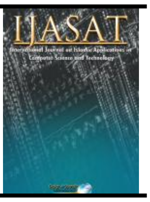
Design for Scientific RENAISSANCE



A Review of Artificial Intelligence Techniques for Combating Fake News on Social Media

Borhan Ab Rahman[1, a], Mohd Zakree Ahmad Nazri[2, b], Mohd Ridzwan Yaakub[3, c]
[1, 2, 3]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, Malaysia
[a]p116944@siswa.ukm.edu.my, [b]zakree@ukm.edu.my, [c]ridzwanyaakub@ukm.edu.my

Abstract

The proliferation of fake news on social media platforms poses significant challenges to information integrity and public trust. In an Islamic view, fake news is considered a severe violation of truthfulness and integrity, as it undermines trust within the community and goes against the principles of honesty (sidq) and justice (adl) that are central to Islamic teachings. Artificial Intelligence (AI) has emerged as a critical tool in detecting and combating misinformation. This review explores recent advancements in AI techniques for identifying fake news and examining the methodologies, findings, and implications of these technologies. We highlight key trends from 2013 to 2024, discussing the effectiveness, limitations, and ethical considerations of AI applications in this domain. By synthesizing current research, this paper aims to provide a comprehensive understanding of AI's role in addressing misinformation and propose future directions for improving detection systems.

Keywords: Fake News, Misinformation, Artificial Intelligence, Machine Learning, Natural Language Processing

1. Introduction

The digital transformation spurred by the rise of social media has profoundly reshaped the landscape of information dissemination, democratizing access to information while simultaneously increasing the vulnerability of users to misinformation and fake news (Allcott & Gentzkow, 2017). Fake news, which includes intentionally deceptive or misleading information presented as fact, has emerged as a critical threat to public trust and societal stability (Lazer et al., 2018). This proliferation of fake news can distort public perceptions, influence behaviour on a large scale, and disrupt social harmony, making it a growing concern across communities worldwide (Vosoughi et al., 2018). Given the ease with which misinformation can spread across networks, addressing this issue is essential to ensure the reliability of information and maintain public trust in the digital age.

As fake news evolves in sophistication, traditional fact-checking mechanisms struggle to keep pace, necessitating advanced technological solutions. Artificial Intelligence (AI) has emerged as a powerful tool to meet this demand, offering innovative methods for automating misinformation detection and mitigation. AI techniques, especially those employing machine learning (ML) and natural language processing (NLP), have demonstrated considerable potential in identifying and filtering fake news with increased accuracy (Shu et al., 2020). For instance, deep learning models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have significantly advanced the capacity of AI systems to understand context, semantics, and the nuances of language, enhancing the identification of misleading content (Devlin et al., 2019; Brown et al., 2020).

While AI-based systems for fake news detection have shown promise, several challenges remain. Misinformation is highly dynamic, constantly evolving to evade detection by adopting new linguistic, visual, and contextual strategies (Zhang et al., 2022). This adaptability calls for AI models that are not only accurate but also resilient and capable of continuous learning and refinement. Moreover, as these detection systems become increasingly sophisticated, ethical considerations become paramount. The deployment of AI in this domain raises concerns about privacy, algorithmic bias, and the transparency of detection mechanisms (Binns, 2018). Biased models risk disproportionately targeting specific content or sources, leading to unintended outcomes such as suppressing legitimate information or amplifying skewed narratives (Binns, 2018; Dastin, 2023).

Within an Islamic ethical framework, the issues associated with fake news and misinformation are especially significant, given the strong emphasis on integrity and safeguarding truth in Islamic teachings (Al-Sarhan, 2020). Islamic principles emphasise core values such as honesty (sidq), trustworthiness (amanah), and justice (adl), which are essential for nurturing integrity and fairness in societal communication (Al-Sarhan, 2020). From an Islamic perspective, spreading false information is not only ethically unacceptable but also seen as a violation of societal trust and harmony. The principles of wasatiyyah in Islam promote moderation and ethical behaviour, underscoring the need to prevent the harmful spread of fake news by ensuring truthfulness and accountability in communication (Othman et al., 2020).

