International Journal Software Engineering and Computer Science (IJSECS)
4 (2), 2024, 409-417
Published Online August 2024 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN : 2776-4869, E-ISSN : 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v4i2.2398.

RESEARCH ARTICLE Open Access

Design Principles for Enhancing AI-Assisted Moderation in Hate Speech Detection on Social Media Platforms

Alex Graf *
Drexul University, Philadelphia, United States.
Corresponding Email: alexgraf@nzqri.co.nz.

Danny Coolsaet
New Zealand Quality Research and Innovation (NZQIR), New Zealand.
Email: coolsaet@nzqri.co.nz.

Received: March 12, 2024; Accepted: July 10, 2024; Published: August 1, 2024.

Abstract: Hate speech on social media poses a growing threat to individuals and society, necessitating technological support for moderators in detecting and addressing problematic content. This article explores the design principles essential for creating effective user interfaces (UIs) in decision support systems that employ artificial intelligence (AI) to aid human moderators. Through a comprehensive study involving 641 participants across three design cycles, we qualitatively and quantitatively evaluate various design options. Our assessment encompasses perceived ease of use, usefulness, and intention to use, while also delving into the impact of AI explainability on users' cognitive efforts, informativeness perception, mental models, and trustworthiness. Notably, software developers affirm the high reusability of the proposed design principles. The findings reveal that well-designed UIs can significantly enhance the effectiveness of AI-based moderation tools, providing clear and understandable explanations that improve user trust and engagement. By addressing both technical and user-centered aspects, this research contributes to the development of more robust and user-friendly AI systems for hate speech detection. Future work should focus on further refining these principles and exploring their applicability in diverse social media contexts to ensure comprehensive and adaptable solutions for content moderation.

Keywords: Hate Speech Detection; Social Media Moderation; AI-based Decision Support; Explainable AI (XAI); Transfer Learning; User Interface Design.

1. Introduction

The ubiquity of social media platforms connecting users globally enables discussions on diverse topics, including politics, finance, and social issues. However, enforcing policies against undesirable content, such as hate speech, remains challenging. Traditionally, human moderators have been relied upon to investigate and review potentially offensive content. The emergence of AI-based decision support systems for hate speech detection has garnered attention. These systems leverage AI models to identify various forms of unwanted content, such as hate speech, racism, or offensive language. Social media platforms connect users worldwide, enabling them to exchange opinions on topics ranging from politics and finance to social issues [1][2]. However, enforcing consistent policies regarding undesirable content, such as hate speech, poses significant challenges [3][4]. Such content can potentially cause psychological harm to affected users [5]. Currently, social media platforms primarily rely on human moderators who investigate and review potentially offensive material [6][7]. Recently, AI-based decision support approaches have gained significant attention in the context of hate speech detection. AI models can identify various forms of unwanted content communicated through speech, including hate speech, racism, and offensive language [8][9]. Additionally, research projects explore decision support via software artifacts. For instance, these artifacts can visualize aggressive comments on a user's timeline [10][11], or treat hate speech as malware by quarantining it and notifying the targeted user [12][13][14].

Beyond researchers, institutions, and social media developers, large companies are also actively addressing hate speech. Intel Corporation, for instance, is developing an AI-based application that detects and redacts audio material based on user preferences, filtering out hate speech and similar content, such as racism or sexism [15]. However, the opacity of modern AI models—often referred to as the "black box" problem—remains a challenge [16][17]. These state-of-the-art models provide powerful predictions but lack transparency. Despite the promising capabilities of AI models, their opaqueness hinders transparency and user understanding of decision-making processes. Explainable AI (XAI) aims to address this issue by introducing transparent models and explanation-generation techniques. Modern AI-based decision support systems can provide robust decision support while offering explanations through user interfaces (UIs). Effective UI design can visualize and support interaction with the inner workings of algorithms, enhancing user comprehension.

