequence of users on a daily path varies from day to day, and that most users tend to travel to more than one city or town in a day. We predicted home of the mobileDNA users and compared smartphone usage at home and outside home and discovered that users spent more time on their smartphone outside their home compared to when they are at home.

1 Introduction

We aim to conduct an exploratory analysis and engineer new features using Python, we used the following packages: pandas; plotly; reverse geocoding and haversine. All insights, patterns and trends form the exploratory analysis are documented in this report. The data for this project will include users that have downloaded the mobileDNA application (app) in Belgium and have been active on the app for more than one month in 2022 (January to June). We were able to successfully determine the city/town, region, province, and country of each Global Position System (GPS) point in our dataset, which aided in identifying individuals who only used the mobileDNA application in Belgium. We were also able to identify how long the user had been actively using the app which aided in identifying users

Coastlands Umhlanga Hotel & Convention Centre
Durban, South Africa
ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

who had used the app for more than a month. Finally, we were able to predict the users' homes and the number of times they moved every day.

Specific focus was given to the visual representation of the GPS location data. This visual representation included visualisations of the number of mobileDNA users in each province in Belgium; top application categories used at home and outside home; sequence of smartphone application usage at home and outside home; density map outlining where user spent most of time using their smartphone and lastly, a cumulative plot showing the movement of the user in a day. The sequence of smartphone usage and the movement of user diagrams can offer us a sense of where, when, and how mobileDNA users use their smartphones daily.

The gaps in the location data recorded by the mobileDNA application is caused by the smartphone being switched off, when the mobile is not used and inconsistence of Huawei to capture GPS data of users.

2 Methodology

The smartphone location data (GPS metadata) collected consists of the following variables: Name of application opened on the smartphone; Duration of the application usage; Receipt of message (or notification); Type of smartphone; Battery charge level; and GPS location where application was used.

The first step was to create the fancyname feature, and the category features from the application feature in Table 1. The "fancyname" represents the commonly used name for an application, while the "category" represents the group to which the application belongs. Each application is uniquely classified into one category.

Table 1: Example dataset

Id	startTime	endTime	application	Battery	Latitude	Longitude
00441003-8202-4ffd-b9aa-7cec9674906f	2022-01-01 03:28:07.982	2022-01-01 03:28:18.483	com.wssyncmldm	28	51.158852	4.155556
00441003-8202-4ffd-b9aa-7cec9674906f	2022-01-01 03:32:29.099	2022-01-01 03:32:30.349	com.instagram.android	28	51.158852	4.155556
00441003-8202-4ffd-b9aa-7cec9674906f	2022-01-01 03:32:34.414	2022-01-01 03:32:50.264	com.facebook.katana	28	51.158852	4.155556

Dates were extracted from the startTime and endTime in Table 1 and labelled startDate and endDate, respectively. The month, day, and hour were then extracted from only the startTime to help in conveniently filtering our data. Table 1's endTime and startTime were also used to create the duration feature. We first created GPS coordinates from our dataset's longitude and latitude variables, and then used reverse geocoding in Python to derive the city, region, province, and country for each GPS coordinate.

In our location data, we engineered activation and deactivation dates from startDate and the date when the user was active for the last time. We discovered that the average time users stay on the application was approximately 4 months.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

To predict a user's home, we assumed the users to be home between 20:00 and 08:00 in the morning the next day on weekdays. We defined home as the most frequent GPS coordinate in this period and allowed for movement within a home.

3 Results

Applications were categorised into groups to get an overview of which group of applications is mostly used at home and outside home. We focused on the following categories: *chat*; *gaming*; *social*; *news and magazines*; and *banking*. *chat* contains applications that are more focused on sending or receiving messages such as WhatsApp; WeChat and Messenger, *gaming* contains applications such as lichess and puzzles, *social* contains applications that are more focused on posting pictures and captions such as Facebook; Instagram and LinkedIn, *news and magazines* category contains applications such as BBC news and lastly *banking* involves applications that more focused on transactions such as Google Pay; Finance Exchange and ING Smart Banking.

We can observe from Figure 1 that *chat* was the most used application category with an average of 361 daily users, while *social* was ranked fifth with an average of 282 daily users. *banking* category was ranked ninth with an average of 208 daily users.

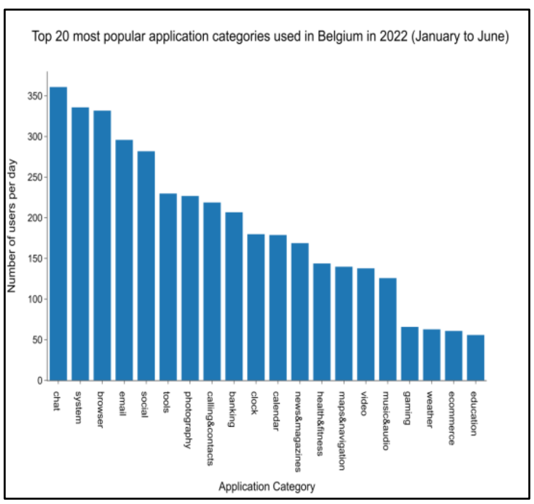


Figure 1: Top 20 most popular application categories

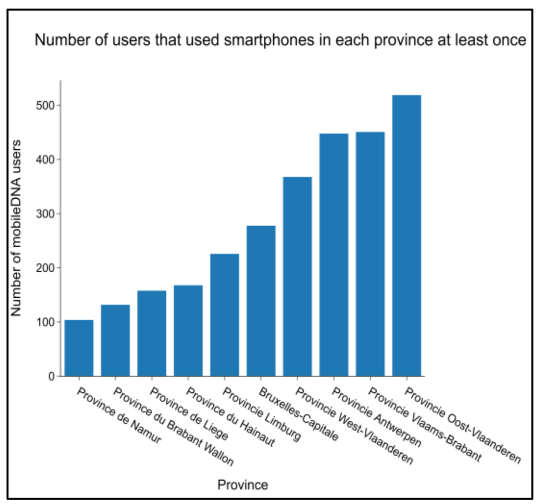


Figure 2: No of unique users in each province

Figure 2 shows that the province of Oost-Vlaanderen had the most unique visitors (519 users), while the provinces of Hainaut, Liege, Branbant Wallon, and Namur were visited by fewer than 200 users, with Namur being visited by the least number of users (104 users).

The average time mobileDNA users spend at home using their smartphones is 112.83 minutes per day while usage outside the home is 68.39 minutes.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

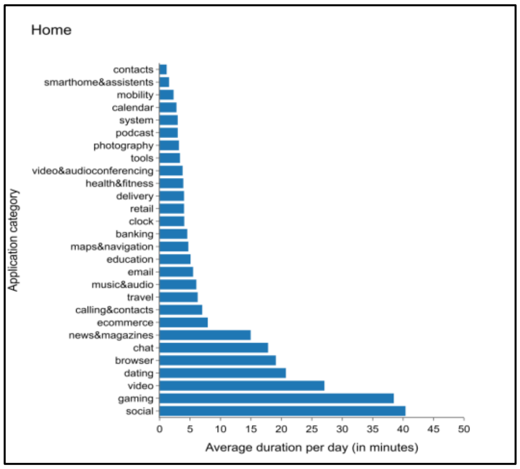


Figure 3: Average duration per application used at home

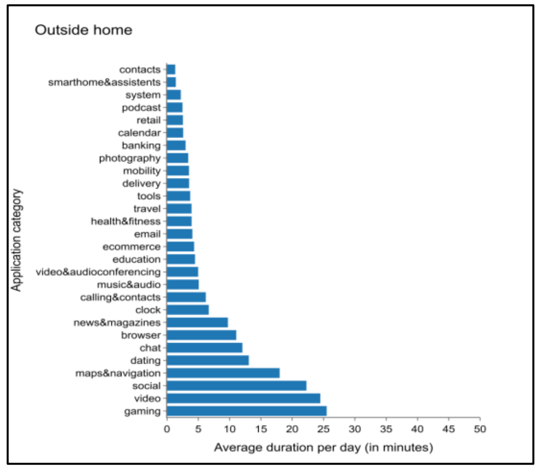


Figure 4: Average duration per application used outside home

Figure 3 shows that *social* is the most popular application category used at home, with an average of 40.41 minutes per day, followed by the *gaming* application with approximately 38.49 minutes. Figure 3 also shows that the applications classified as *video* is ranked third with an average duration of 27.10 minutes per day, followed by *dating* (ranked fourth with an average duration of 20.76 minutes per day), *chat* (ranked sixth with an average duration of 17.83 minutes per day), and *banking* (ranked fifteenth with an average duration of 4.57 minutes per day).

Figure 4 shows that the two most popular application categories used outside the home are *gaming* (ranked first with an average duration of 25.52 minutes per day) and *video* (ranked second with an average duration of 24.50 minutes). We can see that the difference in average duration between *video* and *gaming* is less than 1 minute per day. Figure 4 also shows that *chat* is ranked sixth with an average duration of 12.07 minutes per day, followed by *news and magazines* (ranked eighth with an average duration of 9.75 minutes per day) and *banking* (ranked twenty-second with an average duration of 3.00 minutes per day).

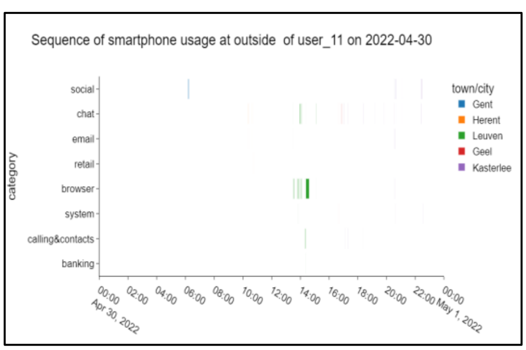


Figure 5: Sequence outside the home on 30 April 2022

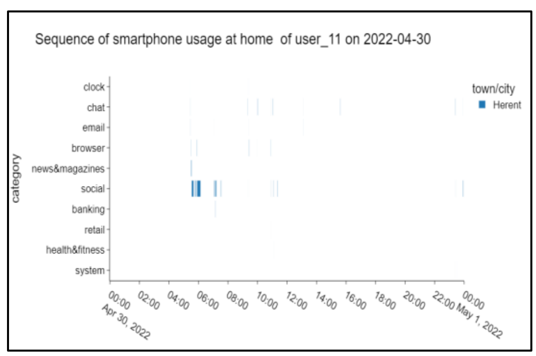


Figure 6: Sequence at home on 30 April 2022

Now we will concentrate on the sequence of smartphone usage of each user, but for the purposes of this report, we will focus on *user 11*. Figure 5 shows that on April 30, 2022, this user visited five cities, namely, Herent, Gent, Leuven, Geel, and Kasterlee. We can also see that *user 11* used their smartphone outside their home only once in the morning, then again after 12:00. Figure 5 also reveals that *user 11*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

mostly used the *browser* application category between 12:00 and 16:00. *User 11* utilised applications related to *email, calling and contacts* only once on April 30, 2022.

Figure 6 shows that *user 11'*s home is in the city of Herent. Figure 6 also shows that *user 11* used their smartphone at home frequently between 5:00 and 6:08 and again between 7:00 and 12:00, the most utilised application categories at home on April 30, 2022, is *social*. *user* 11 also used their smartphone in Gent between 6:10 to 6:13.

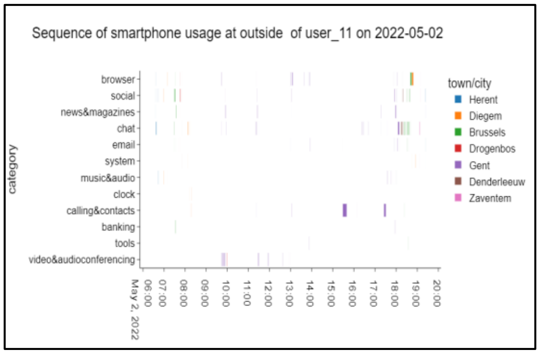


Figure 7: Sequence outside home on 02 May 2022

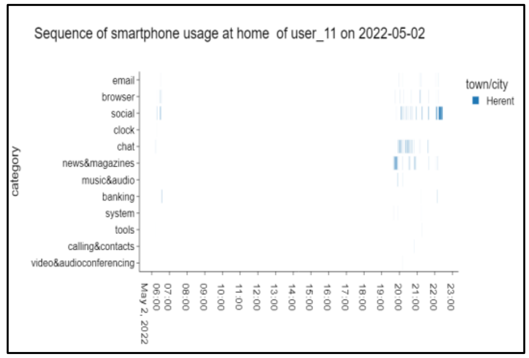


Figure 8: Sequence at home on 02 May 2022

Figure 7 shows that *user 11* travelled to more cities on May 02, 2022 (weekday), than on April 30, 2022. Figure 7 also shows that *user 11* started using their smartphone at 06:00 and stopped using it about 20:00. On May 02, 2022, *user 11* spent more time on *social* and *chat* categories.

According to Figure 8, *user_11* utilised their smartphone at home firstly between 06:00 and 07:00 and between 19:00 to 23:00. We can also observe from Figure 8 that the most utilised application categories at home were *social*; *chat* and *news & magazines* on 02 May 2022.

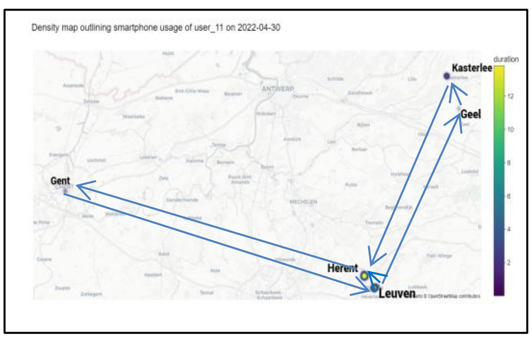


Figure 9: Density map *user 11* on 30 April 2022

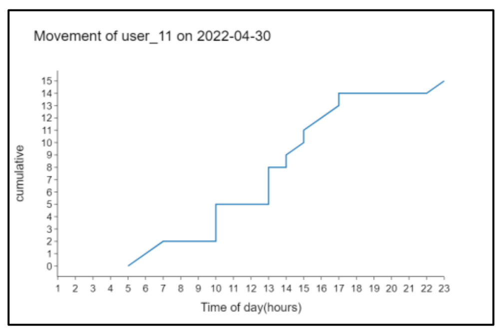


Figure 10: Movement of *user 11* on 30 April 2022

From Figure 9 we see that on 30 April 2022 *user 11* spent more time on his smartphone in the cities of Herent and Leuven since the marking in these locations is yellow, indicating the longest duration there. Figure 9 also shows that this user travelled from Herent (home) to Gent, with distance of about 82 kilometers (km) between the two cities. From Gent, the user travelled to Leuven, then to Herent, which was closer to Leuven (about 5 km between the two cities), and finally to Geel, which was about 50 km from Herent. The user travelled from Geel to Kasterlee and then back to Herent, which is home (Figure 6).

Figure 10 shows that *user 11* begins traveling at 5:00 on April 30, 2022. Figure 10 also shows that the user did not move for roughly three hours between 7:00 and 10:00, then moved three times in the

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

tenth hour and did not move again after 10:00 until 13:00. The user resumed movement from 13:00 to 17:00 and then ceased movement from 17:00 to 22:00 (as we saw from Figure 6 that user was home at this period) and finally, at 23:00, the user moved once. In total *user 11* moved 15 times on this day.

4 Conclusion

Our data was successfully loaded to Python from which we were able to find the city; region; province; and country names for each GPS coordinate in the dataset using reverse geocoding package. We were able to find the distance between GPS coordinates using the haversine formula and this assisted in predicting home of users and how many times did they move per day.

Using our visualisations we were able to successfully create interactive figures that can be filtered with the user and date in our Python script (Figures in Python filtered for only *user 11*).

We were successful in predicting mobileDNA users' homes, and found that the average time spent on a smartphone at home each day 112.83 minutes which is larger than the average time spent on a smartphone outside the home, 68.39 minutes (that is 44.44 minutes less).

We also discovered that the *social* category was utilised more at home (40.41 minutes) than outside the home (22.29 minutes), with a difference in average time spent of around 18 minutes, which was relatively large. The gap between the first and second most utilized application categories at home (*social* and *gaming*, respectively) was large (approximately 2 minutes) relative to the difference between the first and second application categories (*gaming* and *video*, respectively) outside of home which was around 1 minute. Because the *gaming* category was one of the top two most utilised categories both at home and away from home, it is interesting to note that mobileDNA users in Belgium enjoy playing games on their smartphones. *chat* and *news and magazines* categories had the highest average duration per day at home compared to outside home, whereas *banking* and *social* had higher average daily durations outside the home than at home.

Sequence of smartphone usage of a user can give us an idea of where, when and how the users use their smartphones every day. For *user 11*, we showed the number of places this user visited, when the user visited those cities, and which application categories were utilised in those locations by examining the sequence of smartphone usage of *user 11* (Figure 5 and Figure 6). We also determined that *user 11's* home is in the city of Herent (Figure 6) and that this user spent more time on his smartphone at home than outside home on April 30, 2022.

Figure 9, the density map can provide us with an indication of the distance between the locations mobileDNA users visited; hence, it addresses the issue of Figure 5, which simply displays the locations visited rather than where the places are situated.

Movement of a user (Figure 10) can assist in determining the number of places users visit daily.

Some the gaps in Figure 5; Figure 6; Figure 7 and Figure 8 were caused by the inactive time of the mobileDNA users (e.g. users not using their smartphones). We recommended to collect location data every 15 minutes irrespective if the user was using their smartphone to allow more in-depth analysis of movement of users.

We have discovered that most users stay on mobileDNA for roughly 4 months, with some spending less than a month. Hence, we recommend that Media, Innovation and Communication Technology at Gent University conduct a survey to understand why some users deactivate from the app.

Further research is needed to cluster mobileDNA users based on GPS location and then determine whether there are any commonalities in the smartphone usage of those users. This will also help establish where the majority of mobileDNA users' homes are situated and what time of day each location has the most users.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

References

Ali, S. M., Gupta, N., Nayak, G. K. & Lenka, R. K., 2016. Big data visualization: Tools and challenges. *2nd International Conference on Contemporary Computing and Informatic (IC3I),* December.pp. 656-660.

Chopde, N. R. & Nichat, M. K., 2013. Landmark based shortest path detection by using A* and Haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering,* 1(2), pp. 298-302.

Del Rosario, M. B., Redmond, S. J. & Lovell, N. H., 2015. Tracking the Evolution of Smartphone Sensing for Monitoring Human Movement. *Sensors,* 15(8), pp. 18901-18933.

Do, T. M. T., Blom, J. & Gatica-Perez, D., 2011. Smartphone usage in the wild: a large-scale analysis of applications and context. *In Proceedings of the 13th international conference on multimodal interfaces,* pp. 353-360.

El-Rabbany, A., 2002. *Introduction to GPS: The Global Positioning System.* illustrated ed. Norwood: Artech House.

Exelmans, L. & Van den Bulck, J., 2016. Bedtime mobile phone use and sleep in adults. *Social Science & Medicine,* 148(1), pp. 93-101.

Falaki, H. et al., 2010. Diversity in smartphone usage. *In Proceedings of the 8th international conference on Mobile systems, applications, and services,* pp. 179-194.

Ghent University, 2017. *mobileDNA - Application that shows your smartphone usage.* [Online] Available at: https://www.ugent.be/mict/en/research/mobiledna [Accessed 12 10 2022].

Kays, R., Crofoot, M. C., Jetz, W. & Wikelski, M., 2015. Terrestrial animal tracking as an eye on life and planet. *Science,* 348(6240).

Korpilo, S., Virtanen, T. & Lehvävirta, S., 2017. Smartphone GPS tracking—Inexpensive and efficient data collection on recreational movement. *Landscape and Urban Planning,* Volume 157, pp. 608-617.

Li, H. et al., 2015. Characterizing smartphone usage patterns from millions of android users. *In Proceedings of the 2015 Internet Measurement Conference,* pp. 459-472.

McKenzie, G. & Janowicz, K., 2015. Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems,* Issue 54, pp. 1-13.

McKinney, W., 2015. Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit,* Volume 1625.

Okilas, A. F., Siswanti, S. D. & Rachman, M. D., 2014. Location Based Service For Information Publication Using GPS on Android Based Mobile Phone. *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014),* pp. 190-197.

Oshin, T. O., Poslad, S. & Ma, A., 2012. Improving the Energy-Efficiency of GPS Based Location Sensing Smartphone Applications. *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications,* pp. 1698-1705.

Wagner, D. T., Rice, A. & Beresf, A. R., 2014. Device Analyzer: Understanding smartphone usage. *Mobile and Ubiquitous Systems: Computing, Networking, and Services,* 1(978-3-319-11569-6), pp. 195-208.

Wisbey, B., Montgomery, P. G., Pyne, D. B. & Rattray, B., 2010. Quantifying movement demands of AFL football using GPS tracking. *Journal of science and Medicine in Sport,* 13(5), pp. 531-536.

Carlisle, D. (2010, April). *graphicx: Enhanced support for graphics.* Retrieved from http://www.ctan.org/tex-archive/ help/Catalogue/entries/graphicx.html

Voronkov, A. (2004). *EasyChair conference system*. Retrieved from easychair.org

Voronkov, A. (2014). Keynote talk: EasyChair. In *Proceeedings of the 29th ACM/IEEE International Conference on Automated Software Engineering* (pp. 3-4). ACM.

Wikipedia. (n.d.). *EasyChair.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/EasyChair

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Sentiment Analysis of South African tweets about COVID-19 vaccines

Thato Ditsele[1], Wallace Chigona[1] and Malebo Sephodi[1]
[1] University of Cape Town, Cape Town, South Africa
dtstha004@myuct.ac.za,
wallace.chigona@uct.ac.za,gllmal003@myuct.co.za

Abstract

The COVID-19 pandemic and lockdowns have led to a surge in the use of social media for information sharing and learning about the virus and the vaccine. Our study aimed to understand the sentiments of South African Twitter users towards COVID-19 vaccines. Using the Twitter API version 2, we collected a total of 21,084 tweets from 1 January 2021 to 31 December 2021, during the government's rollout of the vaccine. A sentiment analysis was performed using the VADER lexicon-based classifier to categorise the tweets into positive, neutral, and negative sentiments. The results showed that 40% of the tweets were positive towards the vaccine, 32% were neutral, and 28% were negative. The analysis also revealed that people expressed their opinions on vaccinations more frequently during the early months of the year (January-March 2021) in response to the government's announcement for the vaccine rollout. However, the attitudes towards the vaccine changed throughout the year, indicating that people were sceptical of the government's vaccine rollout strategy, which could have affected the overall vaccine adoption. The findings of this study can provide valuable insights for policymakers and healthcare organisations in shaping effective strategies for promoting vaccine adoption.

Keywords— Sentiment analysis, social media, vaccine hesitancy South Africa 2021, Twitter-Sentiment analysis, public opinions, Covid-19 vaccines

1 Introduction

The global impact of the Corona virus disease (COVID-19) since its emergence in China in December 2019 has been significant, leading to the widespread spread of the virus. South Africa reported its first COVID-19 case in February 2020 and has since reported over 3.7 million confirmed cases, with approximately 103,000 deaths (Department of Health, 2022). In response, the South African government employed a 5-level alert system to handle the gradual relaxation of the lockdown, aimed at controlling the spread of the virus (Department of Health, 2022).

Vaccination has been widely recognised as a critical tool in preventing the spread of COVID-19, with human studies indicating that COVID-19 vaccines are 89% safe, with a low to moderate risk of serious disease among older individuals (Bam, 2021). Despite this, there has been growing vaccine

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

scepticism in South Africa, with only 67% of those polled expressing certainty in getting vaccinated and the remaining 33% uncertain or against getting vaccinated (Bam, 2021).

The South African government's failure to prioritise public data in the early days of the COVID-19 outbreak was identified as a mistake that impacted policymakers' ability to deal with the pandemic effectively, leading to a loss of public trust in the government's ability to manage the vaccine rollout (Marivate & Combrink, 2020; Puri et al., 2020). The increasing concern about the impact of social media in increasing anti-vaccination sentiment and low vaccine uptake is a major concern among public health professionals (Rotolo et al., 2022).

Twitter data provides real-time insights into public opinion on a variety of pandemic-related issues, including vaccinations, and has been used to determine South African sentiments about COVID-19 vaccines throughout 2021. Our study aimed to answer the following research questions:

1. What are the sentiments expressed in South African tweets about COVID-19 vaccines?
2. What are the central themes discussed on Twitter about COVID-19 vaccines in South Africa?
3. How did South Africa's Twitter sentiments on COVID-19 vaccines change throughout 2021?

Our study used a methodology that involves opinion mining and sentiment analysis, and its results can be applied in multidisciplinary investigations involving social media data analysis.

2 Literature Review

2.1 Twitter social media platform

Twitter is a real-time microblogging platform that allows individuals or groups to share their opinions and perspectives on various topics through multimedia, text, and other forms of media. The messages shared on Twitter are known as "tweets" and are displayed on a timeline, a collection of tweets in chronological order (Pawar et al., 2015; Shoaei & Dastani, 2020). The availability of data through the Twitter API v2 enables researchers to examine Twitter data in depth, offering valuable insights into public opinions and sentiments (Mohamed Ridhwan & Hargreaves, 2021).

Twitter is a popular source of health care information in South Africa, with around 8 million users among social media networks (van Heerden & Young, 2020; Hoffman et al., 2021). During the COVID-19 pandemic, a tweet about the virus was reported every 45 milliseconds on Twitter, and the hashtag #coronavirus was one of the most frequently used globally in 2020 (Puri et al., 2020). The legitimacy of a Twitter account is determined by whether it is verified or unverified. Verified accounts are those of genuine public interest, while the legitimacy of unverified accounts is unknown (Mir et al., 2022).

2.2 Use of twitter to determine public sentiments on Covid-19 vaccines

Twitter has become an important source of information for individuals seeking information about health care, including information about vaccines (Hoffman et al., 2021). In the context of the COVID-19 pandemic, Twitter has been used to track public sentiment and reactions to vaccine rollout (Hung et al., 2020). Twitter users engage with one another through direct messaging, updates, replies, likes, and retweets, which can indicate user engagement and reveal their attitudes towards (Kang et al., 2017).

Studies have used Twitter data to gain a better understanding of how misinformation about vaccines, including vaccine-induced acquired immunodeficiency syndrome (VAID), has spread on the platform (Shahi et al., 2020a). This misinformation has led to confusion and misperceptions about vaccines, and researchers have used sentiment analysis to track these attitudes and reactions (Maciuszek et al., 2021; Mønsted & Lehmann, 2022).

Twitter data can be collected and analysed through open academic access via API v2, making it a valuable tool for researchers, government organizations, and health care professionals (Luo et al., 2021). A number of studies have analysed tweets using techniques such as logistic regression, deep learning algorithms, topic modelling, Latent Dirichlet Allocation (LDA), and machine learning to better understand the conversations, concerns, sentiments, and reactions elicited by tweets (Li Han Wong et al.,2022; Amjad et al. 2021; Deverna et al. 2021; Anuratha et al., 2021; Alfatease et al., 2021; Piltch-Loeb et al., 2021).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

In conclusion, Twitter plays a crucial role in shaping public perception and understanding of health care, including vaccines. By analysing Twitter data, researchers and health care professionals can gain a deeper understanding of public sentiment and reactions, and use this information to inform vaccine policy and communication efforts. Vaccine manufacturers can also use sentiment analysis to quickly detect negative feelings towards their vaccine and respond accordingly (Anuratha et al., 2021; Hou et al., 2021).

2.3 Challenges to vaccine acceptance

Misinformation or disinformation, refers to misleading information shared intentionally rather than when erroneous information is transmitted accidentally. Misinformation on the internet has also been linked to reduced vaccination uptake. Other studies noted that misinformation was spread, stating that people are testing HIV-positive after vaccination just like COVID-19. Vaccination sceptics continued to dominate the Twitter dialogue, with information centred on government scepticism of boosters and requirements, intellectual property around the vaccine, and who benefits from the vaccine (Department of Health, 2022). Furthermore, misinformation can erode trust in science and public health authorities, leading to a drop-in vaccine uptake (Steffens et al., 2019; Bonnevie et al., 2020b).

Whereas unpleasant effects (31%) and concern about their effectiveness (21%) were common, loss of trust (18%) and anti-vaccination feelings were uncommon (14%) (Burger et al.,2021). Fears about mandated vaccinations appearing on social media can cause fear and scepticism about vaccines, according to the Africa CDC (2021). Conspiracy theories and religious ideas published on social media are two common types of misinformation that influence user perception (Bam, 2021; Chou et al., 2021; Shahi et al., 2020a).

Vaccine hesitation / hesitancy describes the refusal or postponing of vaccination despite its availability (Chadwick et al., 2021; Hoffman et al., 2021; Wilson & Wiysonge, 2020). South Africa's COVID-19 vaccine hesitancy can be addressed through collective efforts to ensure that there are no constraints related to the supply and system that administers vaccines. The data creates a picture of the vaccine debate on Twitter as highly polarised, with individuals who express similar views on vaccinations more likely to communicate with one another and post content from comparable sources. In other studies, developed economies such as Singapore rely heavily on their government to provide accurate information and demonstrate trust in science and their vaccine program (Mohamed Ridhwan & Hargreaves, 2021).

Pro- and anti-vaccine information may spontaneously separate into distinct groups, possibly as a result of social media self-selection that brings together communities with similar viewpoints (Puri et al., 2020). Vaccination opponents and proponents can be classified as social groups rather than by more specific criteria such as gender, nationality, or affiliation. People in a variety of countries voiced scepticism COVID-19 vaccines, as well as serious worries about vaccine safety and distrust in governments and doctors (Hou et al., 2021; Puri et al., 2020). Some researchers use sentiment analysis scores to infer that positive sentiment is pro-vaccine and negative sentiment is anti-vaccination (Lyu et al., 2022; Maciuszek et al., 2021b).

3 Methodology

Twitter sentiment analysis is the process of analysing and categorising opinions expressed in Twitter posts (tweets) into positive, negative, or neutral classes. The goal is to determine the overall public opinion or emotions towards a specific topic or brand (Alsayed, Alharthi, & Adeyemo, 2019). As illustrated in Figure 1, the study went through five phases (Mohamed Ridhwan & Hargreaves, 2021).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

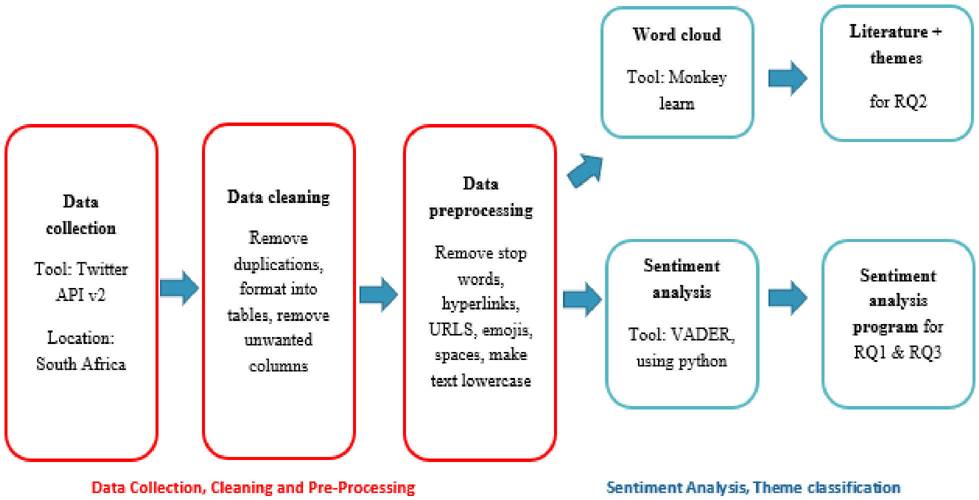


Figure 1: Five phases used in the study

3.1 Phase 1: Data collection

We utilised the Twitter historical tweets downloader, after obtaining developer access to Twitter version 2. Our focus was to search and download historical tweets from South Africa. The keywords selected from trending topics were related to vaccines, Covid-19, and anti-vaccination and were selected based on their relevance and popularity (Li Han Wong et al., 2022). The keywords included:

Vaccineforall, Vaccine, Antivaccine, Vaccinationcovid, Covid19, Vaccine, Shot, covid-19, Coronavirus, AstraZeneca, modernavax, anti-vaccination, anti-vax, anti-vaxxers, pro-vax, covid19jab, vaccinesideeffects, Antivax, Antivaxxer, Antivaxxers, vaccine, Moderna, pfizersideffects, j&jsideaffects, vaccinesideeffects, novaxx

The date range for this study was from January 1st, 2021 to December 31st, 2021.The metadata extracted from the tweets included: tweet content, location, date and time of post, author, language, type of account, account verification status, author friends, retweet count, and likes count. This information was crucial in analysing the tweets and understanding the perspectives of the authors on the topics related to vaccines and Covid-19.

3.2 Phase 2: Data cleaning

The Twitter dataset includes South Africans' location fields, as well as race, gender, age, and languages. We used the keyword 'vacc,' which is a prefix for words like vaccine, vaccinated, vaccinate, and vaccination, to narrow down the results. The tweets were filtered to exclude non-English tweets. Some tweets contain English and other languages but categorised as English by Twitter, which we could not remove entirely. We eliminated 43 656 tweets, leaving 21 084 tweets suitable for data analysis.

3.3 Phase 3: Data selection

The final dataset consisted of 21 084 tweets. We used Python's Panda package to transform the data into a data frame for the study. We used the Natural Language Toolkit (NLTK) to perform tokenization, sentiment analysis (using the NLTK Sentiment Intensity Analyzer), and emotive analysis on the text.

Tokenisation is the process of dividing large text blocks into smaller chunks. In this case, tokens could be letters, numbers, or sub words (Ioana-Andreea et al., 2021). To tokenize, the Word Tokenize package was imported into Python Pandas. The analysis of word sequences aids in understanding the meaning of the text (Mir et al., 2022). All punctuation was removed during standardisation (Kumar & Sebastian, 2012). The tweet id and user id were anonymised.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.4 Phase 4: Sentiment analysis of tweets

VADER is a pre-trained model that employs rule-based values tailored to social media sentiments. It employs dictionary phrases with evaluation capabilities to classify tweets based on the compound score. The compound score was used to decide whether a tweet may be labelled (good, negative, or neutral). A compound score of ≥0.05 satisfies the labels positive, and between -0.05 and 0.05 was neutral and a score of ≤−0.05 was negative (Mir & Sevukan, 2022).We used the Natural Language Toolkit (NTLK) package in the first step of sentiment analysis to tokenize the content data and remove commonly used words. Based on the tokens, a frequency dictionary was created. The Python script then generated a frequency list in descending order. Stop words are frequently used words such as "drink," "please," "guys," or "covid," "corona" are among the top keywords. These words, along with the rest of the NTLK stop words list, were removed to gain insight from the analysis (see Appendix A).

3.5 Phase 5: Word frequency and themes

The data cleansing resulted in 21 084 tweets in English. Using the Python df.sample random sample formula (n=200, random state=1), we obtained a random sample of 200 tweets. Previous research was divided into two categories: pro-vaccination and anti-vaccination. We added a third type of vaccination hesitancy (Maciuszek et al., 2021b; Monsted & Lehmann, 2022b). The central themes were determined through literature and word cloud topics. The word cloud was created using the Monkeylearn web API.

The choice of a sample size of 200 tweets was because of time and resource constraints, and the need for a representative sample that accurately captures the overall themes of the population.

4 Findings and discussion

4.1 Sentiments about COVID-19 vaccines

An overall sentiment analysis using the VADER lexicon-based method and was performed on 21 084 tweets. Figure 2 shows the most frequent words generated from the dataset without stop words to find the 30 most frequently used words in the tweets. The top three keywords used by Twitter users were “vaccine,” “covid19”, and “people”. This was expected given that the filter list used to filter the tweets was specifically looking for vaccine discussion; this validates our Twitter data. Figure 3 shows the distribution of sentiments in percentages. The graph depicts the overall sentiment distribution of tweets from 1st January 2021 to 31st December 2021. Most tweets were positive 41% (8543), seconded by neutral tweets 32% (5894). Finally, negative tweets account for only 28% of all tweets (6647).

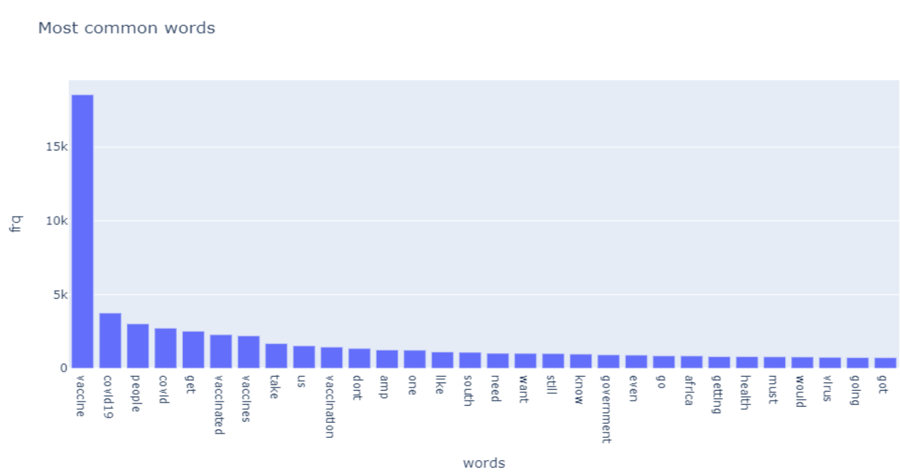


Figure 2: 30 frequent words without stop words

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

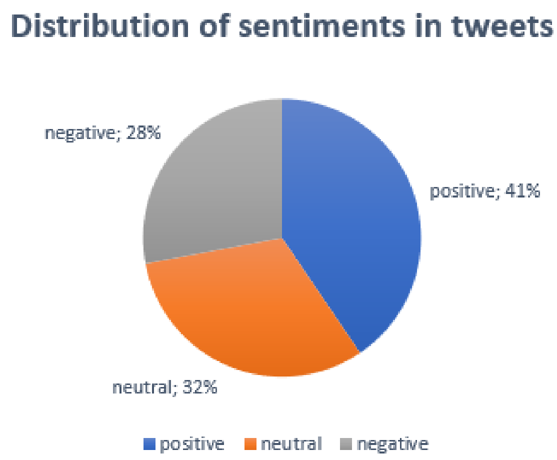


Figure 3: The distribution of sentiments in tweets

Table 1 shows a selection of tweets and their sentiment classifications. We categorised the hesitant tweets with neutral tweets.

Positive tweets (n = 8543)	**Neutral tweets (n = 5894)**	**Negative tweets (n = 6647)**
"Please get sputnik V vaccine is the only I will take"	*"Why are we not educating people on the vaccine the same way people who advocate for Prep etc"*	*"IM NOT TAKING THAT VACCINE!!!! 🤌🏿"*
"I wanna be getting my vaccine too 😭"	*"Why is the vaccine roll out so low in Northern Cape?"*	*"Say No to the vaccine"*
"Finally took my vaccine today. No pain, no effect whatsoever"	*"Why should people be forced to take the Covid-19 vaccine 💉"*	*"I have no interest in all your vaccine lies! God sees you! Petition Against Mandatory Vaccinations in South Africa"*

Table 1: Examples of tweets for the different sentiments

4.2 Central themes discussed on Twitter about COVID-19 vaccines in South Africa

Figure 4 depicts a word cloud of the diversity of multilingual hashtags used to discuss COVID-19 vaccinations in a Twitter dataset. The larger the size of the word in the cloud, the more frequently it occurs in the corpus (Aguilar-Gallegos et al., 2020; Kang et al., 2017; Mir & Sevukan, 2022). The word cloud shows that the most common words were 'vaccine', 'government', 'covid19', 'country', 'vaccine information'. Additional topics that were in the corpus were: "country has no confidence", "puppet master bill gates", "south african vaccine rollout strategy", "severe symptom" amongst topics relating to our central themes.

