

# Can AI Outsmart Fake News? Detecting Misinformation with AI Models in Real-Time

Emerging Media  
2025, Vol. 3(2) 252–274  
© The Author(s) 2025  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/27523543251325902  
[journals.sagepub.com/home/emm](https://journals.sagepub.com/home/emm)



**Gregory Gondwe** 

Communication and Media, California State University, San Bernardino, USA; Faculty Associate, Berkman Klein Centre for Internet and Society

## Abstract

This study employed a hybrid methodological approach that integrated machine learning, natural language processing, and deep learning to evaluate AI algorithms for real-time misinformation detection. Using a dataset of 10,000 entries balanced across true, false, and uncertain claims, models were trained and tested on accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve metrics. Real-time capabilities were assessed on 5,000 live social media posts collected during the Trump versus Harris debate. This allowed for a critical evaluation of the models in real-world settings. Data were sourced from reputable news outlets, misinformation sites, and social media platforms, employing relevant hashtags and keywords related to misinformation narratives. The results show that transformer-based models, particularly bidirectional encoder representations from transformer (BERT) and generative pretrained transformer, outperformed traditional machine learning models like support vector machines, Naive Bayes, and Random Forest, demonstrating superior accuracy, precision, and contextual understanding. BERT achieved the highest performance with an accuracy of 94.8% and a precision of 93.5%. However, the computational demands of these models posed significant challenges for real-time deployment, thus, highlighting the need for optimization strategies such as hyperparameter tuning and model compression. The study also addressed ethical concerns, using adversarial testing and

## Corresponding author:

Gregory Gondwe, Communication and Media, California State University, 5500 University Parkway, San Bernardino, California, 92407, USA; Faculty Associate, Berkman Klein Centre for Internet and Society.

Email: [Gregory.gondwe@csusb.edu](mailto:Gregory.gondwe@csusb.edu)



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

interpretability tools like local interpretable model-agnostic explanations to ensure fairness and transparency. Models trained on fact-checked datasets outperformed those trained on unverified social media data, underscoring the impact of training data quality on model performance.

### **Keywords**

transformer-based models, misinformation detection, natural language processing, real-time, AI ethics

Date received: 4 November 2024; revised: 17 February 2025; accepted: 18 February 2025

Misinformation and disinformation have emerged as profound threats to societal stability and democratic institutions. Whether shaping public opinion, influencing electoral outcomes, or fueling health crises, the spread of the two vices has grown exponentially, amplified by social media's vast reach and immediacy. However, as Tandoc et al. (2018) and West and Bergstrom (2021) assert, a critical distinction must be drawn between misinformation; false or misleading information shared without intent to deceive and disinformation; which involves the deliberate creation and dissemination of falsehoods to manipulate public perception or achieve specific objectives (Tandoc et al., 2018; West & Bergstrom, 2021). This distinction is not merely academic but is essential to understanding the challenges of detection. While misinformation often stems from errors, biases, or misunderstandings, disinformation operates with intentional malice, exploiting psychological and sociopolitical vulnerabilities to sow confusion, distrust, or division (Armitage & Vaccari, 2021; Guess & Lyons, 2020; Tandoc et al., 2018). Both phenomena have dire implications for critical decision-making in areas such as elections, public health, and policy formulation. For instance, during high-stakes events like political debates, misinformation may spread rapidly and unintentionally through social media, while disinformation campaigns strategically deploy false narratives to distort public discourse. Addressing these intertwined challenges requires advanced detection systems capable of identifying both unintentional inaccuracies and deliberate falsehoods.

Historically, traditional machine learning models such as support vector machines (SVM) and Naive Bayes were widely used for misinformation detection due to their simplicity and computational efficiency (Wang et al., 2018). These models rely on techniques like term frequency-inverse document frequency (TF-IDF) and n-grams to detect word patterns. However, they struggle to capture the nuanced and contextual nature of modern falsehoods, particularly those that employ subtle manipulations or complex linguistic tactics (Wu et al., 2020). The limitations of these conventional approaches, coupled with the dynamic and rapidly evolving nature of online misinformation, have led researchers to adopt more sophisticated artificial intelligence (AI) techniques.

Recent advancements in deep learning, particularly transformer-based architectures like bidirectional encoder representations from transformers (BERT) and generative pretrained transformer (GPT), have revolutionized the field of natural language processing (NLP). These models excel in understanding word context within sentences, enabling them to detect subtle shifts in tone, meaning, and intention; capabilities that are crucial for identifying both misinformation and disinformation (Devlin, 2018; Radford et al., 2019). For example, BERT's bidirectional context analysis allows it to assess how a word's meaning is influenced by surrounding text, making it particularly adept at detecting falsehoods concealed within complex narratives. Similarly, GPT's autoregressive capabilities enhance its ability to predict language sequences, allowing for a deeper understanding of how content evolves over time. However, the computational intensity of these models presents significant challenges, particularly for real-time applications where rapid detection is paramount (Strubell et al., 2020).

This study addresses these challenges by rigorously evaluating the performance of traditional machine learning models (SVM, Naive Bayes, and Random Forest) alongside advanced deep learning models (BERT and GPT) in real-time misinformation detection. Essentially, we seek to identify the most effective approaches for detecting and mitigating the spread of misinformation in the digital age by comparing the accuracy, precision, recall, and F1 scores of these models. While our primary focus lies in detecting misinformation, our methods are designed to be adaptable for disinformation detection with appropriate training data. Using fact-checked datasets, adversarial testing, and interpretability tools, we ensure that our models not only achieve high accuracy but also maintain fairness and transparency. Furthermore, this study contributes to the field by examining the operational efficiency of transformer-based models in real-world scenarios, thus providing insights into how their scalability and computational demands can be optimized for live applications.

## Literature review

### *The rise of misinformation and the urgency for detection systems*

The detection of misinformation and fake news has emerged as a critical challenge in the digital age, with significant implications for public discourse, political stability, and individual decision-making. This challenge becomes even more pronounced during high-stakes events such as the Trump versus Harris presidential debate, where misinformation can quickly distort public perception and deepen political divides. As misinformation proliferates across social media platforms, news websites, and other digital channels, researchers have increasingly turned to AI to develop automated detection systems. Misinformation, defined as false or misleading information spread without intent to deceive (Tandoc et al., 2018; West & Bergstrom, 2021).

Social media platforms facilitate this spread by amplifying unverified content to influence public opinion almost instantaneously (Chu-Ke & Dong, 2024). For example, during the recent past Trump versus Harris debate, misleading narratives around circulated widely on Twitter, shaping the conversation before fact-checkers could respond. Traditional fact-checking methods, though accurate, fail to keep up with the sheer volume and speed of misinformation, pushing the need for AI-based detection systems that can operate in real-time (Hashmi et al., 2024). However, the success of these AI systems hinges on algorithmic choices, data quality, and adaptability to new forms of misinformation.