Moreover, XAI plays a crucial role in monitoring system fairness, transparency, and maintenance. Large companies, including Intel Corporation, are actively engaged in developing AI-based applications to filter hate speech and similar content. However, the black box nature of AI models raises concerns regarding transparency. This research addresses the lack of user evaluation studies in the XAI field, specifically focusing on the perception and effects of explanations on stakeholders. Two interconnected research gaps are identified: a lack of applicable UI design knowledge in the hate speech domain and insufficient focus on users' and decision-makers' evaluations and perceptions of XAI-based explanations. To bridge these gaps, the research poses two key questions:

1) What are the essential design principles for XAI-based UIs supporting moderators in detecting hateful content on social media platforms?
2) How are such UIs perceived by relevant stakeholders, and what is the influence of local explanations?

This research employs a design science research (DSR) approach with three design cycles to answer these questions comprehensively. The proposed design principles are evaluated qualitatively and quantitatively, emphasizing reusability through practitioner assessments. The article concludes with discussions on results, theoretical implications, limitations, and future research opportunities. Social media platforms play a pivotal role in today's digitized world, connecting users across the globe [11]. The data generated on these platforms fuel data analytics and hold immense value for companies, institutions, and individuals [18][19][20]. However, social media platforms also harbor risks. Notably, scholars have underscored their role in disseminating hate speech [11].

Hate speech poses significant harm to individuals and societies, and it even threatens the very fabric of social media platforms [11][21][22][23]. The United Nations (2019) defines hate speech as follows: "[…] any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language concerning a person or a group based on who they are—specifically, their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factors" [24]. Unfortunately, such content is far from rare on social media platforms and in the broader digital sphere. A survey conducted in the US revealed that 37% of surveyed students (aged 12 to 17 years) had personally experienced hate speech [25]. Frequent exposure to hate speech can significantly impact users emotionally and behaviorally [26]. Disturbingly, hate speech has been linked to emotional harm and even suicide among young individuals [27][28]. Many social

media users have reported encountering and being affected by hate based on factors such as ethnicity, religion, politics, or gender [11]. However, users can also play a crucial role in combating hate speech. By reporting offensive content and engaging in counter-speech, they contribute to reducing harmful material online [22].

1.1 Decision Making for Hate Speech Detection on Social Media Platforms

The automated detection of hate speech using AI-based models has garnered significant attention in research, as evidenced by a growing body of scientific literature [29][21][30]. However, despite this interest, there remains a dearth of research on the design of Decision Support Systems (DSSs) that cater to both end users and human moderators. While scholars have primarily focused on end users (i.e., social media users), there is a need to address the challenges faced by human moderators as well. For instance, Modha *et al*. (2020) experimented with a software artifact—a web-browser plugin—based on deep learning techniques. This plugin visualizes various nuances of aggressiveness within a user's timeline. Although it serves as decision support for end users, human moderators still need to manually screen the content. Unfortunately, the evaluation of this web-browser plugin was limited to technical metrics from the field of machine learning-based AI, without considering end-user feedback or perceptions [30][31][23]. Consequently, the impact of its design on end users remains unclear, especially when faced with classification results that lack transparency. In contrast, Paschalides *et al*. (2020) developed MANDOLA, a system that monitors, detects, visualizes, and reports the spread of hateful content online. MANDOLA provides visualizations to users, allowing them to explore detected hate speech using filters based on time, context, and location. Through this approach, users can identify correlations—for instance, between hate speech development and triggering events. The system's intriguing methodology bridges the gap between AI-driven detection and user-centric insights. As we delve deeper into designing effective DSSs for hate speech detection, understanding the interplay between AI models, visualizations, and end-user experiences becomes crucial. By addressing these challenges, we can create more transparent, user-friendly systems that empower both social media users and human moderators in combating hate speech.