A total of 183 vaccine-related tweets were analysed and 9 themes were manually coded based on previous literature and the word cloud, as shown in Figure 5. These themes include misinformation, vaccine hesitancy, government, conspiracy, religion, pro-vaccination (pro-vaxx) and anti-vaxx vaccination (anti-vaccination) and humour. We identified the themes by comparing the words from the word cloud to what previous studies had found.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

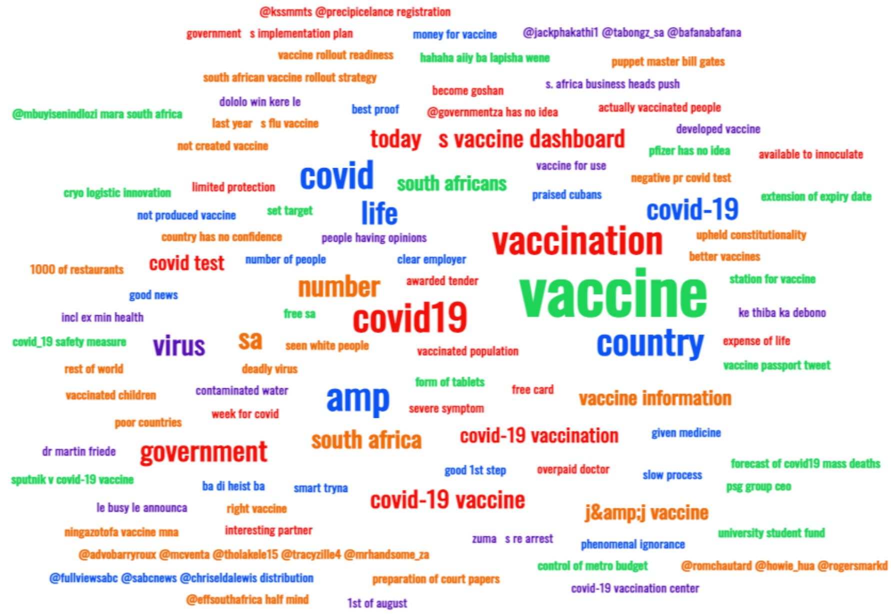


Figure 4: Word cloud of Top 100 words amongst tweets about COVID-19 vaccines

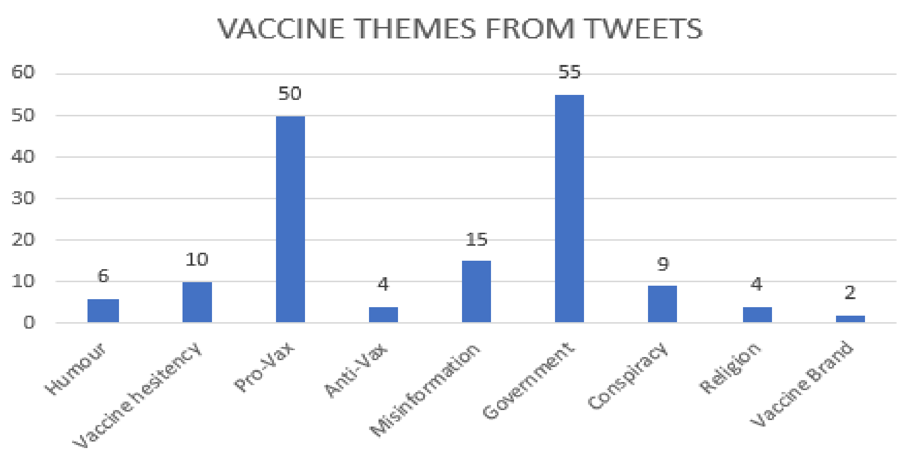


Figure 5: Vaccine themes that emerged from tweets

4.2.1 Government

"Government" was the most common theme (28%, n=55) in vaccine-related discussions. This contrasts with a previous study in Singapore where the main focus was Singapore's government initiatives (Mohamed Ridhwan & Hargreaves, 2021). One of the factors leading to vaccine hesitancy is the possible a link with political issues and disillusionment.

4.2.2 South African governments vaccine rollout plan

Our findings indicate that people had little trust in the South African government's handling of the vaccine rollout. Many people believed that the country can and should produce its own vaccines, but was unwilling to do so (Aguilar-Gallegos et al., 2020; Kang et al., 2017; Mir & Sevukan, 2022).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

“@FloydShivambu Cde Floyd, Africa has resources to produce own vaccines but we don't have leaders. Let's stop blaming the West for our own ineptitude. The least we could have done was to fund research into vaccine discovery. We deserve to be sidelined, West taxpayers can't fund our us forever”.

Many tweets in the corpus came from verified accounts. This resulted in higher tweet engagement than non-verified tweets. The hashtag #VaccineRolloutSA was more widely used to inform the public about the plan to administer vaccines in order to achieve the anticipated herd immunity.

“The first batch of the J&J #COVID19 vaccine arrived at @ortambo_int last night. Frontline healthcare workers will be vaccinated during first phase of the #VaccineRolloutSA”

Government corruption leads to mistrust of authorities: Some people lost confidence in the government’s involvement in the rollout of vaccines. This is supported by evidence of tender fraud involving COVID-19 personal protective equipment (PPE) by senior government officials (Mpofu, 2021, p.52). According to reports, corruption did not anticipate the use and allocation of COVID-19 funds (Mpofu, 2021, p.52).

“ANC has taught us that corruption is the way, we gonna buy those fake vaccine certificates or whatever y'all call it, anything for majwals ??”

“Medical aids don't pay a cent more than they have too why would they allow corruption in the vaccine roll out. @DrZweliMkhize @SundayTimesZA allow for market forces to deliver the drug just regulate the cost per jab.”

The word corrupt appears in 401 tweets in the corpus. People were concerned about political corruption and theft, and they questioned the legitimacy of the vaccine rollout strategies. Since most previous studies came from the developing countries, Tweeter users did not complain as much about corruption.

4.2.3 Vaccine polarity

The second most common theme was "Pro-Vax" (25%, n=50), followed by "vaccine hesitancy" (5%, n=10) and "Anti-Vax" (2%, n=4). These themes were combined because they are the most frequently occurring themes in literature and our tweets, and they frequently overlap with positive, negative, and neutral sentiments. When the vaccine for COVID-19 becomes available, the tweets express a willingness to vaccinate against it. These tweets expressed gratitude for vaccines and how they are made available to save people's lives.

Pro-Vaccination: The Pfizer vaccine was given in two doses. Twitter users used hashtags like #IChooseVaccination to express their desire to get vaccinated and to encourage other Twitter users to do the same. These tweets demonstrate a willingness to receive the vaccine.

“Get your second dose to be fully vaccinated 😉 #IChooseVaccination #staysafe”

A global study comparing people's willingness to vaccinate noted that lower and middle-income countries like South Africa had a higher acceptance rate (Hou et al., 2021). This is similar to our findings; Figure 3 shows only 41% positive sentiments. The majority of South Africans, however, were eager to receive their first or second dose of the vaccine.

Anti-Vaccination: Anti-vaxxers are people who do not want to take or are opposed to vaccines (Hou et al., 2021). Some authors suggest that one of the things that distinguishes vaccine supporters from vaccine opponents is their attitude toward research and scientific knowledge (Maciuszek et al., 2021a).

The user stated that they were against mandatory vaccinations and cited religious texts to back up their position. There are numerous reasons why people are opposed to the vaccine, but two stand out: "freedom" and "choice."

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

“@RevMeshoe @CyrilRamaphosa He must not force us to be vaccinated, some of us our religion and beliefs doesn't allow us to take any vaccine”

“Would love to join a space discussing the science behind Vaxx and Anti-vaxx controversy. My submission would be around why it is scientifically absurd for people who have contacted covid, mild/severe to vaccinate”

The literature frequently mentions how anti-vaccine persons tend to deny scientific data demonstrating the safety and effectiveness of vaccinations and cast doubt on the authenticity of pertinent research, knowledge, and medical authority (Maciuszek et al., 2021a).

4.2.4 Vaccine Hesitancy

There were some comments about the uncertainty surrounding vaccine acceptance due to safety concerns. Some of these vaccine-related concerns include the claims that Johnson and Johnson vaccines were not thoroughly trailed/tested. Though people were concerned about receiving the vaccines, their concerns were not necessarily that the vaccines were unsafe, but that the vaccine development themselves was rushed.

“Do you honestly trust that the vaccine is safe?”

4.2.5 Vaccine brands safety and effectiveness

This theme is more neutral. It contains tweets that are for or against a specific vaccine brand. The vaccine brands were debated as people were choosing their preferred vaccine.

“Personally I want the pfizer. Vaccine Cocktail immunity”

“Me and my younger sister were vaccinated in May with Johnson & Johnson vaccine and didn't experience any side effects. My mom had a headache and was swollen where she was vaccinated”

“Is not because of Bill Gates' depopulation project... No one is infect advocating for Russian or China vaccine they all pushing for Bill Gates vaccine covering their faces with Russia and China!!”

4.2.6 Misinformation

The third most common theme was “Misinformation” (8%, n=15), “Conspiracy” (4.5%, n=200). This theme looked at tweets that were making misleading statements that had no evidence. Some of the contents of the tweets, included comments such as:

“I heard the Vaccine kills your entire family in one swoop, but I ain’t no doctor”.

“All the best with the J&J vaccine. I think it’s a matter of probability & you might be the one to get the blood clot. As they say, “you might be history”.

To manage false and dangerous media content and to convey reliable information intended to encourage the uptake of a potential COVID-19 vaccine to end the pandemic, researchers suggest media partnerships should be expanded (Niemiec, 2020).

4.2.7 Conspiracy theories

Accounted for 4.5% of the sample tweets. Bam (2021) associates these individuals with anti-scientific beliefs (believing in knowledge based on evidence). The corpus contained retweets with phrases like "they are killing people because they want to reduce population". There were a few tweets in the corpus that mentioned the depopulation agenda. These conspiracy theories were linked to one of the most frequently occurring words in Figure 4, “Bill Gates”.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

"Some countries are protesting against the vaccine, they claim Bill Gates is pushing population control publicly. Australia, Brazil and some EU countries are protesting against the vaccine."

There are several conspiracy theory subcategories in our dataset and in other literature, including "microchips in vaccines," "vaccines cause infertility," and "vaccines are for profit" (Nuzhath et al., 2021).

4.2.8 Religious beliefs

Religious opinions made up 2% of the tweets. This number may appear to be insignificant at first glance. However, famous religious leaders have a large Twitter following with high tweet engagement, which may sway their followers' opinions. Some users were leaning toward wanting to practice their own religious beliefs:

"If im entitled to right to a right to practice my religion, and for whatever i believe that the vaccine "generates from some form of Satanism" i see no reason why not to enforce my right to religion."

According to tweets like the one above, some people are unwilling to vaccinate for religious reasons. It also corresponds with literature indicating that social media users from all over the world hold similar religious beliefs about vaccines (Nuzhath et al., 2021; Wilson & Wiysonge, 2020).

4.2.9 Humour

Less than a percent of tweets were jokes. The use of humour in dealing with a serious matter could have been a form of coping mechanism (Mpofu, 2021). Example of jokes that were found were:

"Lol I saw one yesterday. These people had a sign saying "God is my vaccine" lol I laughed. Crazy yt people"

4.3 Changes of South Africa's Twitter sentiments on COVID-19 vaccines

Figure 6 depicts the average sentiments throughout the year. The frequency of each sentiment from 1st January 2021 to 31st December 2021 shows that positive tweets dominate. Positive, negative, and neutral tweets are closer together and overlap on specific days. The high spikes indicated that engagement was higher in the first three months of 2021, from January 2021 to March 2021. Figures 6 and 7 show an increase in Twitter engagement as the year began on the 3rd January 2021 (A), with 116 positive tweets, 88 negative tweets, and 73 neutral tweets. Minister Zweli Mkhize's public briefing statement revealed South Africa's COVID-19 vaccine strategy. On the first day, there was a high volume of tweet engagement because of an important event announced by an official with a verified account. Positive sentiments came from people who were optimistic about vaccination because of their personal experiences.

"*The lackluster manner in which ANC govt is rolling out vaccination processes is as if Covid 19 is yet to happen when in actual fact the pandemic has been with us for a year. It's like embarking on a long trip riding a donkey which has no sense of urgency inspite of eminent perils"*

This is understandable given that prominent individuals and media organisations frequently comment on and cover breaking news in real time using Twitter as a platform to amplify their messaging (Chen et al., 2020). The first substantial rise occurred on 11th January 2021 (B). On this day, South African President Cyril Ramaphosa declared during a national address that the goal of the parliament in designing the path to recovery was to develop a vaccination drive (Business Solutions, 2021). With 198 neutral tweets, 209 positive tweets and 154 negative tweets. The 1st February 2021 (C) had 361 counts of neutral tweets, 282 positive and 197 negative tweets. 8th February (D) had 229 tweets that were neutral, 228 positive, 208 negatives. After that period, there has been a significant drop in engagement then a slight rise in tweets on the 17th February there was a media statement that stated that the total number of

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

COVID-19 cases in South Africa was 1 496 439, and since the last report, 2 320 more cases had been found ("Update on COVID-19 (17th February 2021)," 2021).

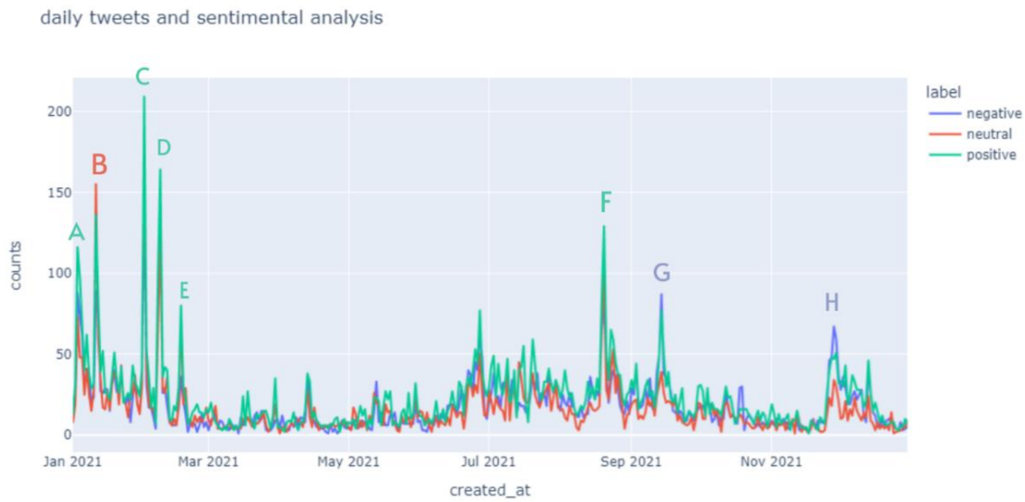


Figure 6: Line graph showing change in tweet sentiment from January 2021 to December 2021

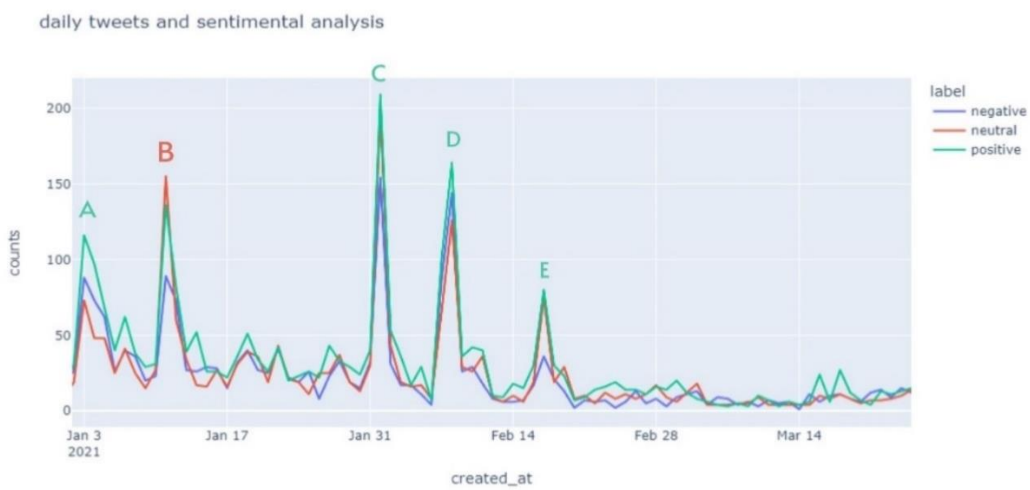


Figure 7: Line graph showing change in tweet sentiment from January 2021 to March 2021.

From April 2021 to August 2021, there was no significant event that caused a high spike in tweets in this period. The counts were averaging below 50 tweets per day. No major announcements were made by prominent public figures except on the 27th June when a government gazette was released on the adjusted level of lockdowns to level 3 (Gazette et al., 2021). The last significant spike for the year occurred on 20th August, 2021 (F). The National Institute for Communicable Diseases (NICD), a component of the National Health Laboratory Service, continues to monitor and supervise COVID-19 in order to inform the public health response ("Latest confirmed cases of COVID-19 in South Africa (20 August 2021)," 2021). They issued a report that detailed the COVID-19 cases' statistics, such as hospital admissions and vaccination rollout.

"People in 18 years and older will be eligible to get their Covid-19 vaccinations from today, 20 August 2021."

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The 20th August 2021 was the day when people 18 years and older could get the vaccine. There was a drop in engagement from that day on until the 14th September 2021 (G). This was the first spike that contained high negative sentiments with 87 tweets, 77 positive tweets and a low 39 neutral tweets.

"Dear President @CyrilRamaphosa and @PresidencyZA Before you force us to be vaccinated, please force us to get the RDP's, Food parcels and force your people to bring back R500 billions. Then the vaccine will be the last."

"That man his telling the truth about vaccines go n make full research you will see vaccine is from luciferase enzyme wich mean lucifer devil"

On this day, topics associated with conspiracies, misinformation, and government distrust were more prevalent. The year's final spike may indicate that 2021 ended on a negative note. The spike on the 28th November 2021 (H), was lower than the previous spikes, with 67 negative tweets, 47 positive tweets, and 34 neutral tweets. President Cyril Ramaphosa issued a national declaration announcement (Gov,2021). According to the president's public announcement, the vaccine rollout had reached approximately 41% of the total South African population, with 25 million vaccine doses administered.

"Then why are infections on the rise for vaccinated people? Answer - compromised immune systems "from vax"

In this period there was higher misinformation and hesitancy with tweets such as the one above questioning the need for vaccines. The rest of the year remains with tweets averaging less than 50 tweets a day. This section saw the change in tweet sentiment over time.

5 Conclusion

More than half of the tweets examined contained positive sentiments, with the remaining two-thirds neutral or negative. The findings do not represent a complete picture of the South African demographic, but rather a snapshot of sentiments. Because social media is more popular among younger people, the results may be skewed toward them. We recommend the inclusion of African languages in NTLK dictionaries because it will allow more people and cultures to be accommodated in analysing data.

The use of Twitter data to understand public sentiment can help advance Information Systems as a discipline by providing new data collection techniques and insights into the impact of big data on organisations and society as a whole. Such research can provide valuable information to governments, organisations, and businesses to make informed decisions and respond to the needs and opinions of their stakeholders.

The study was completed prior to Elon Musk's acquisition of Twitter; changes in the way Twitter runs may occur and provide a distinct dimension while performing research on Twitter. For a government to understand public sentiment on Twitter, they can collect relevant tweets, perform sentiment analysis to classify them as positive, negative or neutral, and analyse the results to identify common themes and opinions. This information can help the government make informed decisions and understand public sentiment on important issues. In this study people were sceptical of the South African government's vaccine rollout strategy, which could have hampered overall vaccine adoption. The study concludes that vaccination attitudes changed throughout 2021, particularly in response to government announcements addressing the public on the status of vaccination rollout.

References

Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávila, J. (2020). Dataset on dynamics of Coronavirus on Twitter. Data in Brief, 30. https://doi.org/10.1016/j.dib.2020.105684

Alfatease, A., Alqahtani, A. M., Orayj, K., & Alshahrani, S. M. (2021). The impact of social media on the acceptance of the covid-19 vaccine: A cross-sectional study from saudi arabia. Patient Preference and Adherence, 15, 2673–2681. https://doi.org/10.2147/PPA.S342535

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Alsayed, A., Alharthi, S., & Adeyemo, D. A. (2019). Twitter Sentiment Analysis Techniques: A Survey. Journal of King Saud University - Computer and Information Sciences, 31(3), 189-201.

Amjad, A., Qaiser, S., Anwar, A., Ijaz-Ul-Haq, & Ali, R. (2021, September 7). Analysing Public Sentiments Regarding COVID-19 Vaccines: A Sentiment Analysis Approach. 2021 IEEE International Smart Cities Conference, ISC2 2021. https://doi.org/10.1109/ISC253183.2021.9562904

Anuratha, K., Sujeetha, S., Nandhini, J. M., Priya, B., & Paravthy, M. (2021). #Vaccine: Using hashtags from Indian Tweets to Capture and Analyse the Sentiments of People on Vaccination for Covid'19 Pandemic. Advances in Parallel Computing, 39, 88–92. https://doi.org/10.3233/APC210183

Bam, N. E. (2021). Understanding the motivations for covid-19 vaccine hesitancy in south africa: narrative literature review.

Burger, R., Buttenheim, A., English, R., Maughan-Brown, B., Köhler, T., & Tameris, M. (2021.). WAVE 4 3 12 May 2021 COVID-19 vaccine hesitancy in South Africa: Results from NIDS-CRAM Wave 4.

Chadwick, A., Kaiser, J., Vaccari, C., Freeman, D., Lambe, S., Loe, B. S., Vanderslott, S., Lewandowsky, S., Conroy, M., Ross, A. R. N., Innocenti, S., Pollard, A. J., Waite, F., Larkin, M., Rosebrock, L., Jenner, L., McShane, H., Giubilini, A., Petit, A., & Yu, L. M. (2021). Online Social Endorsement and Covid-19 Vaccine Hesitancy in the United Kingdom. Social Media and Society, 7(2). https://doi.org/10.1177/20563051211008817

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. JMIR Public Health and Surveillance, 6(2). https://doi.org/10.2196/19273

Chen, X., Association for Computing Machinery. Special Interest Group on Information Retrieval, Association for Computing Machinery. Special Interest Group on Hypertext, H. and W., & ACM Digital Library. (2012.). CIKM'12 : the proceedings of the 21st ACM International Conference on Information and Knowledge Management : October 29 - November 2, 2012, Maui, Hawaii, USA.

Chou, W. Y. S., Gaysynsky, A., & Vanderpool, R. C. (2021). The COVID-19 Misinfodemic: Moving Beyond Fact-Checking. Health Education and Behavior, 48(1), 9–13. https://doi.org/10.1177/1090198120980675

Deverna, M. R., Pierri, F., Truong, B. T., Bollenbacher, J., Axelrod, D., Loynes, N., Torres-Lugo, C., Yang, K.-C., Menczer, F., & Bryden, J. (2021). CoVaxxy: A Collection of English-Language Twitter Posts about COVID-19 Vaccines. https://doi.org/10.5281/zenodo.4526494

Gazette, G., Gazette, R., & Notice, G. (2021). Disaster Management Act: Regulations: Alert level 3 during Coronavirus COVID-19 lockdown: Amendment. www.gpwonline.co.za

Gul, S., Mahajan, I., Nisa, N. T., Shah, T. A., Jan, A., & Ahmad, S. (2016). Tweets speak louder than leaders and masses: An analysis of tweets about the Jammu and Kashmir elections 2014. Online Information Review, 40(7), 900–912. https://doi.org/10.1108/OIR-10-2015-0330

Ioana-Andreea, G., Popa, S., & Ioana, M. (2021). *COVID-19 Vaccine infodemic: Sentiment analysis of the twitter content*. https://global.oup.com/academic/content/word-of-the-year/?cc=ro&lang=en&

Kang, G. J., Ewing-Nelson, S. R., Mackey, L., Schlitt, J. T., Marathe, A., Abbas, K. M., & Swarup, S. (2017). Semantic network analysis of vaccine sentiment in online social media. Vaccine, 35(29), 3621–3638. https://doi.org/10.1016/j.vaccine.2017.05.052

Kumar, A., & Sebastian, T. M. (2012). Sentiment Analysis on Twitter. www.youtube.com

Li Han Wong, B., Vybornova, O., Grima, S., & Nia, M. Z. (2022.). Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts.

Luo, C., Chen, A., Cui, B., & Liao, W. (2021). Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: Semantic network analysis of two social media platforms in the United States and China. Telematics and Informatics, 65. https://doi.org/10.1016/j.tele.2021.101712

Lyu, H., Wang, J., Wu, W., Duong, V., Zhang, X., Dye, T. D., & Luo, J. (2022). Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination. Intelligent Medicine, 2(1), 1–12. https://doi.org/10.1016/j.imed.2021.08.001

Marivate, V., & Combrink, H. M. (2020). Use of available data to inform the COVID-19 outbreak in South Africa: A case study. Data Science Journal, 19(1), 1–7. https://doi.org/10.5334/dsj-2020-019

Mir, A. A., Rathinam, S., & Gul, S. (2022). Public perception of COVID-19 vaccines from the digital footprints left on Twitter: analyzing positive, neutral and negative sentiments of Twitterati. Library Hi Tech, 40(2), 340–356. https://doi.org/10.1108/LHT-08-2021-0261

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Mir, A. A., & Sevukan, R. (2022). Sentiment analysis of Indian Tweets about Covid-19 vaccines. Journal of Information Science. https://doi.org/10.1177/01655515221118049

Mohamed Ridhwan, K., & Hargreaves, C. A. (2021). Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. International Journal of Information Management Data Insights, 1(2). https://doi.org/10.1016/j.jjimei.2021.100021

Mønsted, B., & Lehmann, S. (2022). Characterizing polarization in online vaccine discourse—A large-scale study. PLoS ONE, 17(2 February). https://doi.org/10.1371/journal.pone.0263746

Niemiec, E. (2020). COVID-19 and misinformation . EMBO Reports, 21(11). https://doi.org/10.15252/embr.202051420

Pawar, K. K., Shrishrimal, P., & Deshmukh, R. R. (2015). Twitter Sentiment Analysis: A Review. International Journal of Scientific & Engineering Research, 6(4). http://www.ijser.org

Puri, N., Coomes, E. A., Haghbayan, H., & Gunaratne, K. (2020). Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Human Vaccines and Immunotherapeutics, 2586–2593. https://doi.org/10.1080/21645515.2020.1780846

Rotolo, B., Dubé, E., Vivion, M., MacDonald, S. E., & Meyer, S. B. (2022). Hesitancy towards COVID-19 vaccines on social media in Canada. Vaccine, 40(19), 2790–2796. https://doi.org/10.1016/j.vaccine.2022.03.024

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2020a). An Exploratory Study of COVID-19 Misinformation on Twitter. http://arxiv.org/abs/2005.05710

Mpofu, S. (2021). Digital Humour in the Covid-19 Pandemic. Palgrave Macmillan.

Shoaei, M. D., & Dastani, M. (2020). The role of twitter during the COVID-19 crisis: A systematic literature review. In Acta Informatica Pragensia (Vol. 9, Issue 2, pp. 154–169). University of Economics - Prague. https://doi.org/10.18267/J.AIP.138

van Heerden, A., & Young, S. (2020). Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS ONE, 15(10 October). https://doi.org/10.1371/journal.pone.0239304

Appendix A: Stop words for text processing (100 words)

['setsidiki1','meant','covid19','vaccine','booster','irvinjimsa','sihlewasembo','arw','forcing','thing','kuthi','respect','covid','care','putting','pressure','us','vaccinate','pls','siyekeni','sizaziyela','ngelethu','xesha','vaccinated','curfew','lifted','thank','jesus','vaccinefree','curfew','cabinet''hhaybo','guys','drink','won’t','take','vaccine','doctoriamthe','happened','natural','immunity','infection','successful','recovery','deltaomricon','still','discriminated','medical','institutions','advocating','vaccine','employer','theres''robot','vaccine','pfizerwhistleblower','yerlydave','petersweden7','moosefucker','logic','go','get','vaccinated','stop','blabbing','tweeter','vaccinate','dammit',’tell','partner','protect','times','goes','cause','shes','double','jabbed','dont','make','supergirl','must','still','follow','covid''guidelines','period','pfizerwhistleblower','decide','vaccinate','vaccine','get','jampj','pfizer','ceo','pfizer','taken','vaccine','pfizerwhistleblower','guys','believe']

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Instructors' perception of the competencies required to teach DS in HLI

Nkosikhona Theoren Msweli [1*†]
[1] EasyChair
[2] University of Pretoria, South Africa.
Nkosikhona.msweli@gmail.com

Abstract

While data science education (DSE) may be a solution for democratising data science in higher learning institutions (HLI), challenges on achieving this goal remain. One of the main challenges is the shortage of qualified academic staff members who can deliver multidisciplinary curricula effectively. Teaching data science (DS) is a developing topic that represents a gap in the literature as the education sector embarks on a journey of unearthing the knowledge required to teach DS. This study aimed to gain insight into instructors' perceptions of their skills and competencies in teaching DS. Twenty-six (26) DS instructors were surveyed. The sample included instructors from various disciplines possessing various levels of teaching experience. The study used a 16-construct questionnaire based on the pedagogical content (PCK) framework to understand the scope of DS PCK. The collected data were analysed using IBM SPSS AMOS version 21. Follow-up interviews were conducted to identify competencies needed in DSE and pedagogical beliefs. This qualitative data was analysed thematically. As expected, the study revealed lower female DS instructor participation. The study further showed that the majority of DS instructors have pedagogical and content knowledge of DSE. Business understanding is often taught at the post-graduate and undergraduate levels. Data understanding is taught at the undergraduate level but not as often at the postgraduate level. Short learning programs do not target model evaluation and deployment as a course. The study recommends that curriculum developers consider including the model evaluation and deployment in DS curricula and how these concepts can be taught as part of the curricula. Furthermore, the study recommends the adoption of a DS framework that will guide the development and structuring of DS curricula to ensure standardization and interdisciplinary pedagogies that support content delivery. Research should also be conducted on how industry partnerships can be forged to keep instructors engaged with ongoing DS developments. Lastly, societal interests in data continue to put additional pressure on DSE. Given the multidisciplinarity of the field and the involvement of government in the usage of data from various sources, government, and industry should play an active role in ensuring the availability, access, and

[*] Masterminded EasyChair and created the first stable version of this document
[†] Created the first draft of this document

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

ethical use of data, especially for educational purposes. It is envisaged that this will trigger the need to equip society at large with data skills and policies to govern the process.

1 Introduction

DSE is an emerging and challenging area. It is also a growing topic of interest in Information System (IS) research and education. Several disciplines and complex concepts are blended into DS, and therefore require specific teaching practices to make learning and understanding easier for students. For instance, Machine learning (ML) has many complex and difficult-to-teach algorithms (Sulmont, Patitsas & Cooperstock, 2019) which require DS instructors to empathize with student with diverse backgrounds and expectations (Kross et al., 2022). Irrespective of the complexity of the content or concepts involved, instructors need to ensure that learning outcomes are achieved. This means that DSE instructors need to be agile and open to multiple teaching practices.

DSE programs are becoming more openly available, however, only a few instructors are qualified and knowledgeable enough to teach in this field. This implies that data scientists are generally not properly skilled (Attwood et al., 2019; Yu & Hu, 2019). In fact, the lack of solid background and knowledge in DS among instructors ranks as the biggest limiting factor for integrating DS skills into the curriculum (Emery et al., 2021; Saddiqa et al., 2021). As a result, students enrolled in DS programs are often confronted with challenges emanating from and associated with variable quality of content delivery (Fox & Hackerman 2003; Sunal et al. 2004). It is quite evident that instructors need to find a way to assist students to develop an accurate, adequate, and generally better understanding of DS (Qian et al., 2017). Despite such an obvious gap in DSE, literature that focuses on the design and adoption of strategic approaches for delivering DS programs to a diverse group of students in various domains is scant (Sulmont, Patitsas & Cooperstock, 2019; Twinomurinzi et al., 2022). In addition, research on how DS should be taught or what type of competencies are required from data science academic staff is still lacking. Such research opportunities have the potential to support academia as it struggles to position itself within the data science field (Cao, 2019; Engel, 2017; Mike, 2020) and lead to the creation of new research topics in IS (Cao, 2017) such as how instructors approach DS teaching practices (Lau et al., 2022). Another challenge confronting the field of DSE includes data science tools and techniques that are continuously evolving. Consequently, it is important to examine how instructors tackle this challenge and ultimately unearth instructors' practices as well as their underlying reasoning for teaching DS.

To improve instructors' knowledge of DS topics, it is important to establish the instructors' base knowledge, experience, perceptions, and knowledge gap and variation (Saeli et al., 2011; Shulman, 1986). Instructors need a specific type of knowledge to teach DS concepts effectively, and this knowledge is totally different from content knowledge or general pedagogy. Such knowledge is described as PCK (Pedagogy Content Knowledge), that is, a type of knowledge that represents the blending of content (e.g., algorithms, modeling, business scenarios, etc.) and pedagogy (e.g., how to teach algorithms or business cases, etc.). PCK includes the understanding of how instructors will take a specific topic, rearrange it to fit the diverse interests and abilities of learners, and present it for learning purposes (Shulman, 1987).

Understanding how instructors teach DS may also depend on how familiar they are with using DS skills (Emery et al., 2021). Research must address questions such as "what works in DSE" and "what conceptual frameworks guide the practice of DS instructors and enable them to recognise and discuss effective teaching practices?". Therefore, this study formulated the following research questions:

How does the PCK framework resonates with the competencies of DS lecturers in HLI?

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

This study aims to gain insight into instructors' perceptions of their skills and competencies in teaching DS. The PCK framework was adopted to capture some of the essential attributes of knowledge required by facilitators for scholarly integration in their teaching.

The remainder of the paper is structured as follows: following this introduction, the study takes a brief look at the study's theoretical framework. Thereafter, we give an account of the method used to collect the empirical data used in the study. After presenting the results and discussing the contribution and limitations of the study, the paper concludes by addressing the implications of the results for DSE and suggesting an agenda for further research.

2 Literature Review

2.1 DS Instructional Programs

At the postgraduate level, DS programs have been proliferating across the globe (Hosack and Sagers, 2015; Raj et al., 2019; Li, Milonas & Zhang, 2021). On the other hand, undergraduate DS programs are still being investigated (Zhang, Huang & Wang, 2017; Mikalef et al., 2018; Çetinkaya-Rundel & Ellison, 2021; Davenport & Malone, 2021). DS short-learning programs are often well designed and commoditized to mainly address DS technical skills such as predictions, data analytics, ML, and statistical programming. However, the integration of these technical concepts within a full DS course still needs to be explored (Qiang et al., 2019; Silva et al., 2014). In particular, the development of teaching guidelines for DSE and training has not been adequately researched.

As attested by Demchenko et al. (2019), DSE must reflect multi-disciplinary knowledge and competencies to afford data scientists insights into other domains. DSE should further afford skills and competencies to work with various forms of data and interpret the analytical results, especially for those who lack DS literacy (Dichev & Dicheva, 2017). Therefore, it is important to establish instructors' confidence in teaching DS (Mike, 2020), or how best to prepare instructors to teach DS in various domains (Emery et al., 2021). For instance, data scientists in biomedicine need to be trained in computer science, statistics and mathematics, and biomedicine (Garmire et al., 2017; Hassan & Liu, 2020). Stephenson et al. (2018) alluded to the fact that offering DS skills in computer science courses only could lead to the under-representation of other disciplines. Research on how to address teaching practices in DSE are necessary. For instance, Emery et al. (2021) investigated ways of preparing instructors to teach DS in undergraduate biology and environmental science courses.

2.2 Teaching DS

Fayyad and Hamutcu (2021) raised an important question "how do we teach data scientists while there is so much debate on who they are?". There appears to be a growing concern among those who provide training and employment to data scientists. Apart from unclear roles of data scientists, teaching DS faces other challenges, including teaching multidisciplinary content, the misconception of DS concepts (Jafar, Babb & Abdullat, 2016), student diversity, and student cognitive skills (Sulmont, Patitsas & Cooperstock, 2019; Donoghue et al., 2021). These challenges may lead to low student registrations and throughput in DS programs. Instructors need to know their students and their characteristics to apply appropriate pedagogy (Gudmundsdottir & Shulman, 1987). Sentance and Csizmadia (2017) found that various pedagogies can improve students' ability to solve a problem. However, instructors need support in terms of professional development on how to work with different pedagogies in a multidisciplinary setting to effectively teach DS concepts (Emery et al., 2021; Lau et al., 2022). Table 1 provides examples of pegagogies used for specific concepts.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Data science concept	*Teaching pedagogy/strategy*	*Source*
Data preparation, Visualisation	Project-based learning	Saltz and Heckman (2015)
Machine learning, modeling	Similes, gamification, storytelling	Song and Zhu (2016; Garcia-Algarra (2020)
Model deployment	Story-telling Experiential learning	Jaggia *et al.* (2020) Anslow *et al.* (2016)

Table 1:Pedagogies in DSE

Effective teaching occurs when learning and understanding are achieved (Blair et al., 2021). To determine whether student learning has been enhanced, teachers must evaluate their teaching practices. DS instructors interested in assessing their teaching practices can apply various frameworks to inform their questions and teaching strategies (Kim, Ismay & Chunn, 2018; Hassan & Liu, 2019). It is worth noting that the conceptual framework that is applicable and useful to higher education; the focus has instead been on secondary schools (Saeli et al., 2011; Başaran, 2020; Taopan, Drajati & Sumardi, 2020).

While some DS concepts appear to have existed in various domains, strategies for their content delivery are lacking (Sulmont, Patitsas & Cooperstock, 2019; Dill-McFarland et al., 2021; Lau et al., 2022). For instance, data frames that were originally designed by statisticians for exploratory data analysis, are now viewed by data scientists as data sets. Instructors can discuss how each discipline or any domain for that matter views and uses a particular concept (Jafar, Babb & Abdullat, 2016; Lau et al., 2022). Technical concepts, especially the ones listed in Table 2, have been put forward as key competencies of DS, and instructors are expected to demonstrate these competencies to teach DS.

Table 2: DS competencies

Data science competency	**Source**
Data visualisation	Shirani (2016)
Apache Hadoop and programming languages	Demchenko (2019; Price & Ramaswamy (2019; Yadav & Debello (2019)
Machine learning	Garcia-Algarra (2020)
Big Data and ethics	Saltz, Dewar & Heckman (2018); Mike (2020)

2.3 The PCK theoretical framework

The concept of PCK was introduced by Shulman (1986), after pointing out a lack of research that targets the course content taught to students. Shulman, (1986) defined pedagogical content knowledge as teachers' interpretations and transformations of subject-matter knowledge in the context of facilitating student learning. As shown in Figure 1, integrating content knowledge (CK) and pedagogy knowledge (PK) enables an understanding of how particular topics are presented to students with different backgrounds. Instructors should be able to transform the knowledge to be taught to the students in a way that is easily understood.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

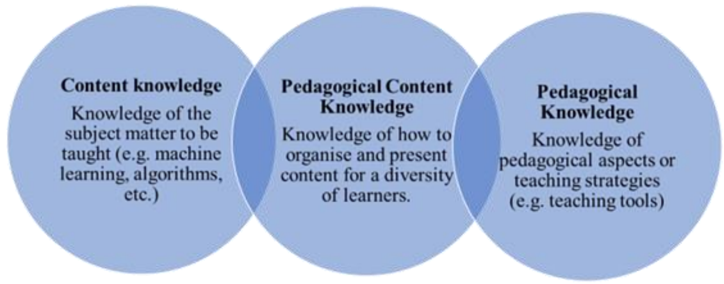


Figure 1: An illustration of pedagogical content knowledge *Source: Shulman (1986)*

PCK outlines the instructor's domain knowledge, pedagogical knowledge, and knowledge of the environment in which the subject is being taught. The framework is useful when differentiating teaching- and non-teaching specialists, for instance, a data science instructor from a data scientist. The argument lies in the capacity of the instructor to transform the CK into forms that are pedagogically powerful and yet adaptive to student backgrounds and abilities (Shulman, 1987). The key elements of pedagogical content knowledge proposed by Shulman, (1987) are as follows:

1. **Content Knowledge** - Knowledge of representations of subject matter – it is about the actual subject matter that is to be taught. Instructors must know and understand the subjects that they teach, including knowledge of methods, tools, concepts, theories, and techniques within a given field. For example, in DSE, instructors know that students must learn ML, algorithms, analytics, data visualization, and so on. Therefore, CK is required for knowledge and understanding of how these concepts or areas can be and are being taught, and the advantages and disadvantages of each teaching practice.

2. **Student Knowledge** – Understanding of students on the subject and the learning and teaching implications that are associated with the specific subject matter. For example, students may confuse data mining and data wrangling as the same concept. Alternatively, students assume business/domain knowledge is not so important.

3. **General pedagogical knowledge (Pedagogical Knowledge)** – Understanding of the practices or methods of teaching and learning and how this understanding encompasses among other things, overall educational purposes, values, and aims.

   Additional Elements:

4. **Curriculum Knowledge** – Knowledge of what should be taught to a particular group of students. Instructors need to know students' learning potential, syllabuses, and program planning documents and how assessments will be conducted.

5. **Pedagogical content knowledge** – Instructors' understanding of how to teach the subject matter, including the use of examples and illustrations to make a particular topic understandable across all students.

6. Knowledge of students and their characteristics and how these may affect their learning,

7. Knowledge of educational contexts, the political, social, and religious workings of groups or of the classroom in which the teaching takes place.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

This study adopted PCK to understand the instructor's perceptions of their knowledge of teaching data science. Prior PCK research into computer science education examined topics such as science (Fraser, 2016), programming (Qian et al., 2017; Rahimi et al., 2018; Saeli et al., 2011), and design of digital artifacts (Rahimi, Barendsen, & Henze, 2016). However, there is little scientific understanding and reporting of teachers' PCK for teaching algorithms, except for Sulmont, Patitsas and Cooperstock (2019) who investigated the difficulties of teaching ML to non-STEM (non-science, technology, engineering, and mathematics) students.

3 Methodology

The study focused on instructors' competence to teach DS by using a 16-construct questionnaire that was based on the PCK framework as a way to understand the scope of what DS PCK could be. Data was collected via an online questionnaire. Quantitative data were analyzed using simple descriptive statistics.

3.1 Participants demographics

A total of 26 instructors participated in the study. These participants were selected using purposeful sampling method. Table 3 shows the demographical information of the study participants.

Category	**Sub-category**	**Frequency**	**Percentage**
Sex	Male	21	80.8
	Female	5	19.2
Highest Qualification	Master's Degree	8	30.8
	Honours Degree	7	26.9
	Bachelor's Degree	10	38.5
	Others	1	3.8
Years of Experience	Less than 1 year	4	15.4
	1 to 5 years of Experience	6	23.1
	+5 years of Experience	4	15.4
	+10 years of Experience	12	46.2
Age	18 to 35 years	4	16.7
	36 to 55 years	20	83.3
Level of data science qualification teaching at	Short Learning Programme (e.g., MOOCs, Badges, micro-credentials)	8	30.8
	Undergraduate level (e.g., Bachelor's Degree, Diploma, Higher Certificate)	11	42.3
	Postgraduate level (Honours, Masters, Doctorate)	7	269

Table 3: Demographics of study participants

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

4 Results and findings

This section provides details on how the data was captured, described, analysed, and interpreted systematically.