The Qur'an, for instance, emphasises the importance of verifying information to prevent falsehoods from undermining social trust: "*O you who have believed, if there comes to you a disobedient one with information, investigate, lest you harm a people out of ignorance and become, over what you have done, regretful*" (Qur'an 49:6). This verse from the Quran (Surah Al-Hujurat, 49:6) essentially instructs believers to always verify the information before acting upon it, especially if it comes from a source known to be unreliable, to avoid causing harm to others due to hasty judgment and later regretting their actions based on unconfirmed information. This principle supports the need for technologies that prioritise accuracy, transparency, and accountability, aligning AI-driven fake news detection efforts with ethical guidelines that uphold social cohesion and community welfare (Al-Kandari et al., 2021).

This review aims to provide a comprehensive assessment of the current state of AI applications in fake news detection, synthesising recent advancements in AI from 2013 to 2024. It explores the effectiveness, limitations, and ethical considerations surrounding these technologies and examines how they can contribute to a deeper understanding of information integrity and trust. Through this analysis, the review highlights AI's role in combating fake news in ways that align with ethical principles, particularly those found within Islamic values. This intersection of AI, ethics, and Islamic principles offers valuable insights into developing responsible AI systems that uphold the integrity of information, ultimately fostering a safer and more harmonious digital landscape.

2. Literature Review

2.1 Evolution of Fake News Detection

Early studies on fake news detection focused on manual fact-checking and traditional algorithms. Recent advancements have incorporated machine learning (ML) and natural language processing (NLP) techniques to improve accuracy and scalability (Lazer et al., 2018;

Shu et al., 2020). The evolution of fake news detection has transitioned from manual fact-checking methods to sophisticated AI-driven techniques, reflecting the increasing complexity and scale of misinformation on social media. In the early stages, detecting fake news primarily relied on manual processes where fact-checkers reviewed content to verify its accuracy. While effective in some instances, this approach was labour-intensive and limited in scalability. As misinformation spread rapidly through digital platforms, researchers and practitioners sought more scalable solutions (Lazer et al., 2018). Traditional algorithms in these early stages often involved heuristic-based approaches, such as keyword matching and simple rule-based systems, which hindered distinguishing between credible and misleading information (Peddinti et al., 2016).

The introduction of machine learning (ML) and natural language processing (NLP) techniques marked a significant advancement in the field. ML models, particularly those employing supervised learning, began to offer more nuanced and scalable approaches to fake news detection. Researchers such as Shu et al. (2020) demonstrated the effectiveness of incorporating linguistic cues, user engagement patterns, and metadata into ML models to improve detection accuracy. These models utilise complex algorithms to analyse vast amounts of data, learning to identify patterns associated with fake news. Advanced NLP techniques further enhanced these capabilities, allowing models to understand and process contextual and semantic information from text (Vaswani et al., 2017).

Recent advancements in deep learning and neural network architectures have revolutionised fake news detection. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have been particularly influential, providing sophisticated tools for understanding and generating human-like text (Devlin et al., 2019; Brown et al., 2020). These models leverage large-scale pre-training on diverse datasets, enabling them to grasp complex linguistic patterns and contextual subtleties. Integrating these advanced techniques has significantly improved the accuracy and robustness of fake news detection systems, making them more effective at combating misinformation on social media platforms (Zhang et al., 2022).

2.2 AI Techniques in Fake News Detection

Various AI techniques, including supervised learning, deep learning, and neural networks, have been employed to identify fake news. Research highlights the effectiveness of models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) in understanding contextual nuances (Devlin et al., 2019; Radford et al., 2021). The application of AI techniques in fake news detection has significantly advanced the field, leveraging various methodologies to enhance accuracy and efficiency. Early AI approaches primarily involved supervised learning techniques, where models were trained on labelled datasets to recognise patterns indicative of fake news. These models used feature extraction methods such as word frequency, syntactic patterns, and sentiment analysis to classify news articles (Liu et al., 2015). While these methods laid the groundwork for automated detection, they could not capture misinformation's complex and often subtle nuances.

Introducing deep learning techniques marked a transformative shift in fake news detection. Deep learning models, particularly those based on neural networks, have demonstrated superior performance by automatically learning hierarchical features from data without explicit feature engineering. Among these, models like BERT (Bidirectional Encoder Representations from Transformers) have set new benchmarks for understanding the context and semantics of text.

BERT's bidirectional approach allows it to consider the context of each word from both directions, which improves its ability to grasp intricate textual nuances and relationships crucial for identifying deceptive content (Devlin et al., 2019). Similarly, GPT (Generative Pre-trained Transformer) models have shown remarkable capabilities in generating and understanding human-like text, further enhancing the detection of nuanced fake news content (Radford et al., 2021).