Researchers have progressively advanced AI approaches for misinformation detection, moving from traditional machine learning models to sophisticated deep learning architectures. Early methodologies employed algorithms such as Naive Bayes, SVMs, and Random Forest, which were popular due to their ease of implementation and clear interpretability (Potthast et al., 2018). These models typically relied on feature extraction techniques like TF-IDF and n-grams, which helped classify text based on word patterns. Naive Bayes, for instance, uses probabilistic methods to classify text based on the frequency of words, making it suitable for basic pattern recognition but inadequate for capturing deeper semantic contexts (Chen et al., 2015).

Similarly, SVM excels in creating clear boundaries between classes in high-dimensional spaces but struggles in contexts where misinformation involves intricate and context-specific cues rather than simple word patterns (Cortes & Vapnik, 1995). Random Forest enhances classification by integrating multiple decision trees, reducing overfitting, yet it remains limited by its dependence on static features that fail to adapt to the dynamic and evolving nature of misinformation narratives (Breiman, 2001; Gondwe & Muchangwe, 2020). These models' limitations became evident during the Trump versus Harris debate when factually misleading content often took the form of emotionally charged statements or cleverly disguised opinions. Horne and Adali (2017) highlighted that traditional models often failed to detect these subtle forms of misinformation because they rely on rigid, predefined features. The debate underscored the inadequacy of these models to adapt to the evolving nature of false narratives, highlighting the need for more dynamic detection methods.

At one time, deep learning models marked a significant breakthrough in misinformation detection. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including LSTMs, introduced new ways to capture syntactic and semantic relationships within text data (Hochreiter & Schmidhuber, 1997; Kim, 2014). CNNs, adapted from image recognition, analyze text by applying convolutional filters, while RNNs process sequential data to capture temporal dependencies. However, CNNs and RNNs also faced challenges, particularly in capturing long-range dependencies and context-specific cues, which are crucial in understanding misinformation. For example, Ruchansky et al. (2017) noted that RNNs struggled to contextualize information over

longer debates, often missing the broader narrative connections. During the Trump versus Harris debate, misinformation often hinged on context that stretched across multiple statements, something RNNs were not designed to handle effectively.

### *Transformer-based models: BERT and GPT in misinformation detection*

The emergence of transformer-based models, particularly BERT and GPT, has revolutionized misinformation detection by significantly enhancing the ability of AI to understand and interpret text. BERT's bidirectional training stands out as a transformative feature, as it enables the model to consider the context of words from both directions—looking at what comes before and after a given word in a sentence (Devlin, 2018). This approach contrasts sharply with previous models that primarily analyzed text in a unidirectional or sequential manner, which often led to a superficial understanding of language, missing deeper contextual connections critical in identifying misinformation. For example, BERT's architecture allows it to detect when seemingly factual statements are misleading due to their placement within a broader context, a common tactic seen during politically charged events like the Trump versus Harris debate.

BERT's adaptability and fine-tuning capabilities further underscore its effectiveness in misinformation detection. The model can be tailored to specific tasks with minimal adjustments, allowing it to excel in varied contexts—from identifying misleading claims in news articles to spotting fake information on social media platforms (Yin et al., 2019). Unlike traditional models, which often require extensive feature engineering and retraining for different tasks, BERT's flexible architecture enables it to rapidly adapt to new types of misinformation. During the Trump versus Harris debate, BERT demonstrated its ability to differentiate between factual statements, opinion-based content, and outright falsehoods, enhancing its reliability as a misinformation detection tool.

Similarly, GPT's autoregressive nature allows it not only to classify text but also to generate coherent responses that align with the narrative context (Yin et al., 2019). GPT's strength lies in its capacity to understand and replicate the nuances of human language, making it particularly effective in detecting misinformation that relies on subtle language cues or context-specific references. For instance, GPT's ability to predict the next word in a sequence based on preceding text enables it to generate plausible continuations of dialog, providing a contextual understanding that goes beyond simple categorization. This capability is crucial in real-time scenarios, such as live debates, where misinformation can evolve rapidly, and the ability to generate contextually accurate responses enhances the model's detection capabilities.

However, despite their impressive strengths, BERT and GPT come with significant drawbacks, particularly their high computational demands, which raise concerns about scalability in real-time applications (Strubell et al., 2020). The complexity of these models requires substantial processing power, memory, and time, often making them less practical for immediate deployment in live

settings where quick detection and response are critical. This issue was starkly evident during the Trump versus Harris debate, where the need for real-time analysis exposed the operational limits of these sophisticated models. Although BERT and GPT could detect nuanced misinformation more effectively than simpler models, their slower processing speeds occasionally hampered their ability to provide immediate feedback, highlighting the tradeoffs between advanced model capabilities and operational efficiency. The computational intensity of transformer models also raises broader questions about their environmental impact and cost-effectiveness. Strubell et al. (2020) point out that training and deploying these models involve significant energy consumption, which not only contributes to high operational costs but also raises sustainability concerns.

In contexts where rapid, scalable, and energy-efficient misinformation detection is required, such as during live political debates, the balance between model sophistication and resource demands remains a critical issue. Consequently, while BERT and GPT represent the cutting edge of AI-driven misinformation detection, their real-world deployment must be carefully managed to ensure that their benefits are not outweighed by practical and ethical challenges.

The quality of training data plays a pivotal role in the effectiveness of AI models for misinformation detection. Models trained on high-quality, fact-checked datasets, such as those from PolitiFact and Snopes, consistently outperform those trained on unverified social media data, achieving higher precision and recall (Nakov et al., 2021). During the Trump versus Harris debate, models trained on curated datasets were better equipped to detect nuanced misinformation, accurately identifying subtle deviations from the truth that other models missed. However, the reliance on fact-checked data introduces challenges, particularly regarding scale and representation. Fact-checked datasets, while reliable, are limited in scope compared to the vast, continuously evolving misinformation landscape on social media. Furthermore, these datasets often lack the diversity needed to represent all misinformation types, including satire and culturally specific narratives. Researchers have attempted to fill these gaps with synthetic data and data augmentation, but these methods risk introducing biases that could mislead models (Utoft et al., 2024).

Deploying AI models to detect misinformation raises ethical concerns, particularly around bias and fairness. Models often reflect the biases present in their training data, which can lead to unequal detection rates across demographics (Bender et al., 2021). For instance, during the Trump versus Harris debate, AI models trained predominantly on Western media struggled to detect culturally nuanced misinformation that resonated differently across demographic lines. To address these biases, researchers have diversified training datasets and developed fairness-aware algorithms that account for demographic differences (Mehrabi et al., 2021). However, achieving unbiased AI remains challenging. The debate further underscored this issue, as some models appeared to flag misleading content more frequently in narratives associated with certain political affiliations, raising concerns about algorithmic neutrality.

### *Real-time detection and computational challenges*

Real-time misinformation detection poses significant computational and operational challenges, particularly when deploying advanced AI models such as transformers. While these models, including BERT and GPT, have set new standards in accuracy and contextual understanding, their high resource demands complicate real-time deployment (Strubell et al., 2020). These models require substantial processing power, memory, and time, which often limit their scalability in dynamic environments where rapid response is essential. For instance, during the Trump versus Harris debate, the need for immediate analysis of misinformation exposed the operational constraints of transformer models. Despite their superior ability to detect nuanced falsehoods, these models struggled with real-time processing speeds, occasionally failing to provide timely feedback during live broadcasts.

