Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

AI MODELS FOR MISINFORMATION DETECTION IN SOCIAL MEDIA: A SURVEY OF TECHNIQUES, APPLICATIONS, AND CHALLENGES

Dr. S. Sathiyapriya, Dr. T. Sumathi
Assistant Professor, Department of Computer Applications, Nallamuthu Gounder Mahalingam College, Pollachi : s.sathipriya@gmail.com, tsumadhijk@gmail.com

Abstract
In the era of ubiquitous digital communication, social media platforms have become primary channels for information dissemination. However, this unprecedented connectivity has also facilitated the rapid spread of misinformation, posing significant risks to public safety, democracy, and social cohesion. Detecting and mitigating misinformation is a critical challenge for governments, media platforms, and technology providers alike. In recent years, artificial intelligence (AI) models, including machine learning (ML) and deep learning (DL) techniques, have shown promising capabilities in automatically identifying and flagging false or misleading content. This survey paper systematically reviews the current state of AI-driven misinformation detection on social media, examining conventional machine learning models, state-of-the-art deep learning architectures, hybrid frameworks, and ensemble techniques. The study also explores applications, publicly available datasets, evaluation metrics, and highlights key challenges such as data imbalance, context dependency, and multilingual misinformation. Finally, the paper proposes future research directions emphasizing explainability, cross-platform detection, multimodal misinformation identification, and ethical considerations in AI deployment for content moderation.

Keywords: Artificial Intelligence, Misinformation Detection, Social Media, Machine Learning, Deep Learning, Natural Language Processing, Fake News.

1. Introduction

Social media has fundamentally transformed the way individual access and share information, offering immediate, low-cost, and far-reaching communication channels. Platforms like Facebook, Twitter (now X), Instagram, WhatsApp, and YouTube have grown into dominant information hubs where news, opinions, and multimedia content are consumed by billions daily. While these platforms have democratized information exchange, they have also inadvertently enabled the proliferation of misinformation — broadly defined as false, misleading, or deceptive information spread intentionally or unintentionally.

The consequences of misinformation can be severe, ranging from public panic during health crises (e.g., COVID-19 infodemics) to influencing electoral outcomes, inciting violence, and eroding trust in public institutions. The speed and scale at which misinformation travels on social media surpasses the capabilities of traditional fact-checking mechanisms, necessitating the adoption of automated and scalable solutions.

Artificial Intelligence (AI) offers significant promise in addressing this issue. AI models, particularly those based on Natural Language Processing (NLP) and Deep Learning (DL), have been increasingly applied to detect patterns, linguistic cues, and semantic inconsistencies indicative of misinformation. From supervised machine learning algorithms like logistic regression and support vector machines to advanced transformer-based models such as BERT and RoBERTa, AI-driven solutions have achieved noteworthy success in misinformation detection tasks.

This survey paper aims to comprehensively review the latest AI-based approaches to misinformation detection on social media, analyze their strengths and limitations, discuss their practical applications, and propose future research directions to enhance their effectiveness and fairness.

2. Literature Review

2.1 Role of AI in Social Media Content Analysis

The emergence of social media as a dominant information-sharing medium has introduced complex challenges in monitoring and moderating vast volumes of user-generated content. Traditional rule-based moderation techniques struggle to keep pace with the dynamic nature of misinformation, necessitating AI-based automated systems. AI models, particularly those powered by Natural Language Processing (NLP), computer vision, and deep learning techniques, have demonstrated significant success in content classification, hate speech detection, rumor identification, and fake news detection.

Recent studies by Zhang et al. (2022) and Alsmadi & O'Brien (2021) emphasized that AI-driven text classification models can effectively distinguish between legitimate and deceptive content by analyzing linguistic patterns, emotional tone, and source credibility. Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (A Robustly Optimized BERT Pretraining Approach) have shown state-of-the-art performance in understanding contextual semantics, making them valuable tools for misinformation detection.

2.2 Misinformation Types and Characteristics

Misinformation on social media manifests in various forms, including fake news articles, doctored images, manipulated videos (deepfakes), rumors, and clickbait headlines. These messages may be intentionally created (disinformation) or spread inadvertently by misinformed users. The detection complexity increases with the introduction of satire, parody, or partial truths embedded within misinformation.

