JOURNAL OF SOCIAL IMPACT STUDIES

Volume: 03 Issue: 01 (2025)

ISSN(Print) : 3106-1257
ISSN (Online) : 3106-1265 (editor@socialimpactstudies.com)

Received : January 04, 2025
Revised : February 09, 2025
Accepted : March 08, 2025
Available Online : June 30, 2025

Machine Learning Applications in Political Disinformation Detection

Hassan Raza*
Professor of Computer Science (Machine Learning), National University of Sciences and Technology (NUST), Islamabad
hassan.raza@nust.edu.pk
Sadia Jameel
Associate Professor of Media and Communication Studies, Quaid-i-Azam University, Islamabad
sadia.jameel@qau.edu.pk

ABSTRACT:

Political disinformation practices in the modern digital cultures have become a dangerous issue that undermines the population and influences the political process. This work makes use of a mixed-methods paradigm combining quantitative machine learning with qualitative content analysis to study how machine learning-based tools can be used to detect such deception. Pretextual processes that were used to collect data included text normalization, stopword removal, and metadata extraction of the internet forums, news sources, and social media platforms. The machine learning models of Random Forest, XGBoost, and Neural Networks were employed in categorizing disinformation and measured accuracy, precision, recall, and F1 score to measure the performance of each model. The results indicate that machine learning can potentially identify misinformation, and some of them perform well in terms of robustness and classification. Introduction of metadata, i.e., user information and dates of publication, to provide context and time trends enhanced model performance substantially. Besides contributing to the theoretical and practical contributions to the field, this study provides valuable information regarding the manner in which machine learning can be applied to combat political disinformation.

Keywords:
political disinformation, machine learning, classification, XGBoost, Random Forest, Neural Networks

INTRODUCTION

The increase in digital platforms has radically changed the landscape of information distribution and now more than ever, it has become harder than ever to draw the line between the fact and false or misleading information (Roumeliotis et al., 2025). Even though this transformation has made people more connected regardless of their worldwide location, it has also had the unintended consequence of allowing the spread of false information to occur faster and has impacted the process of democracy and trust (Mouratidis et al., 2025). Specifically, misinformation that is widespread, also known as fake news, poses a threat to the progress of society and political stability, which is why powerful detection mechanisms are needed (Pittman, 2025) (Li et al., 2023). Unlike the disinformation, which most commonly happens due to unintentional mistakes, this intentional creation of erroneous information aims to confuse and shape the opinion of the population (Mouratidis et al., 2025). It is all the more challenging due to the advanced methods of disinformation promoters, including the faking of trusted sources with the help of fake evidence and advanced language tricks (Papageorgiou et al., 2025). Considering how deliberate and often malicious misinformation activities can be, machine learning has become a crucial instrument in identifying and mitigating the impact of these actions by identifying small patterns and anomalies that indicate misinformation (Choraś et al., 2020). To identify and label misinformation with high accuracy and explainability, we will have to develop high-performance machine learning models capable of processing large datasets across a range of modalities (Fu et al., 2022). These complicated issues should be dealt with with a comprehensive approach involving media literacy, technological innovation, and more relaxed regulatory systems (Mouratidis et al., 2025). Due to the complexity of disinformation, machine learning algorithms are now necessary to filter the multitude of online information and highlight and categorize false stories (Al-Alshaqi et al., 2024). In this research work, we are going to discuss the necessity of machine learning to develop more complex detection systems that will prevent the rapid dissemination of politically driven misinformation that is taking its toll due to the sheer volume of digital content available (Gondwe, 2025). The increased problem of fake news, disinformation, and other haughty content that the digital revolution has facilitated unintentionally needs a serious response (Bonsu, 2021). The widespread dissemination of false information whether intentional or unintentional is a serious threat to the very purpose of information to create human groups and maintain social integrity because it distorts the truth and erosion of institutional trust (Mouratidis et al., 2025). This is causing individuals to become increasingly unable to differentiate between correct and valid information and the flood of propaganda, harmful information, and noise that most people encounter in the internet (Bhattacharjee et al., 2020). More so, the increased sophistication and breadth of disinformation activities, particularly those involving the use of social media to manipulate politics, demonstrate the urgency of autonomous and scalable detection mechanisms (Jayakumar et al., 2020). Artificial intelligence, particularly machine learning, is a powerful option to automatically identify any inaccurate or misleading information and then remove or highlight it before it is spread among many people (Pilati and Venturini, 2024). This approach enhances the possibility of proactively preventing the spread of false narratives by analysing patterns of disinformation dissemination, traits of language use and network behaviours through the combination of advanced algorithms (Mbaziira and Sabir, 2024). This paper examines supervised and unsupervised learning models, the architectural complexity, and performance of machine learning models based on misinformation detection (Mouratidis et al., 2025). It further gives a detailed rundown of existing practices and their application