4.1 Descriptive statistics

Central tendency measures

Central tendency measures were conducted to assess how centred the distribution of the constructs involved in the study is. A five-point Likert scale where the value 1 corresponds to “Strongly disagree” and the value 5 corresponds to “Strongly agree” was applied to measure the following constructs: Content Knowledge (CK), Pedagogical Knowledge (PK), and Pedagogical Content Knowledge (PCK). Table 4 summarises the responses of the participants with only high frequency in each category being reported.

Category	Questions	measure	Frequency	Percentage
Content Knowledge	I am familiar with data science tools, processes, and technique	To a great extent	11	42.3
	I do understand various concepts of data science	To a great extent	11	42.3
	I know what data science students should be taught in terms of content	To a great extent	11	42.3
	I have knowledge and understanding of data science and what it entails	To a great extent	12	46.2
	I am familiar with the data science curriculum and syllabus	To a great extent	10	38.5
	I can create materials that map to a specific level of proficiency among students in teaching data science	To a great extent	10	38.5
Pedagogical Knowledge	I know of the different processes and practices of teaching e.g. establishing learning objectives	To a large extent	13	50
	I know how to organize a classroom and manage students during instruction	To a great extent	15	57.7
	I can differentiate between various instructional strategies (teaching pedagogies)	To a large extent	12	46.2
	I can use various teaching pedagogies	To a great extent	12	46.2
	I know how to align learning outcomes and assessment opportunities with the teaching pedagogy	To a great extent	14	53.8
	I know how to teach in a multidisciplinary setting	To a large extent	15	57.7

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

	I can identify different strategies for evaluating student understanding	To a large extent	12	46.2
Pedagogical Content Knowledge	I know how to teach data science topics/concepts using appropriate teaching pedagogies	To a great extent	13	50
	I know how to pair teaching pedagogy with data science concepts when preparing and delivering the content	To a large extent	12	46.2
	I know the teaching pedagogies that are appropriate for data science education	To a large extent	14	53.8

Table 4: Summary of descriptive statistics on PCK

- ***Content Knowledge***

Instructors were asked about their content knowledge of DS. Six questions were posed to participants to establish their level of knowledge. The results indicate that, to a great extent, 46,2% of instructors have the required knowledge and understanding of DS. In the same vein, 42,3% of the instructors were found to be familiar with DS tools, processes, and techniques; understand various concepts of DS; and know what students should be taught in terms of content. The overall mean (4,12) indicates that most instructors have, to a large extent, the requisite content knowledge.

- ***Pedagogical Knowledge***

Instructors were asked about their pedagogical knowledge of DS. Seven questions were asked to establish their level of knowledge. The results indicate that the majority (57,7%) of instructors know to a large extent how to teach in a multidisciplinary setting. Following in the same pattern, 57,7% of the instructors, to a great extent, know how to organize a classroom and manage students during instruction. Furthermore, 53,8% of the instructors know, to a great extent, how to align learning outcomes and assessment opportunities with the teaching pedagogy. Whereas 50% of the instructors were found to be, to a large extent, knowledgeable of the different processes and practices of teaching e.g., establishing learning objectives,; 50% of the instructors could, to a great extent, identify different strategies for evaluating student understanding.

The overall mean results (4,35) indicate that most instructors, to a large extent, have Pedagogical Knowledge.

- ***Pedagogical Content Knowledge***

Instructors were questioned about their pedagogical knowledge of DS, and 3 questions were posed to establish the level of their knowledge. Based on the results generated from, Results of this part of the study revealed that slightly over half (53,8%) of the instructors, to a large extent, know the teaching pedagogies appropriate for DSE. Analogously, 50% of the instructors, to a great extent, know how to teach DS topics/concepts using appropriate teaching pedagogies. Lastly, only 46,2% of the instructors (to a large extent) know how to pair teaching pedagogy with data science concepts when preparing and delivering the content. The mean score (4,13) suggests that most of the instructors, to a large extent, have pedagogical content knowledge.

Cross tabulations

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Cross-tabulation enables quantitative analysis of the data to understand the relationship or correlation between multiple variables. Several relationships were studied and the results obtained are discussed below:

- The relationship between business understanding and the level of DS qualification they teach - the results revealed that 80% of instructors that teach at the postgraduate level often teach business requirement (business understanding). In comparison, 50% of instructors that teach at the undergraduate level always teach business understanding.
- The relationship between data understanding and the level of DS qualification they teach - The results indicate that 38.5% of the instructors that teach at the postgraduate level often teach data understanding. In contrast, a substantial majority of the instructors (83.3%) that teach at the undergraduate level always teach data understanding.
- The relationship between data preparation and the level of DS qualification they teach - the results indicate that most (63.6%) of the instructors that teach at the undergraduate level always teach data preparation.
- The relationship between data modeling and the level of DS qualification they teach - the results indicate that less than half (45.5%) of the instructors that teach at the undergraduate level often teach data modeling.
- The relationship between model evaluation and the level of DS qualification they teach - the results show that half of the instructors that teach short learning programmes rarely teach model evaluation. In comparison, a mere 40% of the instructors that teach at the undergraduate level often teach model evaluation.
- The relationship between deployment and the level of DS qualification they teach - the results indicate that 80% of instructors that teach DS at the undergraduate always teach how models are deployed.

5 Discussion

DSE demands an interdisciplinary curriculum (Twinomurinzi et al., 2022), and instructors are compelled to employ multiple pedagogies to deliver this curriculum (Asamoah, Doran & Schiller, 2020). When applying PCK, it is envisioned that instructors incorporate their interdisciplinary pedagogical knowledge into teaching data science. It is expected that instructors with over 10 years of teaching experience have the requisite experience to use different pedagogies to teach the curricula. However, their experience may not pertain to teaching data science considering that it is fairly new and emerging discipline. It should be borne in mind that data science instructors are responsible for researching, preparing, conducting, and reviewing educational programs. That being the case, they are also responsible for developing new skills for data scientists. Essentially, instructors may need to enhance their knowledge to fit current trends and familiarize themselves with the content and its application in the real world. This implies that HLIs need resources to capacitate and develop instructors as new trends emerge. For instance, instructors who may not be acquainted with Auto ML or have expertise in Hadoop and Spark could form part of a continuous development or life-long learning program.

Reasons behind gender-based differences in the adoption and use of technology continue to be a challenge that is not addressed by research. Work on the dominance of technology used by males as compared to females is abound (Dichev & Dicheva, 2017). Shahbaz *et al.* (2020) reported that, when compared with females, males feel data analytics is powerful and more useful. It is evident from the low participation rate of female DS instructors in this study that gender gaps still persist in the field of

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

technology, especially DSE. Partnerships with different communities need to be explored to improve the
under-representation of women (Gundlach & Ward, 2021). In education, PCK is necessary to determine gender factors impacting the adoption and use of technology (Saeli et al., 2011).

Model evaluation is not often taught as part of the curriculum. While this step is often overlooked, its significance has triggered the need to standardize it (Baier, Jöhren & Seebacher, 2019). Castellanos *et al.* (2019) have noted that while the industry has shown more interest in data science models, the deployment rate of these models is still very low. One of the reasons could be that the models are not evaluated to establish whether they satisfy all business objectives to qualify for deployment. There is also a possibility that the deployment phase and skills applied are not DS-focused (Ackermann *et al.*, 2018; Baier, Jöhren & Seebacher, 2019). This accentuates the importance of a DS project framework to guide the development of DS programs. Furthermore, such an approach will contribute towards addressing technical and non-technical challenges that are often experienced during the deployment stage (Baier, Jöhren & Seebacher, 2019). The results reported herein revealed the gaps in the inclusion of model deployment in DS teaching and learning. Previous work has also identified these factors (Davenport & Malone, 2021). While it is important to build a working model, it is also important to determine how the industry receives these models and how they are deployed to improve their adoption (Ackermann et al., 2018). Previous work indicates that it is difficult to teach the deployment of models in an educational setting (Jaggia et al., 2020). This could imply inexperience or a lack of skills in DS infrastructure (Castellanos et al., 2019).

While computer science undergraduate programs are the popular preference for DS (Mike, 2020), incorporating DS into domain programs can be of immense benefit (Castellanos *et al.*, 2019; Davenport & Malone, 2021). It is important to understand that the nature of DS has distinct needs and significance depending on the organisation or domain. The same can be said of the way DS is taught to science and non-science students. Such an understanding can be achieved through pedagogical advancements which provide new ways to teach DS concepts and thus build a workforce that is industry relevant. In addition, DS offerings at postgraduate level has the potential to elevate the skills levels of students when they are busy with their studies (Hosack & Sagers, 2015). Having knowledgeable instructors has the ability to support the attainment of advanced analytical skills.

Business understanding and data understanding appeared to be more common in undergraduate programs. These two components are crucial in any DS project. One of the important data challenges is that it moves very fast in different cycles thus leading to data skills being outdated rapidly. It is rather necessary to keep up with new trends and establish a relationship with leading industries for joined initiatives on DSE (Mikalef et al., 2018). PCK is considered powerful in influencing the pedagogical thinking that is necessary for DSE. To illustrate this point, instructors' unfamiliarity with data could be alleviated through continuous professional development in a form of short courses (such as micro-credentials or MOOCs) and workshops (Saddiqa et al., 2021). In a fast-paced industry and ever-changing technologies, micro-credentials offer a better flexible solution for skilling individuals (Msweli, Twinomurinzi & Ismail, 2022).

Therefore, academic training, related industry experience, licensure, prior training, or lecturing experience are required. Knowledge of statistics, programming, data visualization, big data, and building models (machine learning) is also required.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

6 Conclusion, Implications, Limitations, and Areas for further research

PCK for DS instructors is important since DS programs are becoming easily available and are thus accepting students from various backgrounds. Teaching in this field comes with opportunities and challenges. The purpose of this study was to investigate the extent to which the PCK framework resonated with the competence of DS lecturers and how it influences the teaching practices in DSE. The study examined how the PCK framework can be adopted to assess the instructor's knowledge of teaching DS. Therefore, this paper contributes to the body of knowledge by understanding the confidence and competence of those that are teaching or aspire to teach DS. This contribution presents a gap in DS offerings where there is no guidance on how and at what level the specific DS content should be presented, and how it should be presented. The study immediately found fewer female data science instructors. This prompt the need to study gender differences to find the moderating factors behind the low representation of females in data science educational contexts.

A need still exists to clearly define the roles of data scientists so that they can be trained accordingly. For example, their involvement in model evaluation and deployment is not clear. Research is needed in this regard to establish the data scientist's involvement in data science projects. This study has revealed that PCK is useful to instructors when addressing the following knowledge questions: What are the reasons behind teaching a specific DS program?; What DS concepts should be taught by instructors?; What challenges or misconceptions do students encounter within these concepts; and, How should these concepts be taught? Furthermore, it is useful to understand the target audience of the intended course. The overall data science field is unstable and advancing rapidly. Therefore, the findings and recommendations of this present study include emphasizing the importance of continued learning. Such an approach will enable instructors to acquire new data science knowledge and competencies that currently do not exist, thus making it easier to disseminate the purported new knowledge during lesson delivery. This continued learning can be undertaken or achieved through training and workshops, research, or collaborative partnerships with industry. Faculty heads may need to invest resources to improve the quality of teaching and the confidence of data science instructors. This will further require the cooperation of both HLI management and instructors to ensure effective teaching. This research advocates for the use of the PCK framework to improve the teaching and learning of multidisciplinary data science curricula.

LIMITATIONS AND AREAS FOR FUTURE WORK

The study only considered PCK without exploring other factors that might influence the knowledge of teaching DS. The PCK is content general, it does not apply to any specific subject. Therefore, where there are challenges or complex concepts, PCK might not be relevant. The other limitation is that the sample size used in this study was too small and based on a single under-developed region. Future studies can increase the sample size and focus on other regions that are, for example, well off.
Further work needs to consider various factors and how they affect knowledge.

The study further suggests the following for future research:

- Alignment of DS programs with CRISP-DM or the adoption of other similar frameworks.
- Investigate ways/methods of teaching evaluation and deployment as part of the DS curriculum.
- Investigate which pedagogies work better with DS concepts.
- Enquiry on importance of industry knowledge and experience among DS instructors.
- Gender inequality in a DS discipline.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

This research recommends the use of the PCK model in the educational field to plan, organize and carry out DSE activities.

ACKNOWLEDGEMENTS
None.

References

Ackermann, K., Walsh, J., De Unánue, A., Naveed, H., Rivera, A. N., Lee, S.-J., Bennett, J., Defoe, M., Cody, C., Haynes, L., & Ghani, R. (2018). Deploying Machine Learning Models for Public Policy: A Framework *. *KDD*, 15–22. https://doi.org/10.1145/3219819.3219911

Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016). Datathons: An Experience Report of Data Hackathons for Data Science Education. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, 615–620. https://doi.org/10.1145/2839509.2844568

Asamoah, D. A., Doran, D., & Schiller, S. (2020). Interdisciplinarity in Data Science Pedagogy: A Foundational Design. *Journal of Computer Information Systems*, *60*(4), 370–377. https://doi.org/10.1080/08874417.2018.1496803

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., & Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, *20*(2), 398–404. https://doi.org/10.1093/bib/bbx100

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the Deployment and Operation of Machine Learning in Practise. *Twenty-Seventh European Conference on Information Systems (ECIS2019)*. https://www.researchgate.net/publication/332996647

Başaran, B. (2020). nvestigating science and mathematics teacher candidate's perceptions of TPACK-21 based on 21st century skills. *Elementary Education Online*, *19*(4), 2212–2226. https://doi.org/10.17051/ilkonline.2020.763851

Blair, J. R. S., Jones, L., Leidig, P., Murray, S., Raj, R. K., & Romanowski, C. J. (2021). Establishing ABET Accreditation Criteria for Data Science. *SIGCSE 2021 - Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 535–540. https://doi.org/10.1145/3408877.3432445

Cao, L. (2019). Data Science: Profession and Education. *IEEE Intelligent Systems*, *34*(5), 35–44. https://doi.org/10.1109/MIS.2019.2936705

Castellanos, C., Pérez, B., Varela, C. A., Del Pilar Villamil, M., & Correal, D. (2019). A Survey on Big Data Analytics Solutions Deployment. *European Conference on Software Architecture*, 195–210.

Çetinkaya-Rundel, M., & Ellison, V. (2021). A Fresh Look at Introductory Data Science. *Journal of Stactistics and Data Science Education*, *29*(1), 16–26. https://doi.org/10.1080/10691898.2020.1804497

Davenport, T., & Malone, K. (2021). Deployment as a Critical Business Data Science Discipline. *Harvard Data Science Review*, *3*(3). https://doi.org/10.1162/99608f92.90814c32

Demchenko, Y. (2019). Big data platforms and tools for data analytics in the data science engineering curriculum. *ACM International Conference Proceeding Series*, 60–64. https://doi.org/10.1145/3358505.3358512

Dichev, C., & Dicheva, D. (2017). Towards Data Science Literacy. *Procedia Computer Science 108C*, 2151–2160.

Dill-McFarland, K. A., Konig, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y., & Hallam, S. J. (2021). An integrated, modular approach to data science education in microbiology. *PLoS Computational Biology*, *17*(2). https://doi.org/10.1371/JOURNAL.PCBI.1008661

Donoghue, T., Voytek, B., … S. E. S. and D. S., & 2021, U. (2021). Teaching creative and practical data science at scale. *Taylor & Francis*, *29*(S1), 27–39. https://doi.org/10.1080/10691898.2020.1860725

Emery, N., Supp, S. R., Kerkhoff, A. J., Farrell, K. J., Bledsoe, E. K., Mccall, A. C., & Aiello, M. (2021). Training Data: How can we best prepare instructors to teach data science in undergraduate biology and environmental science courses? *BioRxiv*. https://doi.org/10.1101/2021.01.25.428169

Engel, J. (2017). Statistical Literacy for Active Citizenship: a Call for Data Science Education. *Statistics Education Research Journal*, *16*(1), 44–49. http://iase-web.org/Publications.php?p=SERJ

Fayyad, U., & Hamutcu, H. (2021). How Can We Train Data Scientists When We Can't Agree on Who They Are? *Harvard Data Science Review*, *3*, 1–7. https://doi.org/10.1162/99608f92.0136867f

Fraser, S. P. (2016). Pedagogical Content Knowledge (PCK): Exploring its Usefulness for Science Lecturers in Higher Education. *Res Sci Educ*, *46*, 141–161. https://doi.org/10.1007/s11165-014-9459-1

Garcia-Algarra, J. (2020). Introductory Machine Learning for Non STEM Students. *Proceedings of the European Conference on Machine Learning*.

Garmire, L. X., Gliske, S., Nguyen, Q. C., Chen, J. H., Nemati, S., Van Horn, J. D., Moore, J. H., Shreffler, C., & Dunn, M. (2017). The Training of next Generation Data Scientists in Biomedicine. *Pacific Symposium on Biocomputing 2017*. www.worldscientific.com

Gudmundsdottir, S., & Shulman, L. (1987). Pedagogical Content Knowledge in Social Studies. *Scandinavian Journal of Educationl Research*, *31*(2), 59–70. https://doi.org/10.1080/0031383870310201

Gundlach, E., & Ward, M. D. (2021). The Data Mine: Enabling Data Science Across the Curriculum. *Journal of Statistics and Data Science Education*, *29*(S1), S74–S82. https://doi.org/10.1080/10691898.2020.1848484

Hassan, I. B., & Liu, J. (2019). Embedding data science into computer science education. *IEEE International Conference on Electro Information Technology*, *2019-May*, 367–372. https://doi.org/10.1109/EIT.2019.8833753

Hassan, I. B., & Liu, J. (2020). A Comparative Study of the Academic Programs between Informatics/BioInformatics and Data Science in the U.S. *Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020*, 165–171. https://doi.org/10.1109/COMPSAC48688.2020.00030

Hosack, B., & Sagers, G. (2015). Applied Doctorate in IT: A Case for Designing Data Science Graduate Programs. *Journal of the Midwest Association for Information Systems (JMWAIS)*, *1*(1), 1. http://aisel.aisnet.org/jmwais/vol1/iss1/6

Jafar, M. J., Babb, J., & Abdullat, A. (2016). Emergence of Data Analytics in the Information Systems Curriculum. *Proceedings of the EDSIG Conference*. http://iscap.info

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM Framework for Teaching Business Analytics. *Decision Sciences Journal of Innovative Education*, *18*(4), 612–634. https://doi.org/10.1111/dsji.12222

Kim, A. Y., Ismay, C., & Chunn, J. (2018). The fivethirtyeight R package: "Tame Data" Principles for Introductory Statistics and Data Science Courses Chester Ismay Recommended Citation. *Technology Innovations in Statistics Education*, *11*(1). https://scholarworks.smith.edu/mth_facpubs/47

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. T. (2022). The Democratization of Data Science Education. *The American Statistician*, *74*(1), 1–7. https://doi.org/10.1080/00031305.2019.1668849

Lau, S., Nolan, D., Gonzalez, J., & Guo, P. J. (2022). How Computer Science and Statistics Instructors Approach Data Science Pedagogy Differently: Three Case Studies. *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education*, *1*, 29–35. https://doi.org/10.1145/3478431.3499384

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Li, D., Milonas, E., & Zhang, Q. (2021, August 20). Content Analysis of Data Science Graduate Programs in the U.S. *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--36841

Mikalef, P., Giannakos, M. N., Pappas, I. O., & Krogstie, J. (2018). The human side of big data: Understanding the skills of the data scientist in education and industry. *IEEE Global Engineering Education Conference, EDUCON, 2018-April*, 503–512. https://doi.org/10.1109/EDUCON.2018.8363273

Mike, K. (2020). Data Science Education: Curriculum and pedagogy. *ICER 2020 - Proceedings of the 2020 ACM Conference on International Computing Education Research*, 324–325. https://doi.org/10.1145/3372782.3407110

Msweli, N. T., Twinomurinzi, H., & Ismail, M. (2022). The International Case for Micro-Credentials for Life-Wide And Life-Long Learning: A Systematic Literature Review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *17*, 151–190. https://doi.org/10.28945/4954

Price, R., & Ramaswamy, L. (2019). Challenges and approaches to teaching data science technologies in an information technology program with non-traditional students. *Proceedings - 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC 2019*, *Cic*, 49–56. https://doi.org/10.1109/CIC48465.2019.00015

Qian, Y., Lehman, J., Qian, Y., & Lehman, J. (2017). Students' Misconceptions and Other Difficulties in Introductory Pro-gramming: A Literature Review. *ACM Transactions on Computing Education*, *18*(1). https://doi.org/10.1145/3077618

Qiang, Z., Dai, F., Lin, H., & Dong, Y. (2019). Research on the Course System of Data Science and Engineering Major. *2019 IEEE International Conference on Computer Science and Educational Informatization, CSEI 2019*, 90–93. https://doi.org/10.1109/CSEI47661.2019.8938944

Rahimi, E., Henze, I., Hermans, F., & Barendsen, E. (2018). Investigating the pedagogical content knowledge of teachers attending a MOOC on scratch programming. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11169 LNCS*, 180–193. https://doi.org/10.1007/978-3-030-02750-6_14/TABLES/2

Raj, R. K., Parrish, A., Impagliazzo, J., Romanowski, C. J., Aly Ahmed, S., Bennett, C. C., Davis, K. C., McGettrick, A., Susana Mendes Pereira, T., & Sundin, L. (2019). Data Science Education: Global Perspectives and Convergence. *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*, *19*. https://doi.org/10.1145/3304221.3325533

Saddiqa, M., Magnussen, R., Larsen, B., & Pedersen, J. M. (2021). Open Data Interface (ODI) for secondary school education. *Computers and Education*, *174*, 104294. https://doi.org/10.1016/J.COMPEDU.2021.104294

Saeli, M., Perrenet, J., Jochems, W. M. G., & Zwaneveld, B. (2011). Teaching Programming in Secondary School: A Pedagogical Content Knowledge Perspective. *Informatics in Education*, *10*(1), 73–88.

Saltz, J., & Heckman, R. (2015). Big Data science education: A case study of a project-focused introductory course. *Themes in Science & Technology Education*, *8*(2), 85–94. https://www.learntechlib.org/p/171521/

Saltz, J. S., Dewar, N. I., & Heckman, R. (2018). Key Concepts for a Data Science Ethics Curriculum. *ACM Reference*. https://doi.org/10.1145/3159450.3159483

Sentance, S., & Csizmadia, A. (2017). Computing in the curriculum: Challenges and strategies from a teacher's perspective. *Education and Information Technologies*, *22*, 469–495. https://doi.org/10.1007/s10639-016-9482-0

Shahbaz, M., Gao, C., Zhai, L., Shahzad, F., & Arshad, M. R. (2020). Moderating Effects of Gender and Resistance to Change on the Adoption of Big Data Analytics in Healthcare. *Complexity*, 1–13. https://doi.org/10.1155/2020/2173765

Shirani, A. (2016). Identifying Data Science and Analytics Competencies Based on Industry Demand. *Issues in Information Systems*, *17*(IV), 137–144. https://doi.org/10.48009/4_iis_2016_137-144

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *14*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.

Silva, Y. N., Dietrich, S. W., Reed, J. M., & Tsosie, L. M. (2014). Integrating Big Data into the computing curricula. *SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 139–144. https://doi.org/10.1145/2538862.2538877

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364–373. https://doi.org/10.1111/exsy.12130

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). What is hard about teaching machine learning to non-majors? Insights from classifying instructors' learning goals. *ACM Transactions on Computing Education*, *19*(4), 1–16. https://doi.org/10.1145/3336124

Taopan, L. L., Drajati, N. A., & Sumardi. (2020). TPACK Framework: Challenges and Opportunities in EFL Classrooms. *Research and Innovation in Language Learning*, *3*(1), 1–22. https://doi.org/10.33603/RILL.V3I1.2763

Twinomurinzi, H., Mhlongo, S., Bwalya, K. J., Bokaba, T., & Mbeya, S. (2022). Multidisciplinarity in Data Science Curricula. *African Conference on Information Systems and Technology*.

Yadav, N., & Debello, J. E. (2019). Recommended Practices for Python Pedagogy in Graduate Data Science Courses. *Proceedings - Frontiers in Education Conference, FIE*, *2019-Octob*. https://doi.org/10.1109/FIE43999.2019.9028449

Yu, B., & Hu, X. (2019). Toward Training and Assessing Reproducible Data Analysis in Data Science Education. *Creative Commons Attribution 4.0 International (CC BY 4.0)*, *1*, 381–392. https://doi.org/10.1162/dint_a_00053

Zhang, X., Huang, X., & Wang, F. (2017). The construction of undergraduate data mining course in the big data age. *ICCSE 2017 - 12th International Conference on Computer Science and Education*, *Iccse*, 651–654. https://doi.org/10.1109/ICCSE.2017.8085573

Author

Nkosikhona Theoren Msweli BCom Hons (Business Informatics), MIT (Information Systems) is a Lecturer at the University of South Africa, College of Science, Engineering, and Technology. She is currently a Ph.D. candidate at the University of Pretoria pursuing a Ph.D degree in Information Systems. Her research interests include, among other things, data science education, digital and mobile technologies, and online learning. She is a co-editor of NEMISA research colloquium proceedings.

Appendices

https://bit.ly/DSEInstructorPCK

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Towards improving learning experiences in self-paced online learning courses

Sibukele Gumbo[1] and Hossana Twinomurinzi[1]
[1] University of Johannesburg, Johannesburg, South Africa
222239719@student.uj.ac.za, hossanat@uj.ac.za

Abstract

One of the key goals of higher education institutions (HEIs) is student career success. In HEIs, students are given the necessary subject-specific information, skills, and experience to achieve this. In South Africa, which has historically been one of the technological leaders in Africa, there are currently insufficient Fourth Industrial Revolution (4IR) specialists graduating from higher education to meet the rising need for a skilled workforce in those fields. The completion rate of supplementary online courses to expand this 4IR skills base is also a major concern. In this paper, we report on the strategy implemented to improve the course completion rate of a self-selected sample of students who attended face-to-face Data Science introduction workshops in the Eastern Cape province of South Africa. We achieved a 76% completion rate of two MOOCs courses in these workshops. We examined the contribution of the participant's personal factors and background contextual factors. We also listed any other factors suggested by the students, which collectively contribute to the learning experience construct of the social cognitive career theory (SCCT) and serve as a practical means of improving 4IR skills exposure and the outcome expectations. Therefore, this paper shows the mechanism which can be used to offer 4IR programmes, in HEIs, to raise students' self-efficacy and outcome expectations in fields such as Data Science. In addition, the learning experience inputs can be utilized to increase students' interest in majoring in 4IR courses.

1 Introduction

Massive Open Online Courses (MOOCs) are online programs intended for open (free) enrollment to a large audience. MOOCs offer a cost-effective and flexible method to master new skills, enhance careers, and deliver high-quality educational experiences on a large scale. MOOCs are an example of a self-paced learning strategy, which can be used for career advancement and supplemental learning. Other learning strategies include the use of internet resources, academic databases and discussion forums. The first MOOC was founded by Dave Cormier and was based on the connectivist theory of learning which emphasizes communal over individual learning (Mooc.org, 2023). Sustainable

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Development Goal 4 (SDG4) of the United Nations (UN), which calls for "inclusive and equitable quality education. for everyone," is in accordance with the idea of MOOCs (United Nations, 2023). Given the high young unemployment rate, which positively correlates with limited education, SDG4 should be adopted by developing nations like South Africa (SA). Even though MOOCs have the potential to enhance educational results, countries still struggle to fully accept, integrate, and utilize them. For example, one publication states that the MOOCs current completion rate is 7–13% (Nesterowicz et al., 2022).

South Africa has twenty six public universities distributed throughout its nine provinces (SA Department of Basic Education, 2021). Four of them are located in the Eastern Cape province. None of the universities in the Eastern Cape provide Data Science courses as part of their regular curricula, according to (Twinomurinzi et al., 2022). Data Science is an interdisciplinary field with an emphasis on information extraction from data sets, which are frequently very large in size. The field includes analysis, preparing data for analysis, and presenting results to support organizational choices [5]. We see MOOCs as a viable alternative for building Data Scientists in the Eastern Cape. MOOCs can provide free access to online courses in HEI, in a variety of topic areas, including Data Science, for as many people as is practical. The principle that "information should be given freely and the desire to learn should be satisfied without demographic, economic, and geographical limits" is the foundation of MOOCs (Mooc.org, 2023).

We opted to use the learning experience construct of the social cognitive career theory (SCCT) model to assess the adoption of the Data Science programmes in a pilot workshop conducted at an Eastern Cape HEI. Robert W. Lent, Steven D. Brown, and Gail Hackett developed the SCCT model shown in Figure 1 (Medugorac et al., 2020), in 1994. It is a career development theory that explains how individuals make career choices by considering their personal and environmental factors, including their personal beliefs and attitudes, their self-efficacy, their outcome expectations, and their situational factors. It provides a systematic explanation for career development, respond to development of times as stated by (Buthelezi et al., 2009) and helps on focusing on special group like the youth on 4IR career The theory proposes that individuals make career decisions based on a dynamic interaction between their internal cognitive and affective processes and external environmental factors. SCCT provides a framework for understanding the complex process of career decision making and can be used to guide career counseling and support individuals in making informed career choices. It has gained appeal as a theoretical framework for examining career-related decisions across a range of groups, including jobs in Science, Technology, Engineering, and Mathematics (STEM). It is based on Albert Bandura's social cognition theory. Over the past 20 years, the SCCT model has been applied in numerous contexts to examine both professional and academic behavior. Its premise is that the best explanation for career development and decisions connected to careers is the consequence of a complicated web of interrelated elements.

The theory also offers opportunities for social support by instilling expectations, self-efficacy and through the use of observational learning as well as some reinforcements in order to achieve change in behavior. It further assist on personal judgment as to how well an individual can be able to execute a course of action which they are required to encounter with contemporary situations like the era of 4IR. (Buthelezi et al., 2009) specified that SCCT helps to construct a three factor interaction of careers which is crucial on this aspect too which comprise of

1. Self efficacy (can I do this),
2. Outcome expectations (what will happen if one doe it),
3. Personal goal (this regards to as how much one wants to do it).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

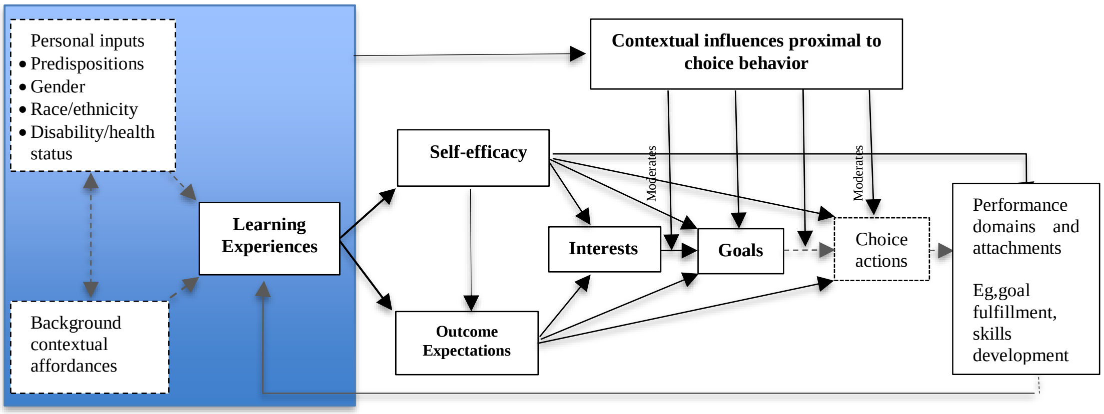


Figure 1: SCCT model (Medugorac et al., 2020) adapted as follows: we study factors highlighted in blue and relationships are indicated by dashed lines, while the remaining variables and factors are represented by solid lines.

According to SCCT (Figure 1), the career-relevant learning experience construct is influenced by individual inputs (e.g. gender and predispositions) and background contextual affordances. Learning experiences also includes performance successes, vicarious learning, social persuasion, and physiological arousal [6]. In order to understand the potential influences on students in HEI and their acceptance of MOOCs, a pilot survey was conducted to identity whether personal and background characteristics influenced students' MOOC learning experiences. We examined these learning experience precursors (highlighted in blue in Figure 1), to determine the extent in which they influenced students' completion of MOOCs courses. The respondents also shed some light on additional factors that affect MOOCs completion and those that could improve these online programs in developing countries. We also discuss 1) the aspects of the Data Science MOOCs that the respondents consider useful or valuable and 2) suggestions by the respondents on how to improve the courses. The following research questions were addressed:

• RQ1: What are the personal and background contextual characteristics of the self-group of HEI students undertaking MOOCs in the Data Science domain?

• RQ2: What other factors can improve MOOCs completion?

The rest of the paper is structured as follows: Section 2 describes the method used to collect data, Section 3 is the results and discussions and, Section 5 is the conclusion.

2 Method

In (Class Central, n.d.), IBM is listed as a provider of MOOCs. The choice to use its platform was based on the availability of free, industry-related courses to increase student competence. We invited first and second year Information Technology (IT) students based at a HEI in the Eastern Cape to a face-to-face workshop to introduce two Data Science courses offered on the IBM online platform. The purpose of the workshop was to introduce students to self-learning technologies and educational opportunities so they update their knowledge, skills, and understanding. IBM offers each participant free access to content, course material, lab exercises, and a certificate. Text, videos, online access to labs, and an examination at the end of each course comprise the content.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The courses chosen were Data Science 101 (IBM, n.d.-a)and Data Science Tools (IBM, n.d.-b). The students were required to bring their laptops and headphones to the workshops. We made use of wireless internet-equipped university facilities. At the end of the workshop, we used an online instrument to conduct a survey where willing workshop participants gave feedback on the following elements:

1. Personal inputs such as gender, age, and ethnicity;
2. Background contextual affordances, including hometown, highest qualification earned, grades in math and English in the matriculation, prior online training, and successful completion of online training;
3. Aspects of the workshop and courses that were considered useful or could be improved.

We also observed all the participants throughout the session and observed whether respondents proceeded with additional online courses after a workshop session. The results from the workshops are described in the next sections.

3 Results and discussions

3.1 Workshop participation

Three Data Science pilot face-to-face workshops took place in October 2022. Table 1 shows the number of students who 1) expressed their interest in the workshops, 2) attended the workshops and 3) completed the online survey. In summary, 320 students expressed their interest in the workshops, 2) 201 students (or participants) attended and 3) 72 students (or respondents) completed the online survey to evaluate their learning experience. The three venues used for the workshops could to accommodate 40, 70 and 100 students, respectively. Due to space restrictions, additional students who had not rsvped were turned away. None of the participants had previously undertaken any IBM online courses; new online IBM platform user accounts were created for all of them.

	Expressed interest	Participated in workshop	Completed online survey
Venue 1	37	26	14
Venue 2	113	73	50
Venue 3	170	102	8
TOTAL	320	201	72

Table 1: Expressed interest vs Participated in workshop vs Completed online survey

3.2 Survey respondents' personal inputs

The survey respondents personal inputs such as gender, age, and ethnicity are summarized in this section. These responses represent key pieces of their demographic data.

39 respondents (or 54.2) were female and 33 respondents (or 45.8%) were male as illustrated in Figure 2. This demonstrates that there is a rise in female students who are interested in technology-related courses at HEIs. The interest in females in technology courses complements the survey by Deloitte Worldwide which states that 'leading global technology organizations will, on average, have nearly 33% of women working for them in 2022, up little over 2 percentage points from 2019'(Hupter et al., 2021).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

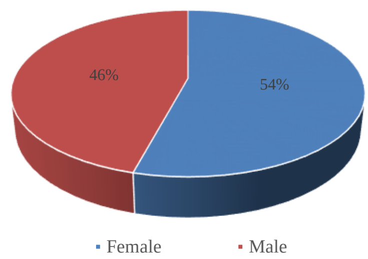


Figure 2: Gender of respondents

Figure 3 illustrates that the respondents' ages ranged from 18 to 34, with a mean age of 22. In the South African perspective, all respondents were considered youth. The introduction of Data Science courses to them contributes towards the advancement and development of a youthful 4IR workforce.

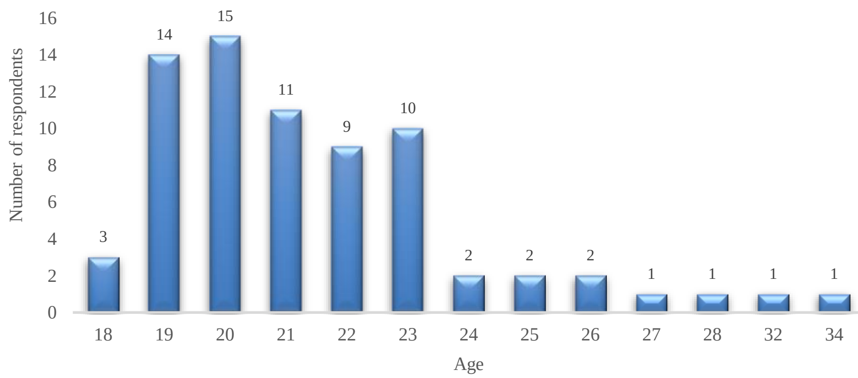


Figure 3: Age distribution of respondents

The four recognized ethnic groups in South Africa are Blacks (or Africans), Whites, Coloreds, and Indians. According to the 2011 census, there are 76.4% of Black South Africans, 9.1% of White South Africans, 8.9% of Colored South Africans, 2.5% of Indian South Africans, and 0.5% of Other/Unspecified South Africans ("Ethnic Groups in South Africa," 2022). Figure 4 displays the distribution of Data Science survey respondents' ethnic groups. One person was Colored, and the rest were African. Therefore, this study was based on non-native English-speaking MOOCs survey respondents.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

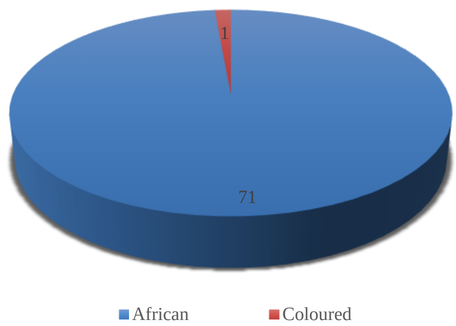


Figure 4: Distribution of ethnic groups

3.3 Survey respondents’ background contextual affordances

The environmental background elements for the respondents are outlined in this section. We summarised the respondents' home areas, their highest degree held at the time they attended the Data Science sessions, their matriculation Math and English final results, and whether they had previously taken or completed a MOOC.

The distribution of the survey respondents’ home areas are illustrated in Figure 5 and Figure 6. The majority of the students came from the Eastern Cape (68%) province where the university is located. 10% of the respondents were from Gauteng and 14% of the respondents came from Kwa-Zulu Natal. The remaining respondents were from the following geographic regions: Limpopo (6%) and 1.4% from Free State, Mpumalanga and North West. There were no Western Cape and Northern Cape respondents.

Figure 6 shows that 70% of the respondents were from urban areas, while 30% were from rural areas. This information demonstrates that the university enrols students from different regions of South Africa. The diversity of provinces and rural complement numbers is encouraging because empowering students from disadvantaged communities in STEM related careers is a significant step towards breaking the cycle of poverty, crime, and unemployment.

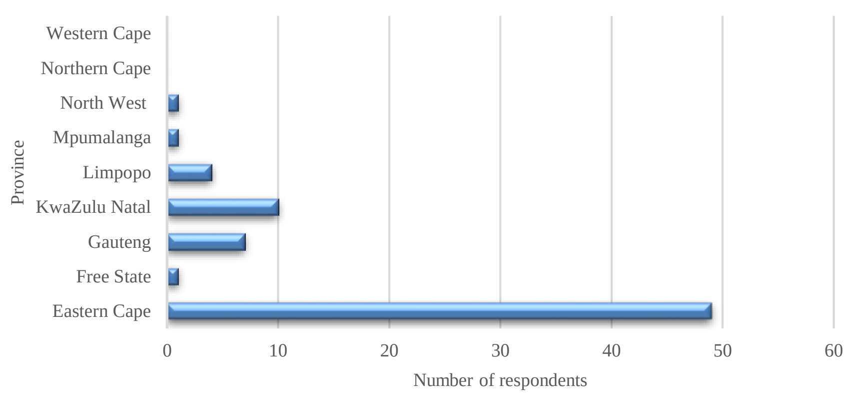


Figure 5: Distribution of respondents' home province

Preprints of Proceedings of
NEMISA Digital Skills Conference (Colloquium) 2023
15-17 February 2023
Coastlands Umhlanga Hotel & Convention Centre
Durban, South Africa
ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

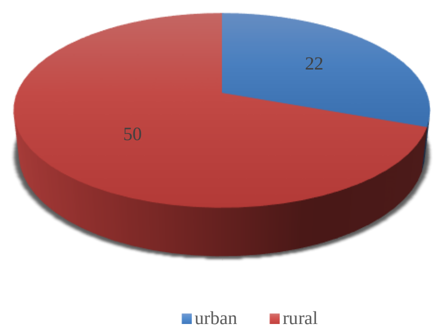


Figure 6: Distribution of respondents' home area

Amongst the respondents, 62 (or 86.3%) noted that Matric (National Senior Certificate) was their highest qualification. Matric, which is the qualification earned after completing the Grade 10 to Grade 12 period of further education and training, is equivalent to NQF level 4. Some of the respondents had alternative HEI credentials such as Bachelor, BTech and Diplomas. The distribution of the respondent's highest qualifications are illustrated in Figure 7.