According to Zhou & Zafarani (2020), misinformation can be categorized based on intent (malicious vs. unintentional), content format (text, image, video), and propagation patterns (viral vs. isolated). Recognizing these attributes is crucial in designing AI models that can adapt to varying misinformation typologies. Multimodal misinformation, combining text with images or videos, presents further detection challenges that AI systems are actively addressing.

2.3 AI Models for Misinformation Detection

A diverse range of AI models has been developed to automate misinformation detection. Machine learning classifiers like Logistic Regression, Support Vector Machines (SVM), and Random Forests were among the earliest approaches, typically trained on handcrafted feature sets derived from linguistic, user-based, and network-based attributes.

The rise of deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks enabled end-to-end learning of semantic representations, reducing reliance on manual feature engineering. Transformer-based models, including BERT, XLNet, and ELECTRA, have further advanced detection accuracy by capturing bidirectional contextual dependencies in text data.

Studies by Shu et al. (2020) and Patwa et al. (2021) highlight that ensemble techniques and hybrid frameworks integrating multiple AI models often outperform individual classifiers in detecting diverse misinformation types on platforms like Twitter and Facebook.

3. Objectives of the Survey

The primary objectives of this survey paper are:

1. **To systematically review AI models applied for misinformation detection in social media environments.**
2. **To categorize existing techniques based on model types, detection strategies, and content formats.**
3. **To analyze publicly available datasets, evaluation metrics, and benchmark results in recent literature.**
4. **To identify key challenges and limitations associated with current AI-driven misinformation detection systems.**
5. **To suggest future research directions for developing more accurate, explainable, and ethical AI solutions for content moderation.**

Methodology

This survey adopts a structured, systematic literature review methodology to examine and analyze existing AI models for misinformation detection on social media platforms. The methodology followed several phases, ensuring the inclusion of credible, high-quality research papers and technical resources.

4.1 Research Question Formulation

The survey seeks to address the following research questions:

- What types of AI models are predominantly used for misinformation detection on social media?
- How do these models handle various misinformation content types (text, image, video, and multimodal)?
- What datasets, evaluation metrics, and benchmarks are commonly employed in existing studies?
- What are the limitations of current AI models, and what future directions are proposed?

4.2 Literature Search Strategy

An extensive search was conducted across reputed academic databases and digital libraries, including:

- **IEEE Xplore**
- **ACM Digital Library**
- **Elsevier's ScienceDirect**
- **SpringerLink**
- **Google Scholar**
- **Scopus**

The search covered peer-reviewed journals, conference proceedings, and survey papers published between **2018 and 2024**. Relevant search keywords included *"misinformation detection"*, *"AI for fake news"*, *"deep learning for rumor detection"*, *"BERT for social media content moderation"*, *"machine learning misinformation classifiers"*, and *"multimodal misinformation detection"*.

4.3 Inclusion and Exclusion Criteria

Only papers meeting the following criteria were considered:

- Focused on AI-based or ML-based misinformation detection.
- Published in reputable, indexed journals or conferences.
- Employed social media datasets (Twitter, Facebook, YouTube, Reddit, WhatsApp).
- Included performance metrics and comparative analysis.

Papers lacking empirical results or those focusing purely on theoretical models without implementation were excluded.

4.4 Data Extraction and Analysis

Selected studies were examined based on:

- AI model type (ML, DL, Transformer, Hybrid)
- Dataset details
- Content format handled (text, image, video, multimodal)
- Evaluation metrics (Accuracy, F1-score, Precision, Recall, AUC)
- Model performance and limitations

A thematic analysis was conducted to classify models and benchmark their effectiveness in real-world social media environments.

5. AI Models for Misinformation Detection

This section provides an in-depth overview of various AI models applied to detect misinformation on social media, categorized into **Machine Learning Models**, **Deep Learning Models**, **Transformer-based Models**, and **Hybrid Models**.

5.1 Machine Learning Models

Traditional ML models were among the earliest methods for misinformation detection due to their interpretability and ease of deployment.