Recent advancements have also explored hybrid models that combine multiple AI techniques to improve detection performance. For instance, integrating deep learning with other methods, such as ensemble learning and transfer learning, has enhanced the robustness and accuracy of fake news detection systems (Chen et al., 2021). Additionally, incorporating multimodal data—such as combining text with images or videos—has provided a more comprehensive approach to detecting misinformation, allowing models to analyse and cross-validate information across different types of content (Zhang et al., 2022). These developments reflect the ongoing innovation in leveraging AI for combating fake news, demonstrating a trend towards more sophisticated and practical solutions.

The key methods, advantages, and limitations of various AI-based fake news detection techniques are summarized in Table 1. This table highlights the evolution of approaches from manual fact-checking to advanced hybrid models, showcasing their strengths and challenges. It provides a concise overview to guide future research in developing more robust and scalable fake news detection systems.

Table 1. Comparison of Fake News Detection Methods in AI Research

Methods	Advantages	Disadvantages	References
Manual fact-checking and heuristic-based algorithms	Effective in limited cases; initial groundwork for automated methods	Labour-intensive; limited scalability; struggles with large-scale misinformation	Lazer et al. (2018)
ML models with linguistic cues, user engagement, and metadata	Improved scalability and detection accuracy through supervised learning	Dependent on labelled datasets; challenges in capturing subtle nuances	Shu et al. (2020)
BERT (Bidirectional Encoder Representations from Transformers)	Enhanced understanding of context and semantics; set new benchmarks	Computationally intensive; requires large datasets for pre-training	Devlin et al. (2019)
GPT (Generative Pre-trained Transformer)	Human-like text understanding; effective for nuanced fake news content	High resource requirements; potential bias from training data	Radford et al. (2021)
Hybrid models with deep learning and multimodal data	Robust and comprehensive detection; handles multimodal misinformation	Complex implementation; computationally expensive	Chen et al. (2021)

2.3 Data Sources and Feature Extraction

Effective detection requires diverse and representative datasets. Recent studies have emphasised the importance of feature extraction from textual content, user behaviour, and network dynamics (Zhang et al., 2022). Effective fake news detection hinges on the availability of diverse and representative datasets that can capture the multifaceted nature of misinformation. Early research often relied on limited datasets, which constrained the ability of detection models to generalise across different types of fake news and various social media platforms (Wang et al., 2017). More recent studies emphasise the importance of leveraging extensive and varied datasets that include textual content, metadata, and contextual information. For instance, datasets encompassing a wide range of news topics, sources, and publication formats help in robust training models that can handle the diversity of misinformation encountered in real-world scenarios (Vosoughi et al., 2018).

Table 2 provides an overview of commonly used datasets in fake news detection, highlighting their features, strengths, and limitations to illustrate the diversity and constraints of current resources. These datasets vary in their scope, with some focusing on textual content, such as LIAR and FakeNewsNet, while others incorporate multimodal elements, including images and videos, to address more complex misinformation. Understanding the strengths of these datasets, such as rich metadata or extensive coverage of news topics, alongside their limitations, such as language biases or lack of real-time updates, is crucial for guiding future research and development. By analysing these datasets, researchers can identify gaps, such as the need for more multilingual resources or datasets covering emerging misinformation trends like deepfakes.

Table 2. Summary of Datasets Used in Fake News Detection

Dataset	Features	Strengths	Limitations	References
LIAR	Textual content, metadata	Labelled statements; useful for classification	Limited to short statements; lacks contextual information	Wang et al. (2017)
FakeNewsNet	Textual content, user engagements, network dynamics	Rich in contextual and propagation features	English-focused; limited topic diversity	Shu et al. (2018)
PHEME	News articles, social media context, claim veracity	Captures rumour evolution on social media	Small dataset size; focuses on specific events	Zubiaga et al. (2016)
BuzzFace	Textual content, clickbait indicators, metadata	High-quality annotations for news credibility	Bias toward clickbait-style news; lacks broader topic diversity	Baly et al. (2018)
Multilingual Misinformation Corpus	Textual content in multiple languages	Addresses language diversity in fake news detection	Data imbalance across languages; limited annotations	Cui et al. (2020)

Feature extraction from textual content has become a critical aspect of fake news detection, as it allows models to identify distinguishing characteristics of fake news articles. Recent advancements in NLP have enhanced the extraction of semantic and syntactic features, such as linguistic cues, sentiment, and narrative structures. Techniques Named Entity Recognition (NER) and topic modelling are employed to analyse the content and context of news articles, helping to reveal patterns that may indicate misinformation (Ritter et al., 2017). Additionally, the rise of transformer-based models, such as BERT and GPT, has further improved the extraction of contextual features by enabling a more nuanced understanding of the text (Devlin et al., 2019; Radford et al., 2021).