1.2 Artificial Intelligence for Hate Speech Detection on Social Media Platforms

Artificial Intelligence (AI) has become a focal point in research, with its applications spanning various domains [27]. In our study, we define AI as machine-learning-based systems capable of interpreting external data, learning from it, and flexibly adapting to achieve specific goals [27]. Our focus lies specifically on text-based hate speech detection—an area frequently addressed using AI models [21][30][31][32]. These AI models can be categorized as either single methods or hybrids [29]. Single methods include models like logistic regression, representing machine learning, while more complex deep learning models, such as convolutional neural networks, fall into the hybrid category [21][30][32]. Hybrid methods combine different machine learning or deep learning models for tasks like text classification [32]. Beyond AI models, integrated data features play a crucial role in hate speech detection. For instance, incorporating users' psychological features into the input for machine learning models enables the detection of related concepts, including cyberbullying. Plaza-del-Arco *et al*. (2021) conducted a study comparing multilingual and monolingual pre-trained language models with traditional machine learning models. Their findings highlighted the superiority of transfer learning. However, most AI-based hate speech detection systems remain black boxes. Efforts have been made to address this opacity in state-of-the-art AI models [29]. This research embarks on designing user interfaces (UIs) for AI-based Decision Support Systems (DSSs) focused on hate speech detection within social media platforms. Through iterative design cycles, guided by the principles of Design Science Research (DSR), the study reveals the significance of transparent and user-centric AI systems. By continuing to bridge the gap between AI and human expertise, we can create safer online spaces for all.

2. Research Method

In our study, we focused on developing practical Design Principles (DPs) for user interfaces (UIs) within AI-based Decision Support Systems (DSSs) intended for human moderators of social media platforms. Our approach was grounded in the Design Science Research (DSR) methodology, which enables scholars to create knowledge that is applicable in real-world settings [29][25]. These DPs serve as nascent design theories, representing operational principles [2]. The instantiated DPs were embedded within UIs of varying maturity levels, which we evaluated through both qualitative and quantitative methods across three consecutive design cycles. During these evaluations, we provided introductory materials on AI-based DSSs for hate speech

detection. Participants engaged with multiple exemplary hate speech cases and completed surveys. Notably, for practitioners' evaluation of the DPs, we deliberately withheld any specific UI examples. Our investigations revealed opportunities for optimization, assessed users' positive perceptions of the UI, explored the impact of local explanations, and examined the reusability of DPs in practical settings. We adhered to the DSR methodology proposed by Peffers *et al*. (2007). The initial activity—problem identification and motivation—was discussed in the preceding sections. Subsequent activities, including defining objectives, design and development, demonstration, evaluation, and communication of results, will be detailed in the following sections. Figure 1 provides an overview of our DSR project, followed by a summary of each individual design cycle [29].