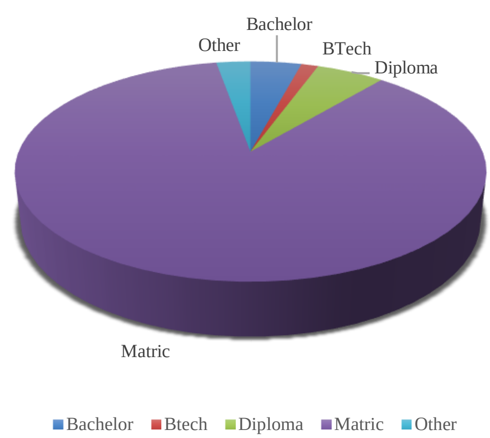


Figure 7: Distribution of highest qualifications

There are numerous concerns about the state of South Africa's educational system, particularly in Science and Mathematics subjects. All subjects have a standard pass rate of 30%, which sets a low pass bar. To make matters worse, more students than ever are choosing to take basic Mathematics Literacy exams, which disqualifies them from pursuing the courses that require Mathematics, at universities. We asked the respondents to state whether they had undertaken Mathematics Literacy or Mathematics examinations at Matric level.

Since none of the respondents were native English speakers (and the MOOCs chosen were taught in English), we also requested their level of English skill. Their English, Mathematics or Mathematics Literacy results are illustrated in Figure 8.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The respondents' English scores ranged from ‘moderate achievement’ (40–49%) to ‘outstanding achievement’ (80–100%). The respondents' average English grade was 60-69% which is considered a ‘significant accomplishment’ according to (Nethononda, 2022). We concluded that the respondents had a good understanding of the language used throughout the IBM course content.

16 respondents reported on their Mathematics Literacy results and 56 respondents reported on their Mathematics results. The fact that more respondents reported on Mathematics results, which is held to a higher level, was encouraging. The Mathematics results of the respondents ranged from ‘moderate achievement’ (30-39%) to ‘outstanding achievement’ (80-100%), while the Mathematics Literacy results ranged from ‘adequate achievement’ (50-59%) to ‘outstanding achievement’ (80-100%). This means that the Mathematic average of 40 – 49% was lower than that of Mathematics Literacy (60-69%).

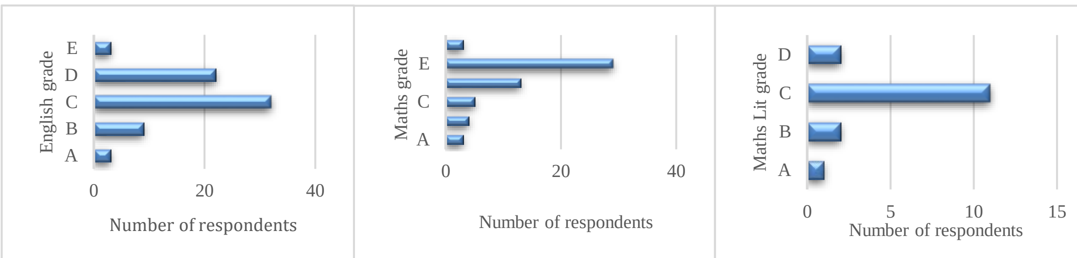


Figure 8: English, Mathematics and Mathematics Literacy matric results, respectively

The respondents were asked if they have even taken any MOOC/s prior to these face-to-face introduction sessions. 41 (or 57%) responded that they have taken a MOOC before, while 31 respondents (or 43%) had never done so. These results are illustrated in Figure 9. Although the participants did not specify which MOOC they had undertaken, they verbally acknowledged that none of the courses they had taken were IBM courses.

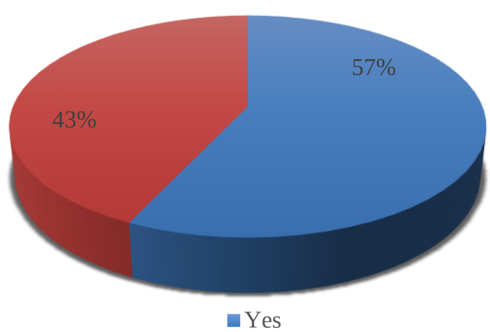


Figure 9: Undertaken online MOOC/s before

As previously stated, the face-to-face sessions were conducted to introduce two IBM courses to the participants. The first course, Data Science Foundations – Level 1, covered aspects such as ‘What is Data Science?’, ‘Why Data Science?’ and ‘Where to start?’. The second course, ‘Data Science Tools’ was a theoretic introduction to Data Science Tools such as Python and R. We observed that all

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

participants were excited to be part of the sessions because there are limited opportunities to study Data Science in the Eastern Cape province.

All 72 survey respondents completed at least one of the two MOOCs over the course of the workshop day. Specifically, only one of the 72 respondents completed one course, while the respondents completed at least two courses.

We also observed how many workshop participants completed at least one course during the workshop session. 141 (or 70%) of the 201 workshop participants submitted at least one verified IBM course completion certificate by the end of the workshop. This MOOC completion rate was substantially higher than the MOOC completion rate noted in (Nesterowicz et al., 2022).

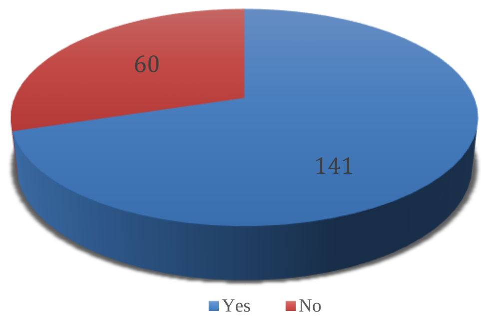


Figure 10: Completed at least one IBM MOOC by the end of the workshop session

We monitored the workshop participants for a month to ascertain whether they would complete additional IBM courses on their own. The total number of certificates received by the end of the month are illustrated in Figure 11. The highest number of IBM course completions was 12, while the lowest number was 0. 12 (or 22%) of the workshop participants were unable to finish any IBM course and 36 participant's finished just one course. This means that 24% of the participants did not fulfill the workshop requirements, while 76% completed the two IBM courses.

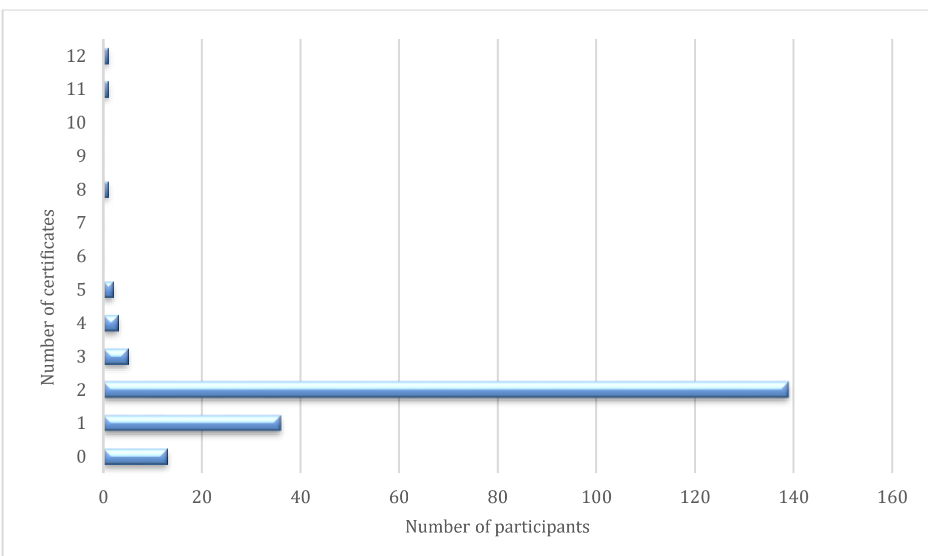


Figure 11: Certificates received between October and November 2022 from all workshop participants

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.4 Other survey feedback

The online survey respondents were asked what aspects of the workshop and course content they found valuable. The respondents appreciated exposure to Data Science concepts, basics and tools and the fact that the content was relevant to students who were not in the IT field. They felt confident that they could do more with the introductory skills which they had learnt.

In terms of the course layout, some respondents noted that it was useful that the MOOCs were in the form of video content which featured interviews of Data Scientist's personal experiences. The introduction of programming languages such as R and Python and the prospect of practically acquiring additional programming skills made them verbally commit to exploring more courses on the IBM platform. They also appreciated that there were quizzes and an examination to test their knowledge, although they would have preferred class discussions. The respondents felt that the course completion certificate was the most valuable component of the workshop as it would assist them in applying for jobs in the future.

The survey respondents provided feedback on what aspects of the workshop, course content and general learning can be improved. The respondents' main point was that coaching sessions, such as this workshop series, were critical in increasing the exposure of Data Science courses in the Eastern Cape. They felt that it was the responsibility of HEI to conduct these coaching sessions. This is in line with the social persuasion component of learning experiences, in the SCCT model. Through coaching, the participants were taught to believe that they could effectively obtain Data Science skills. They further suggested that the workshops and content should be more accessible and be introduced to learners in high school. If possible, these workshops should be incentivized. The face-to-face workshops, through coaching, were an opportunity for the physiological arousal of individuals' about MOOCs.

The respondents noted that more class discussions sessions should be allowed in order for MOOCs courses to be easier. The workshop session only allowed limited interaction between participants. We found that the concept of vicarious learning influenced the completion rate as the participants saw others completing the MOOCs successfully. However, a few respondents felt that learning as individuals was more fruitful, therefore the workshops should be offered online only. Performance success based on the completion of the first MOOC encouraged most participants to complete the second MOOC with more ease. While some respondents stated that more advanced MOOCs content and more practical work (instead of theory) should be introduced in the future; the majority of the respondents were content with the workshop and MOOCs courses chosen as an introduction.

The workshop participants were fortunate to have laptops, smartphones and a connection to the internet. However, we found that due to electricity challenges, the internet connection was unstable, and therefore affected the completion rate of the MOOCs as the online video content could not be viewed or downloaded. In a country like South Africa, where at present electricity is a challenge, it might be necessary to present offline MOOCs to improve the learning experience. There were additional suggestions that MOOC online content should also be printed as this would aid further research and studies and to promote continuity.

4 Conclusions

This paper reports on a MOOCs workshop series to introduce the concept of Data Science to HEI students in the Eastern Cape province of South Africa. We examined traits that are precursors of the SCCT's learning experience construct and provide us with a useful way to increase 4IR skill exposure and outcome expectancies in IT students enrolled at this particular university. The study showed that there is a strong interest in courses such as Data Science in HEI as shown by the number of students

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

who expressed their interest and the course completion rate of 76%. We noted the lack of Data Science programmes in the Eastern Cape province, therefore the findings demonstrated that the skill-base in this field can be increased if more effort is put by HEIs into establishing 4IR training programmes through MOOCs. Encouraging youth to develop skills is one of the first stages in creating the 4IR workforce and can aid in achieving SDGs. The concept of formally integrating MOOCs into HEI curriculum is a mechanism which can be used globally to increase the skills development of students.

Before a conclusion can be drawn regarding their learning experiences effect on student development, more research measurements and analyses are necessary on other constructs such as outcome expectations and self-efficacy in the SCCT model. Our analysis invites further exploration into other constructs in the SCCT model. Future studies should examine the MOOCs results and make use of hypothesis testing to determine how effective 4IR programmes are.

One of the study's limitations was the non-normality of the data in this sample; there were very few variances especially in the respondents demographic data. According to the features of the sample, students who took the Data Science introduction had high levels of interest before data collection, which probably affected the predictive value of these dimensions in subsequent analyses. This study suffers from problems related to voluntary responding, as certain groups are more likely to participate in a research, due to their personal characteristics. The participation of female respondents to male respondents was almost equal.

Students can improve 4IR skills and personal performance standards through ongoing physical exposure to online courses. People are likely to set objectives for maintaining or growing their involvement in an activity as soon as they become interested in it. People are more likely to develop a long-lasting interest in a task when they believe they are capable of executing it and when they anticipate the task will result in positive outcomes. The implementation of finely planned support of community and institutions for career choices in sustainability may be useful in giving rise to careers supporting the transition to more sustainable forms of economy.

References

Buthelezi, T., Alexander, D., & Seabi, J. (2009). Adolescents' perceived career challenges and needs in a disadvantaged context in South Africa from a social cognitive career theoretical perspective. *South African Journal of Higher Education*, *23*(3), Article 3. https://doi.org/10.4314/sajhe.v23i3.51033

Class Central. (n.d.). *IBM Courses*. Retrieved January 15, 2023, from https://www.classcentral.com/institution/ibm

Ethnic groups in South Africa. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ethnic_groups_in_South_Africa&oldid=1122246833

Hupter, S., Mazumder, S., Bucaille, A., & Crossan, G. (2021). *Women in the tech industry: Gaining ground, but facing new headwinds*. Deloitte Insights. https://www2.deloitte.com/za/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/statistics-show-women-in-technology-are-facing-new-headwinds.html

IBM. (n.d.-a). *Data Science 101*. Retrieved February 14, 2023, from https://cognitiveclass.ai/courses/data-science-101

IBM. (n.d.-b). *Data Science Tools*. Retrieved February 14, 2023, from https://cognitiveclass.ai/badges/data-science-tools

Medugorac, V., Šverko, I., & Babarovic, T. (2020). Careers in Sustainability: An Application of Social Cognitive Career Theory. *International Journal for Educational and Vocational Guidance*, *20*(3), 477–499. https://doi.org/10.1007/s10775-019-09413-3

Mooc.org. (2023). *Massive Open Online Courses*. https://www.mooc.org

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Nesterowicz, K., Bayramova, U., Fereshtehnejad, S.-M., Scarlat, A., Ash, A., Augustyn, A. M., & Szádeczky, T. (2022). Gamification Increases Completion Rates in Massive Open Online Courses. *International Journal of Information and Communication Technology Education*, *18*(1). https://doi.org/10.4018/IJICTE.294447

Nethononda, M. (2022). *Matric symbols and level*. Matric College. https://www.matric.co.za/matric-symbols-and-level/

SA Department of Basic Education. (2021). *List of South African Universities*. https://www.education.gov.za/FurtherStudies/Universities/tabid/393/Default.aspx

Twinomurinzi, H., Mhlongo, S., Bwalya, K., Bokaba, T., & Mbeya, S. (2022). Multidisciplinarity in Data Science Curricula. *African Conference on Information Systems and Technology*. https://digitalcommons.kennesaw.edu/acist/2022/presentations/21

United Nations. (2023). *Sustainable Development Goals (SDG 4) Goal 4: Quality Education*. United Nations Western Europe. https://unric.org/en/sdg-4/

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Factors Contributing to Cybersecurity Awareness, Education and Training

xxxxxxxxxxx[1,2*] and xxxxxxxxxxxxxx[1†]
[1] EasyChair
[2] University of xxxx, xxxx, xxxx
xxxx@xxxx.com, xxxx@xx.ac

Abstract

This paper aims to identify the contributing factors for successful cybersecurity awareness, education, training, programs. The study adopted the systematic literature review method and included 58 primary studies. The study explores approaches for cybersecurity awareness, education and training to improve cybersecurity skills and practice in the extant literature. The study offers several recommendations towards effective cybersecurity awareness, education and training programs.

Keywords: Digital Skills, Cybersecurity, Cybercrime, Awareness, Education, Training

1 Introduction

The past few decades have seen an increasing reliance on information and communication technologies (ICT's) and the internet throughout various sectors of society. With this growing reliance on technology there has also been an upsurge in cyber threats and cyber related crimes (Muthuppalaniappan & Stevenson, 2021). Additionally, it is noted that during the recent COVID-19 pandemic an increasing number of organisations shifted to digital modes of operation due to the social distancing requirements (Naidoo, 2020). Consumers have also progressively accepted more online formats of shopping, education, entertainment and digital communication channels as their new normal. It is evident that the internet permeates many areas of society (Akdemir & Lawless, 2020). The context of the COVID-19 pandemic also further exacerbated the challenges faced with cyber related crimes (Naidoo, 2020) (Muthuppalaniappan & Stevenson, 2021). Even as we enter a post-pandemic phase the trend in people falling prey to cybercrimes continues on an upward trajectory (Monteith, et al., 2021). Literature reports that the human elements are a dominant component of cybersecurity since people's behaviours, personality traits, online activities, and attitudes towards technology impact their

[*] Masterminded EasyChair and created the first stable version of this document
[†] Created the first draft of this document

Coastlands Umhlanga Hotel & Convention Centre
Durban, South Africa
ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

vulnerability online (Monteith, et al., 2021). The human aspects are the main facilitator of victimisation in cybercrime (Akdemir & Lawless, 2020) and humans are often referred to as the weakest link in cybersecurity (Aldawood & Skinner, 2018). The literature highlights that one of the ways to fight cyber related crimes and particularly address weaknesses related to the human element is to increase awareness, education and training programmes (Bele, Dimc, Rozman, & Jemec, 2014) (Aldawood & Skinner, 2018). This study sought to understand the contributing factors for successful cybersecurity awareness, education, training, and skills programs as reported in the extant literature. The study adopted the systematic literature review (SLR) method to identify awareness and education approaches that may support safer cybersecurity practices, reduce cybersecurity incidents and cyber related crime.

2 Cybersecurity Overview

Cybersecurity is defined as "the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies" that are used to protect users from cybercrimes (von Solms & van Niekerk, 2013). Cybersecurity can also be understood as securing hardware, software, data and information that exists in an online system from various types of breach (Tirumala, Valluri, & Babu, 2019).

Cybercrime is a broad term and encompasses criminal activity involving computers or computer networks. Examples of cybercrimes range from advanced fee fraud scams and general online confidence schemes to unauthorized access and copyright infringement. Despite continuous intervention by institutions and governments to curb the rise of cyber criminality, due to increasing digital interconnectivity, millions of people worldwide are continuing to be negatively affected by cybercrimes annually. Cybersecurity awareness, education, and training are three areas that can be considered to improve a user's cybersecurity detection and prevention competencies and skills (Holdsworth & Apeh, 2017).

2.1 Impact of Cybercrime

There are various types of cybercrime that organisations may suffer as identified by (Paoli, Visschers, & Verstraete, 2018) and these include: the illegal access to IT systems, cyber espionage, data or system interference, cyber extortion and internet fraud. Individuals are increasingly becoming targets of phishing attacks resulting in billions of dollars lost annually (Jensen, Dinger, Wright, & Thatcher, 2017). Individuals can be attacked by malware, trojans, ransomaware, identity theft and loss of sensitive information such as credit card details or passwords. Organisations on the other hand suffer losses due to cybercrime which can be related to: material costs (e.g. damage to infrastructure), personnel costs (the time spent by employees addressing the cyber incidents), hardware and software replacement, loss of assets, revenues lost, reputational harm and loss of privacy (Paoli, Visschers, & Verstraete, 2018).

Approximately 15 million data records were exposed worldwide through data breaches over the 3rd quarter of 2022 representing a 37% increase from the previous quarter (Statista, 2022). There is concern that these trends will continue to grow. Governments from around the world have observed the destruction caused by cybercrimes and are continuously trying to combat these activities. Government approaches for combating cybercrimes vary but generally involve collaborating with other governments and agencies to better police the wide geographical reach of the crimes; and criminalizing cybercrimes broadly to discourage the behaviour (Broadhurst, 2006). Organisations and institutions also implement special protection and securities to reduce system vulnerabilities, but it is still almost impossible to prevent cyberthreats, especially from a global perspective (Broadhurst, 2006). Although the increasing

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

policing, regulations and security advances assist in mitigating and preventing some of the damage done by cyber criminals, cybercrime continues to be a burdening issue affecting millions of people, organisations, and governments every year. A significant contributing factor to the prevalence of cyber security is vulnerabilities in systems and as well as in people (Platsis, 2019).

2.2 Cybersecurity Awareness, Education and Training

Cybersecurity awareness is the understanding and awareness of cybersecurity threats to enhance a person's ability protect themselves against cyber-attacks in the online context (Muhirwe & White, 2016). It is the ability of being mindful and alert by performing tasks in a secured manner when utilizing a computer or mobile device. The factors that contribute to cybersecurity awareness are the users knowledge, attitude and behaviours (Muhirwe & White, 2016). Efforts towards user awareness and ongoing education are both considered essential in the fight against of cyber threats (Smyth, Curran, & McKelvey, 2019). Awareness, educational programmes and training can drive an effective and appropriate cybersecurity culture (Smyth, Curran, & McKelvey, 2019).

The goal of cybersecurity education is to instil long lasting skills and change behaviours . However, it is noted that organization may find that implementing effective cybersecurity education and training is challenging. Cybersecurity education should assist users in understanding various concepts (Javidi & Sheybani, 2018) and additionally it should be targeted such that it helps users stop old behaviours and adopt new practices aligned with internet safety (Smyth, Curran, & McKelvey, 2019).

3 Research Method

The study adopted the SLR method and aimed to address the following research question: "what are the contributing factors for effective cyber security education, training, and awareness programs?". The SLR approach is explained below.

3.1 Search, Screening and Selection Process

The journal articles and conference papers for this systematic literature review were searched using a search string which included the following key terms (and applicable synonyms):

- Critical success factors
- Cybersecurity
- Awareness
- Education
- Training

The search was conducted by using the following databases: IEEE Xplore Digital Library; WorldCat Discovery Service; Emerald Insight; ProQuest; and ScienceDirect. The search resulted in 3927 articles that were returned. These were evaluated against the predefined inclusion and exclusion criteria:

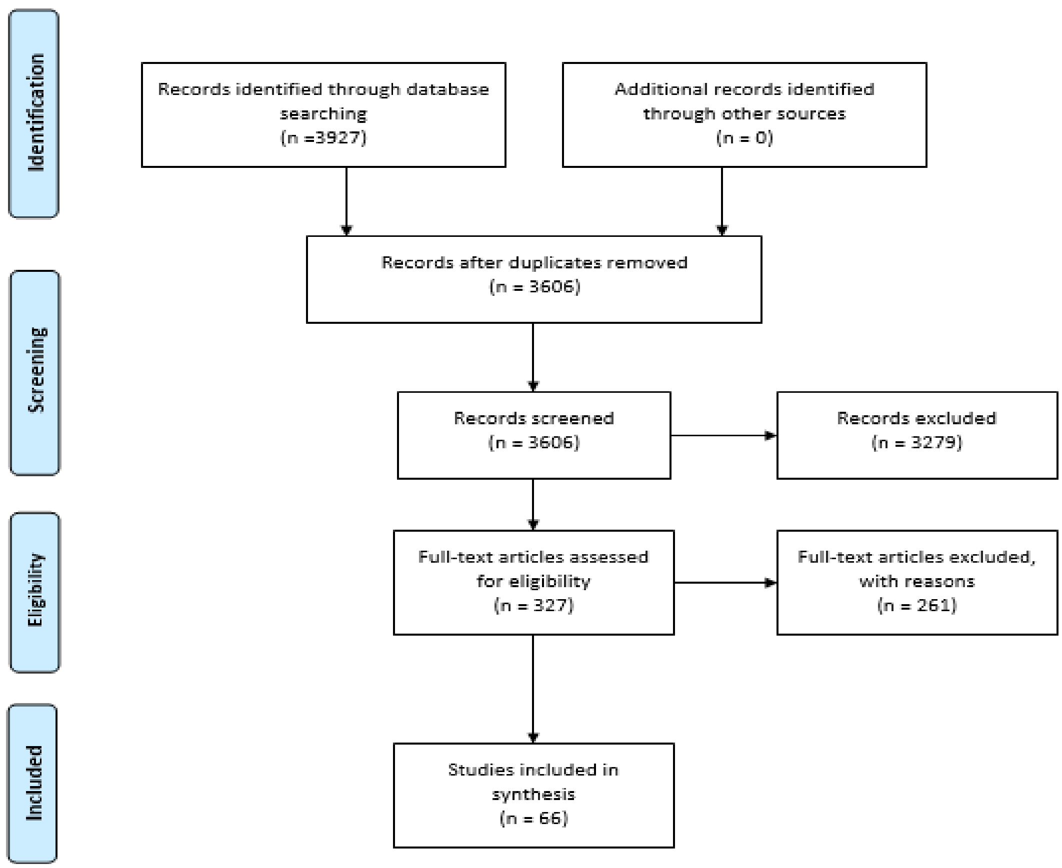
Table 1: SLR Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
1. Publications written in the English language.	1. Publications where only the abstract but not the full text is available.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

2. Publications that presents a method, technique, or process for educating, training, or raising awareness regarding cyber security and cybercrime. 3. Publications that show how education, training, and awareness programs have worked. 4. Publications explaining cyber security and cybercrime. 5. Publications that were published between 2005 and 2020.	2. Duplicate papers. 3. Publications not relating to cybercrime or cyber security. 4. Publications focused on technological applications to deter cybercrimes, not relating to education, training and awareness.

Figure 1: PRISMA Chart



After evaluating the articles 66 eligible articles remained. These were then taken through a quality assessment process.

3.2 Quality Assessment and Data Analysis

A further evaluation of the articles was conducted using the following quality assessment questions:

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

1. The study details research about either cyber security education, training and / or awareness programs.
2. The study discusses factors necessary for the successful implementation of cyber security education, training, and awareness programs clearly?
3. Does the article specify/discuss the research design and justify the appropriateness of the research design/methodology?

A scoring system for each question was adopted as follows: 1 point for yes, 0.5 points for partly and 0 points for no. The scoring of the articles was done to assist in weighing the importance of the studies identified during the primary study selection and ensuring that the results of the selected studies would be appropriate to answer the research question. This resulted in a final of 58 papers that were used in the SLR analysis and discussion. The data from the studies (articles) was analysed using the thematic analysis approach (Maguire & Delahunt, 2017) with the aim of identifying factors contributing to cybersecurity awareness, education and training to support appropriate skills and behaviour.

4 Findings and Discussion

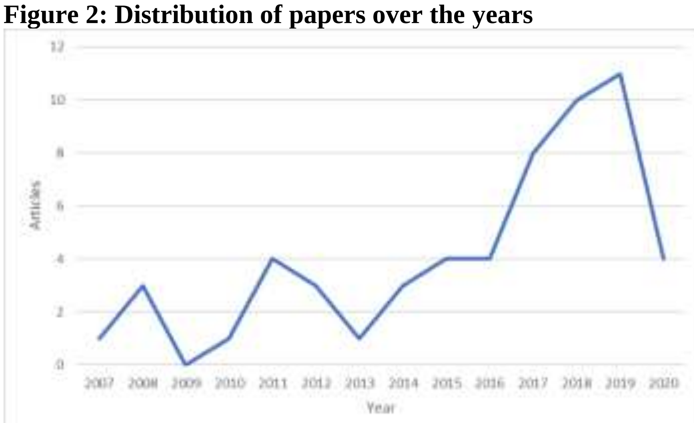
4.1 Overview of the Studies

The types of papers that were included for analysis are highlighted below:

Table 2: Study Type

Study Type	Number of Articles
Review	21
Case Studies	6
Survey	21
Experiment	10
Total	**58**

Figure 2: Distribution of papers over the years



ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The following sections discuss the factors that may be considered for successful cybersecurity awareness, education, training, and skills programs.

4.2 Assessing Cybersecurity Awareness

The first theme and contributing factor is related to assessing the cybersecurity awareness of users. It is vital to ensure that the users for the awareness programs are categorized so that the correct awareness message is directed towards the right participants (Rahim, Hamid, Mat Kiah, Shamshirband, & Furnell, 2015) (Shah, Jones, & Choudrie, 2019) (Takata & Ogura, 2019). Thus, before developing a cybersecurity education, training, and awareness program, it is important to assess individuals and tailor the programs for the users perceived vulnerabilities. By changing the behaviour of individual users, the entire organisational approach to cybersecurity can improve (Chen, Dawn Medlin, & Shaw, 2008).

There are multiple methods used to assess cybersecurity awareness including: survey- based questionnaires, value-focused approaches, vocabulary tests, observations, gaming tools, focus groups, direct classroom training, text-based training, discussions in teams, posters, educational presentation, online lessons, video-based training, newsletter articles, interviews, accessing user responses to emails, document reviews, and elements of situational awareness, which uses scenario-based content to bring security threat awareness to the users (Chen, Dawn Medlin, & Shaw, 2008). These methodologies should be used together as it will better determine user's awareness and allow for a more focused education and training (Bishop , et al., 2017) (Sari & Prasetio, 2017) (Tschakert & Ngamsuriyaroj, 2019).

Users should be assessed based on their vulnerabilities in cybersecurity awareness and education as unified training has its drawbacks. The education and training should be based on the person or groups actual responsibilities and use of the internet (Aldawood & Skinner, 2019)

4.3 Pedagogical Approaches

There are two broad types of training which are computer-based training and instruction- led training. Computer-based training can be conducted by training videos, guided instructions, interactive applications, web-based courses, and use assessments, quizzes, and mini-challenges to access the trainee's knowledge (Ghafir, et al., 2018) . The advantages of computer-based training are that it is a cost-effective training method, that is easy to deliver and has a flexible structure that provides easy access to information. Unfortunately, computer-based training does not provide sufficient support or help and can also feel redundant for a skilled trainee (Ghafir, et al., 2018). Instruction-led training has proven to be a very effective method for behaviour development. This method of training can be as training events or workshops (Ghafir, et al., 2018). This approach is effective because it is tailored towards the security needs of the organisation or trainees and there is an opportunity for immediate direct feedback and face to face communication. Instruction-led training is however costly and takes long to conduct.

Training can also be introduced by a phased approach consisting of two phases. The first phase involves creating awareness and giving users knowledge of common and modern cybersecurity threats they may face, the ability to distinguish between these cybersecurity threats, and the ability to identify phishing emails and fake websites (Frauenstein & von Solms, 2014). The second phase is to train the users and provide them with the competencies needed to use the technological controls that assist in combatting phishing attacks (Frauenstein & von Solms, 2014).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Individual security awareness training programs help users to be able to improve the way they protect themselves from cyber threats and report any security violations they witness (Hagen, Albrechtsen, & Ole, 2011) (McCrohan, Engel, , & Harvey, 2010). These awareness programs do not last indefinitely in the users' memory, so they should be performed continuously as opposed to once-off. It is also effective to incorporate cybersecurity education in schooling system and institutions of higher learning as it allows young students to be aware of cybersecurity issues (Ahmed, et al., 2019) which can be taken forward into their personal lives and careers.

Another option is gamification. Gamification, in the context of cybersecurity training, is a technique that uses game design and principles to provide cybersecurity education. Gamification is a tool that allows e-learning to be interesting and engaging thus capturing the user's full attention and leading to better retention (Holdsworth & Apeh, 2017). The benefits of this technique are that there is direct communication, in that, the need for user participation and feedback are highlighted in the game design, thus improving the results with regards to changes in users cybersecurity awareness and behaviour in an appealing and enjoyable manner (Alotaibi, Furnell, Stengel, & Papadaki, 2017) (Tioh & Mina, 2015). This approach of active engagement greatly improves the retention of knowledge amongst the users learning it and the game design additionally allows for flexibility and inclusivity in any topic (Alotaibi, Furnell, Stengel, & Papadaki, 2017) (Labuschagne, Veerasamy, Burke, & Eloff, 2011). Gamification also improves richness of the information. The richness of security awareness information refers to the various forms of media that can be used, such as hypermedia, multimedia, and hypertext. The use of these media types allows for training material to be presented visually and highlight critical concepts and the relationships between the concepts clearly (Labuschagne, Veerasamy, Burke, & Eloff, 2011).

Gamification usually works by simulating cybersecurity threats. The scenarios typically include topics such as identity theft, password management, dealing with worms and trojans, filters, patches, and working with links (Labuschagne, Veerasamy, Burke, & Eloff, 2011). In the training environment, anytime a risk is identified the application shows short eye catching tutorials that explain the problem shortly and concisely (Kirlappos & Sasse, 2012). Game-based learning is more effective than traditional and hands-on training as it combines elements from both types of training. The effect of hiding the knowledge and education in the appearance of a game, makes it easier for adoption (Tioh & Mina, 2015). Games generally have a short feedback cycle and so users would get their punishments or rewards quickly, reinforcing the concepts (Tioh & Mina, 2015). Gamification is a valuable and critical technique to incorporate into a cybersecurity awareness, training and education program.

4.4 Curriculum Considerations

These programs should teach concepts that come from a large range of cybersecurity threats and the curriculum should be tailored to the user's needs. Awareness topics are classified into novice/beginner, intermediate, and advanced levels and users are assessed on each level. Awareness training should not only be limited to computers but also mobile phones and any interaction users may have on the internet (Zeybek, Yilmaz, & Alper Dogru, 2019).

The following cybersecurity training topics should be included in programs:

- Password usage and management. Techniques to crack passwords are getting more sophisticated and teaching proper password management and frequent changing is essential (Nagarajan, Allbeck, Sood, & Janssen, 2012).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

- Protection from malware and spam. A complete program must cover topics of the use and maintenance of anti-virus and anti-malware tools (Nagarajan, Allbeck, Sood, & Janssen, 2012).
- Teaching patch management. Patch management is very critical as patches fix security faults in the software (Nagarajan, Allbeck, Sood, & Janssen, 2012).
- Social engineering prevention. Social engineering techniques are a big threat to security systems because they focus on the human element and do not get prevented by tools and software. Awareness can only be achieved through regular cybersecurity training programs that focus on social engineering techniques that are employed by cyber criminals (Nagarajan, Allbeck, Sood, & Janssen, 2012).
- Behavioural cybersecurity, game theory and risk management is also recommended to be in the curriculum (Patterson, Winston, & Fleming, 2016).
- Online self-efficacy is an individual's belief in their ability to detect and mitigate cybersecurity threats. Online self-efficacy can be increased by training and is thus a key topic (Teimouri, Benrazavi, Griffiths, & Hassan, 2018).

It is very important that users get continuous information about the topics on a regular basis, as continuous exposure creates true understanding of the cybersecurity topics (McCrohan, Engel, , & Harvey, 2010). Thus the curriculum needs to be updated regularly to reflect current threats.

4.5 Organizational and Demographic Aspects

To have a successful cybersecurity awareness programs, there are a few key attributes that must be incorporated into the program. These attributes are ensuring support from top management, trying to ensure a creative, fun, and interesting learning process, ensuring that content is relevant to the audience that is receiving it. Additionally clear content and explanations of the impact of the cybersecurity threats, and the use of different kinds of media, reward systems for users that are learning the most, continuously reminding the user of what they have learnt, and making sure there is a way to measure the effect and impact that the program (Sari & Prasetio, 2017).

The language, gender, lifestyles, and ethnicity of the users also need to be considered as this impacts the adoption of the cybersecurity initiatives (Alarifi, Tootell, & Hyland, 2012) (Aldawood & Skinner, 2019). The user's computer experience also has a significant impact on their cybersecurity awareness (Rocha Flores, Holm, Svensson, & Ericsson, 2014). Special consideration must be given to these factors, especially language, as users have higher cybersecurity awareness when they are assessed in their mother tongue (Kruger, Flowerday, Drevin, & Steyn, 2011). In the case of the adoption of security controls, particularly in mobile phones, one study note that female users tend to have a lower awareness and adoption than male users (Parker, Ophoff, Van Belle, & Karia, 2015).

5 Conclusion

Several success factors needed for cybersecurity awareness, training, and education were identified in this study. This systematic literature review search resulted in 3927 articles, and through a screening and data quality assessment, identified 58 articles that discuss key considerations for developing and implementing cybersecurity awareness, training, and education programs. The key considerations focused on the importance of assessing awareness of users, selection of pedagogical approaches, design

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

of the curriculum and supporting organizational and demographic aspects. The factors identified in this SLR may be used to inform the creation of cybersecurity awareness, training, and education programs towards building the skills and competencies of internet users. By incorporating the factors identified in this review, educators can create programs that benefit users of the cyberspace.

References

Ahmed, N., Islam, M., Kulsum, U., Islam, M., Haque, M., & Rahman, M. (2019). Demographic Factors of Cybersecurity Awareness in Bangladesh. . *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)* (pp. 685–690). Dhaka,: IEEE.

Akdemir, N., & Lawless, C. J. (2020). Exploring the human factor in cyber-enabled and cyber-dependent crime victimisation: a lifestyle routine activities approach. *Internet Research, 30*(6), 1665-1687.

Alarifi, A., Tootell, H., & Hyland, P. (2012). A study of information security awareness and practices in Saudi Arabia. *2012 International Conference on Communications and Information Technology* (pp. 6-12). Hammamet: IEEE.

Aldawood, H., & Skinner, G. (2018). Educating and Raising Awareness on Cyber Security Social Engineering: A Literature Review. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 62-68). Wollongong, Australia: IEEE.

Aldawood, H., & Skinner, G. (2019). Challenges of Implementing Training and Awareness Programs Targeting Cyber Security Social Engineering. *2019 Cybersecurity and Cyberforensics Conference (CCC)* (pp. 111-117). Melbourne: IEEE.

Alotaibi, F., Furnell, S., Stengel, I., & Papadaki, M. (2017). Enhancing cyber security awareness with mobile games. *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)* (pp. 129–134). Cambridge: IEEE.

Bele, J. L., Dimc, M., Rozman, D., & Jemec, A. S. (2014). Raising Awareness of Cybercrime--The Use of Education as a Means of Prevention and Protection. *10th International Conference Mobile Learning 2014* (pp. 281-284). Madrid: International Association for the Development of the Information Society.

Bishop , M., Burley , D., Buck , S., Ekstrom , J., Futcher , L., Gibson , D., . . . Parrish , A. (2017). Cybersecurity Curricular Guidelines. In L. F. M. Bishop (Ed.), *IFIP World Conference on Information Security Education. 503*, pp. 3-13. SPRINGER.

Broadhurst, R. (2006). Developments in the global law enforcement of cyber-crime. *Policing: An International Journal of Police Strategies & Management*.

Carlisle, D. (2010, April). *graphicx: Enhanced support for graphics.* Retrieved from http://www.ctan.org/tex-archive/ help/Catalogue/entries/graphicx.html

Chen, C. C., Dawn Medlin, B., & Shaw, R. S. (2008). A cross-cultural investigation of situational information security awareness programs. *Information Management & Computer Security, 16*(4), 360–376.

Frauenstein, E., & von Solms, R. (2014). Combatting phishing: A holistic human approach. *Information Security for South Africa 2014* (pp. 1-10). Johannesburg: IEEE.

Ghafir, I., Saleem, J., Hammoudeh, M., Faour, H., Prenosil, V., Jaf, S., . . . Baker, T. (2018). Security threats to critical infrastructure: The human factor. *The Journal of Supercomputing : An International Journal of High-Performance Computer Design, Analysis, and Use, 74*(10), 4986–5002.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Hagen, J., Albrechtsen, E., & Ole, J. (2011). The long-term effects of information security e-learning on organizational learning. *Information Management & Computer Security, 19*(3), 140–154.

Holdsworth, J., & Apeh, E. (2017). An Effective Immersive Cyber Security Awareness Learning Platform for Businesses in the Hospitality Sector. *25th International Requirements Engineering Conference Workshops (REW),* (pp. 111–117). IEEE.

Holdsworth, J., & Apeh, E. (2017). An effective immersive cyber security awareness learning platform for businesses in the hospitality sector. *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)* (pp. 111-117). IEEE.

Jensen, M. L., Dinger, M., Wright, R. T., & Thatcher, J. B. (2017). Training to Mitigate Phishing Attacks Using Mindfulness Techniques. *Journal of Management Information Systems*, 597-626.

Kirlappos, I., & Sasse, M. (2012). Security Education against Phishing: A Modest Proposal for a Major Rethink. *Security & Privacy, 10*(2).

Kruger, H., Flowerday, S., Drevin, L., & Steyn, T. (2011). An assessment of the role of cultural factors in information security awareness. *2011 Information Security for South Africa (ISSA)* (pp. 1-7). Johannesburg: IEEE.

Labuschagne, W., Veerasamy, N., Burke, I., & Eloff, M. (2011). Design of cyber security awareness game utilizing a social media framework. . *2011 Information Security for South Africa (ISSA)* (pp. 1-9). Johannesburg: IEEE.

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education, 9*(3).

McCrohan, K., Engel, , K., & Harvey, J. (2010). Influence of Awareness and Training on Cyber Security. *Journal of Internet Commerce, 9*(1), 23–41.

Monteith, S., Bauer, M., Alda, M., Geddes, J., Whybrow, P. C., & Glenn, T. (2021). Increasing Cybercrime Since the Pandemic: Concerns for Psychiatry. *Current Psychiatry Reports, 23*(18), 1-9.

Muhirwe, J., & White, N. (2016). Cybersecurity Awareness and Practice of Next Generation Corporate Technology Users. *Issues in Information Systems, 17*(2).

Muthuppalaniappan, M., & Stevenson, K. (2021). Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. *International Journal for Quality in Health Care, 33*(1), 1-4. doi:10.1093/intqhc/mzaa117

Nagarajan, A., Allbeck, J., Sood, A., & Janssen, T. (2012). Exploring game design for cybersecurity training. . *2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 256– 262). Bangkok: IEEE.

Naidoo, R. (2020). A multi-level influence model of COVID-19 themed cybercrime. *European Journal of Information Systems, 29*(3), 303-321.