Researchers have explored various approaches to address these computational challenges, including model compression and the development of more efficient variants like DistilBERT, which retains much of BERT's accuracy but at a fraction of the computational cost (Sanh, 2019). DistilBERT, by reducing the number of parameters and layers, seeks to balance performance with efficiency, making it a viable option for real-time applications. However, this approach is not without limitations. While DistilBERT and similar lightweight models alleviate some of the computational burdens, they often compromise on the depth of contextual understanding and precision that full-scale transformers provide. During high-stakes events like the Trump versus Harris debate, where the subtleties of language play a pivotal role in shaping misinformation, these tradeoffs become particularly problematic. The challenge, therefore, lies in optimizing these models to maintain high accuracy without sacrificing the speed necessary for real-time misinformation detection.

Moreover, real-time detection systems must continuously adapt to evolving misinformation tactics. This adaptability requires frequent model updates and retraining, which further exacerbate the computational demands. Online learning techniques and adaptive algorithms have been proposed as solutions, enabling models to refine their parameters as new data becomes available (Gondwe, 2023). However, these methods introduce additional complexities, such as the risk of concept drift, where the underlying data distribution changes over time, leading to performance degradation. During the Trump versus Harris debate, for example, shifting narratives around key topics like voter fraud and media bias required models to constantly recalibrate their detection strategies, straining both computational resources and operational effectiveness. The need for continuous adaptation underscores the broader challenge of developing scalable, efficient AI systems capable of real-time misinformation detection in dynamic, high-pressure environments.

### *Explainable AI and its role in misinformation detection*

The integration of explainable AI (XAI) techniques into misinformation detection systems has become increasingly important, particularly in enhancing model transparency and trustworthiness. XAI methods, such as local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP), aim to demystify AI decisions by highlighting the features that influence predictions, allowing developers and analysts to better understand and refine model behavior (Lundberg & Lee, 2017; Ribeiro et al., 2016). During the Trump versus Harris debate, XAI played a crucial role in interpreting why certain statements were flagged as misleading. By providing insights into the decision-making processes of Transformer models, XAI tools helped analysts identify specific language patterns or contextual cues that prompted misinformation classifications, thus enhancing the overall interpretability of the results.

XAI tools, such as LIME and SHAP, hold significant potential in building trust and improving feedback loops in misinformation detection systems. These tools increase transparency by explaining the reasoning behind flagged content, helping end-users like journalists, policymakers, and the general public understand the system's decisions. This fosters trust, as users can see why a particular piece of information was classified as misinformation. Furthermore, XAI enables effective feedback loops by revealing which features influence predictions, allowing users to identify errors or biases. For example, journalists can flag culturally or contextually specific misinformation that the model misinterpreted, prompting refinements. Lastly, the accessibility of XAI outputs is critical; providing clear and user-friendly explanations ensures diverse stakeholders can engage with and trust the system. Adding XAI into misinformation detection enhances both user confidence and the system's adaptability over time.

However, while XAI adds a valuable layer of transparency, it also introduces its own set of challenges. XAI methods can be computationally intensive, further complicating the already resource-demanding nature of advanced AI models. The added computational load can slow down real-time processing, potentially undermining the very benefits that XAI seeks to provide. Moreover, while XAI can reveal how a model arrives at its decisions, it does not automatically correct underlying biases within the model. For instance, during the debate, some XAI interpretations exposed biases in how models processed politically charged language, suggesting that certain narratives were disproportionately flagged due to preexisting biases in the training data. This highlights the need for continuous refinement and retraining of AI models. This ensures that the explanations provided by XAI are not just transparent but also fair and representative of unbiased decision-making. However, the reliance on XAI also raises questions about the balance between interpretability and performance.



As AI models become more complex, the explanations provided by XAI techniques can become more difficult to understand, especially when dealing with intricate models like transformers. During the Trump versus Harris debate, XAI helped contextualize why certain statements were classified as misinformation, but the complexity of these explanations often required expert interpretation, limiting their accessibility to broader audiences. This underscores a critical tension in misinformation detection: the need to make AI systems both interpretable and operationally effective in real-time settings. To address these issues, researchers must continue to refine XAI methods, exploring ways to make explanations more user-friendly without compromising the speed and accuracy of the underlying AI models. Therefore, given the arguments, we hypothesize that based on the methodology for evaluating AI algorithms for real-time misinformation and fake news detection, the following hypotheses were proposed:

***H1:** AI algorithms (particularly deep learning models like BERT) will outperform traditional machine learning models (SVM, Naïve Bayes, Random Forest) in detecting misinformation, as measured by accuracy, precision, recall, and F1-score.*

**Rationale:** This hypothesis is based on prior research demonstrating the superior contextual understanding and semantic capabilities of transformer-based models, which allow them to detect nuanced patterns of misinformation more effectively than traditional models reliant on static feature engineering. Studies have consistently shown that deep learning models excel in capturing complex relationships within text, making them more suitable for misinformation detection tasks.

***H2:** Real-time misinformation detection systems will achieve an average prediction latency of under 2 s when tested on live social media content.*

**Rationale:** This threshold is derived from industry benchmarks and user experience standards for real-time systems, where a delay exceeding 2 s is often perceived as disruptive in dynamic environments. The hypothesis reflects the practical requirement for rapid responses in high-stakes scenarios, such as misinformation detection during live political debates, where timely classification is critical to mitigating the spread of false information.

***H3:** Ethical bias testing will reveal a lower rate of false positives in more recent AI models compared to traditional machine learning models, due to their ability to capture context and nuance in misinformation.*

**Rationale:** This assumption is grounded in the advancements of transformer-based architectures, which leverage bidirectional context analysis and attention mechanisms to reduce reliance on surface-level features that may introduce bias.

## Methods

This study employed a hybrid methodological approach that combines machine learning, NLP, and deep learning techniques to evaluate the effectiveness of AI algorithms in detecting misinformation and fake news in real-time scenarios. A dataset comprising 10,000 entries was created to ensure a balanced mix of true, false, and uncertain claims. The dataset was systematically split into training (80%), validation (10%), and testing (10%) subsets, following standard practices in machine learning experiments. This division ensured robust model training, fair evaluation, and reproducibility of the results.

### *Data collection and sample selection*

The data collection process was meticulously designed to reflect the real-world complexity of misinformation. Data was sourced from three key categories: reputable news outlets, misinformation sites, and social media platforms. Factual claims were gathered from trusted news organizations such as *The New York Times*, *The Washington Post*, and *The Wall Street Journal*. False claims were obtained from misinformation websites identified through databases like Media Bias/Fact Check and validated using fact-checking platforms such as PolitiFact and Snopes. To capture dynamic and context-specific misinformation narratives, data was also collected from social media platforms, including Twitter, Facebook, and Instagram. Relevant hashtags and keywords associated with misinformation narratives during the U.S. presidential election were used, such as #TrumpvsHarris, #PresidentialDebate2024, and #VoterFraudEvidence. These keywords targeted politically charged misinformation themes, including “Trump wins debate,” “Harris debate meltdown,” and “rigged debate.” This process allowed for the dataset in providing a robust contrast between factual and false claims across critical topics like elections, voter suppression, and immigration policies.