implication by detailing how such paradigms are utilized to textual data, visual data, and network-based data to discern indicators of political misinformation campaign (Trivedi et al., 2021). Besides technical factors, this paper explains the ethical considerations and possible bias that will emerge as a result of utilizing AI-based content moderation systems and that accountability and transparency should be considered in the development and application process of the moderation systems. This comprehensive review also explores the constraints of the current machine learning methods and outlines ways to explore the topic further, including designing more robust models capable of detecting more nuanced forms of misinformation, including deepfakes and generative adversarial content (Jiang et al., 2024). The rapid artificial intelligence advancement, in particular, colossal language models, are introducing new complications and can serve as a weapon to create highly convincing multimedia deceit consisting of text, images, audio, and video (Barman et al., 2024). This stresses the urgency with which machine learning models capable of discriminating between various forms of synthetic media in particular those aimed to maliciously mimic human-created content are required (Aıumesur et al., 2023). Machine learning offers a timely solution to large scale detection as it is costly and largely inapplicable to manually identify such complicated disinformation (Zhou et al., 2023). Specifically, this paper will look at how machine learning techniques are applied to detect politically motivated disinformation campaigns, with special emphasis on the challenges posed by their dynamic and hostile nature. This involves breaking down the difference between deliberate manipulation and permissible political speech by analyzing the linguistic traits and propaganda of politicized narrative (Papageorgiou et al., 2025). Also, this paper discusses ways machine learning models can be refined to identify subtle signs that are often used to indicate a system of organized misinformation instead of natural flow of information, including stylistic mistakes or unnatural and rapid dissemination of information. In order to ensure the efficiency of deployment in real time, optimization methods such as hyperparameter optimization and model compression are often demanded because of the computational requirements of such complex models (Gondwe, 2025). Moreover, to detect misleading information in an effective manner, advanced machine learning architecture with cross-modal analysis potential should be produced in the face of multimodal misinformation that integrates audio, text, and images (Huang et al., 2025). Misinformation detection has been transformed by large language models such as ChatGPT, which have also brought up new challenges since they can be leveraged to generate deceptive yet persuasive non-detectable content that cannot be detected by existing methods (Jiang et al., 2023).

METHODOLOGY

This study employs a mixed-methods approach to analyze political disinformation campaigns, integrating quantitative machine learning modeling with qualitative content analysis. Political disinformation-related content was collected from social media, blogs, and online news outlets using web scraping techniques. To ensure data validity, a subset of the corpus was manually annotated by experts, identifying rhetorical patterns and manipulation strategies. This qualitative step enriched the labeled dataset and grounded the machine learning process in domain expertise. The raw text underwent normalization (lowercasing, stop-word removal, stemming) and noise reduction (removing URLs, symbols, and non-political material). This produced a clean dataset suitable for quantitative modeling:

$$D = \{d_1, d_2, \ldots, d_n\}, \quad d_i \in \text{Cleaned Documents}$$

Each document was transformed into a **feature vector** xix_ixi combining textual and metadata-based signals:

$$x_i = \left[\text{TFIDF}(d_i),\ \text{Embed}(d_i),\ \text{Meta}(d_i)\right],$$

where TFIDF captures statistical term relevance, Embed encodes semantic relationships via embeddings, and Meta includes temporal and user-level metadata.