Paoli, L., Visschers, J., & Verstraete, C. (2018). The impact of cybercrime on businesses: a novel conceptual framework and its application to Belgium. *Crime, Law and Social Change, 70*(4), 397-420.

Parker, F., Ophoff, J., Van Belle, J.-P., & Karia, R. (2015). Security awareness and adoption of security controls by smartphone users. *2015 Second International Conference on Information Security and Cyber Forensics (InfoSec)* (pp. 99–104). Cape Town: IEEE.

Patterson, W., Winston, C., & Fleming, L. (2016). Behavioral Cybersecurity: A needed aspect of the security curriculum. *SoutheastCon 2016* (pp. 1-7). Norfolk: IEEE.

Platsis, G. (2019). The human factor: Cyber security's greatest challenge. *Cyber Law, Privacy, and Security: Concepts, Methodologies, Tools, and Applications*, 1-19.

Rahim, N. H., Hamid, S., Mat Kiah, M. L., Shamshirband, S., & Furnell, S. (2015). A systematic review of approaches to assessing cybersecurity awareness. *Kybernetes, 44*(4), 606-622.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Rocha Flores, W., Holm, H., Svensson, G., & Ericsson, G. (2014). Using phishing experiments and scenario-based surveys to understand security behaviours in practice. *Information Management & Computer Security, 22*(4), 393–406.

Sari, P., & Prasetio, A. (2017). Knowledge sharing and electronic word of mouth to promote information security awareness in social network site. *International Workshop on Big Data and Information Security (IWBIS),* (pp. 113-117).

Shah, M. H., Jones, P., & Choudrie, J. (2019). Cybercrimes prevention: Promising organisational practices. *Information Technology & People, 32*(5), 1125–1129.

Smyth, S. J., Curran, K., & McKelvey, N. (2019). The Role of Education and Awareness in Tackling Insider Threats. In I. Vasileiou, & S. Furnell, *Advances in Information Security, Privacy, and Ethics (AISPE) Book Series* (pp. 33-51). IGI Global.

Statista. (2022, November 29). *Number of data records exposed worldwide* . Retrieved from https://www.statista.com/statistics/1307426/number-of-data-breaches-worldwide/

Takata, T., & Ogura, K. (2019). Confront Phishing Attacks—From a Perspective of Security Education. *2019 IEEE 10th International Conference on Awareness Science and Technology (ICAST)* (pp. 1-4). IEEE.

Teimouri, M., Benrazavi, S., Griffiths, M., & Hassan, M. (2018). A Model of Online Protection to Reduce Children's Online Risk Exposure: Empirical Evidence From Asia. *Sexuality & Culture : An Interdisciplinary Quarterly, 22*(4), 1205–1229.

Tioh, J., & Mina, M. (2015). Digital defenders: Computer security literacy via game- based learning. *2015 IEEE Frontiers in Education Conference (FIE)* (pp. 1-5). IEEE.

Tirumala, S. S., Valluri, M. R., & Babu, G. A. (2019). A survey on cybersecurity awareness concerns, practices and conceptual measures. *2019 International Conference on Computer Communication and Informatics.* Coimbatore: IEEE.

Tschakert , K., & Ngamsuriyaroj, S. (2019). Effectiveness of and user preferences for security awareness training methodologies. *Heliyon, 5*(6).

Zeybek, M., Yilmaz, E., & Alper Dogru, I. (2019). A Study on Security Awareness in Mobile Devices. . *2019 1st International Informatics and Software Engineering Conference (UBMYK)* (pp. 1-6). Ankara: IEEE.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Towards a Conceptual Framework for Investigating the Impact of Organisational Resources, Entrepreneurial Orientation Dimensions and Big Data Analytics on Business Performance of South African E-commerce SMMEs

Marungwane Leshego Mogale
University of Cape Town, Cape Town, South Africa
mglmar007@myuct.ac.za

Abstract

Data Analytics (BDA) is a crucial component of high-performing e-commerce businesses. In South Africa, where SMMEs are key contributors to economic growth, it is crucial to understand the resources and capabilities they should have in place to adopt BDA and positively impact their business performance. This paper aims to review current literature detailing organisational resources, BDA, and its impact on the business performance of South African e-commerce Small, Medium and Micro Enterprises (SMMEs) and develop a conceptual framework. While various studies show the impact of BDA on business performance in large organisations, and developed and developing economies, the impact it has on business performance, as well as that of organisational resources and entrepreneurial orientation dimensions on South African e-commerce SMMEs, is not well explored. This research adopted a deductive approach, using a systematic approach to literature, 411 journal and conference proceedings articles were retrieved from 2016 to date. 15 articles were selected after analyses and synthesis through a narrative literature approach. Key organisational resources that enable the use of BDA were identified from the literature and used to develop a conceptual framework that can be used for future studies using empirical data. This research is in progress and the preliminary findings were derived through a literature review, that highlights organisational resources such as IT infrastructure, IT human resources, financial resources, risk-taking, innovativeness and proactiveness as relevant to the usage of BDA and its influence on business performance.

Keywords- Big Data Analytics, SMMEs, E-commerce, Resource-based View, Dynamic Capabilities View, Entrepreneurial Orientation

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

1 Introduction

Big data analytics (BDA) is a growing area of interest, and businesses are becoming aware of its value to productivity, profitability, and competitive advantage. BDA is a process needed to understand large volumes of data and extract meaningful information and knowledge to gain actionable insights to establish a competitive advantage (Ferraris et al., 2019). Studies show the predictive value from processing large volumes of data has a strategic potential to transform business processes and create capabilities for tackling key business issues (Batistič & van der Laken, 2019). BDA plays a significant role in business innovation through the creation and introduction of new products and services. It enables data-driven decision-making and creates innovative ways of managing processes and activities that lead to better business performance (Ferraris et al., 2019; Seseni & Mbohwa, 2021). Small, Medium and Micro Enterprises (SMMEs) are among businesses that should take advantage of BDA, though faced with many challenges including the lack of financial resources, IT infrastructure and technical skills. The reduction in memory, storage, and bandwidth prices have made it possible for more businesses to enter the BDA market more economically, which also offers opportunities for SMMEs (Ferraris et al., 2019).

Research has found that the demonstration of risk-taking, innovativeness and proactiveness by SMMEs has positively impacted their business performance (Lomberg et al., 2017; Dubey et al., 2020). Akter & Wamba (2016) drawing from the Resource-based View (RBV) argues that BDA is a key element of high-performing e-commerce businesses. Shan et al., (2019) drawing from Dynamic Capacities View (DCV) found in their study that IT technology capabilities such as IT infrastructure and IT human resources influence competitive advantage directly thus increasing business performance. Idle resources such as financial resources aid businesses in adapting to changing consumer demands and market conditions (Shan et al., 2019). This is particularly crucial in South Africa, where financial resources may have an impact on an SMME's success since it is one of their major challenges (Seseni & Mbohwa, 2021).

This paper posits that organisational resources such as IT infrastructure and IT human resources, risk-taking, innovativeness and proactiveness influence the adoption of BDA which impacts business performance. While various studies show the impact of BDA on business performance in large organisations, and developed and developing economies, to our knowledge, the impact it has on business performance, as well as that of organisational resources and entrepreneurial dimensions on South African e-commerce SMMEs, is not well explored. The objective of this paper is to explore current literature to gain an empirical understanding of the problem that will guide the development of a conceptual framework.

2 Method

To identify existing literature, this paper adopted a systematic review which aims to locate, assess, and synthesise data from related research using a replicable, scientific, and transparent process methodology (Hussain et al., 2020). The primary search was conducted on credible databases with high-ranking journals and conference proceedings, these include Google Scholar, EbscoHost, IEEE Xplore, Science Direct, Emerald and Scopus. The search terms were defined based on the scope of this paper and the defined objectives. The search was carried out using the following search string multiple search strings in Table 1 and was limited to the title only. Multiple search strings were used to obtain base knowledge focusing on BDA and business performance and to contextualise this review to e-commerce and South African SMMEs. The combination of all key search terms does not return any results as the

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

objective of this study is a gap in the literature. The selection was limited to literature published in 2016 onwards, with earlier sources being reviewed to extract vital background information and only journal and conference proceedings papers were included. The secondary search was carried out using citation chaining to find other related articles to those found in the primary search and fill the gaps in the search strategy; both backward and forward searching were used. After removing all duplicate articles and applying all exclusions, a total of 15 papers were selected for analysis and synthesis, and this sample size was considered big enough to provide an overview of BDA, organisational resources, business performance and e-commerce of South African SMMEs. A narrative approach to literature review was used for analysis and synthesis, which is conducted by summarising earlier studies with an emphasis on theories, frameworks, fundamental variables, and their research findings considering hypothesised correlation (Rahman, 2018).

Search String		Results
1.	("big data" OR "big data analytics" OR "data analytics") AND ("business performance" OR "competitive advantage" OR "firm performance")	81
2.	("big data" OR "big data analytics" OR "data analytics") AND "e-commerce"	156
3.	("big data" OR "big data analytics" OR "data analytics") AND ("SME" OR "SMME" OR "startup" OR "entrepreneu*")	76
4.	("SME" OR "SMME" OR "startup" OR "entrepreneu*") AND "South Africa"	91
5.	"e-commerce" AND "South Africa"	7
	Total	411

Table 1: Search String

3 Literature Review

SMMEs are one of the main boosters of economic growth and development, as they reduce the unemployment rate by employing a significant portion of the workforce (Ncube & Zondo, 2022). In South Africa, 98.5% of the total businesses are SMMEs, of which 25.8% of the workforce is employed. These businesses contribute 39% of the gross domestic product (GDP) of the country (Seseni & Mbohwa, 2021). The Coronavirus outbreak changed global trends such as the increase in e-commerce sales in developed and developing countries including South Africa, this demand gave rise and opportunities for SMMEs to enter the e-commerce market (Bhatti et al.,2020).

E-commerce in South Africa generated $6.78 billion in revenue in 2020, $7.56 billion in revenue in 2021 and is estimated to generate $8.74 billion by the end of 2022 (Statista, 2021). E-commerce platforms generate large volumes of data, and the ability to extract meaningful information and knowledge to gain actionable insights has been a key driver in enabling competitive advantage and an environment where decisions are made based on insights rather than human intuition (Batistič & van der Laken, 2019; Ferraris et al., 2019). BDA offers e-commerce SMMEs an opportunity to understand their customer needs through behavioural analysis, increase conversions through price optimisation and increase the return on investment (Akter & Wamba, 2016; Seseni & Mbohwa, 2021). BDA adds

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

business value as it gives managers the capability to make data-driven decisions based on evidence rather than intuition and reduces human judgement error due to its precision (Batistič & van der Laken, 2019; Ferraris et al., 2019).

Mikalef et al., (2018) discuss the various characteristics of big data found in literature, including the 3Vs (volume, variety, and velocity), 5Vs (volume, velocity, variety, veracity, value) and 7Vs (volume, velocity, variety, veracity, value variability, visualisation). (1) Volume refers to the quantity of data that increases exponentially daily, increasing the need for IT infrastructure which can store and process this data; (2) Velocity refers to the speed at which data is created, collected and processed in real-time which enables a faster decision-making process and allows businesses to be agile; (3) Variety refers to the types of data generated from different digital platforms which could either be structured or unstructured; (4) Veracity represents the reliability and quality of the collected data, which should contain less noise, be complete and dated; (5) Value refers to the strategic and informational benefit of big data that allows businesses to make decisions and create a competitive advantage (Akter & Wamba, 2016; Ferraris et al., 2019). (6) Variability describes the dynamic potential made possible by analysing big data, and (7) Visualisation refers to the representation of insights found in the data (Mikalef et al., 2018). This paper adopts the 7 characteristics, as big data in the e-commerce market presents dynamic opportunities such as the ability to track user behaviour and find ways to convert and retain customers, while the visualisation of insights can assist businesses to optimise for sales, improve decision making and user experience (Akter & Wamba, 2016).

RBV theory defines organisational resources and capabilities as foundations for competitive advantage and to ensure long-term sustainability. RBV focuses on inimitable, valuable, rare, incomparable internal resources. These resources can be tangible assets such as technologies and intangible assets such as employee and management skills which are used to increase performance and gain a competitive advantage. The effective usage of these resources is where the organisational capabilities reside. (Dubey et al., 2020; Horng, et al., 2022). Some researchers criticise RBV for not considering the external environment related to the business and failing to explain the lack of business performance during volatile times. E-commerce SMMEs existing in a highly dynamic environment where technology is constantly evolving therefore being required to operate in a dynamic manner to stay ahead of the competition (Almazmomi et al., 2022).

Due to the RBV criticism, DCV was birthed as an extension. It refers to the business' "ability to integrate, build, and reconfigure internal and external resources/competencies to address, and possibly shape, rapidly changing business environments" (Dubey et al, 2020, p.3). Resources and capabilities are seen as the foundation of DCV with resources including technology, knowledge and human resource and capabilities indicating productivity and performance Dubey et al, (2020). Almazmomi et al., (2022) explain that DCV is better at explaining high-tech SMMEs business than RBV given the dynamic nature of their businesses with the high rate of technological innovation.

Entrepreneurial Orientation (EO) refers to the organisation's processes, practices and decision-making capabilities that enable the organisation to explore new market opportunities and create a competitive advantage (Lomberg et al., 2017; Dubey et al., 2020). EO is a demonstration of risk-taking, innovativeness and proactiveness capabilities. With contextualisation through research, the three dimensions: risk-taking, innovativeness and proactiveness are set to be independent and have their own effect on business performance, however, some studies have contextualised the three dimensions as dependent and all contributing to business performance. Studies employing the three dimensions as dependent may hide or inaccurately attribute effects from variation in one dimension of EO while the dependent dimensional view may hide the effects of covariation between two or all dimensions (Lomberg et al., 2017). This paper considers the three dimensions as independent and each contributing towards business performance.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Shan et al., (2019) highlight in their study the importance of IT technology capabilities such as IT infrastructure, IT human resources and financial resources through the adaptation of DCV and RBV theories, and their implications for managers that leverage BDA to achieve competitive advantages in business. Financial resources are also highlighted as the key resources that enable the adoption of BDA. Batistič & van der Laken (2019) also reiterate that BDA can only add value if the right IT infrastructure and IT human resources (skills) are in place. Lomberg et al., (2017) highlight the EO dimensions of risk-taking, innovativeness and proactiveness as influences on business performance. Dubey et al., (2020) also discuss the EO dimensions and their impact on the adoption of BDA and operational performance.

3.1 IT Infrastructure Resources

IT infrastructure refers to the infrastructural component which enables the business to perform its daily activities digitally (Li & Chan, 2019). The correct infrastructure is important in developing any IT capabilities. This infrastructure must be able to integrate internal and external data sources, store data, and allow the processing and visualisation of insights (Shan et al., (2019). Integration with external partners gives opportunities to mutually benefit from the insights found in the data. Due to the everyday change in technology and data volumes, IT infrastructure needs to be flexible to accommodate the dynamic environment to meet the business objectives and technology requirements (Li & Chan, 2019). For e-commerce businesses, the correct IT infrastructure is a key component of running the business.

3.2 IT Human Resources

IT human resources emphasises the need for employees to gain the necessary knowledge and skills to address problems associated with BDA and to fully utilise its capabilities to achieve high business performance. IT knowledge resources are considered the key resources of any organisation and they are unique to every business (Shan et al., 2019). The IT skills needed include managerial and technical skills relating to the collection, analysis, and presentation of BDA (Batistič & der Laken, 2019). Literature shows that IT skills are difficult to find, compensate and retain especially for SMMEs due to the high demand for the skills in the market, this is no different in the South African context. The available human resource may possess the theoretical knowledge of IT but lack practical experience, this is a challenge for SMMEs while searching for talent given that success is dependent on the correct technical and business knowledge (Behl, 2020). SMMEs often outsource external parties and vendors if they do not have sufficient technical skills to adopt an innovation such as BDA (Maroufkhani et al., 2020).

3.3 Financial Resources

In South Africa where an SME's success can be influenced by financial resources, it is important to also consider this resource and its influence on business performance. To stay ahead of the competition and remain sustainable, enough financial investment is needed to aid the process of innovation and continuous adaptation to change in the market. Financial resources are also key in ensuring that the correct talent is found and retained. Time needs to be invested in areas of innovation, especially in BDA, and continuous learning (Shan et al., 2019).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.4 Business Performance

A business is always judged by its profitability by the market as an indicator of growth. Studies highlight the importance of productivity in operations as a key indicator of business performance. SMMEs need to understand their market share as they do not only compete with other SMMEs but also compete with large enterprises. The ability for businesses to invest in valuable resources and capabilities gives them a market differentiator and a competitive advantage. E-commerce businesses face higher challenges that affect business performance such as the volatility of customer needs and existing competitors in the market. While e-commerce SMMEs need to be innovative, they also need to have a competitive spirit to drive business performance (Behl, 2020). Having the right IT infrastructure, IT human resources, financial resources and BDA capabilities has been proven to positively impact business performance (Akter & Wamba,2016; Ferraris et al., 2019).

4 Conceptual Framework

A review of previous studies has revealed a research gap in the role that organisational resources play in the adoption of BDA and its impact on business performance for South African e-commerce SMMEs. The adaptation of RBV, DCV and EO theories from previous studies has revealed the importance of IT infrastructure, IT human resources, financial resources, risk-taking, innovativeness and proactiveness in the adoption of BDA. Through the exploration of literature and theoretical frameworks, the below conceptual model is proposed as the lens through which future studies can test with empirical data. Its constructs are defined in Table 2.

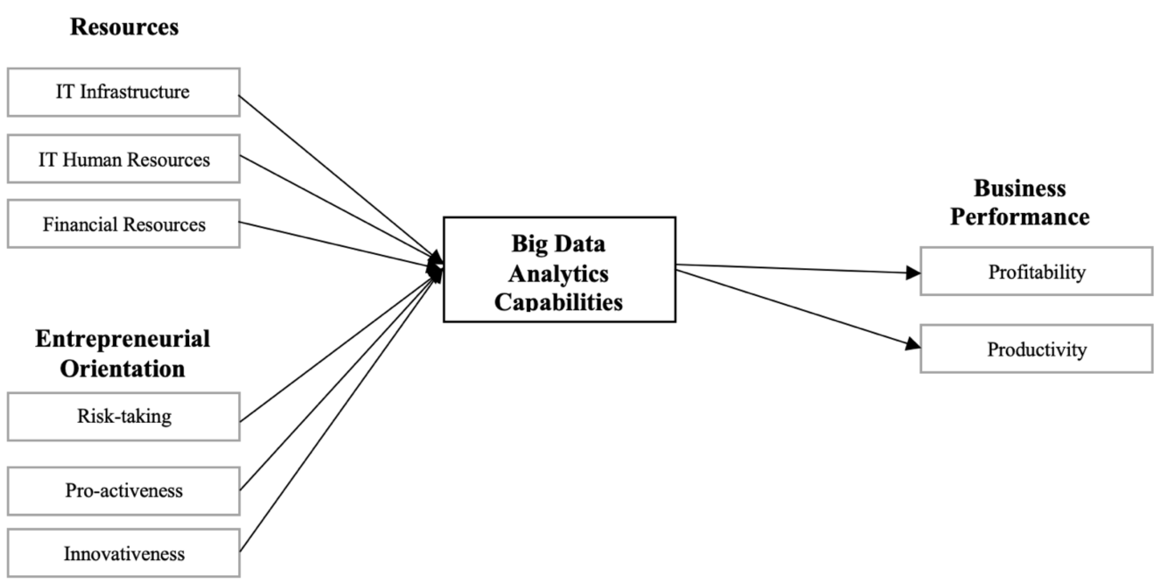


Figure 1: Conceptual Framework

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Construct	Definition	Source
IT Infrastructure	Infrastructural components enable the business to perform daily activities digitally and enable IT capabilities. Infrastructure must be able to integrate internal and external data sources, store data, and allow the processing and visualisation of insights while being flexible to the increasing data volume.	Shan et al.,(2019) Li & Chan (2019)
IT Human Resources	IT skills and knowledge needed include managerial and technical skills relating to the collection, analysis and presentation of Big Data Analytics. These are needed to address problems associated with Big Data Analytics and to fully utilise its capabilities to achieve high business performance.	Shan et al.,(2019) Batistič & der Laken (2019) Behl (2020)
IT Infrastructure	Infrastructural components enable the business to perform daily activities digitally and enable IT capabilities. Infrastructure must be able to integrate internal and external data sources, store data, and allow the processing and visualisation of insights while being flexible to the increasing data volume.	Shan et al.,(2019) Li & Chan (2019)
IT Human Resources	IT skills and knowledge needed include managerial and technical skills relating to the collection, analysis and presentation of Big Data Analytics. These are needed to address problems associated with Big Data Analytics and to fully utilise its capabilities to achieve high business performance.	Shan et al.,(2019) Batistič & der Laken (2019) Behl (2020)
Financial Resources	Financial resources are needed to aid the process of innovation and continuous adaptation to change in the market to stay ahead of the competition and remain sustainable.	Shan et al.,(2019)
Entrepreneurial Orientation (EO): risk-taking, innovativeness and proactiveness	Entrepreneurial orientation (EO) defines the ability to take risks, be innovative and proactive. These dimensions affect the organisations' processes, practices and decision-making capabilities that enable competitive advantage. The three dimensions (risk-taking, innovativeness and proactiveness) are independent and contribute to business performance.	Lomberg et al., (2017) Dubey et al., (2020)
Big Data Analytics Capabilities	Enable the creation of market insights by obtaining internal and external knowledge. These can influence strategy, processes, productivity and profitability. This capability is dependent on tangible and intangible resources.	Akter & Wamba (2016) Mikalef et al., (2017) Batistič & der Laken (2019)
Business Performance	The degree to which a company is successful in accomplishing its goals and objectives. This comprises non-financial performance indicators like productivity in operations as well as financial performance indicators like profitability, revenue growth, and return on investment. These aim at increasing shareholder value.	Behl (2020) Maroufkhani et al., (2020)

Table 2: Definitions of Constructs

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

4.1 Research Propositions

Drawing from RBV, DCV and existing literature, IT infrastructure, IT human resources and financial resources have been identified as key enablers for BDA capabilities as found through the analysis of the selected papers. The researcher, therefore, proposes that:

- Proposition 1a: IT infrastructure has a positive impact on the usage of BDA capabilities
- Proposition 1b: IT Human Resources have a positive impact on the usage of BDA capabilities
- Proposition 1c: Financial Resources have a positive impact on the usage of BDA capabilities

Drawing from EO and its dimensions as identified through the analysis of the selected papers, the researcher proposes that:

- Proposition 2a: EO's risk-taking has a positive impact on the usage of BDA capabilities
- Proposition 2b: EO's proactiveness has a positive impact on the usage of BDA capabilities
- Proposition 2c: EO's innovativeness has a positive impact on the usage of BDA capabilities

Akter & Wamba (2016) highlights that BDA plays a role in improving business performance while Ferraris et al., (2019) highlight the importance of organisational resources such as IT human resources, which positively impacts business performance, Lomberg et al., (2017) also highlights the impact of EO dimension on business performance. Therefore, the researcher proposes that:

- Proposition 3a: Organisational Resources (IT infrastructure, IT human resources and financial resources) under the mediative effect of BDA capabilities positively impact business performance(productivity, profitability)
- Proposition 3b: Entrepreneurial Orientation dimensions (Risk-taking, Pro-activeness, Innovativeness) under the mediative effect of BDA capabilities positively impact business performance(productivity, profitability)

Drawing from existing literature from various researchers including Akter & Wamba (2016), Batistič & van der Laken (2019), Ferraris et al., (2019), Behl (2020), it has been proven that BDA capabilities impact the business performance in the form of profitability and productivity. It is for this reason that the researcher proposes:

- Proposition 4a: BDA capabilities have a positive impact on Productivity
- Proposition 4b: BDA capabilities have a positive impact on Profitability

5 Conclusion and Recommendations

This paper intended to explore current literature to gain an understanding of the role that organisational resources and entrepreneurial dimensions play in the adoption of BDA and its impact on business performance for South African e-commerce SMMEs and develop a conceptual framework. These constructs emerged through the adoption of RBV, DCV and EO theories creating a new lens through which they can be viewed and form the basis of the conceptual framework which can be tested with empirical data in future. This review highlighted BDA as a key element of high-performing e-commerce businesses as it gives them an opportunity to understand their customer needs and the ability to make data-driven decisions. It is also revealed that the right IT infrastructure and IT human resources are key to enabling the adoption of BDA, with the support of financial resources. The influence of risk-taking, innovativeness and proactiveness is also vital as it has a direct impact on business performance.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Further research is required to test the conceptual framework with empirical data that is expected to reveal the key organisational resources that influence the usage of BDA capabilities and the influence on business performance in South African SMMEs.

References

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, *26*(2), 173–194. https://doi.org/10.1007/s12525-016-0219-0

Almazmomi, N., Ilmudeen, A., & Qaffas, A. A. (2022). The impact of business analytics capability on data-driven culture and exploration: achieving a competitive advantage. *Benchmarking: An International Journal*, *29*(4), 1264–1283. https://doi.org/10.1108/bij-01-2021-0021

Batistič, S., & der Laken, P. (2019). History, Evolution and Future of Big Data and Analytics: A Bibliometric Analysis of Its Relationship to Performance in Organizations. *British Journal of Management*, *30*(2), 229–251. https://doi.org/10.1111/1467-8551.12340

Behl, A. (2020). Antecedents to firm performance and competitiveness using the lens of big data analytics: a cross-cultural study. *Management Decision*, *60*(2), 368–398. https://doi.org/10.1108/md-01-2020-0121

Bhatti, A., Akram, H., Basit, M., Khan, A., Mahwish, S., Naqvi, R., & Bilal, M. (2020). E-commerce trends during COVID-19 Pandemic. *International Journal of Future Generation Communication and Networking*. 13(2), 1449-1452. https://www.researchgate.net/publication/342736799

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., & Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, *226*, 107599. https://doi.org/10.1016/j.ijpe.2019.107599

Ferraris, A., Mazzoleni, A., Devalle, A., & Couturier, J. (2019). Big data analytics capabilities and knowledge management: impact on firm performance. *Management Decision*, *57*(8), 1923–1936. https://doi.org/10.1108/md-07-2018-0825

Horng, J. S., Liu, C. H., Chou, S. F., Yu, T. Y., & Hu, D. C. (2022). Role of big data capabilities in enhancing competitive advantage and performance in the hospitality sector: Knowledge-based dynamic capabilities view. *Journal of Hospitality and Tourism Management*, *51*, 22–38. https://doi.org/10.1016/j.jhtm.2022.02.026

Hussain, A. B., Jasim, O. M., & Mahdi, A. A. (2020). Usability Evaluation of Mobile Tracking Applications: A Systematic Review. *International Journal of Interactive Mobile Technologies (iJIM)*, *14*(05), 119 https://doi.org/10.3991/ijim.v14i05.13353

Li, T. C., & Chan, Y. E. (2019). Dynamic information technology capability: Concept definition and framework development. *The Journal of Strategic Information Systems*, *28*(4), 101575. https://doi.org/10.1016/j.jsis.2019.101575

Lomberg, C., Urbig, D., Stöckmann, C., Marino, L. D., & Dickson, P. H. (2017). Entrepreneurial Orientation: The Dimensions' Shared Effects in Explaining Firm Performance. *Entrepreneurship Theory and Practice*, *41*(6), 973–998. https://doi.org/10.1111/etap.12237

Maroufkhani, P., Tseng, M. L., Iranmanesh, M., Ismail, W. K. W., & Khalid, H. (2020). Big data analytics adoption: Determinants and performances among small to medium-sized enterprises. *International Journal of Information Management*, *54*, 102190. https://doi.org/10.1016/j.ijinfomgt.2020.102190

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and E-Business Management*, *16*(3), 547–578. https://doi.org/10.1007/s10257-017-0362-y

Ncube, T.R., & Zondo, D. (2022). Entrepreneurial Attributes responsible for Small and Medium Enterprise Growth in South Africa: Small and Medium Enterprise Owners' perspectives. *International Journal of Special Education* 37(3), 2022-8223. https://www.researchgate.net/publication/359314373

Rahman, K. (2018). A Narrative Literature Review and E-Commerce Website Research. *ICST Transactions on Scalable Information Systems*, *5*(17), 154806. https://doi.org/10.4108/eai.29-5-2018.154806

Seseni, L. & Mbohwa, C.(2021). The Significance of Big Data in the Success of SMEs in Emerging Markets: A Case of South Africa. *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management, Singapore*, 1986-1996. http://www.ieomsociety.org/singapore2021/papers/376.pdf

Shan, S., Luo, Y., Zhou, Y., & Wei, Y. (2018). Big data analysis adaptation and enterprises' competitive advantages: the perspective of dynamic capability and resource-based theories. *Technology Analysis & Strategic Management*, *31*(4), 406–420. https://doi.org/10.1080/09537325.2018.1516866

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Understanding Digital Economy Dilemmas

Nixon Muganda Ochara, PhD.
Wits Business School
University of Witwatersrand, Johannesburg, South Africa
Nixon.Ochara@wbs.ac.za

Abstract

Against the backdrop of escalating contradictions and critiques of the digital economy's trajectory, this study analyzes how the emerging digitalization issues might be philosophically understood from a systems viewpoint. Five systemic digitalization challenges including the circular economy (CE), cyberphysical systems (CPS), sharing economy (SE), digital transformation (DT), and smart systems were identified (SS). To investigate digitalization challenges, the machine, organism, cultural/political, societal/environmental, and interrelationship systems metaphors were used. The machine viewpoint demonstrates that the circular economy challenge may be examined utilizing Hard Systems Thinking (HST) methodologies, with a focus on sufficiency via product design and business model innovation. The organism approach demonstrated how the digital twin notion may be investigated using Socio-Technical Systems (STS) and the Viable System Model (VSM) to diagnose and forecast CPS viability in an increasingly linked Industry 4.5/5.0 environment. In analyzing SE's rentier capitalism, the cultural/political viewpoint demonstrated the applicability of purposeful systems techniques for "people complexity." The societal/environmental viewpoint stressed emancipatory systems approaches to "coercive complexity" as crucial to evaluating the perpetuation of digital exclusion by DT from an emancipatory systems perspective. The interrelationship viewpoint emphasized the significance of systems approaches for researching "structural complexity" in intelligent systems. These viewpoints aid decision-makers in identifying problem-solving strategies based on systems thinking.

1 Introduction

The digital economy is a key metaphor for understanding the Fourth Industrial Revolution (4IR) or Industry 4.0/5.0. Digital economy continues to impact global growth and reconfiguration of various sectors. According to Medynska et al., (2022) the digital economy contributes 15.5% to global GDP, but its effects vary widely across countries, industries, and sectors. At the macro-level, the influence of digitalization is visible in innovation ecosystems, and competitive and relational dynamics that have transformed how commerce (e.g., ecommerce/ e-business), public administration (e.g., digital

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

government), and the current framing of the digital economy based on extensive global and national information infrastructures. Recent thinking and technological advances have 'smartified' markets, value chains, products, services, and material and non-material sectors (Appio et al., 2021). At the mesolevel, digital disruptions have influenced how organizations and agencies have restructured their capabilities, processes, and routines (Appio et al., 2021). How meso-level organizing units (regulatory bodies, stakeholder groups, R&D agencies, etc.) structure their dynamic capabilities in response to disruptive innovations is critical for framing digital transformation in these agencies and at national levels. The micro-level focuses on the individual and how digital transformation has changed life, work, and teams (Verhoef et al., 2021).

The prevalence of discourses on the digital economy suggests a change in socioeconomic values. These value shifts are correlated with the emergence of new digital innovations that now define a utopian vision: the digital world as a key resource for governments, businesses, and individuals where economic, political, and social value can be created equitably for the benefit of society. While these changes were largely inevitable, a lack of understanding of the digital economy's complex and systemic dilemmas and how to address them exacerbated social exclusion. Digital economy artifacts have fuelled digital exclusion, new forms of the digital divide, loss of trust, and discrimination (Harvey et al., 2021; Marshall et al., 2020; Ranchordás, 2020; Vartanova & Gladkova, 2019). The extensive list of failed digital transformation initiatives demonstrates how unprepared decision makers are for 'more recent' digital economy challenges (see Bucy et al., 2016). "Traditional" analysis may be constrained by "global" challenges, and systems thinking may be preferred for understanding digital economy issues. This paper employs systems thinking to understand the complexities of the digital economy better. The paper argues that traditional "reductionist" or "analytical" thought is inadequate for addressing problems in the digital economy.

To advance the argument for embracing systems thinking, the paper is structured as follows: first, synthesis, the systems-thinking mode, is differentiated from analysis, the dominant reductionist mode. After clarifying systems thinking, "complexity" is introduced as a key to understanding digital economy dilemmas. With the introduction of "Systems Thinking," the focus shifts to preparing contemporary decision makers to understand the current global systemic dilemmas and why "Systems Thinking" is "ripe" for the digital economy humanity now finds herself in. The paper concludes with reflections and recommendations for systems thinking.

If you already prepared your paper using a different style, you can find it very hard to conform to the EasyChair style. In this case, read Section 7 of this document.

2 Global Complexity: From Reductionist to Systems Thinking

Reductionism or machine age thinking holds that all things and events, their attributes, and our experience and understanding of them are made up of ultimate constituents, indivisible pieces (Ackoff, 1997). Yi et al., (2022) argues that traditional reductionism is no longer applicable for understanding socially complex innovation systems of the contemporary world. Thus, systems thinking is increasingly recognized to be the appropriate lens for understanding complex systems with emergent properties. While the reductionist worldview is still prominent, Ackoff (2018) argues its limitations have always been acknowledged since the 1930s publication of von Bertalanffy's "General Systems Theory" (Bertalanffy, 1937). Thus, by rejecting reductionism as the dominant way of thought, synthetic or systems thinking emerged as a new worldview focused on systems, their expanding complexity, and the difficulties of controlling them successfully.

Systems are regarded as wholes which lose their essential properties when taken apart, implying that they are wholes that cannot be understood through analysis (Ackoff, 1979). Analysis of a system only reveals its *structure and how it works*; it yields know-how or knowledge, but not *understanding*, which

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

explain why a system works the way it does (Ackoff, 1979). Therefore, in systems thinking, a phenomenon to be explained is viewed as part of a larger whole, a system, and is explained in terms of its role in that system (Daellenbach et al., 2012). Table 1 synthesis ideas of reductionism (*analysis*) and systemic thinking (*synthesis*) and differentiates these two world views in terms of their processes, outcomes of the processes, and the predominant source of explanation when describing the phenomena of interests.

Dimensions	*Analysis (Reductionism)*	*Synthesis (Systems Thinking)*
Process	1. Take system apart	1. Place the system as part of a larger system.
	2. Seek understanding of each part	2. Understand the containing system
	3. Aggregate the parts to the whole	3. Disaggregate the understanding of the whole to the parts
Source of Explanation	Internal - Closed System	External – Open System
Product	Knowledge of Structure	Understanding of Role and Function

Table 1: Reductionism versus Synthesis

The contrasts captured above reveals different modes of thought that are complementary to seeking an explanation to any phenomenon. Given the current age has been christened the "Systems Age" (Ackoff, 1973; Mitroff et al., 1974; Tarasenko, 2020; van Gigch, 1984; Zandi, 2000), the use of systems thinking tools for understanding current digital transformation dilemmas is not only apt but also fits well with the current age.

3 Complex Systemic Digital Economy Dilemmas

Table 2 synthesizes the critical global megatrends that are shaping the global and local discourses surrounding global dilemmas linked to the digital economy. These global megatrends, in the manner of Friedenthal et al., (2021), are being influenced by the global environment, currently shaped by human and societal needs; technologies that underly the evolution of societal systems; and stakeholder expectations, aligned to broader societal and technological trends. The systemic issues coming from humanity's struggle to reorient decision making in the face of these systemic challenges are predicated on digital transformation as a prerequisite for any system-level remedies envisioned by decision makers.

Global Megatrends	*Emergent Systemic Dilemmas*	*References*
Sustainability	Circular Economy	Geissdoerfer et al., 2017;Belmonte-Ureña et al., 2021; Nayal et al., 2022; Ogunmakinde et al., 2022
Industry 4.0/Society 5.0	Cyber-Physical Systems	Dantas et al., 2021; Hughes et al., 2022; Pourmehdi et al., 2022; Suleiman et al., 2022
Interconnected World	Sharing Economy; Social Exclusion	Ganapati & Reddick, 2018a; Reich, 2015; Srnicek, 2017
Digital Transformation	Sustaining versus Disruptive Innovations; Exploration versus Exploitation	Bodrožić & S. Adler, 2022; Clarke, 2019; Dickel & Schrape, 2017; Dunbar-Hester, 2019; Hensmans, 2021

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Smart Systems	Synthetic Thinking, Emergence, Self-Organization, Context-Awareness	Hermann et al., 2016; Pascual et al., 2019; Romero et al., 2020; Valckenaers et al., 2003
System Complexity Growth	Distributed Modularization	Ethiraj & Levinthal, 2004; Gärtner & Schön, 2016; Micheli et al., 2019

Table 2: Systemic Digital Economy Dilemmas

Sustainability is humanity's attempt to right past wrongs to save the planet. Humanity faces environmental, economic, and social challenges under the megatrend of "sustainability" (Zander & Mosterman, 2013). Following the failure of the Millennium Development Goals (see McCloskey, 2015), a new international development agenda, the UN 2030 Agenda for Sustainable Development, was adopted in September 2015 (Johnston, 2016). The challenges humanity faces in achieving the 17 SDGs are less clear. The Circular Economy (CE) and Industry 4.0 are two emerging discourses that have the potential to contribute to the 17 SDGs. CE advocates for an economic system that separates environmental pressure from economic growth by switching from linear to circular production (Belmonte-Ureña et al., 2021; Sanguino et al., 2020). The current system is based on large-scale extraction, use, and disposal of materials, which is the root of intergenerational and intergovernmental issues such as waste disposal in natural areas, resource scarcity, and climate change (de Souza Junior et al., 2020). The CE paradigm is a "regenerative system that minimizes resource input and waste, emission, and energy leakage by slowing, closing, and narrowing material and energy loops." Maintenance, repair, reuse, remanufacturing, refurbishing, and recycling can all help (Geissdoerfer et al., 2017, p.759). Thus, CE has emerged as a strategic response to the challenges of a linear economy's sustainability, promoting a circular rationale centered on a "zero waste economy" (van Langen et al., 2021). But CE's theoretical, practical, and ideological foundations are questioned by critics. On the one hand, CE supporters envision sustainable futures based on planned circularity, circular modernism, bottom-up sufficiency, and peer-to-peer circularity; on the other hand, critics claim it is a myth that "paints" a "utopian" future (Bauwens et al., 2020; Corvellec et al., 2022).

The evolution of ***Industry 4.0/5.0*** is founded on the long-held belief that technological development and sustainable development are inextricably linked; despite criticism, the conceptualization of this link is neoliberal and Eurocentric in nature (Schelenz & Pawelec, 2022). In spite of this criticism, it is deemed essential to implement innovative technologies as a key factor in establishing a proactive, self-configured, and automated system for achieving sustainability objectives. Industry 4.0 refers to the evolution of industries toward cyber-physical systems, and Society 5.0 refers to the evolution of society toward socio-cyber-physical systems (Friedenthal et al., 2021). Industry 4.0 seeks to optimize production in productive sectors with its integrative and interconnected technologies (Berawi, 2019; Dantas et al., 2021; Ramakrishna et al., 2020). However, it remains unclear how to achieve *integration* and *interconnection* of Industry 4.0's core technologies (Hughes et al., 2022; Pourmehdi et al., 2022; Suleiman et al., 2022). Consequently, although there is substantial evidence linking Industry 4.0 and sustainable development, the identified impediments create roadblocks to the agenda's realization, posing difficult choices.

The global community is increasingly being facilitated by Industry 4.0 technologies, resulting in higher levels of socio-political and economic interdependence (Friedenthal et al., 2021). The increased global interdependence subsequently incites the need for *sharing of resources* and the ***interconnection of systems (Interconnectivity)*** in global partnerships (Friedenthal et al., 2021). The interconnected world has seen the rise of new economic business models such as the "sharing economy", whose focus is on the sharing of unused or underutilized capacity. While there is evidence that the ***sharing economy*** exists, critics view it as a questionable phenomenon that threatens existing markets and institutions (Codagnone & Martens, 2016). In this "*share-the-scraps-economy*" (Srnicek, 2017, p. 254), what is emerging is a new "precariat" labor class that is flexible but without work security (Standing, 2018).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The concept of ***"Digital Transformation"*** (DT) has emerged as a key metaphor for describing Industry 4.0/5.0. According to Bodrožić & S. Adler (2022), four competing and conflicting future trajectories of DT development are possible: digital authoritarianism, digital oligarchy, digital localism, and digital democracy (Figure 1). Because of the contradictory effects and benefits, this suggests a systemic dilemma. The desired scenario is a digital democracy, which combines a proactive system-building regime with a management paradigm based on community and collaboration (Bodroi & S. Adler, 2022). However, most DT implementations are based on digital authoritarianism, digital localism, and digital oligarchies, which exacerbate social exclusion. The DT dilemmas envisaged are couched in the Schumpeterian notion of' '***creative destruction***,' which is linked to the instability and emergence of new forms of economic cycles embedded in these contradictory DT trajectories.