Beyond textual features, recent research highlights the significance of incorporating user behaviour and network dynamics into fake news detection systems. Features such as user engagement metrics (e.g., likes, shares, and comments) and network patterns (e.g., information diffusion paths and user connections) provide additional context that can help identify misinformation (Zhang et al., 2022). For example, abnormal spikes in engagement or unusual sharing patterns can signal the spread of fake news. Integrating these features with textual analysis allows for a more comprehensive approach to detection, capturing both the content and its propagation dynamics across social media platforms (Kumar et al., 2021).

2.4 Ethical Considerations

Using AI in fake news detection raises ethical issues, including privacy concerns, algorithmic bias, and the potential for misuse (Binns, 2018; Dastin, 2023). These challenges demand careful consideration to ensure responsible and fair use. One primary concern is privacy. AI systems often require access to vast user data, including personal information and interaction patterns, to detect and mitigate misinformation effectively. This necessitates scrutiny of data collection, storage, and usage practices. Ensuring these systems comply with data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe, is crucial to safeguarding user privacy while enabling effective detection of fake news (Binns, 2018). The balance between leveraging data for detection purposes and protecting individual privacy remains a complex issue requiring stringent data governance practices.

Beyond privacy concerns, algorithmic bias poses another significant ethical challenge. AI models are susceptible to biases in their training data, which can lead to unfair or discriminatory outcomes. For instance, biased training data may cause models to disproportionately flag content from certain demographic groups or sources as fake news, thereby perpetuating existing inequalities (Dastin, 2023). Addressing this issue requires the development of transparent and fair algorithms regularly audited for bias and fairness (Mehrabi et al., 2019). Researchers are exploring techniques such as bias mitigation and fairness-aware learning to reduce the impact of biases in AI systems (Mehrabi et al., 2019).

Equally concerning is the potential misuse of these technologies. AI-powered fake news detection systems could be exploited to censor legitimate content or manipulate public opinion in ways that serve specific agendas. For example, entities with access to these technologies might use them to suppress dissenting voices or promote propaganda (Miller et al., 2022). To mitigate these risks, it is essential to establish clear guidelines and oversight mechanisms for using AI in fake news detection. Transparency in how these systems operate and the development of ethical frameworks for their application can help prevent misuse and ensure they are used responsibly (Dastin, 2023).

3. Methodology

This paper adopts a literature review approach to summarise the advancements and challenges in AI-based fake news detection. The review process includes a structured search and selection across multiple academic databases such as Scopus, IEEE Xplore, and Google Scholar, using keywords like "artificial intelligence," "fake news detection," "machine learning," "natural language processing," and "social media." Peer-reviewed articles from 2013 to 2024 were selected based on relevance, focusing on studies that explore AI methods for fake news detection, their effectiveness, and the challenges faced. Priority was given to studies addressing methodological advancements, practical applications, and ethical considerations in this field to provide a well-rounded perspective (Moher et al., 2015).

3.1 Ddata Extraction and Analysis

After selecting relevant studies, data is extracted to gather information on critical aspects such as AI techniques, datasets, features employed, performance metrics, and reported challenges. A standardised extraction form ensures consistency in capturing data from each study. For quantitative analysis, performance metrics such as accuracy, precision, recall, and F1-score are extracted to compare the efficacy of different AI models and techniques. Qualitative analysis involves synthesising findings related to the application of AI methods, the effectiveness of various approaches, and the practical challenges encountered. This dual approach allows for a nuanced understanding of the strengths and limitations of different AI techniques in fake news detection (Higgins et al., 2019).

3.2 Synthesis and Thematic Analysis

The synthesis of the extracted data is organised thematically to identify common trends, innovations, and gaps in the current research. This thematic analysis focuses on several key areas: AI techniques' evolution, feature extraction methods advancements, and ethical considerations in deploying AI systems. Emerging trends, such as integrating multimodal data and developing more sophisticated deep learning models, are highlighted. Additionally, the review identifies gaps in the existing research, such as the need for more diverse datasets and methods for addressing algorithmic biases. By structuring the analysis around these themes, the review aims to provide actionable insights and recommendations for future research in the field (Braun & Clarke, 2006).