5.1.1 Logistic Regression
Used for binary classification (misinformation vs. legitimate content), logistic regression works on feature sets derived from content attributes (word frequencies, sentiment scores), user metadata (account age, follower count), and propagation features (retweet ratios).
5.1.2 Support Vector Machines (SVM)
SVMs maximize the margin between misinformation and legitimate content classes. They perform well on small to medium datasets but struggle with large-scale, unstructured data.
5.1.3 Random Forest
An ensemble technique that constructs multiple decision trees and outputs the majority class. Random Forest models are known for their robustness against overfitting and ability to handle heterogeneous feature spaces.
5.1.4 Naïve Bayes
This probabilistic classifier is simple and effective for text-based misinformation detection, especially when combined with TF-IDF or Bag-of-Words vectorization.
Limitation: ML models require extensive feature engineering and cannot automatically capture semantic meaning, limiting their scalability for multilingual or multimodal content.
5.2 Deep Learning Models
Deep learning models revolutionized misinformation detection by enabling automatic feature extraction and end-to-end training.
5.2.1 Convolutional Neural Networks (CNN)
Primarily used for image and text classification. In text analysis, CNN extracts n-gram level features and identifies misinformation cues like sensationalist phrases and polarizing words.
5.2.2 Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)
RNNs and LSTMs capture sequential dependencies in text, making them suitable for analyzing long-form social media posts, news articles, or rumor chains.
5.2.3 BiLSTM
Bidirectional LSTMs process text in both forward and backward directions, enhancing contextual understanding and detection accuracy.
5.2.4 Graph Neural Networks (GNN)
GNNs model the propagation structure of misinformation by analyzing social media post-sharing patterns and retweet networks. Studies (Shu et al., 2020) show GNN-based models outperform traditional text-only approaches.
Limitation: Deep learning models demand large labeled datasets, extensive computational resources, and are often black-box models lacking interpretability.
5.3 Transformer-based Models
The transformer architecture revolutionized NLP, with models like BERT setting new benchmarks for misinformation detection.
5.3.1 BERT (Bidirectional Encoder Representations from Transformers)
Pre-trained on large text corpora, BERT can be fine-tuned for misinformation detection tasks. It captures bidirectional dependencies in text, understanding context beyond surface-level features.
5.3.2 RoBERTa
An optimized version of BERT that improves performance by training with larger batch sizes, removing the next-sentence prediction objective, and using more data.
5.3.3 XLNet, ALBERT, ELECTRA
These models build upon BERT's limitations, offering improved training efficiency and accuracy. Studies (Patwa et al., 2021) report that transformer-based models consistently outperform traditional ML and DL models in fake news detection.

Limitation: High memory consumption and longer inference times make deploying these models challenging on resource-constrained devices.
5.4 Hybrid Models
Hybrid models combine multiple AI techniques for improved detection accuracy and adaptability.

- **Text + Image models:** Combine CNNs for image analysis and BERT for text content to detect multimodal misinformation.
- **Ensemble models:** Aggregate predictions from different classifiers (e.g., Random Forest, LSTM, and BERT) to boost performance and reduce model bias.
- **Propagation-aware models:** Integrate text content, user features, and propagation patterns using GNN-based frameworks.

Hybrid approaches address single-model limitations and offer enhanced robustness in real-world, noisy social media environments.

Datasets and Evaluation Metrics

A critical aspect of misinformation detection research is the availability of high-quality, annotated datasets and appropriate evaluation metrics. This section reviews widely used datasets and metrics.

6.1 Popular Datasets

Dataset	Description	Modality	Source
LIAR	12,836 manually labeled short statements from Politifact.	Text	Politifact
FakeNewsNet	Integrates news content, user profiles, and social context from Twitter.	Text, Propagation	Twitter
BuzzFeedNews	Fact-checked political news shared on Facebook during 2016 U.S. elections.	Text, Image	Facebook
COVID19FN	Fake news articles related to COVID-19 pandemic.	Text	Various
Weibo Rumor	Chinese social media posts labeled as rumors or non-rumors.	Text, Propagation	Sina Weibo
Twitter15/16	Public Twitter rumor datasets from 2015-2016.	Text, Propagation	Twitter

Limitations: Most datasets are text-centric, with fewer resources for image, video, and multimodal misinformation.

6.2 Evaluation Metrics

Common metrics for assessing misinformation detection models:

- **Accuracy (Acc):** Proportion of correct predictions.
- **Precision:** Ratio of true positives to predicted positives.
- **Recall (Sensitivity):** Ratio of true positives to actual positives.
- **F1-Score:** Harmonic mean of precision and recall, balancing both.
- **Area Under Curve (AUC):** Measures model's ability to distinguish between classes.
- **Macro / Micro Averages:** For multi-class settings, especially useful in multilingual and multimodal misinformation detection.