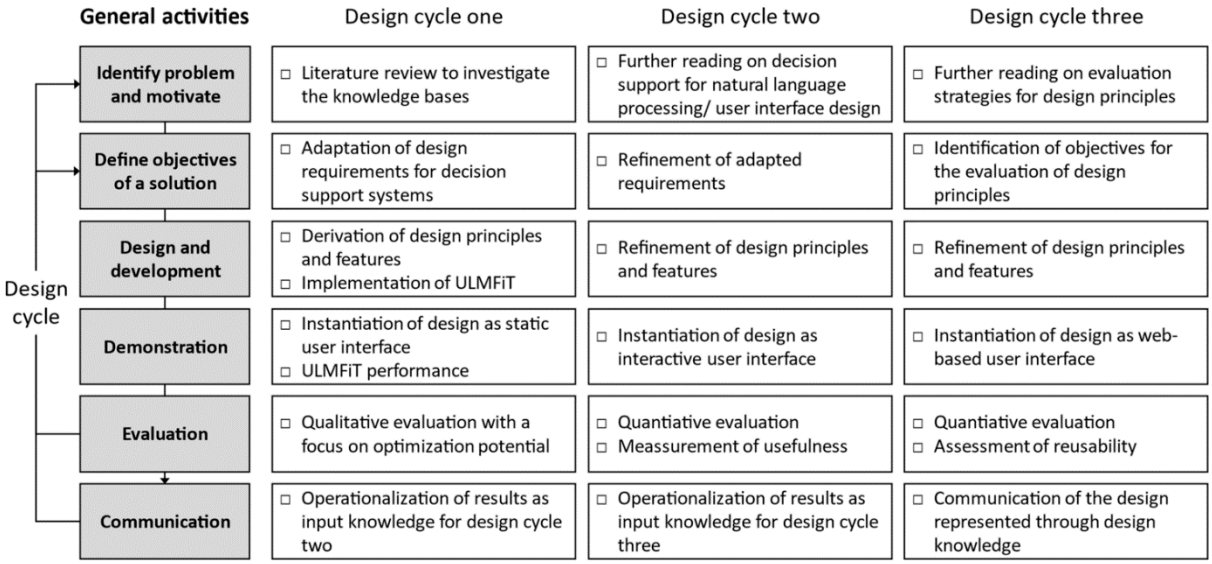

Figure 1. The design science research process adapted from Peffers *et al*. (2007)

Our design science research (DSR) project commenced with a comprehensive literature review. Through this exploration, we identified critical gaps in existing research related to human moderators and their pivotal role in AI-based hate speech detection. Furthermore, we recognized a dearth of prescriptive design knowledge specifically tailored for user interfaces (UIs) within the hate speech detection domain. To address these challenges, we first established generic requirements for Decision Support Systems (DSSs), drawing inspiration from prior work [12]. Subsequently, we meticulously crafted Design Principles (DPs) and Design Features (DFs). Our implementation leveraged the Universal Language Model Fine-Tuning (ULMFiT), a transfer learning model [25]. The second design cycle involved a deep dive into DSSs for natural language processing tasks, coupled with extensive desk research on UI design. Armed with insights from the initial evaluation, we refined our Design Requirements (DRs), DPs, and DFs. The revised design materialized as an interactive UI, which we subjected to quantitative evaluation through an experiment involving 190 participants recruited via CloudResearch and Amazon Mechanical Turk (MTurk) [17] By integrating our prototype into a web survey using Adobe XD, we aimed to assess users' perceived usefulness [24]. The third design cycle focused on strategies for evaluating our DPs. We further refined the artifact and implemented it in a production-ready environment—a web-based UI. Our overarching goals encompassed two key aspects: (i) assessing the impact of explainability (specifically, local explanations) on constructs such as perceived cognitive effort, informativeness, mental model, and trustworthiness through an experiment involving 360 participants [30][27][28]; and (ii) evaluating the reusability of our DPs.

3. Result and Discussion

3.1 Results

In this section, we delve into the derivation process for our Design Principles (DPs), which were subsequently translated into Design Features (DFs) [3][5][6][11][2]. Both DPs and DFs reside within the solution space and must effectively address the Design Requirements (DRs) originating from the problem space [1]. DPs serve as a means to communicate design knowledge in an accessible format [3], while DFs represent specific operational components that can be implemented in a prototype artifact [27]. Furthermore,