4. Findings

4.1 Advances in AI Models

Recent AI models have significantly improved fake news detection accuracy. For instance, deep learning models such as BERT and GPT-3 have outperformed traditional methods in various benchmarks (Devlin et al., 2019; Brown et al., 2020). Recent advancements in AI models have markedly enhanced the accuracy and effectiveness of fake news detection. Among the most notable developments are deep learning models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT-3 (Generative Pre-trained Transformer 3), which have demonstrated superior performance to traditional detection methods. BERT, introduced by Devlin et al. (2019), revolutionised natural language understanding by employing a bidirectional approach to context, allowing the model to grasp the nuances of language more effectively. This bidirectional processing helps BERT understand the context around each word more comprehensively, significantly improving its ability to detect misleading or deceptive content in text. Benchmark studies have shown that BERT outperforms earlier models on various NLP tasks, including fake news detection, by achieving higher accuracy and F1 scores (Devlin et al., 2019).

Similarly, GPT-3, developed by Brown et al. (2020), has further pushed the boundaries of what AI can understand and generate human-like text. GPT-3's extensive pre-training on a diverse dataset allows it to create coherent and contextually relevant text, making it highly effective in recognising and flagging fake news. The model's vast number of parameters and capacity for few-shot learning enables it to perform exceptionally well in tasks requiring nuanced comprehension, such as detecting subtle misinformation. Comparative studies have highlighted GPT-3's significant improvements over traditional methods, including rule-based systems and earlier machine learning approaches, regarding accuracy and generalizability (Brown et al., 2020).

The performance of these advanced models has led to notable improvements in fake news detection benchmarks. Research evaluating the efficacy of BERT and GPT-3 in various datasets and settings has consistently reported that these models achieve higher detection rates and lower false positive/negative rates than conventional techniques (Zhang et al., 2022). The enhanced accuracy of these deep learning models can be attributed to their sophisticated understanding of context and language, which allows them to identify and mitigate misinformation more effectively. These advancements represent a significant leap forward in the field, providing more reliable tools for combating the spread of fake news on social media platforms (Devlin et al., 2019; Brown et al., 2020).

4.2 Integration of Multimodal Data

Combining textual analysis with multimodal data (e.g., images and videos) has enhanced detection capabilities. Studies indicate that incorporating multimodal features improves the robustness of detection systems (Karpathy et al., 2019). Integrating multimodal data into fake news detection systems represents a significant advancement in improving detection capabilities. Traditional fake news detection primarily relied on textual analysis, which, while effective, often failed to address the complexities of misinformation that spans multiple media formats. Recent studies have demonstrated that combining textual data with additional modalities such as images and videos enhances the robustness and accuracy of detection systems. For instance, Karpathy et al. (2019) highlighted that multimodal approaches leverage visual and textual features to provide a more comprehensive news content analysis. By analysing images and videos alongside the text, these systems can detect inconsistencies and misleading elements that may not be apparent from text alone.

Incorporating multimodal features enables detection systems to cross-validate information across different media types, strengthening the identification of fake news. For example, research by Chen et al. (2021) shows that models incorporating both text and image data can achieve higher accuracy in detecting fake news by leveraging the complementary information provided by each modality. Textual analysis might identify linguistic patterns indicative of misinformation, while image analysis can reveal discrepancies between visual content and accompanying text. This synergistic approach allows for a more nuanced understanding of the content, making it more challenging for misleading or deceptive information to go undetected (Zhang et al., 2022).

Further advancements in deep learning techniques have facilitated the development of models capable of efficiently integrating and analysing multimodal data. For instance, recent work by Li et al. (2023) on multimodal transformers demonstrates that these models can simultaneously process and fuse information from text, images, and video to improve detection performance. The use of such models has shown promising results in various benchmarks, highlighting their potential to enhance fake news detection systems' effectiveness. By incorporating multimodal features, these systems improve accuracy and increase their ability to generalise across different types of misinformation, offering a more robust defence against fake news on social media platforms (Li et al., 2023).

