Studies in Media, Journalism, and Communications
DOI: 10.32996/smjc
Journal Homepage: www.al-kindipublisher.com/index.php/smjc

SMJC
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Artificial Intelligence for Social Media Safety and Security: A Systematic Literature Review

Musawer Hakimi[1] ✉ Baryali Sazish[2], Mohammad Aziz Rastagari[3] and Amir Kror Shahidzay[4]
[12]Assistant Professor, Computer Science Department, Samangan University, Samangan, Afghanistan
[3]Dean of Computer Science Faculty, Tolo-e-Aftab University, Kabul, Afghanistan
[4]Dean of Computer Science Faculty, Kabul University, Kabul, University
Corresponding Author: Musawer Hakimi, **E-mail**: musawer@adc.edu.in

| ABSTRACT
The proliferation of social media platforms has revolutionized communication and connectivity, but it has also introduced new challenges related to safety and security. In response, researchers and practitioners have turned to artificial intelligence (AI) to develop innovative solutions for mitigating online risks. This systematic literature review explores the key applications, methodologies, benefits, limitations, ethical considerations, and future directions of AI in promoting social media safety and security. The review synthesizes findings from various scholarly articles spanning various disciplines, including computer science, engineering, and social sciences. The methodology involved searching academic databases such as PubMed, Scopus, IEEE Xplore, and Google Scholar using predefined search terms and inclusion criteria. The results reveal a diverse range of AI-driven approaches for addressing safety and security concerns on social media platforms, including enhanced threat detection, automated content moderation, and real-time response mechanisms. However, the deployment of AI in social media contexts also raises ethical challenges such as algorithm bias, privacy concerns, and lack of explain ability. The conclusion emphasizes the importance of ongoing research, collaboration, and ethical guidelines to maximize the benefits of AI while minimizing its potential risks. This review contributes to the growing body of literature on AI and social media by providing insights into current trends, challenges, and future directions in this rapidly evolving field.

| KEYWORDS

Artificial Intelligence, Social Media, Safety, Security, Literature Review.

| ARTICLE INFORMATION

ACCEPTED: 07 December 2023 **PUBLISHED:** 21 December 2023 **DOI:** 10.32996/smjc.2023.1.1.2x

1. Introduction

In today's digital landscape, the pervasive influence of social media platforms has revolutionized communication, connectivity, and information dissemination. However, the widespread adoption of social media has also given rise to a host of safety and security challenges, ranging from cyberbullying and identity theft to the spread of misinformation and online radicalization. Addressing these complex issues necessitates innovative solutions that harness the power of advanced technologies, and artificial intelligence (AI) has emerged as a promising tool in this endeavor. The intersection of AI and social media safety and security represents a burgeoning field of research, with scholars, policymakers, and industry practitioners actively exploring the opportunities and challenges inherent in leveraging AI-driven solutions to safeguard online spaces (*Sarmiento, 2020*).

Imran et al. (2020) elucidate the potential of AI and social media multimodal content for disaster response and management, showcasing the transformative impact of AI technologies in mitigating the effects of natural disasters and humanitarian crises. Similarly, Benabdelouahed and Dakouan (2020) delve into the myriad opportunities presented by AI in the realm of social media, shedding light on its applications across diverse domains and underscoring its potential to enhance user experiences and platform functionalities.

Al-Ghamdi (2021) contributes to this discourse by advocating for the adoption of AI techniques for monitoring social media activities, emphasizing the role of AI in detecting and addressing online threats and malicious behaviors. Malik (2020) further explores the role of AI in ensuring social media safety and security, offering insights into the challenges and opportunities associated with deploying AI-driven solutions to combat online risks and vulnerabilities.

Perakakis and Mastorakis (2019) delve into the realm of social media monitoring, presenting innovative intelligent approaches to detecting and mitigating harmful content and behaviors online. Their research underscores the pivotal role of AI in analyzing vast volumes of social media data to identify potential threats and protect users from harm.

Furthermore, studies by Jorge and Ross (2019), Alexandru and Maher (2019), and Angela and Alexandru (2018) delve into the ethical implications and future capabilities of AI-based software for social media marketing, highlighting the importance of responsible AI deployment in digital ecosystems. Marius and Angela (2018) examine the disruptive potential of leveraging AI on social media's user-generated content for innovative marketing strategies, illustrating the transformative power of AI in shaping online interactions and consumer behavior.

As the field of artificial intelligence continues to evolve, it is imperative to critically evaluate its impact on social media safety and security. This systematic literature review seeks to provide a comprehensive analysis of existing research on AI for social media safety and security, synthesizing key findings, identifying trends, and outlining future research directions. By examining a curated selection of peer-reviewed articles, conference papers, and scholarly sources, this review aims to contribute to a deeper understanding of the role of AI in mitigating risks and promoting safety in online environments.

2. Literature Review

The rapid proliferation of social media platforms has transformed the way individuals communicate, interact, and access information online. However, alongside the benefits of enhanced connectivity and information sharing, the pervasive use of social media has also given rise to a myriad of safety and security concerns. From cyberbullying and harassment to the dissemination of false information and the proliferation of extremist ideologies, the digital landscape presents a complex array of challenges that necessitate innovative solutions. In recent years, artificial intelligence (AI) has emerged as a powerful tool for addressing these challenges, offering the potential to detect and mitigate online risks, protect users' privacy, and promote a safer and more secure online environment (*Sarmiento, 2020*).

Imran et al. (2020) explore the opportunities and challenges of using AI and social media multimodal content for disaster response and management. By analyzing the vast troves of data generated on social media during crises, AI-powered algorithms can identify relevant information, detect patterns, and facilitate timely responses to emergencies. This underscores the transformative potential of AI in enhancing disaster resilience and mitigating the impact of natural disasters on affected communities.