Example:

BERT-based models achieved an F1-Score of 93% on LIAR, outperforming traditional classifiers (Patwa et al., 2021).

7. Challenges and Limitations

Despite AI's promising advancements in misinformation detection, several challenges persist:

7.1 Data Quality and Availability

- Scarcity of high-quality, multilingual, and multimodal misinformation datasets.
- Annotation is labor-intensive, subjective, and prone to bias.

7.2 Model Generalization

- AI models trained on one dataset or region struggle to generalize across different platforms, languages, and topics.

7.3 Evolving Misinformation Tactics
- Misinformation actors constantly evolve strategies, using memes, deepfakes, and encoded language that evade AI detection.

7.4 Ethical Concerns
- Potential for censorship, suppression of dissent, or biased moderation.
- Privacy issues related to content tracking and user profiling.

7.5 Computational Overheads
- Deep learning and transformer-based models demand high computational resources and energy consumption, limiting real-time deployment on edge devices.

8. Future Directions

Key research trends and directions for AI-driven misinformation detection:

8.1 Multimodal Misinformation Detection
- Development of unified AI frameworks capable of processing and correlating text, images, videos, and metadata.

8.2 Explainable AI (XAI)
- Improving model transparency and interpretability, enabling users and moderators to understand AI decisions.

8.3 Low-Resource Language Support
- Expanding detection capabilities to underrepresented languages and regional misinformation ecosystems.

8.4 Cross-Platform Detection
- Designing AI systems capable of detecting misinformation as it spreads across multiple platforms.

8.5 Real-time Detection and Response
- Optimizing models for lower latency to enable prompt flagging and intervention during misinformation outbreaks.

9. Conclusion

This survey systematically reviewed AI models applied to misinformation detection on social media. From early machine learning models to modern transformer-based architectures, AI technologies have substantially improved the accuracy and scalability of detecting harmful, misleading content. However, challenges such as data scarcity, evolving tactics, and ethical concerns continue to hinder universal, real-time detection.

Future research should focus on explainable, multimodal, multilingual, and cross-platform detection frameworks. Incorporating social context and user behavior analytics, combined with policy interventions, may offer a comprehensive solution to the misinformation epidemic.

10. References

1. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). **Fake news detection on social media: A data mining perspective.** *SIGKDD Explorations*, 19(1), 22–36.
2. Zhou, X., & Zafarani, R. (2018). **Fake news detection: A survey.** *ACM Computing Surveys*, 53(5), 1-40.
3. Patwa, P., et al. (2021). **Fighting an infodemic: COVID-19 fake news dataset.** *Communications in Computer and Information Science*, 209-217.
4. Thorne, J., et al. (2018). **FEVER: A large-scale dataset for fact extraction and verification.** *Proceedings of NAACL-HLT*, 809-819.
5. Volkova, S., et al. (2017). **Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter.** *ACL*, 647–653.
6. Cui, L., Lee, D., & Shu, K. (2019). **Same story, different names: Automated fake news detection using content-based features.** *IEEE BigData*, 822–830.

7. Devlin, J., et al. (2019). **BERT: Pre-training of deep bidirectional transformers for language understanding.** *NAACL*, 4171–4186.
8. Liu, Y., et al. (2019). **RoBERTa: A robustly optimized BERT pretraining approach.** *arXiv preprint arXiv:1907.11692.*
9. Nguyen, D. T., et al. (2020). **Multimodal fake news detection on social media: A survey.** *Information Processing & Management*, 57(6).
10. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). **FakeNewsNet: A data repository with news content, social context, and dynamic information for studying fake news on social media.** *Big Data*, 8(3), 171-188.
11. Wang, W. Y. (2017). **"Liar, liar pants on fire": A new benchmark dataset for fake news detection.** *ACL*, 422–426.
12. Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). **Prominent features of rumor propagation in online social media.** *ICWSM*, 558-561.
13. Jain, A., et al. (2021). **A survey on fake news detection using natural language processing.** *Journal of King Saud University–Computer and Information Sciences.*
14. Zhang, X., et al. (2021). **A survey of multimodal fake news detection.** *Information Fusion*, 76, 1-13.
15. Cinelli, M., et al. (2020). **The COVID-19 social media infodemic.** *Scientific Reports*, 10(1), 16598.