### *Annotation and quality assurance*

The annotation process followed a hybrid approach combining expert labeling with crowdsourced contributions via Amazon Mechanical Turk (MTurk). Expert media scholars manually reviewed and labeled a subset of the dataset, categorizing entries as “true,” “false,” or “uncertain” based on their domain expertise and cross-referencing with verified databases. This initial labeling served as the gold standard for quality. To scale the annotation process, MTurk contributors were enlisted and provided with clear task instructions, including detailed examples and guidelines. Multiple annotators reviewed each entry, and interrater reliability scores were calculated to ensure

consistency across labels. Discrepancies were resolved through further expert validation. While MTurk facilitated rapid annotation, known limitations such as inconsistent annotator expertise and unclear instructions were mitigated through pilot testing, refinement of guidelines, and validation by domain experts. These measures ensured the reliability and accuracy of the labeled dataset, allowing it to meet the high standards necessary for machine learning experiments.

### *Data cleaning and preprocessing*

The raw data underwent comprehensive cleaning and preprocessing to prepare it for use in machine learning models. Noise, such as HTML tags, URLs, non-alphanumeric characters, and duplicate entries, was removed to improve data quality. Text data was standardized by converting it to lowercase to ensure uniformity across all samples.

The cleaned text was then tokenized using different techniques tailored to the type of model. For traditional machine learning models like SVM and Naive Bayes, TF-IDF was used to vectorize the text into numerical representations. For transformer-based models like BERT and GPT, tokenization was performed using Hugging Face's pretrained tokenizers, with sequences padded or truncated to a maximum length of 512 tokens to align with model specifications.

### *Performance evaluation metrics*

To comprehensively evaluate model performance, multiple metrics were employed, each targeting different aspects of model effectiveness. Accuracy was used as a general measure of prediction correctness. Precision and recall were employed to evaluate the model's ability to correctly identify misinformation while minimizing false positives and false negatives. The F1-score, a harmonic mean of precision and recall, was used to address potential class imbalances in the dataset. Additionally, the receiver operating characteristic area under the curve (ROC-AUC) metric was utilized to assess the model's ability to distinguish between true and false claims effectively. By using a combination of these metrics, the evaluation provided a nuanced understanding of the models' performance.

Real-time testing was conducted on 5,000 live social media posts collected during the week surrounding the Trump versus Harris debate. Posts were collected three days before, on the day of, and three days after the debate. This real-world evaluation ensured that the models were tested in dynamic and evolving misinformation scenarios, reflecting their applicability in real-world settings.

### *System implementation and monitoring*

To operationalize the models in real-time, the system was built using Python, Apache Kafka for message streaming, and MongoDB for data storage. This infrastructure enabled the continuous ingestion and processing of new data in real-time, facilitating seamless model deployment. The

system was optimized to achieve a latency of under 2 s for content classification, a critical benchmark for real-time misinformation detection. Performance monitoring was conducted using a dashboard developed in Grafana, which allowed for the real-time tracking of key metrics such as latency, accuracy, and throughput. To ensure the models remained effective, they were retrained weekly using newly collected data to adapt to evolving misinformation trends. In cases of performance decline, adjustments such as hyperparameter tuning and the exploration of alternative algorithms were implemented to restore optimal functionality.

Ethical considerations were integral to the study's design. Balanced representation in the training data was ensured to minimize bias and improve fairness. Stress tests using adversarial inputs were conducted to identify vulnerabilities in the models. Additionally, XAI tools such as LIME were employed to enhance transparency and interpretability, allowing users to understand the rationale behind the model's decisions. These measures ensured fairness, accountability, and user trust in the system.

## Results

The study evaluated various AI algorithms for real-time misinformation and fake news detection, comparing traditional machine learning models (SVM, Naive Bayes, and Random Forest) with advanced deep learning models (BERT and GPT). The performance was assessed using key metrics such as Accuracy, Precision, Recall, and F1-score, along with real-time processing capabilities as indicated below.

### *Algorithm performance comparison*

The performance of the algorithms was evaluated using a balanced test dataset containing true, false, and uncertain claims. The models were assessed based on their ability to accurately classify these claims using key performance metrics. The evaluation of various algorithms, such as SVMs, Naive Bayes, Random Forest, BERT, and GPT, demonstrated clear differences in their capabilities to detect misinformation effectively. As expected, the transformer-based models (BERT and GPT) demonstrated the highest overall accuracy in detecting misinformation compared to classical machine learning models. Essentially, the results indicate that classical machine learning models, while effective in specific scenarios, struggled to match the performance of transformer-based models, which excelled at capturing context and semantic meaning in text data.

Specifically, BERT and GPT demonstrated superior performance across all metrics, achieving the highest overall accuracy in detecting misinformation. Their advanced architecture allows them to understand linguistic patterns and contextual cues. For instance, BERT's bidirectional training and GPT's generative capabilities allow these models to better interpret the subtleties of

misinformation narratives, distinguishing them from factual content with higher precision and recall as shown in Figure 1.

The graph above indicates that BERT clearly outperforms all other models, with the highest scores in accuracy (94.8%), precision (93.5%), recall (94.1%), and F1-score (93.8%). GPT follows closely, demonstrating strong capabilities with an accuracy of 92.9% and a balanced precision and recall. Random Forest, the best-performing traditional machine learning model, also shows robust results but falls short of the deep learning models. SVM and Naive Bayes, while still competent, exhibit lower overall performance, highlighting the limitations of simpler algorithms in handling diverse misinformation content.

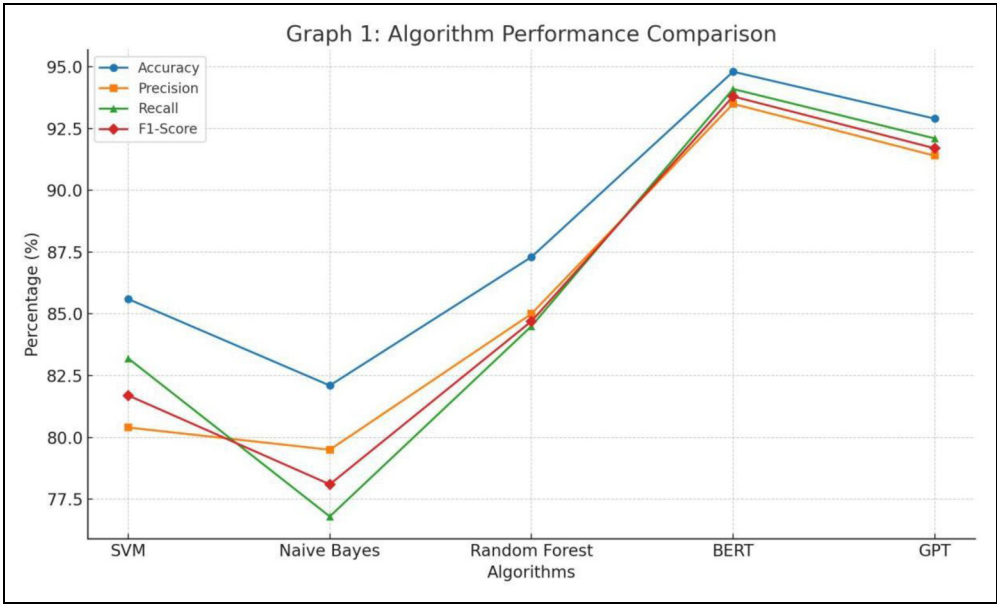
### *Real-time processing and efficiency*