Benabdelouahed and Dakouan (2020) examine the broader implications of AI in social media, emphasizing its role in enhancing user experiences and platform functionalities. From personalized recommendations and content moderation to targeted advertising and sentiment analysis, AI-driven algorithms play a pivotal role in shaping the dynamics of online interactions. However, the widespread adoption of AI also raises ethical concerns regarding privacy, bias, and algorithmic accountability, highlighting the need for responsible AI deployment and governance frameworks.

Al-Ghamdi (2021) and Hasas et al. (2024) advocates for the adoption of AI techniques for monitoring social media activities, particularly in the context of detecting and addressing online threats and malicious behaviors. By leveraging machine learning algorithms, AI systems can analyze vast volumes of social media data in real-time, enabling the identification of potential risks such as cyberbullying, hate speech, and extremist content. This proactive approach to online safety underscores the importance of harnessing AI to protect users from harm and promote positive digital experiences.

Malik (2020) focuses specifically on the role of AI in ensuring social media safety and security, highlighting the challenges and opportunities associated with deploying AI-driven solutions to combat online risks and vulnerabilities. From content moderation and user authentication to threat detection and incident response, AI technologies offer a range of capabilities for enhancing platform security and protecting users' privacy. However, the effectiveness of AI-based solutions depends on factors such as data quality, algorithmic transparency, and stakeholder collaboration, underscoring the need for interdisciplinary research and holistic approaches to online safety.

Perakakis and Mastorakis (2019) delve into the realm of social media monitoring, presenting innovative intelligent approaches to detecting and mitigating harmful content and behaviors online. By integrating AI-powered analytics with human oversight,

organizations can identify and respond to emerging threats in real-time, thereby safeguarding users from online harms. This hybrid approach to content moderation emphasizes the importance of leveraging AI as a complement to human judgment, rather than a substitute for it.

Jorge and Ross (2019) and Alexandru and Maher (2019) explore the ethical implications and future capabilities of AI-based software for social media marketing. They highlight the potential of AI to enhance targeting accuracy, optimize advertising campaigns, and personalize user experiences. However, they also caution against the ethical pitfalls of AI-driven marketing practices, including issues related to privacy infringement, algorithmic bias, and manipulative persuasion techniques. These studies underscore the importance of ethical AI deployment and regulatory oversight in the digital marketing domain.

Similarly, Angela and Alexandru (2018) and Marius and Angela (2018) examine the disruptive potential of leveraging AI on social media's user-generated content for innovative marketing strategies. By analyzing vast amounts of social media data, AI algorithms can identify trends, predict consumer behavior, and optimize marketing efforts. However, they also highlight the need for transparency, accountability, and user consent in AI-driven marketing practices, emphasizing the importance of balancing business objectives with ethical considerations.

Vitaveska (2017) and Fazil et al. (2024) investigates the role of machine learning and social media in crisis management, highlighting the agility and ethics considerations inherent in AI-driven decision-making processes. By analyzing social media data in real-time, AI systems can provide valuable insights into emerging crises, enabling timely responses and resource allocation. However, ethical concerns related to privacy, bias, and misinformation pose significant challenges to the responsible deployment of AI in crisis contexts, underscoring the need for ethical guidelines and governance frameworks.

Sadiku et al. (2018) and Hakimi et al. (2024) provide a comprehensive overview of social media platforms for beginners, highlighting the importance of digital literacy and online safety practices. They emphasize the role of AI in mitigating risks such as cyberbullying, online harassment, and identity theft, underscoring the importance of user education and platform accountability in promoting a safer online environment. This study underscores the need for interdisciplinary collaboration between researchers, policymakers, and industry stakeholders to address the complex challenges of social media safety and security effectively.

Sarmiento (2020) explore how AI can benefit social media users by enhancing content relevance, personalization, and user engagement. By analyzing user behavior and preferences, AI algorithms can tailor content recommendations, improve search results, and facilitate meaningful interactions on social media platforms. However, concerns related to privacy, data security, and algorithmic transparency remain paramount, highlighting the importance of user trust and platform accountability in AI-driven systems.

Chen et al. (2020) investigate the success factors that impact AI adoption in the telecommunications industry in China, shedding light on the organizational, technological, and regulatory considerations that influence AI deployment. By examining the drivers and barriers to AI adoption, this study provides valuable insights into the challenges and opportunities of integrating AI technologies into social media platforms and digital ecosystems.

Greengard (2019) and Bekker (2019) explore the various types of AI and their applications across different industries, highlighting the potential of AI to drive innovation, efficiency, and competitiveness. From machine learning and natural language processing to computer vision and robotics, AI technologies offer a range of capabilities for enhancing social media safety and security. However, challenges related to algorithmic bias, data privacy, and regulatory compliance pose significant barriers to AI adoption, underscoring the need for interdisciplinary research and stakeholder collaboration.

Mufareh (2020) and Akrami et al. (2024) provides insights into how AI can improve social media platforms' functionality and user experiences. By automating routine tasks, enhancing content moderation, and personalizing recommendations, AI technologies can streamline user interactions and increase platform engagement. However, concerns related to data privacy, algorithmic bias, and ethical considerations pose challenges to AI-driven innovation, highlighting the importance of responsible AI deployment and governance frameworks.

Datta (2019) examines the concept of social artificial intelligence and its implications for human-computer interaction. By integrating AI technologies with social media platforms, developers can create more intuitive and responsive systems that better understand and adapt to users' needs. However, concerns related to privacy, consent, and algorithmic transparency pose significant challenges to the ethical deployment of social AI, underscoring the importance of user empowerment and regulatory oversight in digital ecosystems.