4.3 Challenges in Detection

Despite advancements, challenges include detecting sophisticated fake news and balancing false positives and negatives. Ongoing research is focused on addressing these limitations (Zhang et al., 2022). Despite significant advancements in AI and machine learning techniques for fake news detection, several persistent challenges continue to impact the effectiveness of these systems. One of the primary challenges is detecting sophisticated fake news that employs advanced techniques to mimic legitimate content. As fake news creators become more adept at crafting misleading narratives that closely resemble credible information, AI systems must continuously evolve to identify increasingly sophisticated forms of deception. Recent studies, such as those by Zhang et al. (2022), highlight that deep learning models like BERT and GPT-3 have improved detection capabilities. However, they still struggle with highly nuanced and deceptive content that can evade current detection methods. This issue underscores the need for ongoing innovation in AI algorithms to keep pace with the evolving strategies used by misinformation spreaders.

Another significant challenge is balancing false positives and false negatives. In the context of fake news detection, a false positive occurs when a legitimate news article is incorrectly identified as fake. In contrast, a false negative occurs when the system does not detect a piece of fake news. Achieving a balance between minimising both types of errors is crucial, as high rates of false positives can undermine trust in the detection system. In contrast, high rates of false negatives can allow harmful misinformation to spread. Recent research has focused on refining model accuracy through adversarial training and enhanced feature extraction techniques, but achieving an optimal balance remains an ongoing challenge (Müller et al., 2021). This area of research is critical for improving the reliability and user trust in fake news detection systems.

Ongoing research is addressing these limitations through various approaches. Researchers are exploring more advanced model architectures and hybrid techniques that combine multiple AI methods to improve detection performance. For instance, studies are investigating the use of ensemble learning methods that aggregate predictions from multiple models to reduce errors (Chen et al., 2021). Additionally, incorporating more diverse and dynamic datasets, including real-time information and user-generated content, helps improve the system's ability to detect emerging trends in misinformation. By tackling these challenges, the field aims to develop more robust and accurate fake news detection systems capable of effectively combating sophisticated misinformation (Zhang et al., 2022; Wang et al., 2023).

5. Discussion

AI-driven fake news detection has made considerable progress but faces several challenges. Models like BERT and GPT-3 have significantly advanced the state-of-the-art, yet issues such as algorithmic bias and data privacy require ongoing attention. Additionally, the rapid evolution of misinformation tactics necessitates continuous adaptation of detection systems. Future research should improve model generalizability, address ethical concerns, and develop more transparent AI systems. The advancements in AI models, particularly with deep learning architectures like BERT and GPT-3, have significantly enhanced the accuracy and effectiveness of fake news detection systems. These models have demonstrated superior performance in identifying misinformation by leveraging their deep contextual understanding of language. BERT's bidirectional approach allows for a more nuanced interpretation of the text, improving its ability to detect subtle manipulations in news content (Devlin et al., 2019). Similarly, GPT-3's vast pre-training on diverse datasets has enabled it to generate and comprehend text with remarkable fidelity, enhancing its detection capabilities (Brown et al., 2020). These advancements underscore the potential of AI to address complex challenges in misinformation, providing more robust tools for combating the spread of fake news.

5.1 Integration of Multimodal Data

Integrating multimodal data, combining text with images and videos, substantially improves fake news detection. Traditional text-only models often struggle to capture the full context of misleading information, especially when multimedia elements play a significant role. By incorporating multimodal features, detection systems can leverage additional layers of information, such as visual cues and multimedia inconsistencies, to identify fake news more effectively (Karpathy et al., 2019). Studies have shown that models integrating these features achieve higher accuracy rates and are better equipped to handle sophisticated misinformation that spans multiple formats (Zhang et al., 2022). This approach enhances the detection capability and offers a more comprehensive understanding of the content, addressing the limitations of purely textual analysis.

5.2 Challenges in Detection

Despite these advancements, several challenges persist in fake news detection. One of the significant challenges is detecting highly sophisticated fake news that employs advanced techniques to mimic credible information. As misinformation creators become more adept at crafting realistic and convincing content, AI models must continuously evolve to keep pace (Zhang et al., 2022). This ongoing arms race between misinformation creators and detection systems highlights the need for continual innovation and adaptation in AI methodologies. Researchers must develop more sophisticated models that recognise and adapt to new deceptive techniques and content manipulation strategies.

5.3 Balancing False Positives and False Negatives

Another significant challenge is the balance between false positives and false negatives. False positives occur when legitimate news is incorrectly flagged as fake, while false negatives occur when fake news goes undetected. Achieving an optimal balance between these types of errors is crucial for maintaining user trust and ensuring the effectiveness of detection systems (Müller et al., 2021). High rates of false positives can lead to unjustly suppressing legitimate content, whereas high rates of false negatives can allow harmful misinformation to spread. Ongoing research focuses on refining detection algorithms to minimise errors and improve system performance.