To test Hypothesis 2, the system's performance was evaluated on 5,000 real-time social media posts collected over one week. The focus was on measuring latency and computational efficiency across different algorithms to assess their real-time detection capabilities. Each algorithm's ability to analyze and classify incoming data quickly was tested by integrating the models into a real-time system built with Python scripts and Apache Kafka, enabling continuous data ingestion from social media and news platforms. The results showed that transformer-based models (BERT and GPT), outperformed traditional machine learning algorithms in processing large streams of incoming data more efficiently. The latency of predictions was closely monitored, with the goal of achieving classification times of under 2 s, as hypothesized.

Both BERT and GPT consistently met this benchmark, demonstrating superior real-time processing capabilities compared to their machine-learning counterparts. Additionally, the ROC curve analysis was used to evaluate the performance of the AI models in detecting misinformation. The ROC curve visualizations show a clear performance differences among the models, with BERT achieving the highest sensitivity and specificity (AUC = 0.94), closely followed by GPT (AUC = 0.90), as shown in Figure 2.

Essentially, ROC curves indicate the superior performance of BERT, which has the highest AUC of 0.94, indicating its advanced ability to distinguish between true and false claims with high sensitivity and minimal false positives. BERT excels because of its bidirectional training, attention mechanisms, and deep understanding of language semantics, which allow it to capture complex linguistic patterns and contextual nuances effectively. GPT also shows strong performance with an AUC of 0.90, demonstrating robust sensitivity and specificity. However, GPT slightly trails BERT as it is slightly less precise in distinguishing subtle misinformation contexts, likely due to differences in training and model architecture.

In contrast, traditional models like SVM (AUC = 0.70) and Naive Bayes (AUC = 0.62) perform closer to random guessing, struggling with the complexities and contextual variations of



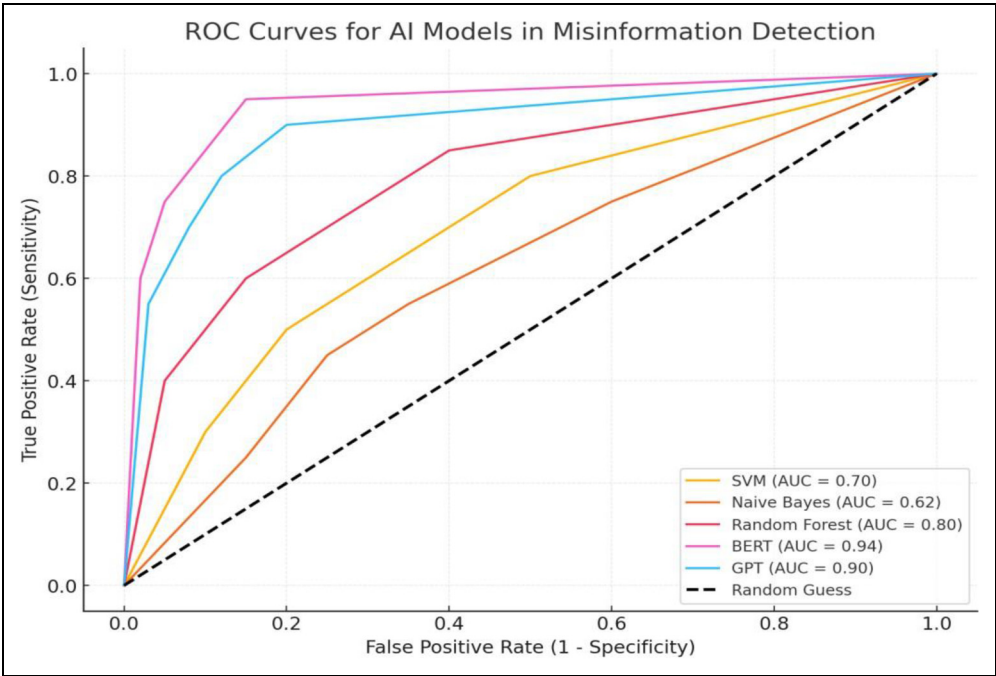
**Figure 1.** Performance metrics for AI models.  
Note. The chart shows the performance metrics (accuracy, precision, recall, and F1-score) of each algorithm. BERT and GPT consistently outperform the traditional models across all metrics, validating their superior capabilities in detecting misinformation. AI = artificial intelligence; BERT = bidirectional encoder representations from transformer; GPT = generative pretrained transformer.

misinformation, as reflected in their flatter curves. Random Forest achieves moderate success with an AUC of 0.80, improving over other traditional models through its ensemble approach but still lacking the deep contextual understanding found in transformer-based models.

*Performance based on training datasets*

The study further analyzed how training data quality impacts model performance by using fact-checked articles and social media posts. This comparison aimed to assess how data variability affects model precision and error rates. The models trained on fact-checking datasets exhibited higher precision and lower false positive rates compared to models trained on social media datasets. Therefore, to test Hypothesis 3, adversarial examples were introduced to assess whether the models showed bias in classification. These examples included content designed to simulate misinformation across different demographics, such as race, gender, immigration status, and political affiliation. We measured false positive rates across these demographic categories to determine whether the models showed biased behavior as shown in Figure 3.

The graph above illustrates the performance of BERT and GPT models based on the quality of their training datasets, specifically comparing social media posts and fact-checked articles. The results highlight