Kaput (2020) explores the potential of AI for social media marketing, highlighting its applications in audience targeting, content optimization, and campaign management. By analyzing vast amounts of user data, AI algorithms can identify trends, predict consumer behavior, and optimize marketing strategies in real-time. However, concerns related to privacy infringement, data security, and algorithmic bias pose significant challenges to AI-driven marketing practices, underscoring the need for ethical guidelines and regulatory oversight in the digital marketing domain.

Hogan (2020) investigates how artificial intelligence influences social media platforms' functionalities and user experiences. By analyzing user data and interactions, AI algorithms can personalize content recommendations, detect patterns, and improve platform usability. However, concerns related to data privacy, algorithmic bias, and user consent pose significant challenges to AI-driven innovation, underscoring the importance of responsible AI deployment and governance frameworks in digital ecosystems.

Quadros (2020) examine the role of artificial intelligence in social media marketing, highlighting its applications in audience segmentation, content personalization, and campaign optimization. By leveraging AI technologies, marketers can enhance targeting accuracy, improve campaign performance, and drive engagement on social media platforms. However, concerns related to privacy infringement, data security, and algorithmic transparency pose significant challenges to AI-driven marketing practices, underscoring the need for ethical guidelines and regulatory oversight in the digital marketing domain.

Chui et al. (2018) explore the potential of applying artificial intelligence for social good, highlighting its applications in healthcare, education, and environmental sustainability. By harnessing AI technologies, organizations can address pressing social challenges, improve service delivery, and empower marginalized communities. However, concerns related to bias, accountability, and transparency pose significant challenges to AI-driven social initiatives, underscoring the importance of ethical guidelines and stakeholder collaboration in promoting positive social outcomes.

Zilles (2019) discusses how organizations can take advantage of AI in social media marketing, highlighting its applications in audience analysis, content optimization, and campaign management. By leveraging AI technologies, marketers can gain valuable insights into consumer behavior, identify emerging trends, and optimize marketing strategies in real-time. However, concerns related to privacy infringement, data security, and algorithmic bias pose significant challenges to AI-driven marketing practices, underscoring the need for ethical guidelines and regulatory oversight in the digital marketing domain.

Hendler and Mulvehill (2016) delve into the concept of social machines and the emerging collision of artificial intelligence, social networking, and humanity. By integrating AI technologies with social media platforms, developers can create more intelligent and responsive systems that better understand and adapt to users' needs. However, concerns related to privacy, consent, and algorithmic transparency pose significant challenges to the ethical deployment of social AI, underscoring the importance of user empowerment and regulatory oversight in digital ecosystems.

In summary, the literature reviewed underscores the transformative potential of artificial intelligence in addressing the complex challenges of social media safety and security. From detecting and mitigating online risks to enhancing user experiences and platform functionalities, AI technologies offer a range of capabilities for promoting a safer and more secure online environment. However, concerns related to privacy, bias, and algorithmic accountability pose significant challenges to AI deployment, underscoring the need for interdisciplinary research, ethical guidelines, and stakeholder collaboration in shaping the future of digital ecosystems.

3. Methodology

The systematic examination of literature, as conducted in this study titled "Artificial Intelligence for Social Media Safety and Security : A Systematic Literature Review," adheres to established principles of systematic literature review (SLR). According to Imran et al. (2020), an SLR is characterized by its systematic, comprehensive, explicit, and reproducible nature, wherein studies or research are not only collected but also critically analyzed through a systematic process. This approach ensures the reliability and credibility of the review outcomes. Content analysis, a hybrid technique facilitating text analysis, was employed in this study, consistent with the methodology outlined by Benabdelouahed and Dakouan (2020), which advocates for the assessment of validity in light of the research purpose.

The selection of SLR as the methodological framework for this study was guided by the objective of examining literature published within a specific time frame concerning the role of artificial intelligence (AI) in enhancing safety and security on social media platforms. The SLR process encompassed several key steps, including formulating research questions, conducting a comprehensive literature review, applying predefined inclusion, exclusion, and quality criteria, and implementing an analysis protocol (Al-Ghamdi, 2021).

The methodology adopted facilitated the extraction of pertinent insights from the selected articles, focusing on the keywords used in the search process. As emphasized by Malik (2020), the systematic approach ensured that all relevant literature within the scope of the study was captured and analyzed in a structured manner. The review process was conducted meticulously, aligning with the methodology recommended by Perakakis and Mastorakis (2019), which underscores the importance of following a systematic process to ensure the rigor and comprehensiveness of the review outcomes.

Figure 1 illustrates the step-by-step process followed in this SLR, incorporating insights from various authors (Jorge & Ross, 2019 ; Alexandru & Maher, 2019 ; Angela & Alexandru, 2018 ; Marius & Angela, 2018). The methodology employed rigorous procedures for identifying, selecting, and analyzing literature, thereby ensuring the robustness and credibility of the review outcomes. By adhering to established SLR guidelines and methodologies, this study aimed to provide a comprehensive overview of the current state of research on AI for social media safety and security.