both DPs and DFs contribute to the overall structure of an Engineering Design Theory (EDT), elucidating how specific DRs can be met [1]. Our first DP centers around transfer learning, a well-established approach in text classification [27]. Given that social media platforms serve as rich repositories of textual data and versatile information [23], the availability of data on these platforms can significantly vary [24]. Transfer learning, even with as few as 100 labeled examples, can achieve state-of-the-art performance [25]. This becomes particularly crucial when addressing social media-specific challenges, such as data scarcity for training AI models [23]. Moreover, the synergy between human intelligence and AI approaches can lead to impactful decisions [3]. Hence, we establish: DP1: Provide the system with transfer learning techniques for classifying unstructured data, enabling users to make informed decisions based on the decision support provided. Users rightfully expect high-quality advice. The second DP focuses on enhancing explainability to address the black box problem inherent in many AI models. Modern AI models, while powerful, often lack transparency, making it difficult for users to understand how decisions are made [27]. Explainable AI (XAI) techniques are crucial in justifying individual outcomes and detecting potential errors or biases [27]. XAI serves to increase user trust and ensure that AI systems are accountable. Therefore, we propose: DP2: Integrate Explainable AI techniques to provide transparency in decision-making processes, thereby increasing user trust and enabling the detection of errors or biases. Our third DP addresses user interaction with the system. Effective user interface (UI) design is essential for facilitating interaction between users and AI models. By visualizing the inner workings of algorithms, UIs can help users understand and engage with the system more effectively [3]. This includes the use of clear visual aids and interactive elements that make complex data more accessible [27]. Consequently, we propose: DP3: Design user interfaces that effectively visualize algorithmic processes and support user interaction, making complex data more accessible and comprehensible. The fourth DP emphasizes the importance of context-specific customization. Social media platforms and their user bases are diverse, with varying needs and preferences. Customizable interfaces allow the system to adapt to different contexts, improving usability and relevance [3]. This involves allowing users to modify settings and features to better suit their specific requirements. Thus, we establish: DP4: Develop customizable interfaces that adapt to the specific needs and preferences of different user groups, enhancing usability and relevance. Finally, our fifth DP highlights the need for continuous feedback and improvement. AI systems should not be static; they must evolve based on user feedback and emerging trends in data [3]. This requires mechanisms for collecting user feedback and incorporating it into regular updates and improvements to the system [27]. Therefore, we propose: DP5: Implement mechanisms for continuous feedback and iterative improvement, ensuring that the system evolves with user needs and emerging data trends. These five DPs form the foundation of our approach to developing effective and user-friendly AI-based Decision Support Systems for hate speech detection on social media platforms. By addressing key challenges such as data scarcity, explainability, user interaction, customization, and continuous improvement, these principles aim to create a robust framework for designing UIs that support human moderators in managing online content effectively.

3.2 Discussion

The primary focus was on developing practical Design Principles (DPs) for user interfaces (UIs) within AI-based Decision Support Systems (DSSs) intended for human moderators of social media platforms. Our approach was grounded in the Design Science Research (DSR) methodology, which enables scholars to create knowledge that is applicable in real-world scenarios [1][2][3]. These principles serve as nascent design theories, representing operational principles [2]. The instantiated DPs were embedded within UIs of varying maturity levels, which we evaluated through both qualitative and quantitative methods across three consecutive design cycles. During these evaluations, we provided introductory materials on AI-based DSSs for hate speech detection. Participants engaged with multiple exemplary hate speech cases and completed surveys. Notably, for practitioners' evaluation of the DPs, we deliberately withheld any specific UI examples. Our investigations revealed opportunities for optimization, assessed users' positive perceptions of the UI, explored the impact of local explanations, and examined the reusability of DPs in practical settings. We adhered to the DSR methodology proposed by Peffers *et al.* (2007). The initial activity—problem identification and motivation—was discussed in the preceding sections. Subsequent activities, including defining objectives, design and development, demonstration, evaluation, and communication of results, are detailed in the following sections. Figure 1 provides an overview of our DSR project, followed by a summary of each individual design cycle. Our DSR project commenced with a comprehensive literature review. Through this exploration, we identified critical gaps in existing research related to human moderators and their pivotal role in AI-based hate speech detection. Furthermore, we recognized a lack of prescriptive design knowledge specifically tailored for UIs within the hate speech detection domain. To address these challenges, we first established generic requirements for DSSs, drawing inspiration from prior work [5]. Subsequently, we meticulously crafted DPs