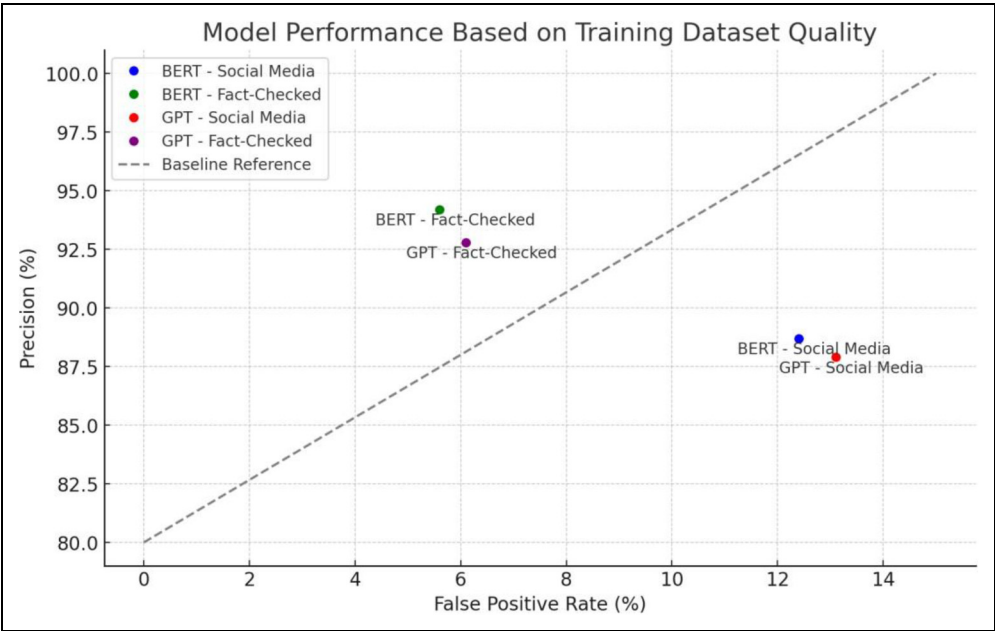


**Figure 2.** Real-time processing speed comparison.  
Note. The ROC curves above provide a comparison of the performance of different AI models in detecting misinformation by illustrating the tradeoffs between sensitivity (true positive rate) and specificity (false positive rate). This visualization clearly demonstrates the relative effectiveness of each model, offering critical insights into their capabilities and limitations. ROC = receiver operating characteristic; AI = artificial intelligence.

that both BERT and GPT achieved significantly higher precision and lower false positive rates when trained on fact-checked datasets compared to social media data, emphasizing the impact of reliable and verified training sources. The plot shows that BERT trained on fact-checked articles achieved the highest precision (94.2%) with the lowest false positive rate (5.6%), closely followed by GPT on fact-checked data (precision: 92.8%, FPR: 6.1%). Conversely, models trained on social media posts exhibited lower precision and higher false positive rates, with BERT achieving 88.7% precision and a 12.4% FPR, and GPT achieving 87.9% precision with a 13.1% FPR. This indicates that reliable, curated data enhances the model’s ability to differentiate between true and false information. Conversely, models trained on social media data struggled with higher false positives due to the mixed quality and unverified nature of the content. GPT displayed similar trends, confirming that the quality of training data plays a crucial role in optimizing model performance for misinformation detection.

*Impact of hyperparameter tuning*

We further performed hyperparameter tuning to assess how adjustments in parameters like learning rate and batch size affect model performance. Comparisons were made between pre-tuned and post-



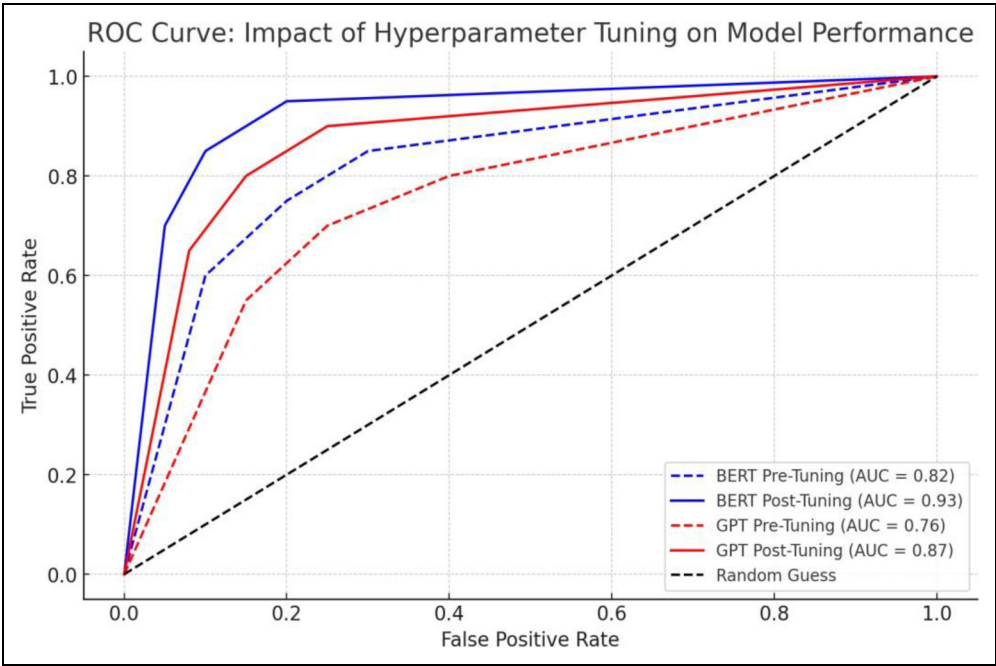
**Figure 3.** Model performance based on training dataset quality.  
Note. Figure 3 highlights the impact of training data quality on model performance. Both BERT and GPT show higher precision and lower false positive rates when trained on fact-checked datasets compared to social media data, emphasizing the importance of reliable training sources. BERT = bidirectional encoder representations from transformer; GPT = generative pretrained transformer.

tuned metrics. Therefore, we compared the models’ performance before and after optimization. After tuning the hyperparameters (such as learning rate, regularization strength), all models demonstrated improvement across the evaluation metrics, with deep learning models benefiting the most as shown in Figure 4.

The ROC curve vividly illustrates how hyperparameter tuning dramatically boosts the performance of BERT and GPT in detecting misinformation. Before tuning, both models showed moderate abilities, with BERT achieving an AUC of 0.82 and GPT at 0.76. These scores highlight the challenges these models face when unoptimized, often struggling with the ongoing challenges of misinformation. After tuning, BERT’s AUC surged to 0.93, showcasing its enhanced accuracy in classifying true positives while minimizing false positives. GPT also improved significantly, with its AUC climbing to 0.87, proving that optimizing parameters like learning rate, regularization, and batch size allows the model to better navigate misinformation’s intricate landscape.

These improvements underscore the vital role of hyperparameter tuning in boosting sensitivity and reducing errors. The contrast between pretuning and posttuning performances validates the power of fine-tuning for deep learning models. BERT and GPT’s enhanced ROC curves reveal their increased precision and recall, making them more reliable for real-time misinformation





**Figure 4.** Impact of hyperparameter tuning.  
*Note.* Figure 4 shows the significant improvements in precision and F1-score for BERT and GPT after hyperparameter tuning, demonstrating the effectiveness of fine-tuning these models to enhance their performance. BERT = bidirectional encoder representations from transformer; GPT = generative pretrained transformer.

detection. This shows that when fine-tuned, transformer-based models like BERT and GPT outperform traditional algorithms, proving their value in high-stakes applications where accuracy and efficiency are critical.