5.4 Ethical Considerations

The ethical implications of AI in fake news detection cannot be overlooked. Privacy concerns arise from the need to collect and analyse large amounts of user data to train and operate detection systems. Ensuring these systems adhere to data protection regulations and respecting user privacy is essential (Binns, 2018). Additionally, algorithmic bias remains a significant issue, as biased models can lead to unfair or discriminatory outcomes. Addressing these ethical concerns requires the development of transparent and fair algorithms regularly audited for bias and fairness (Dastin, 2023). Researchers and practitioners must work together to establish ethical guidelines and practices for deploying AI in this sensitive area.

These ethical challenges can also be addressed through the lens of Islamic ethical principles, particularly the Maqasid al-Shari'ah (higher objectives of Islamic law). This framework emphasizes the preservation of essential values: faith, life, intellect, progeny, and property, providing a balanced approach to tackling modern technological dilemmas (Mohd Saifuddeen et al., 2013). For example, safeguarding user privacy aligns with the Maqasid goal of protecting life and property, underscoring the importance of stringent data governance practices to prevent the exploitation of sensitive information, which is considered a trust in Islamic teachings.

Furthermore, addressing algorithmic bias reflects the Islamic principle of fairness and societal equity. This requires the creation of diverse datasets and the implementation of fairness-aware learning techniques, ensuring AI systems do not perpetuate inequalities or disproportionately affect marginalized communities (Mohadi & Tarshany, 2023). Lastly, Islamic ethics emphasize truthfulness (ṣidq) and transparency to prevent misuse of technology. Deceptive or manipulative practices, prohibited under Islamic teachings, must be countered with ethical guidelines and oversight, ensuring AI systems promote societal harmony and uphold human dignity (Zubair et al., 2019).

By embedding these principles into AI systems, developers can ensure their work promotes ethical and equitable practices while adhering to universal standards of fairness and justice. This commitment to fairness is essential for building trust in AI technologies and fostering equitable detection systems that serve all users without bias.

5.5 Integration of Real-Time Data

Incorporating real-time data into fake news detection systems offers another avenue for improvement. The real-time analysis allows for the immediate identification and response to emerging misinformation, enhancing the system's ability to address rapidly spreading fake news (Wang et al., 2023). However, this approach also presents challenges, such as the need for efficient processing and the ability to handle large volumes of data quickly. Advances in real-time data processing and analysis technologies are crucial for maintaining the effectiveness of detection systems in dynamic and fast-paced information environments.

5.6 Hybrid and Ensemble Methods

Hybrid and ensemble methods have shown promise in improving fake news detection. By combining multiple AI techniques, such as deep learning models with traditional machine learning approaches, researchers can leverage the strengths of each method to enhance overall performance (Chen et al., 2021). Ensemble methods, which aggregate predictions from multiple models, can reduce errors and improve detection accuracy. These approaches provide a more robust solution by addressing the limitations of individual models and incorporating diverse perspectives on the data.

6. Conclusion

AI has significantly advanced the detection and combating of fake news on social media, offering improvements in accuracy, scalability, and real-time processing capabilities. However, each technique has its strengths and weaknesses, highlighting the need for benchmarking to evaluate their suitability for specific applications. Transformer-based models, such as BERT, demonstrate superior performance in textual misinformation detection due to their ability to understand contextual semantics. Meanwhile, multimodal approaches, which combine text and visual analysis, are more effective in detecting multimedia misinformation, though they often require higher computational resources.

Incorporating Islamic ethical principles into fake news detection is essential, as values such as honesty, trustworthiness, and justice emphasize the importance of truthfulness and accountability in communication. These principles guide the development of AI systems that prioritize accuracy, transparency, and community welfare, ensuring they promote social harmony and ethical communication practices.

Future research should aim to refine and expand detection techniques while addressing emerging challenges. First, creating diverse and representative datasets that encompass various types of misinformation will enhance the generalizability of detection models. Second, tackling new trends like deepfakes and synthetic media is critical for staying ahead of evolving deception methods. Lastly, interdisciplinary approaches that combine insights from AI, social sciences, and communication studies can offer a more holistic understanding of misinformation and inform innovative detection strategies. Regular benchmarking will remain crucial to tailoring techniques to specific real-world scenarios and ensuring the development of effective and ethical misinformation detection systems.