and Design Features (DFs). Our implementation leveraged the Universal Language Model Fine-Tuning (ULMFiT), a transfer learning model [19][21]. The second design cycle involved an in-depth examination of DSSs for natural language processing tasks, coupled with extensive desk research on UI design. Armed with insights from the initial evaluation, we refined our Design Requirements (DRs), DPs, and DFs. The revised design materialized as an interactive UI, which we subjected to quantitative evaluation through an experiment involving 190 participants recruited via CloudResearch [18] and Amazon Mechanical Turk (MTurk). By integrating our prototype into a web survey using Adobe XD, we aimed to assess users' perceived usefulness [25]. The third design cycle focused on strategies for evaluating our DPs. We further refined the artifact and implemented it in a production-ready environment—a web-based UI. Our overarching goals encompassed two key aspects: (i) assessing the impact of explainability (specifically, local explanations) on constructs such as perceived cognitive effort, informativeness, mental model, and trustworthiness [27] through an experiment involving 360 participants; and (ii) evaluating the reusability of our DPs [30]. Through these three design cycles, we successfully developed and tested design principles that can enhance the effectiveness of AI-based decision support systems in detecting hate speech on social media platforms. Thus, our research provides both theoretical contributions and practical implications for better DSS design and implementation.

4. Related Work

The proliferation of social media platforms has significantly transformed communication, enabling users to connect globally and engage in discussions on various topics, including politics, finance, and social issues [13]. However, these platforms also serve as venues for the dissemination of undesirable content such as hate speech, which poses significant challenges for moderation and enforcement of content policies [1][15]. Traditional methods of content moderation rely heavily on human moderators who manually review and assess potentially offensive material. This approach, while effective to an extent, is labor-intensive and may not scale well given the volume of content generated on these platforms [8][7]. In response to these challenges, there has been a growing interest in leveraging artificial intelligence (AI) to automate the detection and moderation of hate speech. AI-based decision support systems (DSSs) have been developed to identify various forms of unwanted content, including hate speech, racism, and offensive language [14][19]. These systems utilize machine learning models trained on large datasets to detect patterns and features indicative of hate speech. Notable research in this area includes the work by Fortuna and Nunes (2018), which explores machine learning approaches for hate speech detection, highlighting the potential and limitations of these techniques [22].

Despite the advancements in AI-based hate speech detection, several challenges remain. One major issue is the "black box" nature of modern AI models, which often lack transparency and explainability [20]. This opacity makes it difficult for users to understand how decisions are made, which can undermine trust in the system. To address this, researchers have proposed the use of Explainable AI (XAI) techniques that aim to make AI models more interpretable and their decisions more transparent [23][26]. Arrieta *et al.* (2020) provide a comprehensive review of XAI techniques and their applications in various domains, including hate speech detection [24].

Another significant body of work focuses on the design and evaluation of decision support systems (DSSs) that incorporate AI for hate speech detection. Paschalides *et al.* (2020) developed MANDOLA, a system that monitors, detects, visualizes, and reports the spread of hateful content online. MANDOLA provides users with visualizations that help them explore detected hate speech using filters based on time, context, and location [28]. This approach bridges the gap between AI-driven detection and user-centric insights, emphasizing the importance of user interface (UI) design in enhancing the effectiveness of DSSs. The integration of transfer learning techniques has also been explored to improve the performance of hate speech detection models. Transfer learning allows models to leverage knowledge from related tasks or domains, thus enhancing their ability to generalize from limited training data [30]. Howard and Ruder (2018) demonstrated the efficacy of transfer learning in text classification tasks, achieving state-of-the-art performance even with relatively small datasets [31]. While substantial progress has been made in the development of AI-based DSSs for hate speech detection, there is still a need for comprehensive frameworks that guide the design of these systems. Existing research often focuses on either the technical aspects of model development or the user interface design in isolation. Our study addresses this gap by proposing a set of design principles that integrate both technical and user-centered considerations, thereby providing a holistic approach to the development of effective AI-based DSSs for hate speech detection on social media platforms.

5. Conclusion

In this research, we embarked on designing user interfaces (UIs) for AI-based Decision Support Systems (DSSs) focused on hate speech detection within social media platforms. Our journey unfolded through iterative design cycles, guided by the principles of Design Science Research (DSR). Let's recap our key findings and contributions. We derived a set of Design Principles (DPs) that emphasize transfer learning, explainability, and user-centric features. These principles serve as actionable guidelines for creating effective UIs in hate speech detection systems. Leveraging transfer learning techniques, we fine-tuned our model for hate speech detection. The prototype UI incorporated local explanations, confidence scores, and relevant case histories. Through experiments and surveys, we assessed user perceptions, cognitive effort, and trustworthiness. Our UI aimed to empower human moderators while maintaining transparency.