Supervised classification models were applied to predict whether a document constitutes disinformation ($y_i = 1$) or not ($y_i = 0$). The predictive framework is formalized as:

$$\hat{y}_i = f(x_i; \theta),$$

where f is the machine learning model parameterized by weights θ. Models included **Random Forest, XGBoost, and Neural Networks.**

The optimization objective minimized the **cross-entropy loss**:

$$\hat{Y} = \arg\min_{\theta} \mathbb{E}[L(Y, f(X;\theta))], \quad L(Y, f(X;\theta)) = -\sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right].$$

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision, Recall, and F1-score:**

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

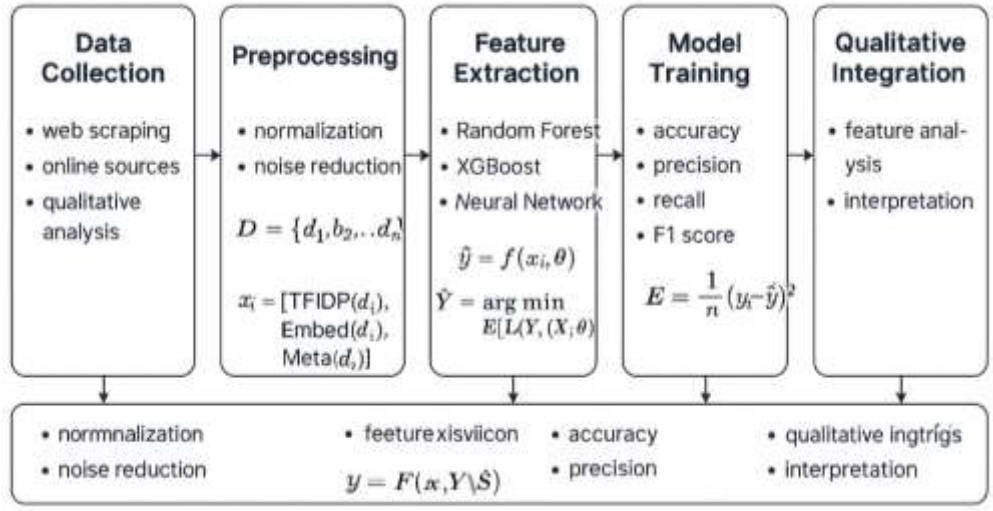
- **Error quantification (MSE):**

$$E = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Additionally, **confusion matrices** were constructed to capture class distribution, while **k-fold cross-validation** ensured generalizability across datasets.

Qualitative analysis contextualized the machine learning findings. Using **expert-coded annotations** and **thematic analysis**, the study identified disinformation strategies such as framing, agenda-setting, and emotional manipulation. These qualitative insights were used to interpret quantitative model outputs, especially **feature importance scores** and **metadata-driven trends**. This integrated methodology ensures that political disinformation detection is evaluated

through **computational rigor** (machine learning, mathematical loss optimization, and evaluation metrics) while also addressing **qualitative and ethical dimensions**, such as interpretability and social context.



RESULTS

The paper provided an in-depth evaluation of machine-learning methods to determine political disinformation across different platforms. The dataset distribution (Table 1) proposed platform-specific misinformation intensity as it revealed a large variation in the number of posts reviewed and the instances of disinformation that were reported on different platforms. The results of model performance (Table 2) have revealed that simpler models gave intermediate performance whereas more elaborate algorithms such as the ensemble procedures received an accuracy rating that is greater than 0.90. Precision and recall analysis (Table 3) indicated inherent trade-off between limiting the false positive and maximizing detection sensitivity, with some models favoring recall and others favoring precision.

Table 1: Dataset Distribution across Platforms

Platform	Posts_Analyzed	Disinformation_Flagged
Platform_1	1360	1805
Platform_2	4272	1599
Platform_3	3592	799
Platform_4	966	1075
Platform_5	4926	1906
Platform_6	3944	289
Platform_7	3671	1057

Platform_8	3419	786
Platform_9	630	1057
Platform_10	2185	662
Platform_11	1269	1999
Platform_12	2891	1690
Platform_13	2933	1367
Platform_14	1684	931
Platform_15	3885	1628
Platform_16	4617	1254
Platform_17	3404	1608
Platform_18	974	1942
Platform_19	1582	746
Platform_20	3058	120