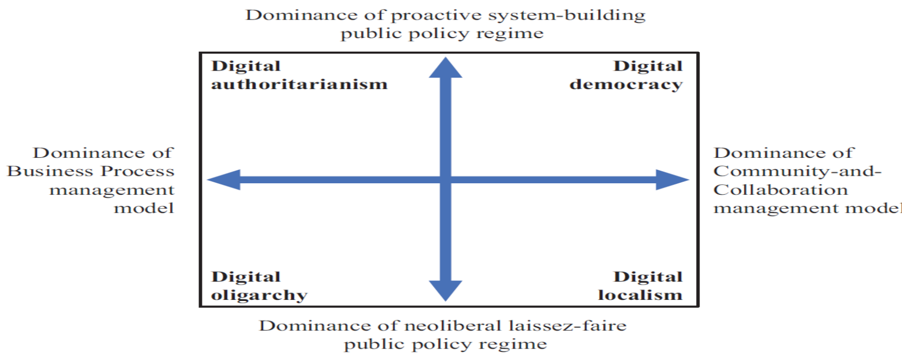


Figure 1: Four Scenarios for Digital Transformation (Source: Bodrožić & S. Adler, 2022)

Smartness entails the intentional development of devices, sociotechnical systems, and fully automated systems using Industry 4.0 technology. Smart systems can update their internal knowledge up to date to make the best decisions(Romero et al., 2020). User mobility, or the ability of the (smart) environment to continuously provide access to computational tasks and resources, is a critical attribute of smart systems. To link the dilemmas associated with smart systems to flexible modularization, an understanding of the *design principles* of Industry 4.0 is critical. Hermann et al., (2016) identifies six critical principles for designing Industry 4.0: *interoperability, virtualization, decentralization, real-time capability, service orientation* and *modularity*. Incorporating these design principles for Industry 4.0 will give rise to smart systems (and its elements) with the following characteristics*: communication capability, embedded knowledge, learning capability, reasoning capability, perception capability* and *control capability*. However, these smart systems are not static, and as the systems elements interact to exhibit the characteristics above, new characteristics and properties emerge. Contemporary methods to system design are still immersed in a "functional top-down design" paradigm founded on the false idea that complex systems design can be "managed." Valckenaers et al., (2003) argue that control in the design of complex systems is illusionary and that 'object-oriented design' moored on 'synthetic' reasoning is appropriate in VUCA situations. Adopting synthetic thinking in the design of complex smart CPS, based on Hermann et al., (2016)'s, systems design principles, is still a global dilemma. The 'Dilemma of Synthetic Design' describes the boundaries of existing smart system design ideas. The "***Dilemma of Synthetic Design***" challenges the limits of analytical thinking in designing complex systems.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

4 Aligning Digital Economy Dilemmas to Systems Thinking

The "System of Systems Methodologies" (SOSM), founded in 1984, is frequently used to comprehend complexity (Jackson & Keys, 1984). The SOSM is presented as a grid (Table 3) with "systems" and "stakeholders" as the primary sources of complexity that decision makers must deal with (Jackson 2020). The vertical axis depicts system complexity as it progresses from simple to complex. A simple system is one with a linear cause-and-effect relationship and predictable outputs. Such a system requires best practices or simple decision-making methods (Jackson 2020). Cause and effect relationships in complex systems are frequently linked in difficult-to-follow chains that are separated in time and space. Decision-making scenarios involving complex systems necessitate the use of experts who can help understand the system's behavior using systems thinking tools such as system dynamics. There are several agents and relationships in complex systems that it is impossible to trace the interactions and outcomes. Despite the numerous cause and effect concatenations, the system typically produces emergent patterns of behavior that can be identified in retrospect (Jackson 2020).

PROBLEM CONTEXTS	*STAKEHOLDERS*		
SYSTEMS	*Unitary*	*Pluralist*	*Coercive*
Complex	Complex–Unitary	Complex–Pluralist	Complex-Coercive
Complicated	Complicated–Unitary	Complicated–Pluralist	Complicated-Coercive
Simple	Simple–Unitary	Simple–Pluralist	Simple-Coercive

Table 3: "Ideal-Type" Grid of Problem Contexts (Jackson, 2020)

The horizontal axis depicts increasing complexity because of conflicting values and/or interests among problem stakeholders. The words "unitary," "pluralist," and "coercive" are used. Values, beliefs, and interests are shared by "unitary" stakeholders. Values and beliefs clash in a "pluralist" relationship, but basic interests are compatible. Among "coercive" stakeholders, there are competing interests as well as irreconcilable values and beliefs. Combining the dimensions of systems and stakeholders yields nine "ideal-type" problem contexts (Table 6). The "Ideal-Type" grid is then used to map aspects of the five previously identified digital economy dilemmas.

All identified digital dilemmas are volatile, uncertain, complex, and ambiguous (VUCA). Morin (2007) considers such dilemmas to be "general complexity" that must be managed by decision makers. CSP, a CST multimethodology, seeks to comprehend dilemmas of "general complexity." CSP embraces systems thinking on the grounds that reductionist thought cannot deal with complexity (Jackson, 2021). Jackson (2020) describes CSP using the acronym EPIC: Exploring the problem situation (E), Producing (P), Intervening flexibly (I), and Checking progress (C). Figure 3 summarizes and describes the EPIC stages of the CSP. The systems thinking approach requires you to use the first stage (Explore) to identify the critical issues of a digital dilemma and the second stage (Produce) to develop an intervention strategy for its resolution. The first stage of CSP demonstrates how to deconstruct a digital transformation dilemma to aid decision makers.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

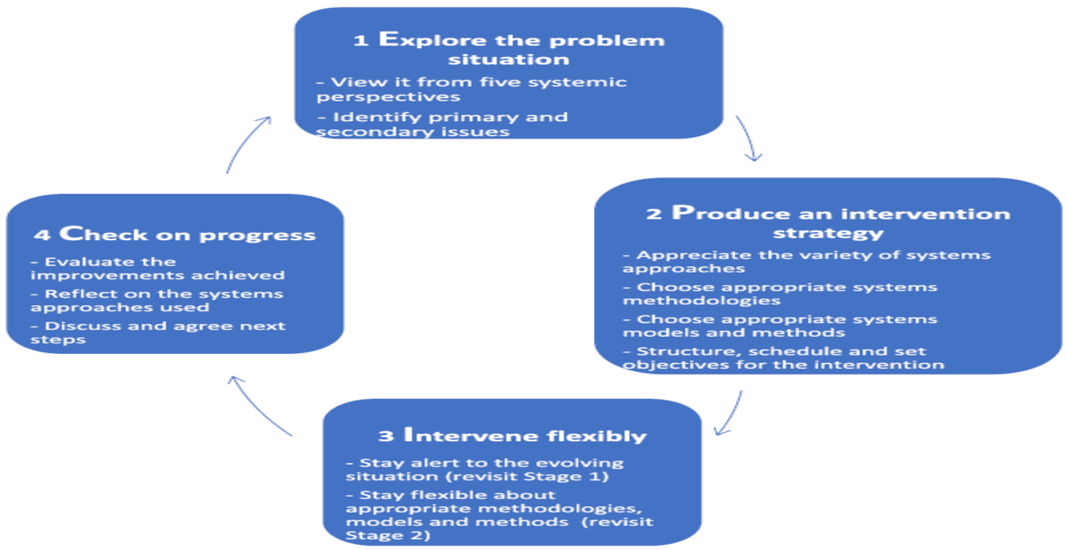


Figure 1:The four key EPIC Stages of CSP (Adopted from (Jackson, 2020a)

In its application for understanding complex problematical situations, CSP makes use of seven integrated and cohesive systemic perspectives that allow decision makers to make sense of specific dilemmas: *machine, organism, cultural, political, coercive system, environmental, interrelationships* (Jackson, 2021). These seven, integrated into five perspectives (See Jackson, 2020a) are the lenses that decision makers can use to 'structure' the dilemmas that they face to foster greater understanding, the goal of systems thinking. To demonstrate, these five systemic perspectives are employed to offer a broad exploration to foster understanding of *aspects* of five digital economies dilemmas identified above. A similar approach can be used for the exploration of any other problematical dilemma that exhibit VUCA characteristics or 'general complexity'.

4.1 Machine View: Exploring Goal Seeking Behavior

The machine perspective is used, exploratively, to identify the causes of faults in an existing problem situation and/or to design a better system (Jackson, 2020). According to this perspective, a problematical situation is like a complex machine made up of parts that work together to achieve a desired outcome. By focusing on how the problematic situation is like a "machine," it is possible to gain insight into the machine perspective. For instance, Morgan (1998) invites us to consider the organization as a machine. Considering the organization as a machine may create valuable insights about how an organization is structured to achieve predetermined results, though the human aspects may be ignored. This mechanical mode of thinking helps decision makers to picture challenging circumstances as interlocking components or elements, each with a clearly defined role in the whole (Kohnen, 1999). This involves identifying the system's aim and its causal connections and effects.

Consider the dilemma of the circular economy, whose potential to disrupt contemporary social institutions built on a linear rationale is being questioned by its detractors. In what ways can the circular economy be conceived using the machine metaphor? To demonstrate, we use Stahel's, (2016) conceptualization, which expands CE to be a performance economy in which items are marketed to consumers as services through rent, leasing, and sharing; while manufacturers and service providers retain ownership of the product and solutions and bear the cost of risks and waste. As a result, the performance economy emphasizes services over products and generates income through sufficiency,

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

such as waste reduction, as well as design, reuse, and business model innovation to circular business models, or CBMs (Stahel, 2016). Using the machine metaphor, the emphasis should be on the structure of the CE to achieve sufficiency through design and business model innovations centred on sustainability yardsticks. Sustainability yardsticks are centred on a dynamic closed-loop system focusing on environmental protection, the full reuse of resources and the recycling of waste (Franco, 2019; Kazancoglu et al., 2021).

Following Jackson (2020), such precision in determining "cause-effect" relationships is linked to 'Hard Systems Thinking' (HST) approaches, notably Systems Analysis (SA), Operations Research/Management Science (OR/MS) and Systems Engineering (SE). For these three 'Hard Systems Thinking' (HST) methodologies, a problematical situation falls within a "simple-unitary" complexity domain (Table 3) where the systems, though having multiple interactions, are "simple"; while the stakeholders agree as to the "goal" of the system. As illustrated above, the goal-oriented nature of the machine metaphor fits well with thinking that underlie HST. However, the machine perspective will only reveal an aspect of the problematical situation where both cases are true, that is, when there is goal agreement and the systems to be designed are 'simple' in the manner of a complete understanding of cause-effect relationship. Prior research demonstrates the application of such HST approaches to understand CE from a machine perspective. Such research is premised on the caveat that the goal of any CE system is sufficiency through product design and business model innovation strategies, with emphasis on quantifiable "cause-and-effect" relationships or factors. For instance, Choi et al.,(2020) synthesizes findings from several research focusing on the sharing and circular economy that employ the use of OR and SE techniques. The machine perspective can help reveal a "slice" of the problematical situation that other perspectives cannot. "***Designing for Sufficiency***" is fundamental to the machine metaphor.

4.2 Organism View: Exploring System Viability

Von Bertalanffy (1937), the founder of General Systems Theory (GST) and the open system concept, is credited with the inception of the organism/organismic perspective. He offered these ideas in opposition to the then-dominant mechanistic and reductionist models in biology and psychology (Hammond, 2010). The organismic perspective shifts from goal-seeking to viability by diagnosing pathologies and suggesting system designs for survival (Jackson, 2020a). The underlying systems approach of the organismic perspective in the manner of von Bertalanffy is moored on the open system concept, in which he highlights the relationship between the organism and environment. The open system approach recognized that organisms are complex entities with many parts (or subsystems) that interact with each other and the environment. When thinking about the interaction of the various parts, the system boundary must be considered because it allows the exchange of materials, energy, and information with the external environment. The system boundary facilitates the exchange of materials, energy, and information with the external environment, according to the open system concept. As an open system that is dependent on its environment, the organism must maintain a dynamic equilibrium with the environment to survive and thrive (Jackson, 2020a). And the sub-systems that comprise the organism only make sense in terms of the functions that the individual parts play in the whole system. However, Sub-systems must have some autonomy or the "brain," which coordinates, will be overwhelmed.

The organism perspective on digital economy dilemmas requires understanding of organizational 'life' images. After the mechanistic view's limitations were realized, the organism vocabulary permeated organizational life discussions. The corporation is a biological concept rooted in a corpus, a body or organism (Ackoff, 1990). C-level executives have adopted organismic language. The CEO is the organization's "head." Other executives have adopted the 'head' vocabulary to represent various business functions, such as 'Head' of Human Capital, CIO as 'head' of Information Systems, etc. The organism metaphor in organizational life confirms Morgan's (1986) claim that "most modern organizational

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

theorists look to nature to understand organizations" (p. 71). Following Morgan's (1986) claim, Prabhu (2022) observed a marked institutional shift on various dimensions (Table 7). The mechanistic perspective seeks efficiency, while the organismic perspective seeks adaptability to the environment (thus a focus on effectiveness). These organismic ideas can be used to understand 'human' systems, which have subsystems with their own purposes, unlike mechanistic systems.

Dimension	Mechanistic Model	Organismic Perspective
Aggregate view	Sum of efficient parts gives an efficient organization	Organization is greater than the sum of its parts
Management style	Analysis-driven (Activities broken down and managed at sub-unit level)	Synthesis-driven
Psychology	Worker is motivated solely by pay	Workers have a hierarchy of needs
Core focus	Efficiency (purpose-driven)	Effectiveness (adaptation-driven)
Hierarchy	Top-down flow of authority. Centralized	Bidirectional/multidirectional flow. Decentralized.

Table 4: Mechanistic versus Organismic Perspective

Consider the integration and interconnection of cyber-physical systems in Industry 4.0 and Society 5.0. CPSs can integrate computation, networking, and physical processes. CPS integration is complicated by heterogeneous components and interactions. While CPSs are promising systems for addressing societal and technical challenges, their success is dependent on a mindset shift to address collaboration and integration of heterogeneous components. The dilemmas of integration and interconnection of CPSs emerge from various sources. For instance, Wang et al., (2022) points to the lack of focus on human beings on initial designs of CPSs, yet they are meant for humans. Instead, they propose a human-cyber-physical systems (HCPS) perspective that elevates the role of human beings to overcome limitations of contemporary CPSs. Bonci et al., (2019) 'zeroes in' on "web technology" as a limiting "language" for the modern design of CPSs and argues for a shared human-to-machine and machine-to-machine language to actualize such a human-centric viewpoint on the architecture of CPSs. According to Potekhin et al., (2022), the viability of CPSs depends on how well physical objects are represented as digital twins to help manage large distributed systems. Using **digital twins**, it is possible to monitor and improve networked services, physical products, machines, and devices (Stary et al., 2022). Because digital twins intertwined with CPSs, evolving open social and technical systems must maintain a dynamic equilibrium with the changing environment (Jackson, 2020a). Adaptability, as the 'goal seeking' behavior, therefore becomes core for these large distributed CPSs, based on the digital twin's concept. Digital twins are seen as integrators of the digital and physical worlds, as well as of the process of creating value both internally and outside (Barth et al., 2020).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

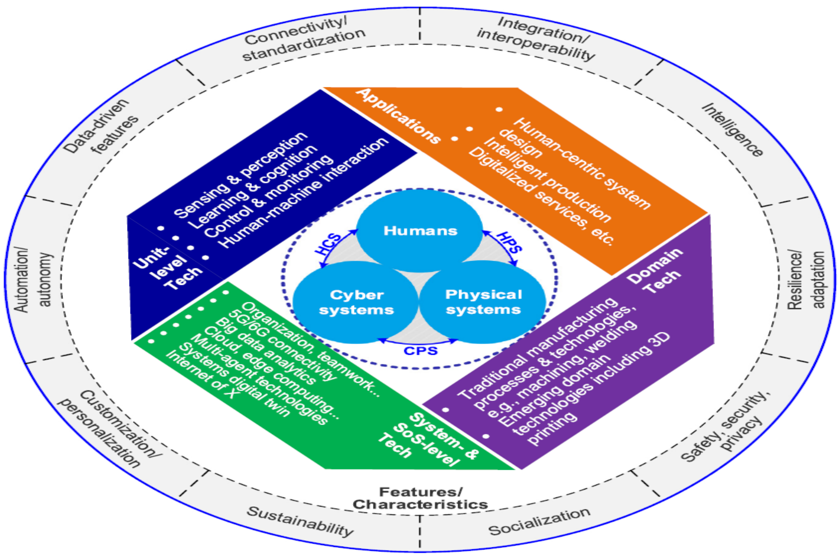


Figure 2: A Human-Cyber-Physical System (Source: Wang et al., 2022)

To overcome the dilemmas associated with distributed CPSs by supporting the proper theorization and design of digital twins to ensure CPSs are adaptable, the relevant systems approaches are Socio-Technical Systems (STS) techniques and the Viable Systems Model (VSM). In the manner of Jackson (2020), STSs and the VSM are associated with problematical situations, in which the systems are "complex", while the stakeholder perspective is "unitary". Complexity is exacerbated due to the heterogeneity of the devices; while the stakeholder perspective is unitary as the discursive practice settles on the "digital twin" concept to characterize distributed CPSs. Exemplars of the use of these two systems thinking approaches include the use of the VSM to create a holistic vision of a system-of-system based on the holon concept to simplify the design and implementation of a CPS architecture that exploits recursivity (Bonci et al., 2018) ; and the use of STS to incorporate the human-centric perspective in the design of distributed CPSs and to enhance the bi-directional and multidirectional control, aggregation and management of such systems (Anumba et al., 2022; Horváth, n.d.; Pessoa et al., 2022). Therefore, "**Designing for Viability**" remains a conundrum, even as CPSs infrastructures take hold in the evolving Industry 4.0/5.0 society.

4.3 Cultural/Political View: Exploring Inscribed Cultural Knowledge

The "sharing" economy, also called "platform capitalism," is gaining traction, revolutionizing product and service production and distribution (Sharif & Huang, 2019). However, "platform capitalism" has a darker side, christened the "*share-the-craps-economy*"(Reich, 2015), which continues to entrench the inequality loop linked to the self-reinforcing effect of digital and social exclusion (Ragnedda et al., 2022). Digital and social exclusion can be viewed through the prism of contemporary global society's dominant *cultural* and *political* narratives. However, it requires posing the question: *what is the dominant cultural/political representation in the current practice of platform capitalism?* A specific value extraction model is the by-product of the dominant cultural and political representations that are embedded in the digital turn towards platform capitalism. This model occurs when the technology surpasses its conventional connective and extractive functions (i.e., data as value) by translating *cultural knowledge as a key business practice* (Dal Maso et al., 2021). By extending the concept of "platform capitalism," Dal Maso et al., (2021) contends that this is a form of "*cultural platform capitalism*". That is, a Western paternalistic perspective underpins the platform business

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

model, in which cultural representations in digital artifacts are based on a 'rentier capitalism' value extraction model.

Rentier capitalism entails creating and extracting value through the techno-economic expansion of ownership and/or control over assets, often due to artificial or natural scarcity, quality, or productivity(Birch, 2020). The hegemonic dominance of Western corporations in rentier capitalism is well established(Christophers, 2022; Klinge et al., 2022), so their 'cultural knowledge' is deeply embedded in the digital economy. Prior research shows that the hegemonic dominance of specific cultural representations via intermediary digital infrastructures has increased digital and social isolation for large segments of the population, as evidenced by the digital divide (Figueiredo & Borges, 2021; Kwet, 2022; Mihelj et al., 2019; Peters, 2022). This form of rentier capitalism, in which digital monopolies create and extract value, is contradictory and has been criticized (Birch et al., 2022). As a result, Western cultural platform capitalism has now become the shared culture, but diversity and differences in human interests continue to fuel resistance, conflict, and contestation of the digital space. Given that Western "cultural knowledge" dominates the digital monopolies shaping the trajectory of the sharing economy, the dilemma of rentier capitalism may be characterized as a "simple-pluralist" structure. Digital monopolies are "simple" because they follow a well-known cultural/political logic of their "cultural knowledge": a "Winner-Takes-All" approach in the design of the digital platform ecosystem. Considering that there are numerous parties involved in the digital platform ecosystem who hold various values, beliefs, and interests, the problem context for rentier capitalism is "pluralist." The cultural/political perspective emphasizes how a diverse groups of individuals become stablished, as they promote processes imprinted with their own interests, to become a “shared culture” (Jackson, 2020). "Winner Take It All" has become the shared cultural knowledge that anchors the sharing economy. To uncover assumptions that underpin such “shared cultures”, possible questions to ask include: What assumptions are inscribed in digital platforms due to Western-oriented cultural platform capitalism? How can these assumptions be 'surfaced' to re-design digital artifacts that reduce social and digital exclusion? Uncovering stakeholders' assumptions in system design emphasizes the 'pluralist' dimension of people complexity. Systems methodologies that align well with problem context of “*Digital Exclusion through Rentier Capitalism*” include *Strategic Assumption Surfacing and Testing (SAST), Interactive Planning (IP),* and *Soft Systems Methodology (SSM)*.

4.4 Societal/Environmental View: Explore ‘Humanization’ of Systems

The discourses of humanity’s societal transition to Industry 4.0/5.0 are predominantly couched in the language of Digital transformation (DT). However, dilemmas associated with the pursuit of ‘undemocratic’ *ideal type DT process* point to outcomes that have exacerbated different manifestations of *social exclusion* (Figure 3). Prior research has determined that social exclusion is by design since ‘smart’ technologies, the bedrock of Industry 4.0/5.0, are currently being embedded into social systems that already have underlying inequalities (Park & Humphry, 2019). There is extensive support from scholarship that reveal that social exclusion is by design (see Goldstraw & Herrington, 2021; Good Gingrich, 2010; Jensen, 2021; Park & Humphry, 2019; Sin, 2019). Smart' technologies and systems, the cornerstone of Industry 4.0/5.0, are therefore being incorporated into social systems with underlying inequities. Addressing social exclusion through DT, which excludes certain segments of humanity, is a persistent problem in the transition to Industry 4.0/5.0. From an Industry 4.0/5.0 standpoint, understanding contemporary social inequities that perpetuate chronic social exclusion requires a systemic lens.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

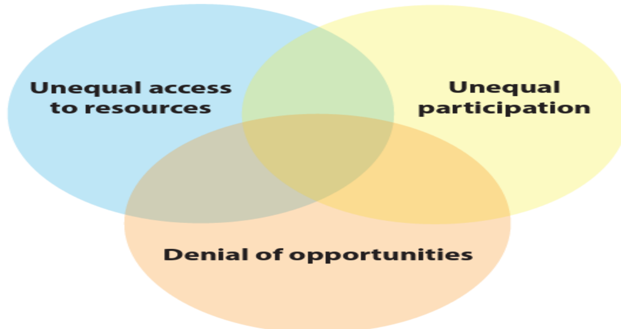


Figure 3:Social Exclusion (Source: Mayer et al., 1958)

The *social/environmental* approach may provide light on parts of the digital exclusion challenge, which is being addressed from a human-centric, sustainable, and resilient standpoint (Figure 4). The *social/environmental perspective*, as a systems thinking lens, fosters the understanding of discrimination and inequality in the context of 'neglected' stakeholders (Jackson, 2020). Social exclusion is a kind of hegemonic tyranny that reveals power dynamics. This kind of hegemonic tyranny dehumanizes by using coercive power via powerful 'digital intermediaries' to influence the DT process' trajectory and outcome. As Industry 5.0 takes shape, the dehumanizing ideas of Industry 4.0 are being disputed (Grabowska et al., 2022), and a new regime of 'humanization' is emerging in which the technological environment of Industry 4.0 is coupled with a human-centric approach to Industry 5.0. Humanism resists the authoritarian control of 'cultural platform capitalism' and strives to include human-centricity, resilience, and sustainability, which define modern debate for human survival. The most probable emancipatory systems lenses for understanding power dynamics in the design of simple, complicated, and complex systems under coercive stakeholder situations are Team Syntegrity (TS), Critical Systems Heuristics (CSH), and Liberating Systems Theory (LST).

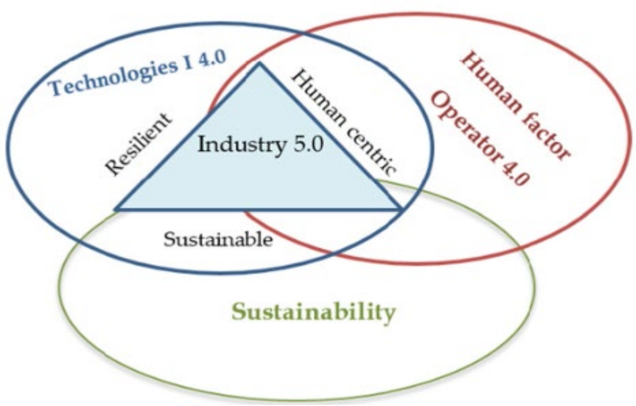


Figure 4: Framework of Industry 5.0 (Source: Grabowska et al., 2022)

4.5 Interrelationship View: Exploring Causality in Systems

As the legitimacy and mobilization for Industry 4.0/5.0 takes hold, it will become clear that the *raison d'être* for the digital transformation process is the development and integration of *smart systems* as the core components of the digital economy. Smart systems are not static, but are dynamic, open,

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

and complex in design, while contemporary approaches to design of systems are still steeped in a ‘functional top-down design’ paradigm premised on the unrealistic assumption that complex systems design can be ‘controlled. Further, smart systems, such as smart cities, smart buildings, and autonomous cars, are each essentially a System-of-Systems (SoS), which are dynamically established as alliances among independent and heterogeneous software systems to offer complex functionalities due to constituent interoperability (Neto et al., 2018). However, recent implementations of smart systems have elicited various challenges that point to a dilemma in how such systems are designed. For instance Mohammadi et al., (2022) cites the difficulty of decision making in smart environments due to the high direct and indirect *dimensional factors*. Following Medina-Borja (2015)’s conceptualization of “smart service systems”, Alter, (2020) characterizes the smartness of systems into four categories with several dimensions: *information processing* (six dimensions), *internal regulation* (five dimensions), *knowledge acquisition* (six dimensions), and *action in the world* (six dimensions). Due to the high dimensionality of these smart systems, computational intensity is extremely high in dimensionality which consequently creates a data storage burden at every smart devices due to their different capabilities (Kamruzzaman, 2021). Also, with the broader push towards Industry 4.0/5.0, there are open security and privacy challenges that have heightened, with an increase in the frequency and sophistication of cyber-physical attacks posing significant threats to organizations globally (Williams et al., 2023). The intrinsic dilemma of designing smart systems arises from the high dimensionality of the concept of ‘smartness’ used to characterize to what extent a system is smart (Figure 5).

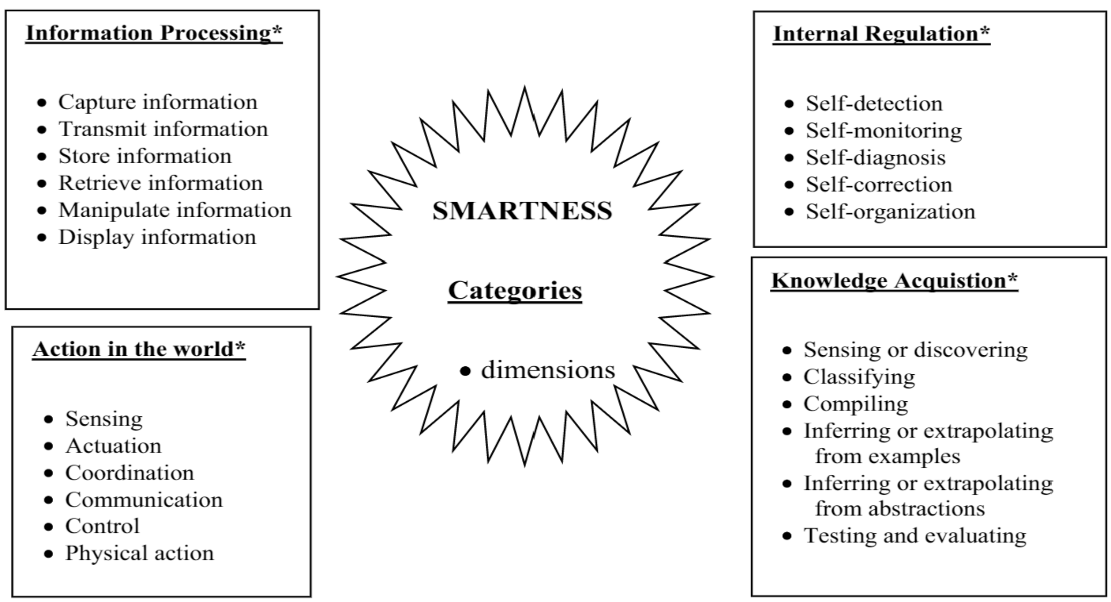


Figure 5: Dimensions of Smartness (Source: Alter, 2020)

Understanding how smart system dimensions interact can shed light on their nature and improve design. The aspects of interest would focus on causal linkages between the dimensions and how such linkages influence the ‘smartness’ of the smart system. The *interrelationship perspective*, a lens of systems thinking, can be an effective approach for structuring such causal linkages of smart systems. The foundation of the interrelationship perspective is that all *issues* (*factors, dimensions*) identified in a problematical scenario are interconnected in chains of reciprocal causation, which suggest the possibility of unintended *consequences* or serve as leverage points for improved design (Jackson, 2020). To understand the 'smartness' of systems, *System Dynamics (SD)* can simulate such known causal

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

relationships. To diagnose smart systems or improve their design, prior research shows how SD may be utilized as a modeling tool. Prior research shows that SD can be used to diagnose or improve smart system design. Khalid Khan et al., (2022) used SD to model the cause-effect relationships and mechanisms of Connected and Autonomous Vehicles (CAVs) cybersecurity and system behavior to create a road map for an optimized, self-regulating, and resilient cyber-safe CAV system. Recent contributions using SD include modeling of smart tourism ecosystems (Sedarati et al., 2022; Shafiee et al., 2022); understanding the relationship between information management barriers and collaborative technologies adoption factors, and sustainability-related policy analysis in the built environment (Amin et al., 2022; Francis & Thomas, 2022); and an exploration of factors that foster smart city success, as well as the cause-and-effect relationships among these determinants (Nunes et al., 2021). Although not exhaustive, these contributions show the proven usage of SD for modeling difficult diagnostic or prognostic cause-effect situations. The *interrelationship perspective* offered through SD makes it an ideal systems methodology for structuring such problematical situations with cause-effect relationships.

5 Conclusions

Systems-based approaches can be used to explore digital dilemmas from five different perspectives: machine, organism, cultural/political, societal/environmental, and interrelationship. The machine perspective emphasizes "design for sufficiency" as an efficiency goal; the organism perspective emphasizes system survival and viability; and the cultural/political perspective reveals how embedded "cultural knowledge" influences system design, use, and control. Understanding discrimination and inequality in the context of 'neglected' stakeholders is facilitated by a social/environmental perspective (Jackson, 2020a). The interrelationship perspective can be used to effectively to structure causal linkages of smart systems with interconnected parts. The synthesis revealed that a 'winner-takes-all' Western-oriented "shared culture" pervades digital artifact design. Systems thinking can assist in bringing to light hegemonic assumptions in digital artifacts of the digital economy. Complementing analytical approaches with systemic perspectives and methodologies can improve complex system diagnosis and design.

References

Ackoff, R. L. (1973). Science in the systems age: Beyond IE, OR, and MS. *Operations Research*, *21*(3), 661–671.

Ackoff, R. L. (1979). The future of operational research is past. *Journal of the Operational Research Society*, *30*(2), 93–104.

Ackoff, R. L. (1990). The management of change and the changes it requires of management. *Systems Practice*, *3*(5), 427–440. https://doi.org/10.1007/BF01064153

Ackoff, R. L. (1997). Systems, messes and interactive planning. *The Societal Engagement of Social Science*, *3*(1997), 417–438.

Alter, S. (2020). Making Sense of Smartness in the Context of Smart Devices and Smart Systems. *Information Systems Frontiers*, *22*(2), 381–393. https://doi.org/10.1007/s10796-019-09919-9

Amin, F., Khan, K. I. A., Ullah, F., Alqurashi, M., & Alsulami, B. T. (2022). Key Adoption Factors for Collaborative Technologies and Barriers to Information Management in Construction Supply Chains: A System Dynamics Approach. *Buildings*, *12*(6), 766.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Anumba, C. J., Lu, Q., & Kan, C. (2022). Beyond building information modeling: the role of cyber-physical systems and digital twins. *Research Companion to Building Information Modeling*, 650–670.

Appio, F. P., Frattini, F., Petruzzelli, A. M., & Neirotti, P. (2021). Digital transformation and innovation management: a synthesis of existing research and an agenda for future studies. *Journal of Product Innovation Management*, *38*(1), 4–20. https://doi.org/10.1111/jpim.12562

Barth, L., Ehrat, M., Fuchs, R., & Haarmann, J. (2020). Systematization of digital twins: Ontology and conceptual framework. *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, 13–23.

Bauwens, T., Hekkert, M., & Kirchherr, J. (2020). Circular futures: what will they look like? *Ecological Economics*, *175*, 106703.

Belmonte-Ureña, L. J., Plaza-Úbeda, J. A., Vazquez-Brust, D., & Yakovleva, N. (2021). Circular economy, degrowth and green growth as pathways for research on sustainable development goals: A global analysis and future agenda. *Ecological Economics*, *185*(March). https://doi.org/10.1016/j.ecolecon.2021.107050

Berawi, M. A. (2019). The role of industry 4.0 in achieving Sustainable Development Goals. *International Journal of Technology*, *10*(4), 644–647.

Bertalanffy, L. von. (1937). The organismic conception. *Perspectives on General Systems Theory: Scientific-Philosophical Studies*, 97–102.

Birch, K. (2020). Technoscience rent: Toward a theory of rentiership for technoscientific capitalism. *Science, Technology, \& Human Values*, *45*(1), 3–33.

Birch, K., Ward, C., & Tretter, E. (2022). Introduction: new frontiers of techno-economic rentiership. *Competition \& Change*, 10245294221082244.

Bodrožić, Z., & S. Adler, P. (2022). Alternative futures for the digital transformation: A macro-level Schumpeterian perspective. *Organization Science*, *33*(1), 105–125.

Bonci, A., Carbonari, A., Cucchiarelli, A., Messi, L., Pirani, M., & Vaccarini, M. (2019). A cyber-physical system approach for building efficiency monitoring. *Automation in Construction*, *102*(January), 68–85. https://doi.org/10.1016/j.autcon.2019.02.010

Bonci, A., Pirani, M., Cucchiarelli, A., Carbonari, A., Naticchia, B., & Longhi, S. (2018). A review of recursive holarchies for viable systems in CPSs. *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, 37–42.

Bucy, M., Finlayson, A., Kelly, G., & Moye, C. (2016). *The 'how'of transformation*.

Choi, T. M., Taleizadeh, A. A., & Yue, X. (2020). Game theory applications in production research in the sharing and circular economy era. *International Journal of Production Research*, *58*(1), 118–127. https://doi.org/10.1080/00207543.2019.1681137

Christophers, B. (2022). *Rentier capitalism: Who owns the economy, and who pays for it?* Verso Books.

Clarke, T. (2019). Creative destruction, technology disruption, and growth. In *Oxford Research Encyclopedia of Economics and Finance*.

Codagnone, C., & Martens, B. (2016). Scoping the sharing economy: Origins, definitions, impact and regulatory issues. *Cristiano Codagnone and Bertin Martens (2016). Scoping the Sharing Economy: Origins, Definitions, Impact and Regulatory Issues. Institute for Prospective*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Technological Studies Digital Economy Working Paper, *1*.

Corvellec, H., Stowell, A. F., & Johansson, N. (2022). Critiques of the circular economy. *Journal of Industrial Ecology*, *26*(2), 421–432. https://doi.org/10.1111/jiec.13187

Daellenbach, H., McNickle, D., & Dye, S. (2012). *Management science: decision-making through systems thinking*. Palgrave macmillan.

Dal Maso, G., Robertson, S., & Rogers, D. (2021). Cultural platform capitalism: Extracting value from cultural asymmetries in RealTech. *Social \& Cultural Geography*, *22*(4), 565–580.

Dantas, T. E. T., de-Souza, E. D., Destro, I. R., Hammes, G., Rodriguez, C. M. T., & Soares, S. R. (2021). How the combination of Circular Economy and Industry 4.0 can contribute towards achieving the Sustainable Development Goals. *Sustainable Production and Consumption*, *26*, 213–227. https://doi.org/10.1016/j.spc.2020.10.005

de Souza Junior, H. R. A., Dantas, T. E. T., Zanghelini, G. M., Cherubini, E., & Soares, S. R. (2020). Measuring the environmental performance of a circular system: Emergy and LCA approach on a recycle polystyrene system. *Science of The Total Environment*, *726*, 138111.

Dickel, S., & Schrape, J.-F. (2017). The logic of digital utopianism. *NanoEthics*, *11*(1), 47–58.

Dunbar-Hester, C. (2019). Challenging digital utopianism. *Media Management and Digital Transformation*.

Ethiraj, S. K., & Levinthal, D. (2004). Modularity and innovation in complex systems. *Management Science*, *50*(2), 159–173.

Figueiredo, T., & Borges, A. L. A. (2021). "And then the pandemic happened...": the digital divide in education in times of Covid-19 and critical information literacy as a route for the exercise of citizenship. *International Review of Information Ethics*, *30*, 8.

Francis, A., & Thomas, A. (2022). System dynamics modelling coupled with multi-criteria decision-making (MCDM) for sustainability-related policy analysis and decision-making in the built environment. *Smart and Sustainable Built Environment*, *Mcdm*. https://doi.org/10.1108/SASBE-09-2021-0156

Franco, M. A. (2019). A system dynamics approach to product design and business model strategies for the circular economy. *Journal of Cleaner Production*, *241*, 118327. https://doi.org/10.1016/j.jclepro.2019.118327

Friedenthal, S., Davey, C., Nielsen, P., Schreinemakers, P., Oster, C., Sparks, E., Matthews, S., Riethle, T., Stoewer, H., Nichols, D., & Roedler, G. (2021). *Systems Engineering Vision 2035*. www.incose.org/sevision

Ganapati, S., & Reddick, C. G. (2018). Prospects and challenges of sharing economy for the public sector. *Government Information Quarterly*, *35*(1), 77–87. https://doi.org/10.1016/j.giq.2018.01.001

Gärtner, C., & Schön, O. (2016). Modularizing business models: between strategic flexibility and path dependence. *Journal of Strategy and Management*.

Geissdoerfer, M., Savaget, P., Bocken, N. M. P., & Hultink, E. J. (2017). The Circular Economy – A new sustainability paradigm? *Journal of Cleaner Production*, *143*, 757–768. https://doi.org/10.1016/j.jclepro.2016.12.048

Goldstraw, K., & Herrington, T. (2021). *Socially Distanced Activism: Voices of Lived Experience of*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Poverty During COVID-19. Policy Press.

Good Gingrich, L. (2010). The symbolic economy of trans-border governance: A case study of subjective exclusion and migrant women from Mexico. *Refugee Survey Quarterly*, *29*(1), 161–184.

Grabowska, S., Saniuk, S., & Gajdzik, B. (2022). Industry 5.0: improving humanization and sustainability of Industry 4.0. *Scientometrics*, 1–28.

Hammond, D. (2010). *The science of synthesis: Exploring the social implications of general systems theory*. University Press of Colorado.

Harvey, M., Hastings, D. P., & Chowdhury, G. (2021). Understanding the costs and challenges of the digital divide through UK council services. *Journal of Information Science*, 01655515211040664.

Hensmans, M. (2021). How digital fantasy work induces organizational ideal reversal? Long-term conditioning and enactment of digital transformation fantasies at a large alternative bank (1963--2019). *Organization*, *28*(1), 132–163.

Hermann, M., Pentek, T., & Otto, B. (2016). Design principles for industrie 4.0 scenarios. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 3928–3937.

Horváth, I. (n.d.). Designing next-generation cyber-physical systems: Why is it an issue? *Journal of Integrated Design and Process Science*, *Preprint*, 1–33.

Hughes, L., Dwivedi, Y. K., Rana, N. P., Williams, M. D., & Raghavan, V. (2022). Perspectives on the future of manufacturing within the Industry 4.0 era. *Production Planning \& Control*, *33*(2–3), 138–158.

Jackson, M. C. (2020a). Critical systems practice 1: Explore—Starting a multimethodological intervention. *Systems Research and Behavioral Science*, *37*(5), 839–858.

Jackson, M. C. (2020b). How we understand “complexity” makes a difference: Lessons from critical systems thinking and the Covid-19 pandemic in the UK. *Systems*, *8*(4), 52.

Jackson, M. C. (2021). Deploying systems thinking to create a triple-crown business school. In *How to lead academic departments successfully*. Edward Elgar Publishing.

Jackson, M. C., & Keys, P. (1984). Towards a system of systems methodologies. *Journal of the Operational Research Society*, *35*(6), 473–486.

Jensen, O. B. (2021). Urban mobilities and power: Social exclusion by design in the city. In *Hybrid Mobilities* (pp. 37–55). Routledge.