As we look ahead, several avenues for future research and development emerge. Enhanced explainability: dive deeper into explainability techniques. Explore novel ways to present local explanations, ensuring that users comprehend AI decisions. Multimodal interfaces: extend our UI to handle not only text but also other forms of media (images, videos, etc.). Multimodal interfaces can enhance hate speech detection accuracy. Dynamic adaptation: investigate adaptive UIs that tailor their presentation based on user preferences, context, and evolving hate speech patterns. Ethical considerations: delve into the ethical implications of AI-driven decision support. Address bias, fairness, and privacy concerns. Human-AI collaboration: explore hybrid systems where AI assists human moderators in real-time decision-making. How can we strike the right balance?. In summary, our journey has laid the groundwork for more effective, transparent, and user-friendly hate speech detection systems. By continuing to bridge the gap between AI and human expertise, we can create safer online spaces for all.

References

[1] Lampropoulos, G., Ferdig, R. E., & Kaplan-Rakowski, R. (2023). A social media data analysis of general and educational use of ChatGPT: Understanding emotional educators. *Available at SSRN 4468181*. https://dx.doi.org/10.2139/ssrn.4468181

[2] Asghar, M. Z., Barbera, E., Rasool, S. F., Seitamaa-Hakkarainen, P., & Mohelská, H. (2023). Adoption of social media-based knowledge-sharing behaviour and authentic leadership development: evidence from the educational sector of Pakistan during COVID-19. *Journal of Knowledge Management*, *27*(1), 59-83. https://doi.org/10.1108/JKM-11-2021-0892

[3] Abdul-Rahman, M., Adegoriola, M. I., McWilson, W. K., Soyinka, O., & Adenle, Y. A. (2023). Novel use of social media big data and artificial intelligence for community resilience assessment (CRA) in university towns. *Sustainability*, *15*(2), 1295. https://doi.org/10.3390/su15021295.

[4] Hatmal, M. M. M., Al-Hatamleh, M. A., Olaimat, A. N., Mohamud, R., Fawaz, M., Kateeb, E. T., ... & Bindayna, K. M. (2022). Reported adverse effects and attitudes among Arab populations following COVID-19 vaccination: a large-scale multinational study implementing machine learning tools in predicting post-vaccination adverse effects based on predisposing factors. *Vaccines*, *10*(3), 366. https://doi.org/10.3390/vaccines10030366.

[5] Ahmed, A. L., & Ahmad, E. (2023). The future of New Zealand tertiary education: Generation alpha.

[6] Miller, D. (2023). Exploring the impact of artificial intelligence language model ChatGPT on the user experience. *International Journal of Technology, Innovation and Management (IJTIM)*, *3*(1), 1-8. https://doi.org/10.54489/ijtim.v3i1.195.

[7] Al-Sa'di, A., & Al-Samarraie, H. (2022). A Delphi Evaluation of User Interface Design Guidelines: The Case of Arabic. *Advances in Human-Computer Interaction*, *2022*(1), 5492230. https://doi.org/10.1155/2022/5492230.

[8] Ahmed, A. L., & Ahmad, E. (2023a). The future of New Zealand tertiary education: Generation alpha.

[9] Ahmed, Z., Bhinder, K. K., Tariq, A., Tahir, M. J., Mehmood, Q., Tabassum, M. S., ... & Yousaf, Z. (2022). Knowledge, attitude, and practice of artificial intelligence among doctors and medical students in Pakistan: A cross-sectional online survey. *Annals of Medicine and Surgery*, *76*, 103493. https://doi.org/10.1016/j.amsu.2022.103493

[10] Al-Sa'di, A., & Litayem, N. (2023, September). User Frustration: Shaping the Experience of Automotive and Computing. In *2023 IEEE 3rd International Conference on Computer Systems (ICCS)* (pp. 56-61). IEEE. https://doi.org/10.1109/ICCS59700.2023.10335584.