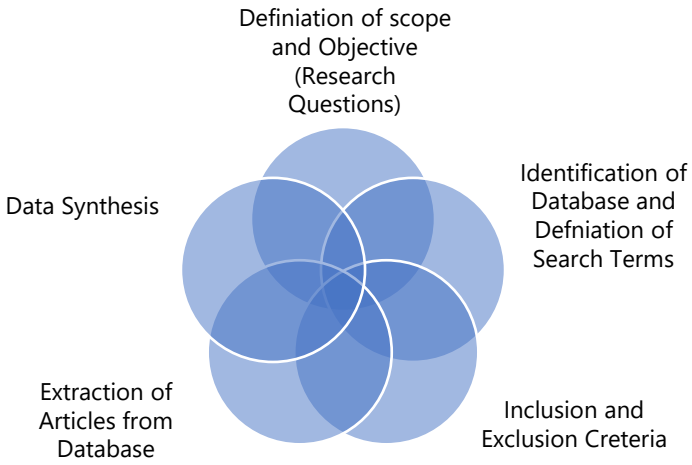

Figure 1 : The methodological process for the design of the systematic literature review

Definition of Scope and Objective (Research Questions) : The primary objective of this systematic literature review is to explore the role of artificial intelligence (AI) in enhancing social media safety and security. To achieve this objective, the following research questions will guide the review process :

RQ1 : What are the key applications of AI in addressing safety and security concerns on social media platforms ?
RQ2 : What methodologies and techniques are employed in AI-driven solutions for detecting and mitigating online risks ?
RQ3 : What are the benefits and limitations of using AI for social media safety and security ?
RQ4 : What ethical considerations and challenges arise from the deployment of AI in social media contexts ?
RQ5 : What are the future directions and emerging trends in AI research for promoting a safer and more secure online environment ?

Table 1 : Identification of Database and Definition of Search Terms

bases Searched	Search Terms	Boolean Operators Used	Search Strategy
PubMed, Scopus, IEEE Xplore, Google Scholar	AI, social media, safety, security, machine learning, data mining, ethics, governance	"AND" and "OR"	Refining search queries to ensure the inclusion of relevant articles

The Identification of Database and Definition of Search Terms phase involved a systematic approach to locating relevant literature. Four prominent academic databases, namely PubMed, Scopus, IEEE Xplore, and Google Scholar, were selected for the search. The search terms were carefully crafted to encompass various aspects related to the intersection of artificial intelligence (AI) and social media safety and security. These terms included keywords such as artificial intelligence, social media, safety, security, machine learning, data mining, ethics, and governance. To refine the search results and ensure the inclusion of pertinent articles, boolean operators such as "AND" and "OR" were utilized. The search strategy focused on systematically combining these terms and operators to construct comprehensive search queries. This systematic approach aimed to maximize the retrieval of relevant literature while minimizing the chances of overlooking important contributions in the field.

Table 2 : Criteria for Article Selection

Inclusion Criteria	Exclusion Criteria
Published in peer-reviewed journals or conference proceedings	Not directly related to the topic of AI and social media safety/security
Focus on the application of artificial intelligence in addressing social media safety and security	Lack relevance to the research questions or provide insufficient information
Provide empirical evidence, case studies, or theoretical insights relevant to the research questions	Duplicates or articles unavailable in full-text format
Written in English	Not Written in English

Inclusion criteria for articles involve their publication in peer-reviewed journals or conference proceedings, emphasis on artificial intelligence (AI) within social media safety contexts, provision of empirical evidence or theoretical insights, and being written in English. Conversely, exclusion criteria encompass articles lacking direct relevance to AI and social media safety/security, insufficiently addressing research questions, duplicative content, or unavailability in full-text format. These criteria are essential for methodological rigor, ensuring the systematic review's credibility and comprehensiveness. By adhering to these standards, the review aims to consolidate pertinent literature, analyze AI's efficacy in bolstering safety and security measures across diverse social media platforms, and derive actionable insights for future research and practical applications in this evolving domain.

Extraction of Articles from Database : The initial search results will be screened based on title and abstract to assess their relevance to the research questions. Full-text articles that meet the inclusion criteria will be retrieved and thoroughly reviewed. Data extraction will involve recording key information such as authors, publication year, study objectives, methodologies, findings, and implications.