7. Acknowledgements

This paper is a research product supported and funded by the Fundamental Research Grant Scheme (FRGS) of the Ministry of Higher Education Malaysia (Code: FRGS/1/2022/ICT02/UKM/02/3).

References

Al-Kandari, A. A., & Al-Hajri, R. (2021). Exploring the role of artificial intelligence in fake news detection: A Maqasid Al-Shariah perspective. *Journal of Islamic Ethics*, 5(2), 67–85.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Al-Sarhan, S. (2020). Ethical frameworks for truthfulness in Islamic teachings: Addressing contemporary challenges of misinformation. *Arab Journal of Ethics*, 12(3), 197–215.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3528–3539. https://doi.org/10.18653/v1/D18-1389

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).*

Chen, J., Zhou, H., & Zhang, Z. (2021). Ensemble learning for fake news detection: A comprehensive review. *Journal of Data Science*, 19(1), 45–65.

Cui, L., Lee, D., & Sun, Y. (2020). CoAID: COVID-19 healthcare misinformation dataset. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2491–2500. https://doi.org/10.18653/v1/2020.emnlp-main.199

Dastin, J. (2023). Ethical implications of AI in social media. *Journal of AI Ethics*, 12(4), 203–215.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Higgins, J. P., Thomas, J., Chandler, J., et al. (2019). *Cochrane handbook for systematic reviews of interventions.* John Wiley & Sons. https://doi.org/10.1002/9781119536604

Karpathy, A., Johnson, J., & Fei-Fei, L. (2019). Visual semantic role labelling for real-world applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Karpathy, A., Li, X. L., & Fei-Fei, L. (2019). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019).*

Kumar, S., West, R., & Leskovec, J. (2021). False information on the web and social media: A review. *ACM Computing Surveys* (CSUR, 54(4), 1–40. https://doi.org/10.1145/3444690

Lazer, D. M., Baum, M. A., Benkler, Y., et al. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Mehrabi, N., Morstatter, F., Saxena, N., et al. (2019). A survey on bias and fairness in machine learning. *ACM Computing Surveys* (CSUR, 52(6), 1–35. https://doi.org/10.1145/3457604

Mohadi, M., & Tarshany, Y. M. A. (2023). Maqasid Al-Shari'ah and the ethics of artificial intelligence: Contemporary challenges. *Journal of Contemporary Maqasid Studies, 2*(2), 79–102. https://doi.org/10.52100/jcms.v2i2.107

Mohd Saifuddeen, S., Chang, L. W., Abdul Halim, I., & Nor Aina, M. K. (2013). Islamic ethical framework to tackle scientific and technological dilemmas. *Journal of Dharma, 38*(4), 373–386.

Moher, D., Liberati, A., Tetzlaff, J., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018). FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *Big Data*, 8(1),101–113. https://doi.org/10.1109/BigData.2018.8622202

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017*).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zubair, T., Raquib, A., & Qadir, J. (2019). Combating fake news, misinformation, and machine learning generated fakes: Insights from the Islamic ethical tradition. *Islam and Civilisational Renewal*, 10(2), 189–204.

Zubiaga, A., Liakata, M., Procter, R., Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3), e0150989. https://doi.org/10.1371/journal.pone.0150989

Biodata

<table>
<tr><td> </td><td><b>BORHAN BIN AB RAHMAN</b>

Borhan is pursuing a PhD at Universiti Kebangsaan Malaysia, specializing in Artificial Intelligence. His research focuses on detecting false information in texts related to Hadith. He aims to advance the field by developing machine-learning algorithms to enhance the accuracy and efficiency of these detection systems.</td></tr>
<tr><td> </td><td><b>ASSOCIATE PROF. DR. MOHD ZAKREE BIN AHMAD NAZRI</b>

Dr. Mohd Zakree is an Associate Professor at the Center for Artificial Intelligence and Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (FTSM, UKM). His areas of expertise include business intelligence, analytics, and decision support systems. His current research focuses on developing algorithms and methodologies for processing unstructured data.</td></tr>
<tr><td> </td><td><b>ASSOCIATE PROF. DR. MOHD RIDZWAN BIN YAAKUB</b>

Dr. Mohd Ridzwan is an Associate Professor at the Center for Artificial Intelligence and Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (FTSM, UKM). His expertise includes sentiment analysis, opinion mining, and online social network analysis. His current research primarily focuses on sentiment analysis and social network analysis.</td></tr>
</table>