[11] Celik, S. (2019). Experiences of internet users regarding cyberhate. *Information Technology & People*, *32*(6), 1446-1471.

[12] Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., ... & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems*, *4*, 19-23. https://doi.org/10.1016/j.iotcps.2023.06.002.

[13] Al-S'di, A. (2022, September). User Centred design Methods in Animal Centred Design: A systematic review. In *2022 International Conference on Theoretical and Applied Computer Science and Engineering (ICTASCE)* (pp. 84-89). IEEE. https://doi.org/10.1109/ICTACSE50438.2022.10009842.

[14] Al-Sa'di, A., & Alsamarraie, H. (2023). Design Thinking Mindset and Self-Awareness of User Experience Practitioners: The Case of New Zealand. *Available at SSRN 4361224*.

[15] Al-Sa'di, A., & McPhee, C. C. A. (2021, August). User-Centred Design in Educational Applications: A systematic literature review. In *2021 International Conference Engineering Technologies and Computer Science (EnT)* (pp. 105-111). IEEE. https://doi.org/10.1109/EnT52731.2021.00025.

[16] von der Heyde, M., Goebel, M., Zoerner, D., & Lucke, U. (2023). Integrating AI tools with campus infrastructure to support the life cycle of study regulations. *Proceedings of European University*, *95*, 332-344.

[17] Litayem, N., & Al-Sa'di, A. (2023, September). Exploring the Programming Model, Security Vulnerabilities, and Usability of ESP8266 and ESP32 Platforms for IoT Development. In *2023 IEEE 3rd International Conference on Computer Systems (ICCS)* (pp. 150-157). IEEE. https://doi.org/10.1109/ICCS59700.2023.10335558.

[18] Yamjal, P., & Ahmed, A. (2022). Strategies for retention and completion in vocational education: Faculty perspectives. *Journal of Management and Business Education*, *5*(4), 247-265. https://doi.org/10.35564/jmbe.2022.0015.

[19] Kapil, P., & Ekbal, A. (2020). A deep neural network-based multi-task learning approach to hate speech detection. *Knowledge-Based Systems, 210*, 1–21. https://doi.org/10.1016/j.knosys.2020.106458

[20] Celik, S. (2019). Experiences of internet users regarding cyberhate. *Information Technology & People, 32*(6), 1446–1471. https://doi.org/10.1108/ITP-01-2018-0009

[21] Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 559, 1–12. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300789

[22] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

[23] FastAI. (2021, March 7). Transfer learning in text. fastai. Retrieved from https://docs.fast.ai/. Accessed 12 Jan 2022

[24] Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons, 63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

[25] Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems, 134*, 1–11. https://doi.org/10.1016/j.dss.2020.113302

[26] Kunst, M., Porten-Chee, P., Emmer, M., & Eilders, C. (2021). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Technology & Politics, 18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

[27] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

[28] Peng, S., Wang, Y., Liu, C., & Chen, Z. (2020). TL-NER: A Transfer Learning Model for Chinese Named Entity Recognition. *Information Systems Frontiers, 22*, 1291–1304. https://doi.org/10.10007/s10796-019-09932-y

[29] Pereira-Kohatsu, J. C., Quijano-Sanchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors, 19*(21), 1–37. https://doi.org/10.3390/s19214654

[30] Plaza-del-Arco, F., Molina-Gonzalez, M., Urena-Lopez, L., & Martin-Valdivia, M. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications, 166*, 1–10. https://doi.org/10.1016/j.eswa.2020.114120

[31] Sundberg, L., & Holmström, J. (2024). Teaching Tip: Using No-code AI to Teach Machine Learning in Higher Education. *Journal of Information Systems*.

[32] Pinzolits, R. (2024). AI in academia: An overview of selected tools and their areas of application. *MAP Education and Humanities*.