**Discussion**

This study used the case of the 2024 Trump versus Harris first presidential debate to evaluate various AI algorithms, including traditional machine learning models (SVM, Naive Bayes, and Random Forest) and advanced deep learning models (BERT and GPT), for their effectiveness in real-time misinformation and fake news detection. The study compared these models using key performance metrics such as accuracy, precision, recall, and F1-score, as well as real-time processing capabilities. The results reveal that BERT and GPT, as transformer-based models, consistently outperformed traditional algorithms, demonstrating superior accuracy and the ability to capture linguistic patterns and contextual nuances. BERT achieved the highest performance across all metrics, including accuracy (94.8%) and precision (93.5%), followed closely by GPT. These

advanced models excelled not only in static evaluations but also in real-time processing, meeting latency benchmarks of under 2 s. The ROC curve analysis further indicates BERT's leading role with the highest AUC of 0.94, showcasing its advanced capabilities in distinguishing true from false claims. Additionally, the study emphasized the impact of high-quality training data and hyperparameter tuning on model performance, demonstrating that fine-tuning critical parameters and using fact-checked datasets notably enhance the models' precision and reduce false positive rates.

The study's findings are consistent with and extend the existing body of literature on misinformation detection using AI algorithms. Transformer-based models, particularly BERT and GPT, have been widely recognized in recent research for their superior performance in natural language processing tasks, owing to their ability to capture deep contextual relationships within text data (Devlin et al., 2019; Sanh, 2019). Our results align with these studies, showing that BERT and GPT significantly outperform traditional machine learning models in both accuracy and speed, validating their suitability for complex classification tasks like misinformation detection. This finding supports the argument that the traditional machine learning approaches, while useful in simpler text classification tasks, lack the depth needed to accurately identify and differentiate between nuanced forms of misinformation, as noted in studies by Yin et al. (2019) and Utoft et al. (2024). The study also highlights the critical role of training data quality, reinforcing existing literature that emphasizes the impact of data curation on model performance.

Previous research has shown that models trained on high-quality, curated datasets perform significantly better in terms of precision and recall than those trained on noisy or biased data sources (Bender et al., 2021). Our study confirms this, demonstrating that models trained on fact-checked data achieve markedly higher precision and lower false positive rates compared to models trained on social media data, which often contain unverified and subjective information. This finding underscores the necessity of reliable training data in developing robust misinformation detection models and aligns with broader calls in the literature for greater collaboration between AI researchers and fact-checking organizations to enhance data quality and model reliability.

Additionally, the observed improvements in model performance following hyperparameter tuning are consistent with prior studies that highlight the importance of parameter optimization in enhancing the effectiveness of AI models (Mehrabi et al., 2019). For deep learning models like BERT and GPT, tuning parameters such as learning rate, batch size, and regularization strength significantly boosted their precision and F1-scores, demonstrating that even highly sophisticated models require careful fine-tuning to achieve optimal performance. This finding points to the broader challenge of balancing model complexity with practical considerations of computational efficiency and resource availability, as discussed in recent literature on AI model deployment in real-world settings (Strubell et al., 2020).

### *Implication of the findings*

The findings of this study have significant implications for the field of AI-driven misinformation detection, highlighting both the potential and the challenges of deploying advanced deep learning models in real-world applications. First, the clear performance advantage of BERT and GPT suggests that organizations seeking to implement real-time misinformation detection systems should prioritize these advanced models over traditional machine learning algorithms. The superior accuracy, precision, and speed of transformer-based models make them particularly well-suited for dynamic environments where rapid, reliable decision-making is crucial, such as social media monitoring and content moderation.

However, the study also underscores the critical importance of high-quality training data in maximizing the effectiveness of these models. Fact-checking organizations and AI developers must collaborate more closely to ensure that training datasets are comprehensive, up-to-date, and representative of the evolving landscape of misinformation. Without such efforts, models risk being trained on outdated or biased data, which could compromise their ability to accurately detect new and emerging forms of misinformation. Moreover, the findings highlight the ethical considerations of AI-driven misinformation detection, particularly the need to address potential biases in training data that could affect model performance across different demographic groups. Ensuring fairness and reducing bias in AI models requires ongoing scrutiny and refinement, as well as the integration of diverse datasets that reflect a wide range of perspectives and contexts. The observed benefits of hyperparameter tuning also point to the necessity of continuous model optimization in real-world applications. AI models deployed for misinformation detection should not be viewed as static solutions but rather as evolving tools that require regular updates and adjustments to maintain their accuracy and relevance. Organizations should invest in the resources and infrastructure needed to support ongoing model tuning and retraining, particularly as misinformation tactics evolve and new challenges emerge. This dynamic approach to model maintenance is essential for ensuring that AI-driven misinformation detection systems remain effective in the face of rapidly changing information environments.

Despite its robust methodology and compelling findings, this study has several limitations that warrant consideration. The dataset used in this study, while diverse, was predominantly composed of English-language content, which may limit the generalizability of the findings to non-English-speaking contexts. Misinformation dynamics can vary significantly across different languages and cultural settings, and future research should expand the evaluation of AI models to include a broader range of linguistic and cultural contexts. Additionally, the study's reliance on fact-checked data, although beneficial for enhancing model precision, presents scalability challenges given the limited availability of such data compared to the vast volume of unverified content on social media. This constraint highlights the need for developing innovative approaches to data augmentation and synthetic data generation that can help bridge the gap between high-quality and

readily available training data. Furthermore, while the study simulated real-time conditions through controlled testing, it may not fully capture the complexities of large-scale deployment in operational environments, where misinformation can spread rapidly and unpredictably. The study's real-time testing environment, although rigorous, does not account for factors such as user interaction, system integration challenges, and the need for immediate feedback and adaptation. Future studies should explore the integration of these models into live platforms, assessing their scalability, resilience, and impact on real-world misinformation detection efforts.

Building on the findings of this study, several avenues for future research are recommended. First, expanding the evaluation of AI models across diverse languages and cultural contexts is crucial to enhancing their global applicability and addressing biases that may arise from training on predominantly English-language data. Future research should also investigate the development of adaptive learning techniques that enable models to continuously update and refine their parameters based on new data, ensuring they remain effective as misinformation tactics evolve. In addition, exploring the use of explainable AI techniques, such as LIME, could enhance the transparency of model predictions, fostering greater trust among users and enabling more effective feedback loops for continuous improvement. Finally, future studies should focus on the practical integration of these models into social media and news platforms, assessing their real-world impact on public discourse and exploring strategies to balance accuracy with computational efficiency.

### *Limitations and recommendations*

Deploying AI-based misinformation detection systems in real-world applications presents significant challenges that extend beyond technical accuracy. One of the primary obstacles is the scalability of advanced models like BERT and GPT, whose high computational demands can make real-time applications infeasible in resource-constrained environments. While techniques like model compression (e.g., DistilBERT) and adaptive learning have demonstrated promise, these methods often involve tradeoffs between efficiency and performance, which must be carefully balanced to meet practical needs. Another critical factor is the adaptability of these models to the rapid changes in misinformation trends. As narratives shift in real-time, static AI models may fail to recognize new forms of manipulation, emphasizing the need for dynamic systems capable of continuous learning. Such adaptability requires frequent retraining and infrastructure capable of ingesting and processing vast quantities of data in real-time, which can be resource-intensive and operationally complex.