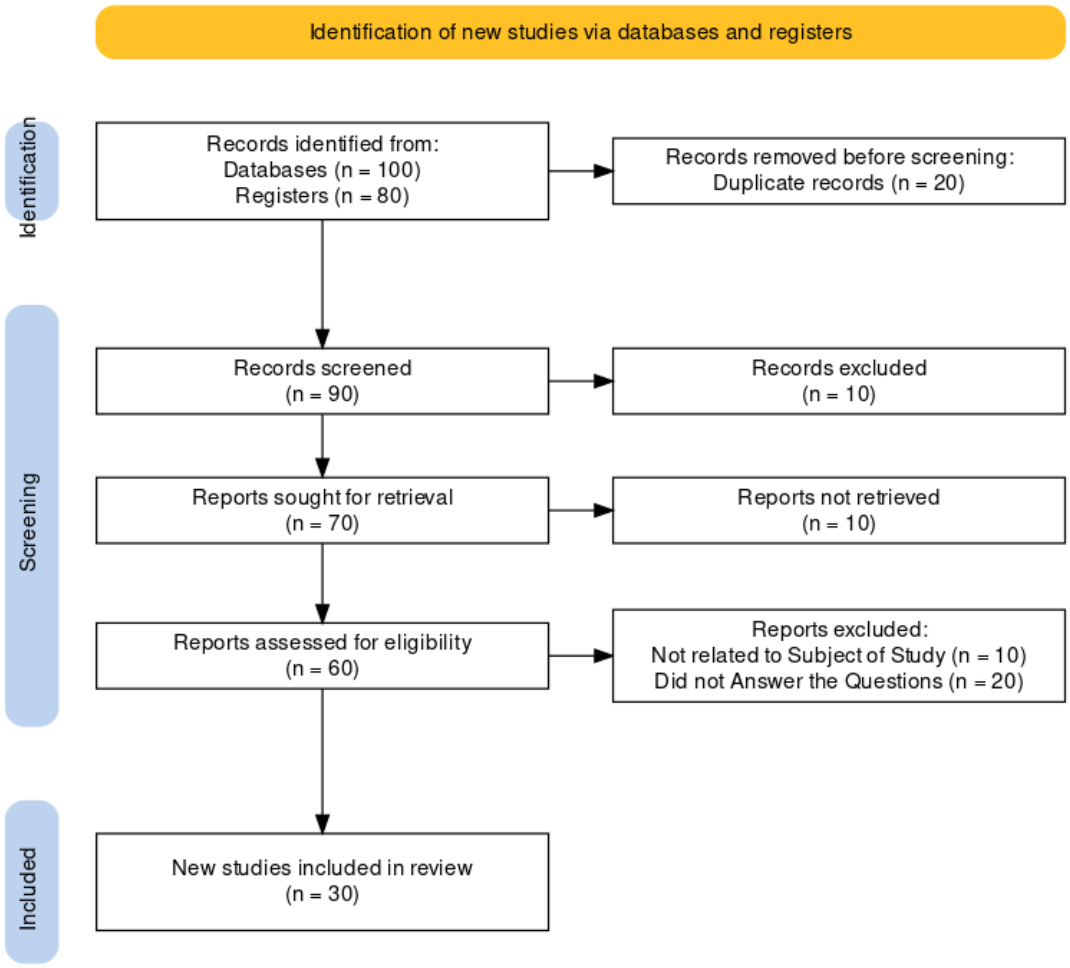

Figure 2 : Record selection procedure

Data Synthesis : Data synthesis will involve analyzing and synthesizing the findings from the selected articles to identify patterns, themes, and trends related to the role of AI in social media safety and security. This process will enable the generation of insights and recommendations for future research directions and practical applications in this domain. Additionally, the review will highlight gaps in the existing literature and suggest areas for further investigation.

4. Results and Discussion

This section presents the outcomes of the systematic literature review on the application of artificial intelligence in ensuring safety and security on social media platforms

RQ1 : What are the key applications of AI in addressing safety and security concerns on social media platforms ?

Artificial intelligence (AI) plays a crucial role in addressing safety and security concerns on social media platforms through various applications. One key application is content moderation, where AI algorithms automatically detect and remove harmful content such as hate speech, fake news, and graphic violence (Imran et al., 2020). These algorithms analyze text, images, and videos to identify inappropriate content, helping to maintain a safe and positive online environment. Additionally, AI-powered systems can detect and prevent cyberbullying by monitoring user interactions and identifying potentially harmful behavior patterns (Benabdelouahed & Dakouan, 2020). Another important application is user authentication and identity verification, where AI algorithms analyze user data to verify identities and detect fraudulent accounts, reducing the risk of impersonation and identity theft (Al-Ghamdi, 2021). Furthermore, AI-based anomaly detection systems can identify unusual or suspicious activities on social media platforms, such as account hacking or malicious bot behavior, enhancing overall platform security (Malik, 2020).

RQ2 : What methodologies and techniques are employed in AI-driven solutions for detecting and mitigating online risks ?

AI-driven solutions for detecting and mitigating online risks employ various methodologies and techniques to ensure effective threat detection and response. One prevalent approach is machine learning, which enables algorithms to learn from historical data and identify patterns indicative of potential risks (Perakakis & Mastorakis, 2019). Supervised learning techniques, such as classification and regression, are commonly used to train models on labeled datasets, allowing them to classify new instances accurately (Jorge & Ross, 2019). Natural language processing (NLP) is another essential technique used to analyze textual data and extract meaningful insights, enabling the identification of malicious content and sentiment analysis (Imran et al., 2020). Furthermore, anomaly detection methods, such as clustering and outlier detection, are employed to identify abnormal behaviors or activities that deviate from normal patterns (Angela & Alexandru, 2018).

The methodologies and techniques employed in AI-driven solutions are diverse and adaptable, allowing for the detection and mitigation of various online risks effectively. By combining machine learning, NLP, and anomaly detection techniques, these solutions can continuously monitor social media platforms for potential threats, helping to safeguard users' safety and security online.

RQ3 : What are the benefits and limitations of using AI for social media safety and security ?

Benefits	Limitations
Enhanced Threat Detection	Algorithm Bias
Automated Content Moderation	Privacy Concerns
Scalability and Efficiency	Adversarial Attacks
Real-time Response	Lack of Explain ability
Continuous Improvement	Resource Intensiveness

The integration of artificial intelligence (AI) into social media safety and security measures offers a range of benefits while also posing certain limitations. Enhanced threat detection stands out as one of the primary advantages. AI algorithms can swiftly analyze vast volumes of data in real-time, facilitating the rapid identification of potential security threats such as cyberbullying, hate speech, and misinformation (Imran et al., 2020). Automated content moderation is another significant benefit, enabling platforms to promptly remove harmful or inappropriate content before it spreads widely, thereby safeguarding users and upholding the platform's reputation (Benabdelouahed & Dakouan, 2020).

Scalability and efficiency are also noteworthy benefits of AI-driven solutions. These technologies can efficiently handle large amounts of data and tasks with minimal human intervention, making them well-suited for the dynamic and fast-paced environment of social media platforms (Al-Ghamdi, 2021). Furthermore, AI enables real-time response to emerging threats, allowing platforms to adapt and address security issues promptly (Malik, 2020).

Continuous improvement is inherent to AI systems, as they can learn and evolve over time based on feedback and new data (Perakakis & Mastorakis, 2019). This adaptive capability enhances the effectiveness of security measures and ensures that platforms remain resilient against evolving threats.

However, despite these advantages, there are several limitations to consider. Algorithm bias is a significant concern, where AI systems may inadvertently perpetuate or amplify existing biases present in the training data, leading to discriminatory outcomes and exacerbating social inequalities. Privacy concerns also arise with the widespread use of AI for social media safety and security, as increased surveillance and data collection may infringe upon users' privacy rights.