Johnston, R. B. (2016). Arsenic and the 2030 Agenda for sustainable development. *Arsenic Research and Global Sustainability - Proceedings of the 6th International Congress on Arsenic in the Environment, AS 2016*, 12–14. https://doi.org/10.1201/b20466-7

Kamruzzaman, M. M. (2021). New Opportunities, Challenges, and Applications of Edge-AI for Connected Healthcare in Smart Cities. *2021 IEEE Globecom Workshops, GC Wkshps 2021 - Proceedings*, *2022*. https://doi.org/10.1109/GCWkshps52748.2021.9682055

Kazancoglu, Y., Ekinci, E., Mangla, S. K., Sezer, M. D., & Kayikci, Y. (2021). Performance evaluation of reverse logistics in food supply chains in a circular economy using system dynamics. *Business Strategy and the Environment*, *30*(1), 71–91. https://doi.org/10.1002/bse.2610

Khalid Khan, S., Shiwakoti, N., & Stasinopoulos, P. (2022). A conceptual system dynamics model for

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

cybersecurity assessment of connected and autonomous vehicles. *Accident Analysis and Prevention*, *165*(June 2021), 106515. https://doi.org/10.1016/j.aap.2021.106515

Klinge, T. J., Hendrikse, R., Fernandez, R., & Adriaans, I. (2022). Augmenting digital monopolies: A corporate financialization perspective on the rise of Big Tech. *Competition \& Change*, 10245294221105572.

Kwet, M. (2022). The Digital Tech Deal: a socialist framework for the twenty-first century. *Race \& Class*, *63*(3), 63–84.

Marshall, A., Dezuanni, M., Burgess, J., Thomas, J., & Wilson, C. K. (2020). Australian farmers left behind in the digital economy--Insights from the Australian Digital Inclusion Index. *Journal of Rural Studies*, *80*, 195–210.

Mayer, K. B., Bureau of Social Affairs, Secretariat, U. N., Office, I. L., Food and Agriculture Organization, & Organization, W. H. (1958). Report on the World Social Situation. *American Sociological Review*, *23*(2), 238. https://doi.org/10.2307/2089039

McCloskey, S. (2015). From MDGs to SDGs: We need a critical awakening to succeed. *Policy \& Practice-A Development Education Review*, *20*.

Medina-Borja, A. (2015). Smart things as service providers: A call for convergence of disciplines to build a research agenda for the service systems of the future. *Service Science*, *7*(1), ii–v. https://doi.org/10.1287/serv.2014.0090

Medynska, T., Sai, L., Akimkin, O., Kruchak, L., & Doroshenko, T. (2022). Characteristics of the Problems of Socio-Economic Development in the Conditions of Digital Economy. *IJCSNS*, *22*(2), 297.

Micheli, G. J. L., Trucco, P., Sabri, Y., & Mancini, M. (2019). Modularization as a system life cycle management strategy: Drivers, barriers, mechanisms and impacts. *International Journal of Engineering Business Management*, *11*, 1847979018825041.

Mihelj, S., Leguina, A., & Downey, J. (2019). Culture is digital: Cultural participation, diversity and the digital divide. *New Media \& Society*, *21*(7), 1465–1485.

Mitroff, I. I., Betz, F., Pondy, L. R., & Sagasti, F. (1974). On managing science in the systems age: two schemas for the study of science as a whole systems phenomenon. *Interfaces*, *4*(3), 46–58.

Mohammadi, F. G., Shenavarmasouleh, F., Amini, M. H., & Reza Arabnia, H. (2022). Data analytics for smart cities: Challenges and promises. *Cyberphysical Smart Cities Infrastructures: Optimal Operation and Intelligent Decision Making*, 13–27.

Morgan, G. (1986). Images of organization London. *UK Sage*.

Morgan, G. (1998). Images of organization: The executive edition. *Thousand Oaks, CA*.

Morin, E. (2007). Restricted complexity, general complexity. *Science and Us: Philosophy and Complexity. Singapore: World Scientific*, 1–25.

Nayal, K., Kumar, S., Raut, R. D., Queiroz, M. M., Priyadarshinee, P., & Narkhede, B. E. (2022). Supply chain firm performance in circular economy and digital era to achieve sustainable development goals. *Business Strategy and the Environment*, *31*(3), 1058–1073. https://doi.org/10.1002/bse.2935

Neto, V. V. G., Garcés, L., Guessi, M., Paes, C. E. B., Manzano, W., Oquendo, F., & Nakagawa, E. (2018). ASAS: An approach to support simulation of smart systems. *Proceedings of the Annual*

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Hawaii International Conference on System Sciences, *2018-January*, 5777–5786. https://doi.org/10.24251/hicss.2018.724

Nunes, S. A. S., Ferreira, F. A. F., Govindan, K., & Pereira, L. F. (2021). "Cities go smart!": A system dynamics-based approach to smart city conceptualization. *Journal of Cleaner Production*, *313*, 127683.

Ogunmakinde, O. E., Egbelakin, T., & Sher, W. (2022). Contributions of the circular economy to the UN sustainable development goals through sustainable construction. *Resources, Conservation and Recycling*, *178*(October 2021), 106023. https://doi.org/10.1016/j.resconrec.2021.106023

Park, S., & Humphry, J. (2019). Exclusion by design: intersections of social, digital and data exclusion. *Information, Communication \& Society*, *22*(7), 934–953.

Pascual, D. G., Daponte, P., & Kumar, U. (2019). *Handbook of Industry 4.0 and SMART Systems*. CRC Press.

Pessoa, M. V. P., Pires, L. F., Moreira, J. L. R., & Wu, C. (2022). Model-Based Digital Threads for Socio-Technical Systems. In *Machine Learning for Smart Environments/Cities* (pp. 27–52). Springer.

Peters, M. A. (2022). Digital trade, digital economy and the digital economy partnership agreement (DEPA). In *Educational Philosophy and Theory* (pp. 1–9). Taylor \& Francis.

Potekhin, V. V., Selivanova, E. N., & Katalinič, B. (2022). Development of a digital transformation model for industrial cyber-physical systems. *AIP Conference Proceedings*, *2456*(January). https://doi.org/10.1063/5.0075038

Pourmehdi, M., Paydar, M. M., Ghadimi, P., & Azadnia, A. H. (2022). Analysis and evaluation of challenges in the integration of Industry 4.0 and sustainable steel reverse logistics network. *Computers and Industrial Engineering*, *163*(October 2021), 107808. https://doi.org/10.1016/j.cie.2021.107808

Prabhu, V. N. (2022). Organ, Organism, Organization: A Study in the Evolution of the Organismic Metaphor. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4031673

Ragnedda, M., Laura, M., & Addeo, F. (2022). Telematics and Informatics The self-reinforcing effect of digital and social exclusion : The inequality loop. *Telematics and Informatics*, *72*(May), 101852. https://doi.org/10.1016/j.tele.2022.101852

Ramakrishna, S., Ngowi, A., Jager, H. De, & Awuzie, B. O. (2020). Emerging industrial revolution: Symbiosis of Industry 4.0 and circular economy: The role of universities. *Science, Technology and Society*, *25*(3), 505–525.

Ranchordás, S. (2020). Connected but still excluded? Digital exclusion beyond internet access. *The Cambridge Handbook of Life Sciences, Informative Technology and Human Rights (Cambridge University Press, 2021, Forthcoming), University of Groningen Faculty of Law Research Paper*, *40*.

Reich, R. (2015). The share-the-scraps economy. *Robert Reich*, *2*(2), 2015.

Romero, M., Guédria, W., Panetto, H., & Barafort, B. (2020). Towards a Characterisation of Smart Systems: A Systematic Literature Review. *Computers in Industry*, *120*, 1–23. https://doi.org/10.1016/j.compind.2020.103224

Sanguino, R., Barroso, A., Fernández-Rodríguez, S., & Sánchez-Hernández, M. I. (2020). Current trends in economy, sustainable development, and energy: a circular economy view.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Environmental Science and Pollution Research, *27*(1), 1–7. https://doi.org/10.1007/s11356-019-07074-x

Schelenz, L., & Pawelec, M. (2022). Information and Communication Technologies for Development (ICT4D) critique. *Information Technology for Development*, *28*(1), 165–188. https://doi.org/10.1080/02681102.2021.1937473

Sedarati, P., Serra, F., & Jere Jakulin, T. (2022). Systems approach to model smart tourism ecosystems. *International Journal for Quality Research*, *16*(1), 285–306.

Shafiee, S., Jahanyan, S., Ghatari, A. R., & Hasanzadeh, A. (2022). Developing sustainable tourism destinations through smart technologies: A system dynamics approach. *Journal of Simulation*, 1–22.

Sharif, N., & Huang, Y. (2019). Introduction: Innovation and Work in East Asia. In *Science, Technology and Society* (Vol. 24, Issue 2, pp. 193–198). SAGE Publications Sage India: New Delhi, India.

Sin, J. (2019). Designing the Divide: A Book Report on the Ultimate Handbook on Designing the Digital Divide. *At CHI4Evil Workshop at the 2019 CHI Conference on Human Factors in Computing Systems. ACM.*

Srnicek, N. (2017). The challenges of platform capitalism: Understanding the logic of a new business model. *Juncture*, *23*(4), 254–257.

Stahel, W. R. (2016). Circular economy - A new relationship with our goods and materials would save resources and energy and create local jobs. *Nature*, *531*, 435–438. https://www.nature.com/articles/531435a.pdf

Standing, G. (2018). The precariat: Today's transformative class? *Development*, *61*(1), 115–121.

Stary, C., Elstermann, M., Fleischmann, A., & Schmidt, W. (2022). Behavior-Centered Digital-Twin Design for Dynamic Cyber-Physical System Development. *Complex Systems Informatics and Modeling Quarterly*, *30*, 31–52. https://doi.org/10.7250/csimq.2022-30.02

Suleiman, Z., Shaikholla, S., Dikhanbayeva, D., Shehab, E., & Turkyilmaz, A. (2022). Industry 4.0: Clustering of concepts and characteristics. *Cogent Engineering*, *9*(1), 0–26. https://doi.org/10.1080/23311916.2022.2034264

Tarasenko, F. P. (2020). The Current Stage of Development of Systems Thinking: The Transition from the Ideology of the Machine Age to the Ideology of the Systems Age. In *Applied Systems Analysis* (pp. 167–173). CRC Press.

Valckenaers, P., Van Brussel, H., Hadeli, Bochmann, O., Germain, B. Saint, & Zamfirescu, C. (2003). On the design of emergent systems: An investigation of integration and interoperability issues. *Engineering Applications of Artificial Intelligence*, *16*(4), 377–393. https://doi.org/10.1016/S0952-1976(03)00080-0

van Gigch, J. P. (1984). The meaning of truth in the systems age. *Human Systems Management*, *4*(3), 220–225.

van Langen, S. K., Vassillo, C., Ghisellini, P., Restaino, D., Passaro, R., & Ulgiati, S. (2021). Promoting circular economy transition: A study about perceptions and awareness by different stakeholders groups. *Journal of Cleaner Production*, *316*, 128166.

Vartanova, E. L., & Gladkova, A. A. (2019). New forms of the digital divide. In *Digital Media Inequalities: Policies against Divides, Distrust and Discrimination* (pp. 193–213).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. *Journal of Business Research*, *122*, 889–901.

Wang, B., Zheng, P., Yin, Y., Shih, A., & Wang, L. (2022). Toward human-centric smart manufacturing: A human-cyber-physical systems (HCPS) perspective. *Journal of Manufacturing Systems*, *63*(March), 471–490. https://doi.org/10.1016/j.jmsy.2022.05.005

Williams, B., Soulet, M., & Siraj, A. (2023). A Taxonomy of Cyber Attacks in Smart Manufacturing Systems. *6th EAI International Conference on Management of Manufacturing Systems*, 77–97.

Yi, S., Zaoli, Y., Xuemei, X., & Harish, G. (2022). Complex System Models and Their Application in Industrial Cluster and Innovation Systems. In *Complexity* (Vol. 2022). Hindawi.

Zander, J., & Mosterman, P. J. (2013). *Computation for humanity: Information technology to advance society*. CRC Press.

Zandi, I. (2000). Science and engineering in the age of systems. *What Is Systems Engineering*.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Negotiating Mobile Phone Usage for MHealth by Maternal Healthcare Clients Who Do Not Own Mobile Phones in rural Malawi

Priscilla Maliwichi[1,2*] Wallace Chigona[1], Address Malata[2] and Karen Sowon[1]

[1] University of Cape Town, Cape Town, S.A

[2] Malawi University of Science and Technology, Limbe, Malawi.

pmaliwichi@must.ac.mw, wallace.chigona@uct.ac.za, amalata@must.ac.mw, csowon@gmail.com

Abstract

In poor-resource settings, owning a mobile phone could be an advantage to using developmental interventions based on mobile phones. However, maternal mHealth interventions in these settings are challenged due to low mobile phone ownership among women. Women are less likely to own a mobile phone than their male counterparts. Therefore, for maternal healthcare clients to use maternal mHealth intervention, it is expected that these clients negotiate mobile phone access and usage from owners of mobile phones in their communities. We employed qualitative research methods to understand how maternal healthcare clients who do not own mobile phones negotiate usage of mobile phones for maternal healthcare. Data was collected using semi-structured interviews with maternal healthcare clients and mobile phone owners, and focus group discussions with the maternal healthcare clients. The study found that maternal healthcare clients used cooperative negotiating tactics such as issue-based, compromising, and accommodating to negotiate mobile phone usage. Negotiating mobile phone usage has the potential to enhance digital skills for mobile phone users who do not own mobile phones. The study may inform mHealth implementers on how they may sensitise beneficiaries of mHealth who lack prerequisite technologies on how to negotiate access of mobile phones for mHealth.

Keywords: Negotiating tactics, negotiation, maternal health, mHealth, Malawi

[*] Masterminded the study and created the first stable version of the paper

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

1 Introduction

Maternal health is defined as the health of a woman during pregnancy, delivery, and the postpartum period (WHO, 2012). Maternal health is one of the priority areas of Sustainable Development Goal (SDG) 3.1, which aims to decrease the global maternal mortality ratio (MMR) to below 70 deaths per 100,000 births. MMR refers to the number of maternal deaths per 100,000 live births (Comfort, 2016; Girum & Wasie, 2017). Sub-Saharan Africa has the highest MMR. The region had 179,000 out of 289,000 global deaths in 2014 (WHO, 2014). As a result, developmental organisations have introduced maternal mHealth interventions to provide maternal-related information to maternal healthcare clients in rural areas of Sub-Saharan Africa where most of these deaths occur (Blauvelt et al., 2018). For instance, mHealth interventions may send maternal healthcare clients tips on maternal healthcare and reminders to visit antenatal clinics through short message service (SMS) and voice messages. mHealth interventions have the potential to promote health seeking behaviour and health facility usage for prenatal care, delivery and postnatal care, which is essential for better maternal health outcomes (Sowon & Chigona, 2021).

mHealth involves the use of mobile devices such as mobile phones for health (Noordam et al., 2011). In rural areas of Sub-Saharan Africa, mHealth is challenged due to low mobile phone ownership (Sondaal et al., 2016). In rural areas, there is disparity in mobile phone ownership. For example, in Malawi and Ethiopia there is a disparity of 26% and 46%, respectively (Rheault & Mccarthy, 2016). In Malawi, low literacy levels and poverty in rural areas affects access and use of Information and Communication Technologies (ICTs). In the villages of Malawi, traditional methods of communication, such as word-of-mouth, village messengers, and village headmen are the preferred way to convey messages (Marron et al., 2020). Low levels of literacy and lack of ICTs, such as mobile phones, make these the traditional methods of communication common in rural areas (Marron et al., 2020),

In Sub-Saharan Africa, women are 15% less likely to own a mobile phone than their male counterparts (Handforth & Wilson, 2019). Consequently, women who do not own mobile phones may not have direct access to mHealth interventions, as well as other mobile phone mediated developmental interventions. Therefore, for maternal healthcare clients to access maternal mHealth intervention, it is likely that these clients negotiate mobile phone access and usage from owners of mobile phones in their communities. Some studies have shown that people negotiate with one another all the time in every aspect of life (Patton & Sundar Balakrishnan, 2012; Zohar, 2015). Other studies have found that social norms may be routinised due to the negotiated agency of those processes in institutions (Giddens, 1984; Ling et al., 2020), such as in health service provision institutions (Ling et al., 2020). In other circumstances, marginalised women have used mobile phones to negotiate harassment issues in the work environment (Pei et al., 2022).

Negotiation is referred to as “a form of decision making in which two or more parties talk with one another in an effort to resolve their opposing interests” (Lewicki et al., 2016). One of the reasons why negotiation occurs is to agree on how to share limited resources (Lewicki et al., 2016), In poor resource settings, where households may have only one family member owning a mobile phone, other family members may negotiate usage of the mobile phone (NSO & MACRA, 2014). For maternal mHealth, negotiating mobile phone usage could be possible with family members and community members for whom the maternal healthcare client finds trustworthy (Maliwichi & Chigona, 2022a). It could be the case that maternal health is associated with cultural norms, which may prohibit the maternal healthcare client from negotiating mobile phone usage for maternal-related issues with anybody. Moreover, pregnancy is associated with taboos, where there could be specific people that maternal healthcare clients may negotiate usage of their mobile phone for maternal-related issues (Nyemba-Mudenda & Chigona, 2018).

A few studies in maternal mHealth have reported on the negotiation process of mHealth use in maternal healthcare, where maternal healthcare clients aimed to make their use of mHealth, a

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

culturally appropriate behaviour (Sowon & Chigona, 2021). Most studies in maternal and reproductive mHealth explained the importance of negotiating cost of mHealth solution with mobile network operators when implementing mHealth solutions for reproductive health (Mangone et al., 2016; Prinja et al., 2018; WHO, 2015). However, little is known about how maternal healthcare clients who lack prerequisite technologies such as mobile phones negotiate usage of mobile phones for maternal mHealth (Larsen-cooper et al., 2015; Maliwichi et al., 2021). Therefore, this study seeks to answer the following research question: *How do maternal healthcare clients who do not own mobile phones negotiate use of mobile phones for maternal mHealth interventions?*

To answer the research question, we used a case study of Chipatala Cha Pa Foni (CCPF) (which means Health Centre by Phone) maternal mHealth intervention. CCPF was launched in 2011 in Malawi as a maternal and child health mHealth intervention in one of the poorest performing district in maternal health (VillageReach, 2017). The intervention allows maternal healthcare clients who do not own mobile phones to use mobile phones of family members and other community members. In addition, the intervention recruited community volunteers and provided them with mobile phones to facilitate access and usage of the maternal mHealth intervention by maternal healthcare clients who do not own mobile phones (Larsen-cooper et al., 2015).

The rest of the paper is organised as follows: the next section is literature review, which highlight the concept of negotiation, negotiations in mHealth and maternal mHealth interventions. Literature review is followed by case description of Chipatala Cha Pa Foni. In this section, the Malawi profile is described and then the CCPF mHealth intervention. The methodology section presents the data collection and analysis methods used in this study. In this section, ethical considerations are also highlighted, and the findings section follows. The section presents negotiating tactics used by maternal healthcare clients to negotiate mobile phone usage, followed by analysis, discussion and conclusion. Lastly, study limitation and future work are highlighted.

2 Literature Review

2.1 Maternal health and mHealth interventions

Maternal Health is one of the areas that have received support in the development of mHealth interventions to combat maternal deaths. For example, MomConnect in South Africa, Rapid SMS in Rwanda and CCPF in Malawi (Blauvelt et al., 2018; Ngabo et al., 2012; Seebregts et al., 2016). Maternal mHealth interventions use SMSs and toll-free hotlines for maternal healthcare clients to receive tips on maternal-related issues and reminders to visit antenatal clinics for vaccinations, medications and routine checkups (Crawford et al., 2014; Willcox et al., 2019). Furthermore, the toll-free hotline enables the maternal healthcare clients to call and ask maternal related help and advice (Larsen-cooper et al., 2015).

Studies in maternal mHealth have found that maternal mHealth interventions promotes the health-seeking behaviour among maternal healthcare clients (Blauvelt et al., 2018; Sowon et al., 2022). Consequently, maternal mHealth interventions improve maternal health outcomes. Therefore, it is essential to promote the use of maternal mHealth interventions for all maternal healthcare clients, regardless of whether or not they own a mobile phone (Maliwichi & Chigona, 2022b).

2.2 The concept of negotiation

Negotiations happen every day in our societies. Even though people associate the concept of negotiation in the world of business or politics (Zohar, 2015); negotiations also happen in our homes, at school, and even at the hospital. Children negotiate with their parents how chores and assignments ought to be done at home. Students could also negotiate the due dates of their assignments with their

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

lectures. In all these contexts, main aim of negotiation is to reach a consensus where the negotiating partners reach an agreement (Lewicki et al., 2016). This could be due to the fact that people depend on each other and negotiations are used to manage situations. In addition, negotiations can be used to resolve conflicts (Lewicki et al., 2016) and change institutional processes (Ling et al., 2020).

- Negotiating tactics or strategies

In negotiations, there are different strategies or tactics that negotiators could use during negotiations (Lewicki et al., 2016). These include yielding or accommodating, inaction (also called avoiding), and compromising. Negotiators practicing accommodating tactic does not have interest whether they would gain their own outcomes, but rather are interested in whether the other party attains his or her outcomes. This involves lowering one's own aspirations to let the other win. Partners practicing inaction as a tactic show little interest in whether they attain their own outcomes, as well as in how the other party obtains their desired outcomes. This could mean that the parties are doing nothing or they have withdrawn from negotiations. Compromising tactics are a conflict management strategy. Here one negotiating partner exerts little effort to achieve their outcomes, but exerts moderate effort to help the other party achieve their outcomes.

- Importance of negotiating skills

Negotiating skills can be in-born. For example, children negotiate with their parents all the time at home. Others have to learn to negotiate due to the nature of their job. Regardless of whether one has an in-born negotiating skill or not, it is important for people to have negotiating skills considering the interdependent nature of the society. Studies have highlighted the importance of enhancing negotiating skills in people (Roloff et al., 2003; Sadki & Bakkali, 2015). Others have published practical guides on how to enhance negotiating skills. For example, there are guidelines which aid implementers of mobile application solutions on how to negotiate with mobile network operators on the cost of implementing affective mHealth solutions (Crawford et al., 2014; WHO, 2015). Therefore, it is important to negotiate in every aspect of life.

2.3 Negotiations in mHealth

Negotiations in mHealth happen at various stages of the development of the mobile applications. For instance, mHealth developers and stakeholders of mHealth negotiate different aspects of mHealth. mHealth developers may negotiate with potential users of mHealth; for example, what functions ought to be present in mobile health application (Giunti, 2018). On one hand, these negotiations happen during the planning and development phase of the mHealth application (Vickery, 2015). On the other hand, negotiations could also happen during mHealth policy formulations, considering the sensitiveness of health data (Sadki & Bakkali, 2015). Hence, policy makers ought to negotiate health data usage with mHealth implementers.

Others have highlighted the importance of negotiating how Smartphone companies may use personal data derived from health application. Users use mobile application for health on their Smartphones, for instance, measuring their blood pressure without knowledge that this data may be sold by mobile phone manufactures (Faulkner, 2018). This is not acceptable, and there ought to be a clear distinction between health (which is a personal right and also things that benefit society as a whole) and consumer goods bought and sold for profit-making purposes (Faulkner, 2018). Therefore, there ought to be a negotiation-based approach to resolving issues concerning the privacy policy conflicts in mHealth (Sadki & Bakkali, 2015).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

2.4 Negotiations in maternal mHealth intervention

In maternal mHealth intervention, the concept of negotiation has been highlighted twofold: on the one hand, mHealth implementers have negotiated the cost of delivering short message services (SMSs) to maternal healthcare clients (Larsen-Cooper et al., 2015; WHO, 2015). In other circumstances, mHealth implementers have negotiated the cost of calls with mobile service providers for maternal mHealth interventions (Crawford et al., 2014; Mangone et al., 2016). These negotiations have led to the provision of toll-free hotlines and free SMS delivery to maternal healthcare clients (Crawford et al., 2014). Moreover, collaborations with different stakeholder in an mHealth project could have an impact in successful negotiations, which could also lead to scaling-up of interventions (Blauvelt et al., 2018).

On the other hand, maternal healthcare clients have negotiated maternal mHealth use. In contexts where pregnancy is surrounded by social-cultural norm, maternal healthcare clients have to negotiate usage of maternal mHealth interventions. This could be due to the fact that the use of mHealth ought to be culturally appropriate given the interdependent nature of maternal healthcare-seeking behaviour (Sowon & Chigona, 2021). Theories have also highlighted on the importance of appropriate technology use, for instance, the compatibility concept in the diffusion of innovation theory (Rogers, 2003). Therefore, it is important to negotiate appropriate use of maternal mHealth considering the sensitivity of maternal issues. However, there is still dearth of literature on how those who lack prerequisite technologies negotiate use of maternal mHealth interventions (Maliwichi & Chigona, 2022a).

3 Case Description of Chipatala Cha Pa Foni

3.1 Malawi Profile

CCPF is an mHealth intervention running in Malawi. Malawi is a country in Southern Africa with a population of about 17,563,749 million people (National Statistics Office (NSO), 2018). About 85.6% of this population resides in rural areas of the country (National Statistics Office (NSO), 2018). CCPF was piloted in 2011 in four rural catchment areas of Balaka District as a maternal and child health intervention. Balaka District was selected due to its lowest maternal and child health indicators in the southern region of Malawi.

The healthcare system in Malawi is organised into four levels, namely: community, primary, secondary, and tertiary levels (MoH, 2022). All these levels are linked through an established referral system. The District Health Officer (DHO) is in-charge of community, primary, and secondary levels, which are at the district level. The health system in Malawi is challenged with limited human resource for health, such as doctors, nurses, and clinical officers. This has been attributed to the migration of health professionals, and death due to HIV/AIDS. The doctor-patient ratio, stands at 1 to 48000, which is one of the lowest in the region (Kawale et al., 2019). This low ratio extends to other health professionals, such as nurses and HSAs (Lutala & Muula, 2022), contributing to poor health outcomes. Consequently, this also contributes to poor maternal health outcomes.

3.2 Chipatala Cha Pa Foni mHealth intervention

CCPF is a toll-free health hotline in Malawi that creates a link between the health centre and remote communities. It has trained health workers, who provide information and referrals over the phone (Innovation Working Group (IWG), 2016). Initially, users could also opt for personalised voice and SMS health messages tailored to women regarding their pregnancy or the health needs of their children (Innovation Working Group (IWG), 2016). The personalised voice and SMS health messages

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

have now been replaced with on demand IVR messages. Remote rural communities in Malawi, as in other developing countries, face challenges such as long distances to the health facility, which prevents people from seeking healthcare. The CCPF maternal mHealth intervention was introduced to improve healthcare-seeking behaviour and healthcare utilisation. By 2018, CCPF was scaled-up to all districts in Malawi and is now fully owned by the Government of Malawi (VillageReach, 2018).

3.3 Community volunteers of CCPF

The implementing agency recruited community volunteers in communities of Balaka District to act as agents of CCPF. These community volunteers were trained on how to use CCPF and provided with a mobile phone. The volunteers were encouraged to sensitise communities by visiting maternal healthcare clients door-to-door, as well as during social gathering in their communities (Watkins et al., 2013). Maternal healthcare clients who do not own mobile phones were encouraged to use mobile phones of community volunteers, other community members, and family members. This made CCPF accessible to maternal healthcare clients regardless of their mobile phone ownership status.

4 Methodology

The study employed qualitative research methods, a single case study research design and interpretive paradigm. Qualitative research methods were more appropriate to explain how maternal clients who do not own mobile phones negotiated mobile phone access and usage of the mHealth intervention. Furthermore, qualitative research methods proved significant in preventing the loss of context. The study bounded the case using the context and the activity, thereby making the case study more appropriate.

4.1 Data collection

Data was collected using semi-structured interviews and focus group discussions (FGDs). We interviewed seven mobile phone owners and seven maternal healthcare clients. These mobile phone owners and maternal health care pairs were geographically dispersed and it was not economically feasible to visit each pair separately. Hence, we opted for telephone interviews, and telephone calls recorded using callX mobile application. Table 1 lists the mobile phone owners and the maternal healthcare clients who used their mobile phones.

Mobile phone owners	Maternal healthcare client
Husband 1	Client 1
Husband 2	Client 2
Husband 3	Client 3
Husband 4	Client 4 and other maternal clients
Mother-in-law 1	Client 5
Community Volunteer 1	Client 6 – Client 13
Community Member 1	Client 14 – Client 20

Table 1: List of mobile phone owners and maternal healthcare clients who used their mobile phones

We conducted two FGDs (of eight and seven maternal healthcare clients, respectively) in two catchment areas, with the help of a community volunteer and a community member. These mobile phone owners had several maternal healthcare clients using their mobile phone for CCPF. Due to

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

covid-19 travel restrictions, we were unable to travel and conduct the interviews in June 2020. Therefore, we opted for telephonic FGDs with the help of the maternal healthcare clients, community volunteer and community member, as the owners of the mobile phones. We asked the mobile phone owners to find a quiet place and put the mobile phone on loud speaker and be moderators of the discussion. The discussion was recorded using callX mobile application.

4.2 Data Analysis

The collected data was transcribed in Chichewa and was later translated into English. English transcripts were then imported into Nvivo 12 for data analysis. We used inductive thematic data analysis to analyse the data (Braun & Clarke, 2006). Table 2 shows the sample themes and their associated codes.

Main theme	Sub-theme	Sub-theme	Sample code
Negotiating tactics	Issue-based	Actual-issue based	"When we reach the volunteer's house, we tell her that we would like to use CCPF, and then she calls 54747, and we listen to our messages, or we talk to the hotline worker..." [Client 12].
		False-issue-based	"I do not know who used my mobile phone for maternal issues. But the one who was pregnant in this house is my daughter-in-law, so I think my son was using my mobile phone to call CCPF because he does not have a mobile phone and uses my mobile phone for so many things…" [Mother-in-law 1].
	Compromising		"I wait for my husband to come from work for me to use the mobile phone... you know, I cannot tell my husband that I want to use the mobile phone when he is going to work... " [Client 1].
	Accommodating		"It is not difficult to ask the neighbor for a mobile phone because normally it is an emergency and we want to seek advice from CCPF so that we know what to do..." [Client 15].

Table 2: Main themes, sub-themes and their associated codes

4.3 Ethical considerations

We obtained permission to conduct the study in Balaka District from the Balaka District Health Office, as well as the Malawi Ministry of Health and the mHealth implementing agency, VillageReach. We also got ethical clearance from the National Health Sciences Research Committee in Malawi.

Consent was sought to interview the participants and issues of privacy and confidentiality were discussed. It was clarified to participants of FGDs that it was difficult to ensure that what will be discussed remain confidential, due to the nature of the discussion. Even, though the study was more about negotiating usage of mobile phone, we did not recruit pregnant women. This was the case in order to mitigate the risks associated with pregnancy, in such case where a participant recalls a traumatic experience. We anonymised the participants using the code *Client x* and *Husband x*, *Community member* or *Volunteer x*.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

5 Findings

In maternal mHealth intervention, access to mobile phones by the intended users is crucial for the success of the intervention. Even though the maternal healthcare clients in this study did not own mobile phones, they accessed mobile phones through different mobile phone owners, such as family relations (husbands and mothers-in-law), community volunteers, and community members.

Maternal healthcare clients in this study negotiated mobile phone usage from mobile phone owners to use maternal mHealth intervention. The findings of this study suggest that maternal healthcare clients used cooperative tactics to negotiate usage of the mobile phone, rather than competitive tactics. Cooperative tactics are tactics that create value by satisfying the interests of all the parties involved, for example willingness to compromise, and having flexible usage terms. These tactics could be categorised further as: 1) issue-based tactics; 2) compromising tactics; and 3) accommodating tactics.

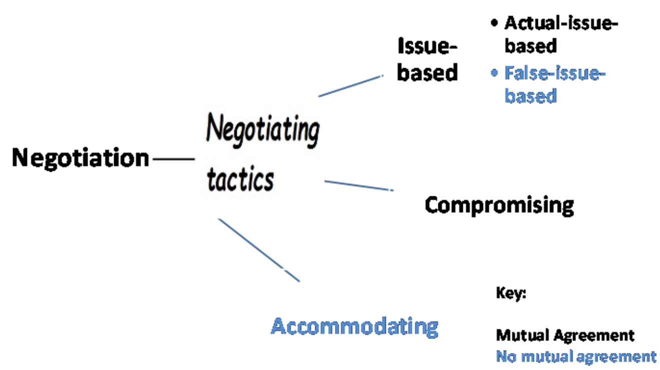


Figure 1: Conceptual framework for negotiating technology use

5.1 Issue-based tactic

Maternal healthcare clients in this study used issue-based tactic to negotiate mobile phone usage. Issue-based negotiation tactics refer to a way of handling one or more negotiation issues in pursuit of a joint or individual goal (Geiger, 2017). Maternal healthcare clients in this study used two ways of issue-based tactic: 1) by using the actual issue, which is to call CCPF for maternal-related issue; and 2) using a different issue to negotiate usage of a mobile phone. These two tactics have been conceptualised further as actual-issue-based tactics, and false-issue-based tactics.

- Actual-issue-based tactic

Most maternal healthcare clients in this study used the actual-issue-based tactics to negotiate usage of a mobile phone. This was apparent when the maternal healthcare client and mobile phone owner had a mutual agreement that the maternal healthcare client could use their mobile phone for maternal-related issue.

> “When we reach the volunteer’s house, we tell her that we would like to use CCPF, and then she calls 54747, and we listen to our messages, or we talk to the hotline worker...” [Client 12].

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

This form of agreement was common between maternal healthcare clients and a dedicated mobile owner. In this case, the maternal healthcare clients trusted the mobile owner, because the mobile owner showed interest in the maternal healthcare client's health and welfare, and took privacy issues about the maternal healthcare client seriously. In Malawi, as well as other African countries, maternal healthcare clients do not just talk about their pregnancy-related issues with ease. Pregnancy is a sensitive topic that is discussed only with someone who can be trusted (Ngomane & Mulaudzi, 2012).

- False-issue-based tactic

Some maternal healthcare clients in this study used false-issue-based tactics of negotiation. We define false-issue-based tactics of negotiation as bargaining of mobile phone usage for an issue, but using the mobile phone for another purpose.

> "I do not know who used my mobile phone for maternal issues. But the one who was pregnant in this house is my daughter-in-law, so I think my son was using my mobile phone to call CCPF because he does not have a mobile phone and uses my mobile phone for so many things…" [Mother-in-law 1].

It is possible that the maternal healthcare clients negotiated usage of the mobile phone for something else, while simply using the mobile phone for maternity-related issues.

5.2 Compromising tactic

The findings of this study suggest that maternal healthcare clients used compromising negotiation tactics when negotiating mobile phone usage. Compromising negotiation tactics usually happens in a win-lose situation (Schmidt & Cross, 2014). Compromising tactics prove to be one of the most basic negotiation tactics where both parties give up something that they want in order to get something else they want more. In this study, maternal healthcare clients had to use concession in order to reach mobile phone usage times with their husbands.

> "I wait for my husband to come from work for me to use the mobile phone... you know, I cannot tell my husband that I want to use the mobile phone when he is going to work ..." [Client 1].

One of the husbands narrated that sometimes, he agreed with his wife on the times both could listen to the messages about pregnancy, or on the times both could talk to the doctor. Compromising negotiation tactics were also more prominent between maternal healthcare clients and community volunteers. A community volunteer mentioned that, when maternal healthcare clients were more advance in their pregnancy and could not travel long distances to access mobile phones, a community volunteer would instead visit them at their homes.

When using compromising tactics, normally negotiating partners work together to find an acceptable middle ground that works for both of them (Schmidt & Cross, 2014). In maternal mHealth, compromising tactics are very important, as both maternal healthcare clients and husbands or community volunteers want to achieve better maternal outcomes.

5.3 Accommodating tactic

The findings of this study suggest that accommodating negotiation tactics was prevalent when maternal healthcare clients were using community members such as neighbours to negotiate mobile phone access.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

> "It is not difficult to ask the neighbor for a mobile phone because normally it is an emergency and we want to seek advice from CCPF so that we know what to do..." [Client 15].

In this case, no formal agreement was necessary for long-term use, but the maternal healthcare client trusted that the neighbour would allow her to use the mobile phone and the neighbour usually accommodated their request. This could be a social norm in villages where people share resources in times of need (Vickery, 2015).

6 Discussion

The findings of this study suggest that the mobile phone owners played a greater role when maternal healthcare clients were negotiating the use of the mobile phone. This could be attributed to the importance of the issue being negotiated for, in this case, maternal health. In addition, health-related issues could be prioritised over personal issues when negotiating mobile phone usage. Therefore, this study highlights: 1) the importance of accommodating beneficiaries of mHealth through technology sharing; 2) when to use bogey or false issue issue over the actual issue; and 3) the role of communities of purpose in promoting access to technologies.

6.1 The importance accommodating beneficiaries of mHealth through technology sharing

In poor-resource settings, beneficiaries of mHealth who do not own mobile phones tend to negotiate mobile phone usage when an emergency occurs. For maternal health, this could be that a maternal healthcare client is in pain and has no idea why this is the case. Sometimes, this could mean saving someone's life. Therefore, it is important for mobile phone owners to accommodate beneficiaries of mHealth or their guardians request to use their mobile phones.

In this study, mobile phone owners used accommodating tactics when maternal healthcare clients negotiated mobile phone usage. An accommodating tactic tends to be more concerned with the cooperativeness of the parties involved (McCracken et al., 2008). Basically, this is a win-lose situation (Coburn, 2013), where "One way to generate alternatives is to allow one person to obtain their objectives and compensate the other person for accommodating their interests" (Lewicki et al., 2016). In this study, community members played the losing partner for accommodating tactics to work for maternal healthcare clients. In maternal health, community members may sacrifice their time in favour of the maternal healthcare client, who may want help at that particular time. Timely access to maternal health information has the potential to serve the lives of the mother and her unborn child. For some community members, this proved to be a way of winning over maternal healthcare clients to use healthcare service during pregnancy.

In this study, accommodating tactics may have worked due to *umunthu* philosophy, which is practiced in most poor-resource communities of Malawi (Zamani & Sbaffi, 2020), where communities value mutuality and reciprocity (Bandawe, 2005). Therefore, it is important for mHealth implementers to sensitise communities to the importance of sharing resources in times of need. Other studies have found that communities that practice a sharing culture, it is a norm in these communities to accommodate beneficiaries of mHealth who lack prerequisite technologies such as mobile phones (James, 2011; Tran et al., 2015). In turn, the beneficiaries feel free to negotiate mobile phone usage from any mobile phone owner.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

6.2 False-issue or actual-issue negotiating tactic: which is appropriate?

Health is a complex phenomenon. Health issues tend to be private. In most cases, mHealth beneficiaries of diabetes or sexually transmitted diseases, for example, would not want anyone to know their health concerns. Maternal health is even more complex, since it is associated with cultural beliefs. On the one hand, beneficiaries of mHealth who do not own mobile phones in this study tend to use false-issue negotiating tactic when negotiating mobile phone usage. This was more prevalent when the maternal healthcare client did not trust the mobile phone owner, or the mobile phone owner was young.

Some studies have found that people may employ “bogey” tactics, which involve exaggerating the importance of a given issue (Lewicki et al., 2016). False-issue-based tactics and bogey tactics could be successful, as long as the truth is not revealed (Geiger, 2017). In maternal health, using false issue-based could happen due to the sensitive nature of maternity-related issues in a rural setting. Maternal healthcare clients treat maternity-related issues as private (Fagbamigbe & Idemudia, 2015).

On the other hand, beneficiaries of mHealth use the actual-issue-based tactic to negotiate usage of the mobile phone. This was due to the fact that the beneficiary of mHealth and the mobile phone owner had a mutual agreement with the mobile phone owner or the mobile phone was a project mobile phone or government-issued mobile phone, for the purposes of maternal and child health intervention (Larsen-cooper et al., 2015; Ngabo et al., 2012). Furthermore, actual-issue-based tactics are easier to use when a mobile phone belongs to maternal healthcare clients’ significant other (Carr et al., 2021).

Regardless of the circumstances of the beneficiaries of mHealth, using an actual-issue-based tactic could be more appropriate, due to the fact that, when false-issue-based tactics or bogey tactics are discovered, people will question the behaviour of the negotiators in this regard. A study in negotiation has also suggested that it is important to negotiate using an important issue first, when using issue-based negotiation tactics (Geiger, 2017). For example, it is important to negotiate mobile phone usage for maternity-related issues first, and later for personal or other concerns. However, certain studies have found that simultaneous bargaining could be beneficial (Patton & Balakrishnan, 2010; Patton & Sundar Balakrishnan, 2012).

6.3 The role of communities of purpose in promoting access to technologies

A community of purpose (CoP) is understood as a community of people who are going through a similar process or are trying to achieve a similar objective (Bhattacharyya et al., 2020). In this study, CoP comprised of family members, community members, community volunteers, and community health workers. CoP played a greater role in providing access of mobile phones to mHealth beneficiaries. The findings of this study suggest that CoP used compromising negotiating tactics when mHealth beneficiaries were negotiating the use of their mobile phones. A community volunteer mentioned that when maternal healthcare clients were more advance in their pregnancy and could not travel long distances to access the mobile phone, where a community volunteer would instead visit them at their homes.

When using compromising tactics, normally negotiating partners work together to find an acceptable middle ground that works for both of them (Schmidt & Cross, 2014). Studies on negotiation have found that prior relationships between negotiating parties influence compromising tactics (McCracken et al., 2008; Schmidt & Cross, 2014). Moreover, this relationship makes negotiation a bit easier for the maternal healthcare client themselves (Zohar, 2015).

In maternal mHealth, compromising tactics are very important, as both maternal healthcare clients and CoP members want to achieve better maternal outcomes; thereby reducing MMR (Sowon & Chigona, 2021). In addition, CoP plays a greater role in reducing the digital divide, which is greater

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

among women (Maliwichi et al., 2021). Moreover, CoP has the potential to enhance the digital skills of mHealth beneficiaries who lack prerequisite technologies by training them on how to use the technology (Larsen-cooper et al., 2015). Access to information has the potential to promote healthy living styles, which could promote sustainable development in rural areas.