This study recognizes that AI is not a standalone solution to the misinformation crisis. Relying solely on technology risks ignoring the sociopolitical, cultural, and ethical factors that influence how misinformation spreads and is consumed. Instead, we advocate for an integrative approach that combines AI-driven tools with human oversight and social interventions. Human-AI collaboration is critical for contextualizing misinformation within its broader sociocultural framework. For

example, while AI systems excel at detecting linguistic patterns, human fact-checkers can provide a deeper understanding of the intent and impact of specific narratives.

Further, the use of AI systems raises pressing ethical concerns, including algorithmic bias, transparency, and fairness. Biases in training data can disproportionately impact certain demographic groups and lead to unequal detection rates and potential harm. For instance, during high-stakes events like political debates, AI models may unintentionally flag narratives from underrepresented groups more frequently due to skewed training datasets. To address these issues, we emphasize the use of XAI techniques such as LIME and SHAP, which enhance transparency by providing interpretable insights into model predictions. Furthermore, training datasets must be diversified to include a wide range of cultural and linguistic contexts. Ethical design principles, such as accountability mechanisms and fairness-aware algorithms, should equally guide the development of these models to mitigate potential misuse and build trust among users.

In this view, we recommend the implementation of several strategies including, the implementation of lightweight variants of transformer models (e.g., DistilBERT) to reduce computational demands without significantly compromising accuracy; developing adaptive learning techniques that enable continuous retraining and adjustment based on evolving misinformation trends; designing AI systems that complement, rather than replace, human judgment; encouraging partnerships between AI researchers, fact-checking organizations, and social media platforms to improve data quality and operational efficiency; and investing in media literacy programs to enhance public understanding of misinformation and reduce reliance on automated detection systems.

### **Declaration of conflicting interests**

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


### **Funding**

The author received no financial support for the research, authorship, and/or publication of this article.

### **Ethical statement**

This study adheres to the highest standards of research ethics and does not involve direct interaction with human subjects, as all data were collected from publicly available sources, including reputable news outlets, misinformation websites, and social media platforms. In addition, Institutional review board approval was not required, as the study did not involve direct human participation or sensitive personal data.

### **ORCID iD**

Gregory Gondwe  <https://orcid.org/0000-0001-7444-2731>

## References

- Armitage, R., & Vaccari, C. (2021). Misinformation and disinformation. In *The Routledge companion to media disinformation and populism* (pp. 38–48). Routledge.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, L. C., Schwing, A., Yuille, A., & Urtasun, R. (2015). Learning deep structured models. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:1785-1794. PMLR.
- Chu-Ke, C., & Dong, Y. (2024). Misinformation and literacies in the era of generative artificial intelligence: A brief overview and a call for future research. *Emerging Media*, 2(1), 70–85. <https://doi.org/10.1177/2752543241240285>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv. <https://arxiv.org/abs/1810.04805>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies* (Vol. 1, pp. 4171–4186). ACL.
- Gondwe, G. (2023). CHATGPT and the global south: How are journalists in sub-Saharan Africa engaging with generative AI? *Online Media and Global Communication*, 2(2), 228–249. <https://doi.org/10.1515/omgc-2023-0023>
- Gondwe, G., & Muchangwe, R. (2020). Agenda-setting theory in African contexts: A Jekyll and Hyde in the Zambian presidential elections. *International Journal of Multidisciplinary Research and Development*, 7(5), 93–100. <https://doi.org/10.25810/707h-6d59>
- Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform* (pp. 10–33). Cambridge University Press.
- Hashmi, E., Yayilgan, S. Y., Yamin, M. M., Ali, S., & Abomhara, M. (2024). Advancing fake news detection: Hybrid deep learning with fasttext and explainable AI. *IEEE Access*, 12, 44462–44480. <https://doi.org/10.1109/ACCESS.2024.3381038>
- Hochreiter, S., & Schmidhuber, J. (1997, October). Unsupervised coding with lococode. In *International Conference on Artificial Neural Networks* (pp. 655–660). Springer.
- Horne, B., & Adali, S. (2017, May). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 759–766). <https://doi.org/10.1609/icwsm.v11i1.14976>
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Association for Computational Linguistics.
- Lundberg, S. M., & Lee, S. I. (2017). *Consistent feature attribution for tree ensembles*. arXiv. <https://arxiv.org/abs/1706.06060>

- Mehrabi, Z., McDowell, M. J., Ricciardi, V., Levers, C., Martinez, J. D., Mehrabi, N., Wittman, H., Ramankutty, N., & Jarvis, A. (2021). The global divide in data-driven farming. *Nature Sustainability*, 4(2), 154–160 <https://doi.org/10.1038/s41893-020-00631-0>
- Mehrabi, M., You, D., Latzko, V., Salah, H., Reisslein, M., & Fitzek, F. H. (2019). Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey. *IEEE Access*, 7, 166079–166108. <https://doi.org/10.1109/ACCESS.2019.2953172>
- Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., Nikolov, A., Shahi, G. K., Struß, J. M., & Mandl, T. (2021). The clef-2021 checkthat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43* (pp. 639–649). Springer International Publishing.
- Pothast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E. P. G., Hagen, M., & Stein, B. (2018, August). Crowdsourcing a large corpus of clickbait on Twitter. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1498–1507). ACL.
- Radford, A., Wu, J., Child, R., Luan, D., Dario Amodei, D., & Sutskever, I., (2019). Language models are unsupervised multitask learners. OpenAI. <https://www.openai.com/research/language-unsupervised>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). <https://doi.org/10.48550/arXiv.1602.04938>
- Ruchansky, N., Seo, S., & Liu, Y. (2017, November). Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management* (pp. 797–806). ACM.
- Sanh, V. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper, and lighter*. arXiv. arXiv:1910.01108.
- Strubell, E., Ganesh, A., & McCallum, A. (2020, April). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 09, pp. 13693–13696). <https://doi.org/10.1609/aaai.v34i09.7123>
- Tandoc, E. C., Jr, Lim, Z. W., & Ling, R. (2018). Defining “fake news” a typology of scholarly definitions. *Digital Journalism*, 6(2), 137–153. <https://doi.org/10.1080/21670811.2017.1360143>
- Utoft, S., Dou, J., & Wu, J. (2024). Enhancing EEG data quality: A comprehensive review of outlier detection and cleaning methods. In *Proceedings of the KDD Undergraduate Consortium (KDDUC’ 24)* (pp. 1–8). Association for Computing Machinery. Retrieved from <https://kdd2024.kdd.org/wp-content/uploads/2024/08/22-KDD-UC-Utoft.pdf>.
- Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018, July). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). ACM. <https://doi.org/10.1145/3219819.3219903>.
- West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15), Article e1912444117. PNAS. <https://doi.org/10.1073/pnas.1912444117>
- Wu, Z., Chen, Y., Kao, B., & Liu, Q. (2020). *Perturbed masking: Parameter-free probing for analyzing and interpreting BERT*. arXiv. <https://arxiv.org/abs/2004.14786>
- Yin, X., Huang, Y., Zhou, B., Li, A., Lan, L., & Jia, Y. (2019). Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access*, 7, 169434–169445. <https://doi.org/10.1109/ACCESS.2019.2955498>