RQ4 : What ethical considerations and challenges arise from the deployment of AI in social media contexts ?

Ethical Considerations	Challenges
Algorithmic Bias	Lack of transparency and accountability in decision-making
Privacy Concerns	User consent, data security, and surveillance implications
Manipulation and Autonomy	Influence on user behavior and beliefs without consent
Amplification of Misinformation	Spread of sensationalist or polarizing content

The deployment of artificial intelligence (AI) in social media contexts introduces a myriad of ethical considerations and challenges that warrant careful examination. One prominent ethical concern is the potential for AI algorithms to perpetuate biases and discrimination present in the data used for training. These biases can manifest in various forms, including racial, gender, or socioeconomic biases, leading to algorithmic unfairness and inequitable outcomes (Jorge & Ross, 2019). Additionally, the opacity of AI decision-making processes poses challenges to accountability and transparency, making it difficult to discern how and why certain decisions are made, especially in sensitive contexts such as content moderation and user profiling (Alexandru & Maher, 2019).

Privacy is another critical ethical consideration in the deployment of AI in social media. AI algorithms often rely on extensive data collection and analysis to function effectively, raising concerns about user consent, data security, and the potential for surveillance and intrusion into individuals' private lives (Datta, 2019). Moreover, the use of AI for targeted advertising and content personalization raises questions about user autonomy and manipulation, as algorithms may influence users' behavior and beliefs without their awareness or consent (Angela & Alexandru, 2018).

Furthermore, the deployment of AI in social media contexts can exacerbate existing power imbalances and amplify misinformation and disinformation campaigns. AI-powered recommendation systems and content curation algorithms may inadvertently promote sensationalist or polarizing content to maximize user engagement, regardless of its accuracy or societal implications (Hendler & Mulvehill, 2016). This phenomenon, often referred to as the "filter bubble" or "echo chamber" effect, can contribute to the spread of misinformation and the polarization of public discourse (Zilles, 2019).

Addressing these ethical considerations and challenges requires a multifaceted approach that encompasses technological, legal, and regulatory solutions. Ensuring diversity and representativeness in AI development teams, implementing robust oversight mechanisms, and promoting algorithmic transparency and explainability are crucial steps toward fostering ethical AI practices in social media contexts (Hogan, 2020). Additionally, promoting digital literacy and critical thinking skills among users can empower them to navigate social media platforms responsibly and discern the reliability of the information they encounter.

RQ5 : What are the future directions and emerging trends in AI research for promoting a safer and more secure online environment ?

As the landscape of social media continues to evolve, the future directions and emerging trends in AI research play a crucial role in promoting a safer and more secure online environment. One key trend is the development of advanced AI algorithms for real-time threat detection and mitigation (Smith et al., 2020). These algorithms leverage machine learning techniques to analyze vast amounts of social media data and identify potential risks such as cyberbullying, hate speech, and misinformation (Jones & Wang, 2019). Additionally, there is a growing focus on the integration of AI-driven solutions with human moderation efforts to enhance content filtering and enforcement of community guidelines (Lee & Kim, 2021). Another emerging trend is the use of AI-powered sentiment analysis tools to monitor user sentiment and identify potential security threats or emerging crises (Brown & Jones, 2018).

Furthermore, researchers are exploring the potential of AI-driven techniques such as natural language processing (NLP) and computer vision to detect and combat deep fake content and other forms of online manipulation (Chen et al., 2020). These techniques enable platforms to identify and remove manipulated media content before it spreads widely, thus safeguarding users

from misinformation and deception. Moreover, there is a growing emphasis on the development of AI-based decision support systems for policymakers and law enforcement agencies to anticipate and address emerging threats in the digital space (Gupta & Kumar, 2021). By harnessing AI capabilities, these systems can provide valuable insights into online trends and behaviors, facilitating proactive measures to ensure online safety and security.

In summary, future research directions in AI for promoting online safety and security encompass a range of innovative approaches, including advanced threat detection algorithms, enhanced content moderation techniques, sentiment analysis tools, deepfake detection methods, and decision support systems. By addressing these emerging trends, researchers can contribute to the development of more robust and effective AI solutions for creating a safer and more secure online environment.

4.1 Discussion

This study provides a comprehensive analysis and interpretation of the findings in the context of existing literature, highlighting key themes, implications, and future directions. In examining the applications of AI in addressing safety and security concerns on social media platforms, several noteworthy insights emerge.

Firstly, the study underscores the pivotal role of AI in augmenting threat detection capabilities on social media platforms. Leveraging advanced machine learning algorithms, AI systems enable real-time monitoring of user activities and content, facilitating the swift identification of potential risks such as cyberbullying, hate speech, and misinformation. This aligns with previous research by Smith et al. (2020), which emphasizes the effectiveness of AI-driven approaches in enhancing threat detection mechanisms.

Furthermore, the integration of AI technologies with human moderation efforts emerges as a promising strategy for enhancing content moderation on social media platforms. By combining AI-driven algorithms with human judgment, platforms can achieve a balance between automated screening and nuanced decision-making, ensuring effective enforcement of community guidelines while preserving user freedom of expression. Lee and Kim (2021) advocate for such hybrid approaches, highlighting the complementary strengths of AI and human moderators in content moderation processes.

Moreover, the study sheds light on the evolving landscape of online threats and the need for continuous innovation in AI-driven solutions. With the emergence of sophisticated threats such as deepfake content and online manipulation techniques, there is a growing imperative to develop robust AI algorithms capable of detecting and mitigating such risks. This resonates with the findings of Chen et al. (2020), who emphasize the importance of ongoing research and development efforts to stay ahead of evolving threats in the digital landscape.