7 Conclusions

Negotiating mobile phone usage for health has the potential to promote the health and wellbeing of mHealth beneficiaries who lack prerequisite technologies, such as mobile phones. In addition, negotiating mobile phone usage by mHealth beneficiaries who lack prerequisite technologies has the potential to improve their digital skills. The study noted that mHealth beneficiaries who do not own mobile phones used several negotiating tactics to negotiate mobile phone usage, and these include: 1) issue-based negotiating tactic; 2) compromising negotiating tactic; and 3) accommodating negotiating tactic. This study may inform policymakers and mHealth implementers on how mHealth beneficiaries who lack prerequisite technology may negotiate usage of technologies. Furthermore, mHealth implementers ought to sensitise the roles of CoP in communities, since they play a greater role in reducing the digital divide gap and promoting an inclusive environment.

8 Limitations of the study and future work

The study was limited in using only a single case study. Therefore, future work may test the conceptual framework in similar settings. Furthermore, the conceptual framework could be tested in other sectors, such as education.

References

Bandawe, C. R. (2005). Psychology brewed in an African pot: Indigenous philosophies and the quest for relevance. *Higher Education Policy*, *18*(3), 289–300. https://doi.org/10.1057/palgrave.hep.8300091

Bhattacharyya, S., Basak, J., Bhaumik, P., & Bandyopadhyay, S. (2020). Cultivating online virtual community of purpose to mitigate knowledge asymmetry and market separation of rural artisans in India. In D. Junio & C. Koopman (Eds.), *Evolving Perspectives on ICTs in Global Souths. IDIA 2020. Communications in Computer and Information Science* (Vol. 1236, pp. 201–215). Springer, Cham. https://doi.org/10.1007/978-3-030-52014-4_14

Blauvelt, C., West, M., Maxim, L., Kasiya, A., Dambula, I., Kachila, U., Ngwira, H., & Armstrong, C. (2018). Scaling up a health and nutrition hotline in Malawi: The benefits of multisectoral collaboration. *BMJ*, *363*, k4590. https://doi.org/10.1136/bmj.k4590

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carr, R. M., Quested, E., Stenling, A., Thøgersen-Ntoumani, C., Prestwich, A., Gucciardi, D. F., McVeigh, J., & Ntoumanis, N. (2021). Postnatal Exercise Partners Study (PEEPS): a pilot randomized trial of a dyadic physical activity intervention for postpartum mothers and a significant other. *Health Psychology and Behavioral Medicine*, *9*(1), 251–284. https://doi.org/10.1080/21642850.2021.1902815

Coburn, C. (2013). *Negotiation Conflict Styles*. Harvard Medical School.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

https://karemmieses.com/wp-content/uploads/sites/10042/2019/08/NegotiationConflictStyles.pdf
Comfort, A. B. (2016). Long-term effect of in utero conditions on maternal survival later in life: evidence from Sub-Saharan Africa. *Journal of Population Economics*, *29*(2), 493–527. https://doi.org/10.1007/s00148-015-0581-9
Crawford, J., Larsen-Cooper, E., Jezman, Z., Cunningham, S. C., & Bancroft, E. (2014). SMS versus voice messaging to deliver MNCH communication in rural Malawi: Assessment of delivery success and user experience. *Global Health Science and Practice*, *2*(1), 35–46. https://doi.org/10.9745/GHSP-D- 13-00155
Fagbamigbe, A. F., & Idemudia, E. S. (2015). Barriers to antenatal care use in Nigeria: Evidences from non-users and implications for maternal health programming. *BMC Pregnancy and Childbirth*, *15*(1), 1–10. https://doi.org/10.1186/s12884-015-0527-y
Faulkner, A. (2018). Blood Informatics: Negotiating the Regulation and Usership of Personal Devices for Medical Care and Recreational Self-monitoring. In R. Lynch & C. Farrington (Eds.), *Quantified Lives and Vital Data. Health, Technology and Society*. Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95235-9_9
Geiger, I. (2017). A model of negotiation issue–based tactics in business-to-business sales negotiations. *Industrial Marketing Management*, *64*, 91–106. https://doi.org/10.1016/j.indmarman.2017.02.003
Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. University of California Press.
Girum, T., & Wasie, A. (2017). Correlates of maternal mortality in developing countries: an ecological study in 82 countries. *Maternal Health, Neonatology and Perinatology*, *3*(1), 19. https://doi.org/10.1186/s40748-017-0059-8
Giunti, G. (2018). 3MD for Chronic Conditions, a Model for Motivational mHealth Design : Embedded Case Study. *JMIR Serious Games*, *6*(3), e11631. https://doi.org/10.2196/11631
Handforth, C., & Wilson, M. (2019). *Digital Identity Country Report: Malawi*. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/02/Digital-Identity-Country-Report.pdf
Innovation Working Group (IWG). (2016). *CHIPATALA CHA PA FONI:Healthcare Through Mobile Phones*. https://www.villagereach.org/wp-content/uploads/2016/08/CCPF-Case-Study-UN-Foundation.pdf
James, J. (2011). Sharing mobile phones in developing countries: Implications for the digital divide. *Technological Forecasting and Social Change*, *78*(4), 729–735. https://doi.org/10.1016/j.techfore.2010.11.008
Kawale, P., Pagliari, C., & Grant, L. (2019). What does the Malawi demographic and health survey say about the country's first health sector strategic plan? *Journal of Global Health*, *9*(1), 1–5. https://doi.org/10.7189/jogh.09.010314
Larsen-Cooper, E., Bancroft, E., Rajagopal, S., O'Toole, M., & Levin, A. (2015). Scale Matters: A Cost-Outcome Analysis of an m-Health Intervention in Malawi. *TELEMEDICINE and E-HEALTH*, *22*(4), 1–8. https://doi.org/10.1089/tmj.2015.0060
Larsen-cooper, E., Bancroft, E., Toole, M. O., & Jezman, Z. (2015). Where there is no phone : The benefits and limitations of using intermediaries to extend the reach of mHealth to individuals without personal phones in Malawi. *African Population Studies Special Edition*, *29*(1), 1628–1642. https://doi.org/10.11564/29-1-714
Lewicki, R. J., Barry, B., & Saunders, D. M. (2016). *Essentials of Negotiations* (6th ed.). McGraw Hill Education.
Ling, R., Poorisat, T., & Chib, A. (2020). Mobile phones and patient referral in Thai rural healthcare: a structuration view. *Information Communication and Society*, *23*(3), 358–373. https://doi.org/10.1080/1369118X.2018.1503698

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Lutala, P., & Muula, A. (2022). Brief behaviour change counselling in non-communicable diseases in Mangochi, Southern Malawi: A hypothetical acceptability study. *Pilot and Feasibility Studies*, *8*(68), 1–16. https://doi.org/10.1186/s40814-022-01032-0

Maliwichi, P., & Chigona, W. (2022a). Towards a Framework on the Use of Infomediaries in Maternal mHealth in Rural Malawi. *Interdisciplinary Journal of Information, Knowledge and Management*, *17*, 387–411. https://doi.org/10.28945/5015

Maliwichi, P., & Chigona, W. (2022b). Factors affecting the use of infomediaries in mHealth interventions: Case of maternal healthcare in rural Malawi. *NEMISA Summit and Colloquium 2022*, 1–30. https://uir.unisa.ac.za/bitstream/handle/10500/28698/DigitalSkill2022Proceedings.pdf?sequence=1&isAllowed=y#page=29

Maliwichi, P., Chigona, W., & Sowon, K. (2021). Appropriation of mHealth Interventions for Maternal Healthcare in sub-Saharan Africa: Hermeneutic Review. *JMIR Mhealth Uhealth*, *9*(10), e22653. https://doi.org/10.2196/22653

Mangone, E. R., Agarwal, S., Engle, K. L., Lasway, C., Zan, T., Beijma, H. Van, Orkis, J., & Karam, R. (2016). Sustainable Cost Models for mHealth at Scale : Modeling Program Data from m4RH Tanzania. *PLoS ONE*, *11*(1), e0148011. https://doi.org/10.1371/journal.pone.0148011

Marron, O., Thomas, G., Burdon Bailey, J. L., Mayer, D., Grossman, P. O., Lohr, F., Gibson, A. D., Gamble, L., Chikungwa, P., Chulu, J., Handel, I. G., De C Bronsvoort, B. M., Mellanby, R. J., & Mazeri, S. (2020). Factors associated with mobile phone ownership and potential use for rabies vaccination campaigns in southern Malawi. *Infectious Diseases of Poverty*, *9:62*, 1–11. https://doi.org/10.1186/s40249-020-00677-4

McCracken, S., Salterio, S. E., & Gibbins, M. (2008). Auditor-client management relationships and roles in negotiating financial reporting. *Accounting, Organizations and Society*, *33*(2008), 362–383. https://doi.org/10.1016/j.aos.2007.09.002

MoH. (2022). *The health care system*. Republic of Malawi. https://www.health.gov.mw/index.php/2016-01-06-19-58-23/national-aids

National Statistics Office (NSO). (2018). *2018 Malawi Population and Housing Census: Preliminary Report* (Vol. 39, Issue 1). http://www.nsomalawi.mw/images/stories/data_on_line/demography/census_2018/2018 Population and Housing Census Preliminary Report.pdf

Ngabo, F., Nguimfack, J., Nwaigwe, F., Mugeni, C., Muhoza, D., Wilson, D., Kalach, J., Gakuba, R., Karema, C., & Binagwaho, A. (2012). Designing and implementing an innovative SMS-based alert system (RapidSMS-MCH) to monitor pregnancy and reduce maternal and child deaths in Rwanda. *Pan African Medical Journal*, *13*(31). https://doi.org/10.11604/pamj.2012.13.31.1864

Ngomane, S., & Mulaudzi, F. M. (2012). Indigenous beliefs and practices that influence the delayed attendance of antenatal clinics by women in the Bohlabelo district in Limpopo, South Africa. *Midwifery*, *28*(1), 30–38. https://doi.org/10.1016/j.midw.2010.11.002

Noordam, C., Kuepper, B., Stekelenburg, J., & Milen, A. (2011). Improvement of maternal health services through the use of mobile phones. *Tropical Medicine and International Health*, *16*(5), 622–626. https://doi.org/10.1111/j.1365-3156.2011.02747.x

NSO, & MACRA. (2014). *National Survey on access to and usage of ICT services in Malawi*. http://www.macra.org.mw/wp-content/uploads/2014/09/Survey_on-_Access_and_Usage_of_ICT_Services_2014_Report.pdf

Nyemba-Mudenda, M., & Chigona, W. (2018). mHealth outcomes for pregnant mothers in Malawi : a capability perspective. *Information Technology for Development*, *24*(2), 245–278. https://doi.org/10.1080/02681102.2017.1397594

Patton, C., & Balakrishnan, P. V. S. (2010). The impact of expectation of future negotiation interaction on bargaining processes and outcomes. *Journal of Business Research*, *63*(8), 809–816. https://doi.org/10.1016/j.jbusres.2009.07.002

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Patton, C., & Sundar Balakrishnan, P. V. (2012). Negotiating when outnumbered: Agenda strategies for bargaining with buying teams. *International Journal of Research in Marketing*, *29*(3), 280–291. https://doi.org/10.1016/j.ijresmar.2012.02.002

Pei, X., Chib, A., & Ling, R. (2022). Covert resistance beyond #Metoo: mobile practices of marginalized migrant women to negotiate sexual harassment in the workplace. *Information Communication and Society*, *25*(11), 1559–1576. https://doi.org/10.1080/1369118X.2021.1874036

Prinja, S., Bahuguna, P., Gupta, A., Nimesh, R., Gupta, M., & Thakur, J. S. (2018). Cost effectiveness of mHealth intervention by community health workers for reducing maternal and newborn mortality in rural Uttar Pradesh, India. *Cost Effectiveness and Resource Allocation*, *16*(1), 1–19. https://doi.org/10.1186/s12962-018-0110-2

Rheault, M., & Mccarthy, J. (2016). *Disparities in Cellphone Ownership Pose Challenges in Africa*. GALLUP. https://news.gallup.com/poll/189269/disparities-cellphone-ownership-pose-challenges-africa.aspx

Rogers, E. . (2003). *Diffusion of innovations* (5th ed.). Free Press.

Roloff, M. E., Putnam, L. L., & Anastasiou, L. (2003). Negotiation Skills. In *Handbook of Communication and Social Interaction Skills* (Issue 1, pp. 801–833).

Sadki, S., & Bakkali, H. E. L. (2015). A Negotiation-based Approach to Resolve Conflicting Privacy Policies in M-Health. *Journal of Information & Systems Management*, *5*(3), 91–105.

Schmidt, R. N., & Cross, B. E. (2014). The effects of auditor rotation on client management's negotiation strategies. *Managerial Auditing Journal*, *29*(2), 110–130. https://doi.org/10.1108/MAJ-03-2013-0836

Seebregts, C., Tanna, G., Fogwill, T., Barron, P., & Benjamin, P. (2016). MomConnect: An exemplar implementation of the Health Normative Standards Framework in South Africa. *South African Health Review*, *2016*(1), 125–135. http://hdl.handle.net/10204/8657

Sondaal, S. F. V., Brown, J. L., Amoakoh-Coleman, M., Borgstein, A., Miltenburg, A. S., Verwijs, M., & Klipstein-Grobusch, K. (2016). Assessing the effect of mHealth interventions in improving maternal and neonatal care in low- and middle-income countries: A systematic review. *PLoS One*, *11*(5).

Sowon, K., & Chigona, W. (2021). Conceptualising healthcare-seeking as an activity to explain technology use - a case of mHealth. *1st Virtual Conference on Implications of Information and Digital Technologies for Development*, 836–852.

Sowon, K., Maliwichi, P., & Chigona, W. (2022). The Influence of Design and Implementation Characteristics on the Use of Maternal Mobile Health Interventions in Kenya: Systematic Literature Review. *JMIR MHealth and UHealth, 10*(1), e22093. https://doi.org/10.2196/22093

Tran, M., Labrique, A., Mehra, S., Ali, H., Shaikh, S., Mitra, M., Christian, P., & West Jr, K. (2015). Analyzing the Mobile "Digital Divide": Changing Determinants of Household Phone Ownership Over Time in Rural Bangladesh. *JMIR MHealth and UHealth*, *3*(1), e24. https://doi.org/10.2196/mhealth.3663

Vickery, J. R. (2015). 'I don't have anything to hide, but … ': the challenges and negotiations of social and mobile media privacy for non-dominant youth. *Information Communication and Society*, *18*(3), 281–294. https://doi.org/10.1080/1369118X.2014.989251

VillageReach. (2017). *A national mobile health hotline in Malawi: One step closer*. https://www.villagereach.org/2017/07/22/digital-health-innovation-last-mile-sustainability-partnership/#:~:text=I am pleased to announce,mobile health hotline in Africa.

VillageReach. (2018). *Press Release:National expansion plan for Chipatala Cha Pa Foni (CCPF): A health and nutrition hotline*. http://www.villagereach.org/press-release-national-expansion-plan-chipatala-cha-pa-foni-ccpf-health-nutrition-hotline/

Watkins, S. C., Robinson, A., & Dalious, M. (2013). *Evaluation of the Information and Communications Technology for Maternal, Newborn and Child Health Project*.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

http://www.villagereach.org/wp-content/uploads/2017/07/ICT_for_MNCH_Report_131211md_FINAL.pdf

WHO. (2012). *Trends in maternal mortality: 1990 to 2010*. WHO.

WHO. (2014). *Trends in maternal mortality: 1990 to 2013* (Estimates By WHO, UNICEF, UNFPA, The World Bank, United Nations Population Division). WHO. https://www.who.int/reproductivehealth/publications/monitoring/maternal-mortality-2013/en/

WHO. (2015). *A practical guide for engaging with mobile network operators in mHealth for reproductive , maternal , newborn and child health*. WHO.

Willcox, M., Moorthy, A., Mohan, D., Romano, K., Hutchful, D., Mehl, G., Labrique, A., & LeFevre, A. (2019). Mobile technology for community health in Ghana: Is maternal messaging and provider use of technology cost-effective in improving maternal and child health outcomes at scale? *Journal of Medical Internet Research*, *21*(2). https://doi.org/10.2196/11268

Zamani, E. D., & Sbaffi, L. (2020). Embracing uMunthu: How Informal Caregivers in Malawi Use ICTs. In J. . Bass & P. . Wall (Eds.), *Information and Communication Technologies for Development. ICT4D 2020. IFIP Advances in Information and Communication Technology, vol 587* (pp. 93–101). Springer, Cham. https://doi.org/10.1007/978-3-030-65828-1_8

Zohar, I. (2015). “The Art of Negotiation” Leadership Skills Required for Negotiation in Time of Crisis. *Procedia - Social and Behavioral Sciences*, *209*, 540–548. https://doi.org/10.1016/j.sbspro.2015.11.285

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

Using Dimensionality Reduction Technique in Machine Learning to diagnose heart diseases in South Africa

Abstract. Heart disease is a major health concern in South Africa, and early and accurate diagnosis is crucial for effective treatment. In this context, dimensionality reduction techniques can play an important role. These techniques can help identify patterns and relationships in large and complex datasets, allowing for more efficient and accurate diagnoses. This paper provides an overview of the use of dimensionality reduction techniques, including principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE), in the diagnosis of heart diseases in South Africa. The paper also highlights the importance of considering the interpretability of the results, as well as potential biases in the data and algorithms, when selecting a technique. The purpose of this study is to predict the accuracy of heart disease using Dimensionality technique to determine if there was any enhance in predicting accuracy. Although the SVM show the better accuracy score of 71% over Random Forest with the score of 61% when PCA model is applied the use of dimensional reduction doesn't produce better results.

Keywords: Dimensional reduction, machine learning, heart diseases data mining algorithm, (PCA) Principal Component Analysis, SVM

1 Introduction

Heart related disease remains amongst the major cause of death throughout the work from the past decade [1]. Cholesterol level, high blood pressure and diabetics are amongst the main contributing factors related to heart diseases. Amongst the risk factors associated with heart diseases can be traced back to family medical history, smoking and drinking patterns of a patient and port health diet.

[2] data mining assist in the extraction of patterns that are found in the process of knowledge discovery within the databases in which intelligent methodologies are applied. The emerging discoveries within the data mining domain that promise to provide intelligent tolls and new technologies which will assist the human to do analytics and gain understanding huge volumes of data remains a challenge with unsolved problems. The current functions within data mining domain includes classifications, clustering, regression, and the discovery of association rules, rule generation, summarization, sequence analysis and dependency modelling.

Numerous data mining challenges can be addressed effectively though the use of soft computing techniques. Amongst this technique are neural network, generic algorithms, fuzzy logic that will ultimately leads an intelligent very interpretable and solutions that are low cost compared to traditional techniques. Dimension reduction is regarded amongst the mostly used method for data mining for the extraction of patterns in a more reliable and intelligent manner and has widely been used to find models that data relationships [3]. The rest of the paper is organized as follows. In Section 2, some background on dimensionality Reduction Algorithms are discussed. In Section 3, the methodology and the simulation results obtained are presented. Finally, Section 4 concludes the paper.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

2 Background

The dimensionality reduction can be regarded as an unsupervised learning technique, which may be applied as pre-processing step for data transformation towards machine learning algorithms on both regression predictive modelling and classification datasets with supervised learning algorithms. Furthermore, dimensionality reduction represents a technique which may be adopted for dropping the amount of input variables in training data. Every time one is dealing with a large capacity of dimensional data, it is normally valuable in reducing the dimensionality through the projection of the data to a lower dimensional subspace which captures the “crux” of the data.[4]. Heart disease is a serious health issue everywhere, including South Africa. For successful treatment and better patient outcomes, it is essential to make a timely and correct diagnosis of heart disease. The complexity of heart illness and the volume of data produced by contemporary medical technologies, however, can make a precise diagnosis of the condition difficult.

Machine learning uses the method of "dimensionality reduction" to lower the number of variables in a dataset while preserving the most important data. This facilitates the discovery of patterns and connections in the data, improving analysis and diagnosis. Principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding are three examples of dimensionality reduction techniques that have been created and used in a range of fields.

2.1 Algorithms Used in Dimensionality Reduction

Several algorithms are applicable to be applied for dimensionality reduction. The two key classes of methods are those selected from linear algebra and those selected from diverse learning [5]

Linear Algebra Methods

The role Matrix factorization methodology is chooses from the area of linear algebra which can be utilized for dimensionality.

Manifold Learning Methods

The fundamental role of Manifold learning methods is simply to pursue a dimensional projection which is lower than that of high dimensional input which captures the visible properties of the input data.

Some of the popular and most familiar methods includes :

- The Embedding Isomap
- The Embedding Locally Linear
- The Scaling Multidimensional
- The Embedding Spectral
- The Embedding t-distributed Stochastic Neighbor

Each one of these algorithms suggests an approach which is very diverse towards the challenge of determining natural relationships in data at lower dimensions.

Currently there is no dimensionality reduction algorithm which can be regarded as the best , and no easy way to detect if the algorithms of the best for one’s data without applying the controlled experiments.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

2.3 Principal Component Analysis

The application and the use of Principal Component Analysis (PCA) might be regarded amongst the most prevalent method for dimensionality reduction with the inclusion of dense data (i.e. few zero values). T h e (P C A) Principal Component Analysis , may be further be viewed as a technique seeking to decrease the data dimensionality.
The (PCA) Principal Component Analysis (PCA) technique can be defined and implemented by means of the tools of linear algebra. The (PCA) Principal Component defines a process that is functional to a dataset, which is normally represented by an n x m matrix A ,whereby the results in a projection of A will be the final outcomes [6].

Correlation

Correlation is defined as a quantitative analysis that seeks to calculate the strong point of connotation within one or two variables and calculate i f there i s any direction establishment concerning the relationship. Classically, within the domain of statistics, there are 4 types of measurable correlations i.e. spearman correlation, kendall rank correlation pearson correlation and the point-biserial correlation [7]

Pearson r correlation

The Pearson *r* correlation may be viewed as amongst the furthermost broadly applied statistic tool for correlation in order to calculate the gradation of the relationship between linear associated variables. A typical example can be that of a stock market, if one desires to quantify how can two more stocks get to be correlated to one another, Pearson *r* correlation can s i m p l y b e applied to quantify the notch of the two relationship.

The below formula can be applied in calculating the correlation for Pearson *r* :

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad (1)$$

r_{xy} = signifies the coefficient between x and y within the correlation Pearson r
n = deals with the amount of observations
x_i = simply outline the value of x (for the certain number of observations)
y_i = simply outlines the value of y (for the certain number of observations)

Kendall rank correlation

The correlation of Kendall rank may be deemed as non-parametric test that can be utilized to calculate the strength of dependency within one or two variables. " If one considers models y and z, whereby each model size is *n*, then one becomes cognizant to the fact that the complete amount of pairings with y z is *n*(*n*-1)/2." [8]. The below formula can be applied to measure the correlation of Kendall rank 's value:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \qquad (2)$$

Nc= represents the quantity of concordant
Nd= represent the quantity of discordant

***Spearman rank correlation*:**

Based on (8) the correlation of Spearman rank may be defined as a non-parametric test which can be applied to calculate the gradation of the connotation amongst two or variables. " The tests conducted for Spearman rank doesn't consist of any assumptions relating to the data distribution and is the suitable correlation analysis whenever the variables are calculated within a scale that is at slightest ordinal variable" (8).

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

The below formula can be applied to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ= signifies the correlation Spearman rank
di= the variance between the ranks of corresponding variables
n= outlines the amount of correlation observations

3. METHODOLOGY

The experimental findings of our model categorization are discussed in this section.

Experimental Environment

Python and Google Colab were used to implement the experimental findings, and an Intel (R) Core i7 CPU with 8 GB of memory was also used.

Data set

a retroactive sample of men in the Western Cape, South Africa, a heart disease high-risk area. Per case of CHD, there are around two controls. After their CHD occurrence, several of the men who tested positive for CHD had blood pressure lowering therapy and other initiatives to lower their risk factors. In certain instances, measurements were taken following these procedures. The larger dataset from which these statistics are drawn is given in Rousseauw et al 1983 .'s South African Medical Journal article. Now the dateset are on Kaggle as open-source dataset:

3.1. Data description

Sbp	Systolic blood pressure
Tobacco	Cumulative
Ldl	Low density lipoprotein Cholesterol
adiposity	This is an increasing overweight which might be associated with a growing risk for diseases
Famhist	Family history
typea	Type-A behavior
Obesity	the state of being overweight
Alcohol	current alcohol consumption
Age	Age at one set
Chd	Coronary heart disease

Table 1: Data description

3.1.1. Data types

“Sbp” int64
“Tobacco” object
“ldl” object
“Adiposity” object
“Typea” int64
“Obesity” object
“Alcohol” object
“Age” int64
“dtype:” object

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.1.2 Family history

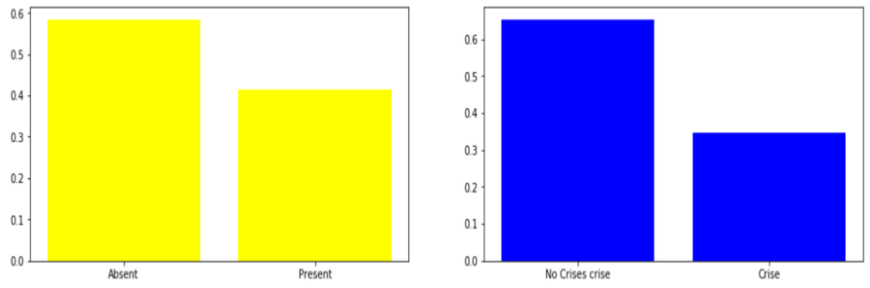


Figure 1. family history of heart diseases

Absent 0.58
Present 0.42
Name: famhist, dtype: float64
0 0.65
1 0.35
Name: chd, dtype: float64

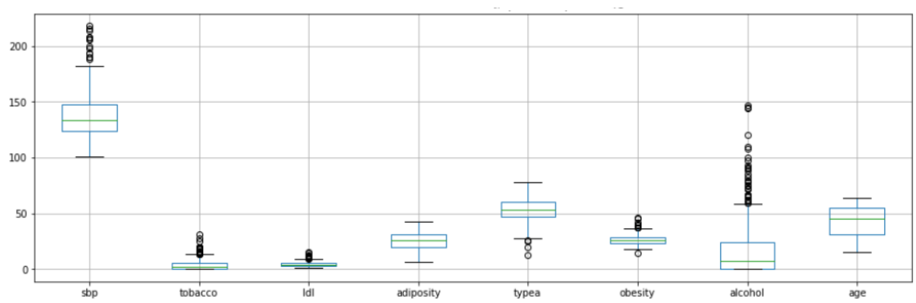


Figure 2: Distribution of the values of potential predictors

In this instance we have observed a large-Scale difference with various variables. These variables will need to be standardize in order avoid those with greater scales been wrongly having too much weight in the calculations. With Regards to aberrant values, their effect will need to be reduced through methods that are not very sensitive towards them.

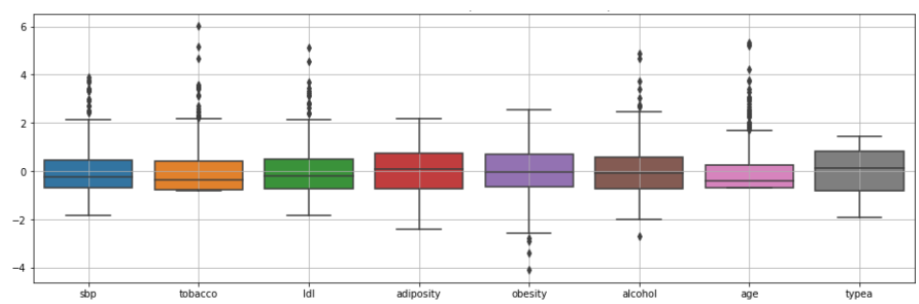


Figure 3: Distribution of the values of all potential standardized predictors

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.2 Dimensionality Reduction

3.2.1 Correlation:

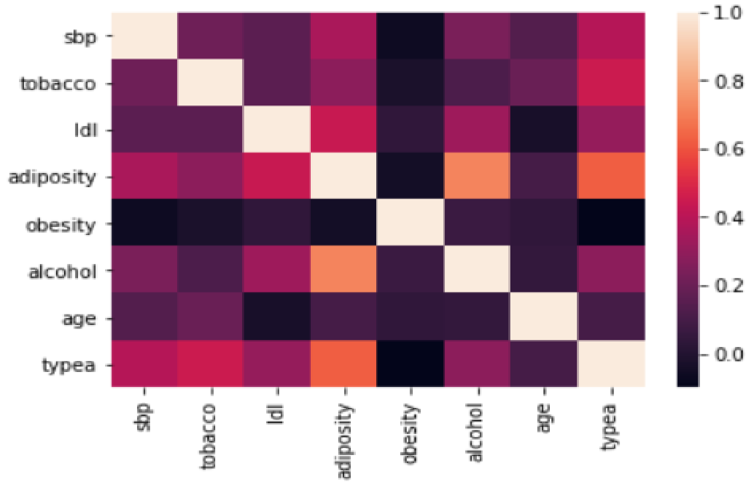


Figure 4: The correlation between various variant

Interpretation

Our observation of the above correlation matrix suggests the following about the coefficient:

- Age has some form of correlation with the consumption of tobacco and level of adiposity.
- There is a strongly correlation between obesity and adiposity;
- Idl is correlated with adiposity.

In summary, the older and obese subjects incline to have additional fats which might be accumulated under the skin.

3.2.2 Bartlett's sphericity test

We conducted the sphericity test and the outcomes value for p was: **True.**
Conclusion of the sphericity test:

Hypothesis: Orthogonality of the variables
Since the p-value is less than the selected threshold, the null hypothesis of orthogonality of variables is rejected.

Therefore: PCA is much relevant within the meaning of this test

3.2.3 Sampling Adequacy for Kaiser’s Measure

We further calculated the Sampling adequacy Kaiser’s measure of (KMO) and the results were kmo: 0.67

Therefore, since the kmo index is between 0.6 and 0.7, a compression relevant index is very possible to be attained. Since some of the values seemed to remain great, we will try a less sensitive approach towards them,

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

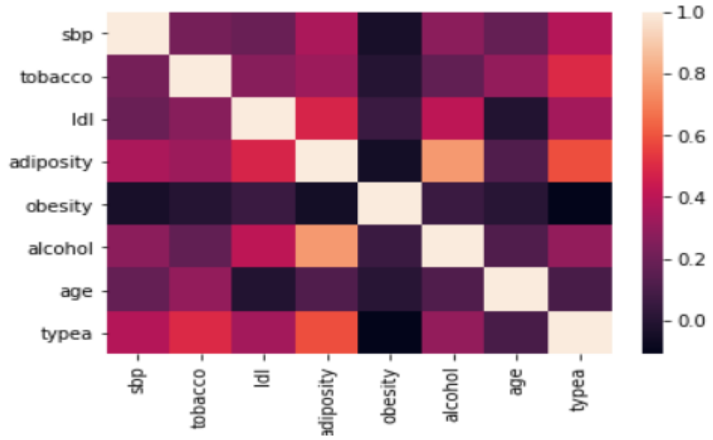


Figure 5: Correlation matrix (Spearman)

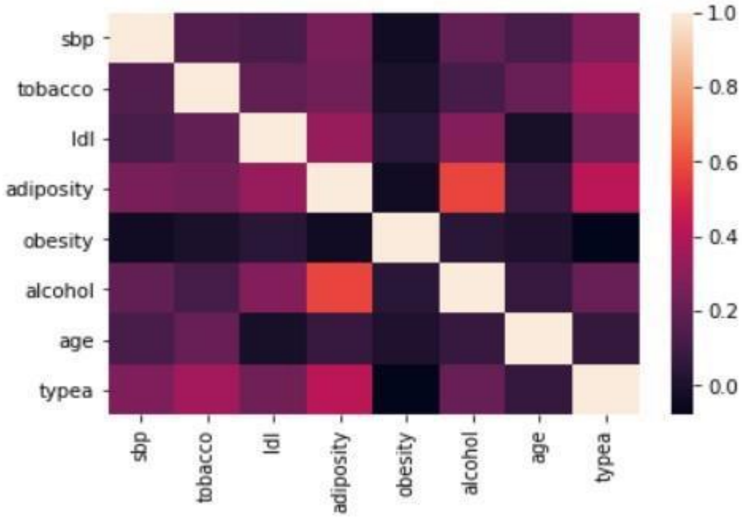


Figure 6: Correlation matrix (Kendall)

A non-parametric PCA based on the Spearman correlation matrix is performed in due to the presence of many atypical values.

Principal Component Analysis (PCA)

PCA is regarded as a model which is applied in covariance structure extended in prepared and established components which carries a declining variance [9]. The inherent data variability is then captured through the features from the linear extraction which are from the novel data sets [10] [11]. The (PCA) Principal components Analysis is normally founded on correlations are normally gets resolute through the application of a mean centered data.

The data is then gets spit into both training and testing, and the shape for training data was (346, 7) and (116, 7) for testing data

The We calculated the variance ration, variance ration cumsum and variance ration sum

[0.36972103 0.16203157 0.15050121 0.1108632 0.10189052]
[0.36972103 0.5317526 0.68225381 0.793117 0.89500752]
0.8950075163890064

The shared inertia explained by the first 5 components is around 90%

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

3.3 Predictive analysis

In this section, we went ahead and trained two applied classifiers i.e. random forest (RF) and the Support Vector Machine (SVM). Firstly, on main components, then secondly on the initial data, afterwards we compared their performance by assessing any the contribution in terms of the reduction in size for the decrement of the noise within the predictive model's construction.

3.3.1 The principal component for Random Forest

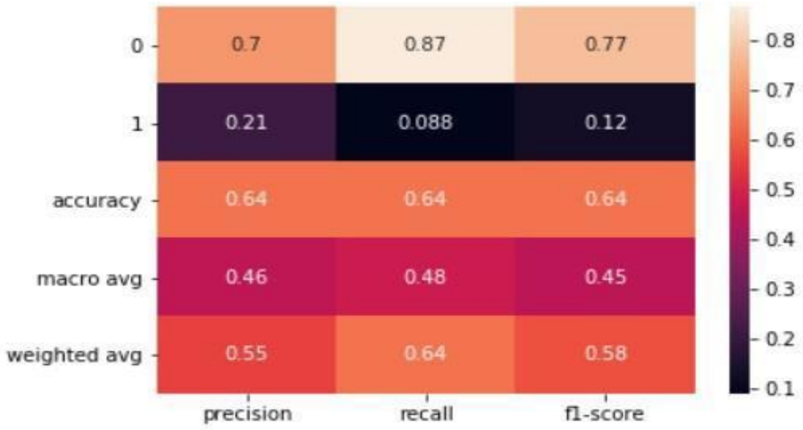


Figure 7: Random Forest Principle component

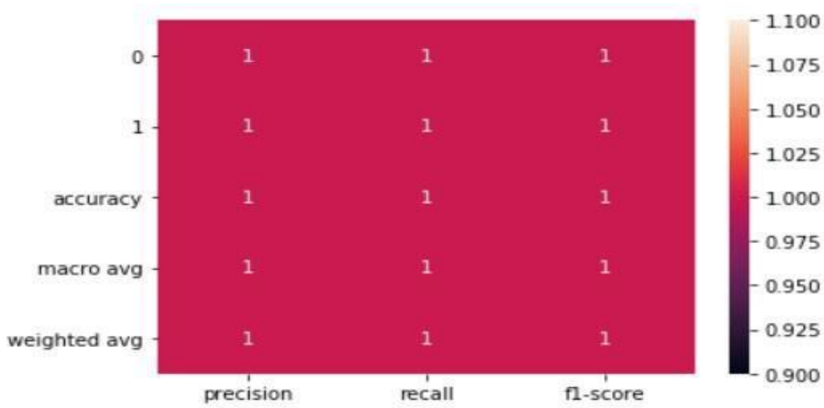


Figure 8: random forest initial Data

3.3.2 The Principal Component for SVM

The principal component for SVM is a technique used in machine learning to reduce the number of features in a dataset by projecting the data onto a new set of orthogonal axes. The goal of this technique is to identify the most important variables in a dataset that capture the majority of the variability in the data. By reducing the number of features, the SVM model can be trained more efficiently and effectively. This results in a more concise and interpretable model that can provide better predictions. In essence, the principal component for SVM is used to maximize the information retention in the reduced dataset while minimizing the loss of information.

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

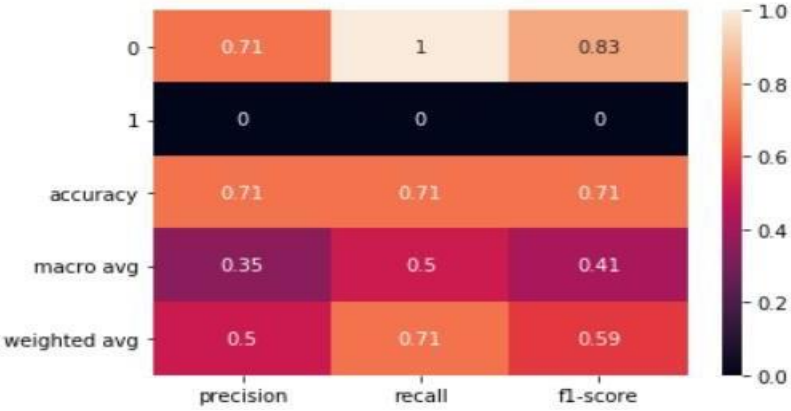


Figure 9. SVM principle component

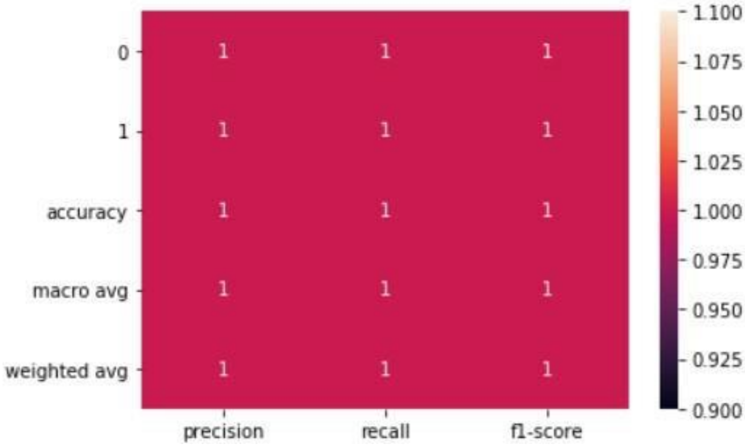


Figure 10: SVM Initial Data

4. Conclusion

In conclusion, it's important to keep in mind that while SVM showed a better accuracy score of 71% compared to Random Forest's score of 61% when PCA was applied, this doesn't always guarantee improved results. Dimensional reduction is just one tool in the machine learning toolkit and there may be other methods and techniques that could lead to even better results. It's essential to consider the limitations of using PCA and to explore other techniques, such as rotating the factorial axis, to achieve the best outcomes. Additionally, it's important to note that accuracy is not the only evaluation metric to consider. Other metrics, such as precision, recall, F1 score, and receiver operating characteristic (ROC) curve, can provide a more comprehensive understanding of a model's performance. Ultimately, it's crucial to choose the best model for the task at hand based on a thorough evaluation of multiple metrics and techniques..

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

References

1. Asha Gowda Karegowda, M.A.Jayaram and A S .Manjunath. Article:Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning. International Journal of Computer Applications 23(2):1–10, (June 2011)

2. Bonett, D. G.. Meta-analytic interval estimation for bivariate correlations. Psychological Methods, 13(3), 173–181(2008). https://doi.org/10.1037/a0012868

3. Coffman, D. L., Maydeu-Olivares, A., & Arnau, J . Asymptotic distribution free interval estimation: For an intraclass correlation coefficient with applications to longitudinal data. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4(1), 4–9(2008) . https://doi.org/10.1027/1614-2241.4.1.4

4. Shan Xu, Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework” (2017) IEEE 2nd International Conference on Big Data Analysis

6. J. Padmavathi . A Comparative Study on Logistic Regression Model and PCA-Logistic Regression Model in Medical Diagnosis. International Journal of Engineering and Technology. 2 - 8, 1372-1378. (2012)

7. Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart Disease Using Hybrid Technique for Selecting Features”, 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).(2017).

8. Swati Shilaskar and Ashok Ghatol. Article: Dimensionality Reduction Techniques for Improved Diagnosis of Heart Disease. International Journal of Computer Applications 61(5):1-8, January (201)3.

9. Jonathon Shlens :A Tutorial on Principal Component Analysis April 22 . Version 3.01, tutorial available at http://www.snl.salk.edu/~shlens/pca.pdf. (2009).

10. R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, S. Speedie. Automatic methods to extract New York heart association classification from clinical notes IEEE Int Conf Bioinformat Biomed (BIBM) (2017). 10.1109/bibm.2017.8217848

11. G. Parthiban, S.K. Srivatsa: Applying machine learning methods in diagnosing heart disease for diabetic patients Int J Appl Inf Syst, 3 (7) (2012), pp. 25-30, 10.5120/ijais12-450593

ISBN: 978-0-6397-0845-4 (e-book)
ISBN: 978-0-6397-0844-7 (print)

ISBN Electronic: 978-0-6397-0845-4 (Preprints)
ISBN Print: 978-0-6397-0844-7 (Preprints)