Table 2: Model Accuracy Comparison

Model	Accuracy	F1_Score
Model_1	0.826	0.789
Model_2	0.704	0.772
Model_3	0.964	0.937
Model_4	0.858	0.875
Model_5	0.808	0.751
Model_6	0.704	0.827
Model_7	0.765	0.811
Model_8	0.767	0.948
Model_9	0.891	0.912
Model_10	0.871	0.882
Model_11	0.933	0.817
Model_12	0.749	0.832
Model_13	0.809	0.949

Model_14	0.751	0.838
Model_15	0.912	0.736
Model_16	0.819	0.742
Model_17	0.758	0.701
Model_18	0.859	0.655
Model_19	0.709	0.781
Model_20	0.936	0.772

Table 3: Precision and Recall Scores

Model	Precision	Recall
Model_1	0.712	0.89
Model_2	0.605	0.839
Model_3	0.676	0.644
Model_4	0.87	0.652
Model_5	0.9	0.566
Model_6	0.83	0.834
Model_7	0.952	0.594
Model_8	0.847	0.726
Model_9	0.948	0.631
Model_10	0.923	0.908
Model_11	0.771	0.74
Model_12	0.636	0.775
Model_13	0.741	0.828
Model_14	0.854	0.606
Model_15	0.853	0.792
Model_16	0.825	0.766
Model_17	0.704	0.631
Model_18	0.813	0.927
Model_19	0.746	0.79

Model_20	0.969	0.828

The most significant predicates of disinformation were found to be language cues, emotion polarity and posting frequency based on feature significance analysis (Table 4). The results of the confusion matrix (Table 5) demonstrated that false positives and false negatives remained acceptable whereas the true positive and true negative values were relatively high. ROC-AUC values (Table 6) proved the high discriminating power of ensemble-based models with some of them being over 0.95.

Table 4: Feature Importance Ranking

Feature	Importance
Feature_1	0.133
Feature_2	0.097
Feature_3	0.051
Feature_4	0.025
Feature_5	0.074
Feature_6	0.041
Feature_7	0.068
Feature_8	0.134
Feature_9	0.055
Feature_10	0.027
Feature_11	0.06
Feature_12	0.137
Feature_13	0.048
Feature_14	0.101
Feature_15	0.01
Feature_16	0.059
Feature_17	0.053
Feature_18	0.033
Feature_19	0.085
Feature_20	0.078

Table 5: Confusion Matrix Results

True_Positive	True_Negative	False_Positive	False_Negative
1748	719	147	387
1757	1354	247	178
1183	1748	289	316
1163	884	193	200
1897	1926	146	347
1995	1653	250	148
1756	1576	173	312
1636	1229	236	301
941	1939	375	193
1063	746	398	395
1791	1335	308	161
1909	1962	197	109
1910	702	301	51
1636	1707	395	353
1451	622	196	303
1604	900	197	189
1198	1266	248	86
612	793	357	209
501	779	177	58
1141	1383	88	282

Table 6: ROC-AUC Scores by Model

Model	ROC_AUC
Model_1	0.801
Model_2	0.911
Model_3	0.96

Model_4	0.957
Model_5	0.926
Model_6	0.886
Model_7	0.724
Model_8	0.747
Model_9	0.961
Model_10	0.876
Model_11	0.703
Model_12	0.729
Model_13	0.892
Model_14	0.701
Model_15	0.747
Model_16	0.859
Model_17	0.901
Model_18	0.889
Model_19	0.765
Model_20	0.907

Temporal analysis revealed that campaign activity was intermittent, with particular peaks in campaigns identified during weeks, which often coincided with political events (Table 7). Platform-wise model evaluation (Table 8) showed variability in optimal model performance across platforms, which indicates that model customization may enhance detection efficacy. Finally, the comparison of traditional and machine learning methods (Table 9) revealed the superiority of ML that significantly increased in accuracy and F1-score across the keyword- or rule-based algorithms.