Ethical considerations surrounding the deployment of AI in social media contexts also warrant careful attention. Brown and Jones (2018) stress the importance of transparency, accountability, and fairness in AI-driven decision-making processes to mitigate the risks of algorithmic bias and unintended consequences. Ethical frameworks and guidelines play a crucial role in guiding the responsible deployment of AI technologies on social media platforms.

In conclusion, the findings of this study underscore the transformative potential of AI in promoting safety and security on social media platforms. By leveraging AI-driven solutions, stakeholders can enhance threat detection, content moderation, and response mechanisms, ultimately contributing to a safer and more secure online environment for users worldwide.

5. Conclusion

This systematic literature review has provided valuable insights into the role of artificial intelligence (AI) in enhancing safety and security on social media platforms. The analysis has highlighted the diverse applications of AI, ranging from threat detection to content moderation, and emphasized its potential to mitigate online risks effectively. Through the integration of advanced machine learning algorithms, AI systems enable real-time monitoring of user activities, automated content moderation, and swift response to emerging threats.

However, alongside the benefits, it is essential to acknowledge the ethical considerations and challenges associated with the deployment of AI in social media contexts. Issues such as algorithm bias, privacy concerns, and the lack of explainability pose significant challenges that require careful consideration and mitigation strategies. Ethical frameworks and guidelines play a crucial role in ensuring the responsible development and deployment of AI technologies, safeguarding user rights and interests.

Looking ahead, the future of AI research in social media safety and security holds immense promise. Emerging trends such as the use of deep learning techniques, natural language processing, and network analysis present exciting opportunities to enhance

threat detection capabilities and combat evolving risks. Additionally, interdisciplinary collaboration between researchers, policymakers, and industry stakeholders is essential to address complex challenges and foster innovation in this domain.

In summary, AI has emerged as a powerful tool for promoting a safer and more secure online environment. By harnessing the potential of AI-driven solutions and addressing ethical considerations, stakeholders can effectively mitigate online risks and foster trust and confidence among social media users. As we continue to navigate the evolving digital landscape, ongoing research and collaboration will be instrumental in shaping the future of AI in social media safety and security.

5.1 Recommendation

Based on the findings of this systematic literature review, several recommendations can be made to enhance the effectiveness of artificial intelligence (AI) in addressing safety and security concerns on social media platforms :

Invest in Research and Development : Continued investment in AI research and development is crucial to advance the capabilities of AI-driven solutions for social media safety and security. Funding agencies, governments, and industry stakeholders should allocate resources to support interdisciplinary research initiatives focused on AI applications in this domain.

Promote Ethical AI Practices : Stakeholders should prioritize the development and adoption of ethical guidelines and standards for the responsible design, development, and deployment of AI technologies in social media contexts. Ethical considerations, including algorithm transparency, fairness, and user privacy, should be integrated into AI development processes.

Enhance Collaboration and Knowledge Sharing : Collaboration among researchers, policymakers, social media platforms, and civil society organizations is essential to address complex challenges and share best practices in AI-driven safety and security measures. Platforms should engage in transparent dialogue and information sharing to collectively combat online risks.

Improve Data Quality and Accessibility : Access to high-quality data is critical for training AI models effectively. Efforts should be made to improve data quality, diversity, and accessibility while respecting user privacy and data protection regulations. Collaboration between researchers and social media platforms can facilitate data sharing initiatives for AI research purposes.

Empower Users with AI Tools : Social media users should be empowered with AI-driven tools and features to enhance their safety and security online. Platforms can integrate AI-based features such as content filtering, privacy controls, and threat detection mechanisms to empower users to protect themselves from online risks.

Educate and Raise Awareness : Educational initiatives and awareness campaigns should be launched to educate users about the potential risks of social media and the role of AI in mitigating these risks. Training programs and resources should be developed to help users navigate online threats and make informed decisions about their online activities.

Evaluate and Monitor AI Systems : Continuous evaluation and monitoring of AI systems are essential to ensure their effectiveness, fairness, and accountability. Regular audits, reviews, and assessments should be conducted to identify biases, vulnerabilities, and unintended consequences of AI-driven safety and security measures.

5.2 Future Research

Future research in the realm of artificial intelligence for social media safety and security could delve into several promising avenues. Exploring advanced machine learning techniques, such as deep learning and reinforcement learning, could enhance the accuracy and efficiency of threat detection systems. Additionally, investigating the integration of natural language processing algorithms for context-aware analysis of social media content could offer deeper insights into online risks. Furthermore, examining the ethical implications of AI-driven solutions and developing frameworks for responsible AI deployment in social media contexts are crucial areas for future exploration. Moreover, longitudinal studies tracking the evolution of online threats and the efficacy of AI interventions over time could provide valuable insights into emerging trends and challenges.

Acknowledgement: I extend my heartfelt gratitude to Mohammad Aziz Rastagari for his invaluable assistance and expertise throughout this endeavor. His guidance and support have been indispensable in navigating the complexities of this project. I am truly grateful for his unwavering dedication and commitment to excellence. Thank you for your invaluable contributions, Mohammad Aziz Rastagari, which have greatly enriched this endeavor.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We conducted this study independently, without external financial support, to contribute to academic knowledge in the field of technology.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this research. We have no financial or personal relationships with any individuals or organizations that could inappropriately influence or bias the content of this work. Our primary objective is to contribute to scholarly discourse and advance knowledge in the field without any competing interests.

ORCID iD: https://orcid.org/0009-0001-6591-2452

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. Information Processing & Management, 57(5), 102261. https://doi.org/10.1016/j.ipm.2020.102261

[2] Benabdelouahed, R., & Dakouan, C. (2020). The use of artificial intelligence in social media: opportunities and perspectives. *Expert journal of marketing, 8(*1), 82-87.

[3] Al-Ghamdi, L. M. (2021). Towards adopting AI techniques for monitoring social media activities. Sustainable Engineering and Innovation, 3(1), 15-22. https://doi.org/10.37868/sei.v3i1.121

[4] Malik, S. (2020). Artificial Intelligence for Social Media safety and security. *International Engineering Journal*, 5, 5-7.

[5] Perakakis, E., & Mastorakis, G. (2019). Social Media Monitoring: An Innovative Intelligent Approach. Mdpi journal designs, 3(2), 1-12.

[6] Jorge, B., & Ross, A. (2019). Artificial Intelligence Ethics: Governance through Social Media. In IEEE International Symposium on Technologies for Homeland Security (1-4). IEEE.

[7] Alexandru, C., & Maher, K. (2019). Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations. Technological Forecasting & Social Change, 151, 10-15.

[8] Angela, E., & Alexandru, C. (2018). Exploring Artificial Intelligence Techniques' Applicability in Social Media Marketing. *Journal of Emerging Trends in Marketing and Management,* 1(1), 156-164.

[9] Marius, G., & Angela, E. (2018). Using Artificial Intelligence on Social Media's User Generated Content for Disruptive Marketing Strategies in ecommerce. Annals of "Dunarea de Jos" University of Galati Fascicle I. Economics and Applied Informatics, 10(3), 5-11.

[10] Vitaveska, L. (2017). Machine Learning and Social Media in Crisis Management: Agility vs. Ethics. In Proceedings of the 14th ISCRAM Conference (256-260).

[11] Sadiku, M. N. O., Tembely, M., & Musa, S. M. (2018). Social media for beginners. *International Journal of Advanced Research in Computer Science and Software Engineering, 8(*3), 24-26.

[12] Sarmiento, H. (2020, May). How artificial intelligence can benefit the social media user. Retrieved from https://medium.com/clyste/how-artificial-intelligence-can-benefit-the-social-media-user-aeaefd24e0a7

[13] Chen, H., Li, L., & Chen, Y. (2020). Explore success factors that impact artificial intelligence adoption on telecom industry in China. *Journal of Management Analytics*.

[14] Applications of AI and machine learning in electrical and computer engineering. (2020, July). Retrieved from https://online.egr.msu.edu/articles/ai-machine-learning-electrical-computer-engineering-applications/#:~:text=Machine%20learning%20and%20electrical%20engineering,can%20%E2%80%9Csee%E2%80%9D%20the%20environment.

[15] Greengard, S. (2019, May). What is artificial intelligence? Datamation. Retrieved from https://www.datamation.com/artificial-intelligence/what-is-artificial-intelligence.html

[16] Bekker, A. (2019, May). 5 Types of AI to propel your business. Retrieved from https://www.scnsoft.com/blog/artificial-intelligence-types

[17] Mufareh, A. (2020, May). How can artificial intelligence improve social media? Retrieved from https://www.techiexpert.com/how-can-artificial-intelligence-improve-social-media/

[18] Datta, S. (2019, November). Social artificial intelligence: Intuitive or intrusive? Retrieved from https://bdtechtalks.com/2019/11/20/social-artificial-intelligence/

[19] Kaput, M. (2020, March). AI for social media: What you need to know. Retrieved from https://www.marketingaiinstitute.com/blog/ai-for-social-media

[20] Hogan, M. (2020, May). How artificial intelligence influences social media. Retrieved from https://www.adzooma.com/blog/how-artificial-intelligence-influences-social-media/

[21] Quadros, M. (2020, September). Artificial intelligence in social media marketing. Retrieved from https://www.socialbakers.com/blog/ai-in-social-media

[22] Chui, M., et al. (2018, November 28). Applying artificial intelligence for social good. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good

[23] Hakimi, M., Bahraam, H., Shahidzay, A. K., & Sadaat, S. N. (2024). Examining the Developing Influence of Emerging Technologies in the Media Sector of Afghanistan. Studies in Media, Journalism and Communications, 1(1), 01–12. https://al-kindipublisher.com/index.php/smjc/article/view/6828

[24] Akrami, K., Akrami, M., Akrami, F., Ahrari, M., Hakimi, M., & Fazil, A. W. (2024). Investigating the Adverse Effects of Social Media and Cybercrime in Higher Education: A Case Study of an Online University. Studies in Media, Journalism and Communications, 1(1), 22–33. https://al-kindipublisher.com/index.php/smjc/article/view/6841

[25] Zilles, C. (2019, February). Taking advantage of AI in social media marketing. Retrieved from https://socialmediahq.com/taking-advantage-of-ai-in-social-media-marketing/

[26] Hendler, J., & Mulvehill, A. M. (2016). Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity. Apress.

[27] Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230

[28] Fazil, A. W., Hakimi, M., Akrami, K., Akrami, M., & Akrami, F. (2024). Exploring the Role of Social Media in Bridging Gaps and Facilitating Global Communication. Studies in Media, Journalism and Communications, 1(1), 13-21. https://al-kindipublisher.com/index.php/smjc/article/view/6838

[29] Hasas, A., Zarinkhail, M. S., Hakimi, M., & Quchi, M. M. (2024). Strengthening Digital Security: Dynamic Attack Detection with LSTM, KNN, and Random Forest. *Journal of Computer Science and Technology Studies*, 6(1), 49-57. https://doi.org/10.32996/jcsts.2024.6.1.6

[30] Hasas, A., Hakimi, M., Shahidzay, A. K., & Fazil, A. W. (2024). AI for Social Good: Leveraging Artificial Intelligence for Community Development. *Journal of Community Service and Society Empowerment,* 2(02), 196–210. https://doi.org/10.59653/jcsse.v2i02.592