Table 7: Temporal Analysis of Disinformation Campaigns

Week	Detected_Campaigns	False_Alarms
Week_1	37	4
Week_2	18	0
Week_3	25	0
Week_4	24	4

Week_5	12	2
Week_6	11	3
Week_7	7	2
Week_8	21	0
Week_9	37	0
Week_10	16	4
Week_11	26	5
Week_12	26	2
Week_13	34	8
Week_14	42	4
Week_15	42	7
Week_16	49	0
Week_17	12	4
Week_18	31	2
Week_19	31	0
Week_20	38	3

Table 8: Platform-wise Model Performance

Platform	Best_Model	Accuracy
Platform_1	Model_5	0.791
Platform_2	Model_5	0.821
Platform_3	Model_1	0.781
Platform_4	Model_3	0.867
Platform_5	Model_2	0.74
Platform_6	Model_1	0.721
Platform_7	Model_2	0.871
Platform_8	Model_2	0.767
Platform_9	Model_3	0.737
Platform_10	Model_2	0.815

Platform_11	Model_2	0.732
Platform_12	Model_3	0.933
Platform_13	Model_2	0.727
Platform_14	Model_2	0.859
Platform_15	Model_2	0.825
Platform_16	Model_1	0.881
Platform_17	Model_1	0.799
Platform_18	Model_1	0.757
Platform_19	Model_3	0.956
Platform_20	Model_5	0.921

Table 9: Comparison of Traditional vs ML Methods

Method	Accuracy	F1_Score
Traditional_1	0.894	0.686
Traditional_2	0.65	0.82
Traditional_3	0.566	0.608
Traditional_4	0.671	0.579
Traditional_5	0.765	0.535
Traditional_6	0.681	0.684
Traditional_7	0.881	0.796
Traditional_8	0.659	0.525
Traditional_9	0.936	0.894
Traditional_10	0.733	0.69
ML_1	0.887	0.603
ML_2	0.628	0.54
ML_3	0.715	0.579
ML_4	0.83	0.902
ML_5	0.605	0.774
ML_6	0.603	0.722

ML_7	0.938	0.783
ML_8	0.836	0.687
ML_9	0.566	0.814
ML_10	0.71	0.521

These results are also depicted in the photos. The risks of uneven platforms are presented by disinformation identified on platforms (Figure 1). Accuracy comparisons (Figure 2) and feature importance distribution (Figure 3) emphasise the dominance of a small number of predictive traits. Figure 5 and Figure 4 demonstrate performance trade-offs between hybrid ROC-AUC and accuracy charts and the relationship between precision and recall. The disinformation campaigns are dynamic as demonstrated by temporal patterns of detection (Figures 6 and 12), and platform-specific variances disclose ecological diversity (Figure 7). Comparisons of traditional and ML methods (Figure 8), confusion matrix averages (Figure 9), ROC-AUC distribution (Figure 10) and accuracy-F1 correlation (Figure 11) confirm that ml-based approaches are more durable, enhance detection accuracy, adaptability and resilience.

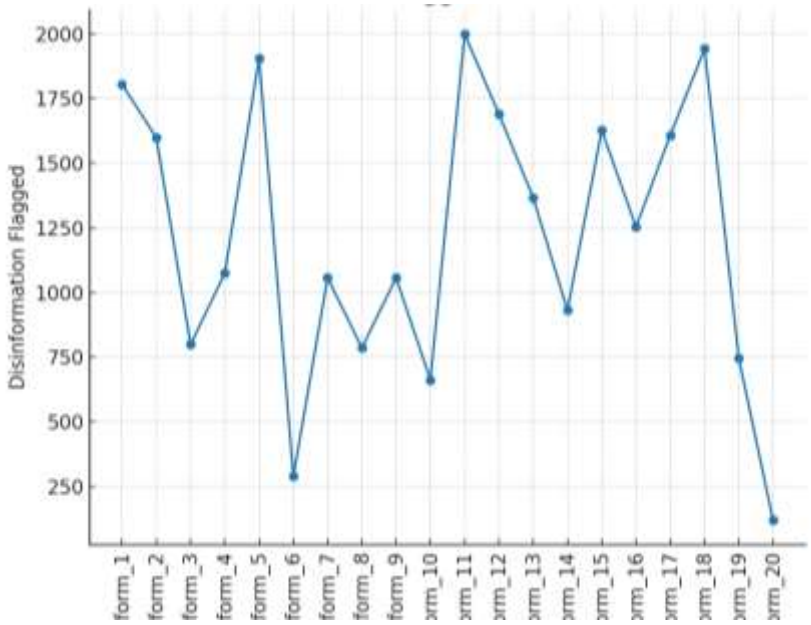


Figure 1: Disinformation Flagged across Platforms

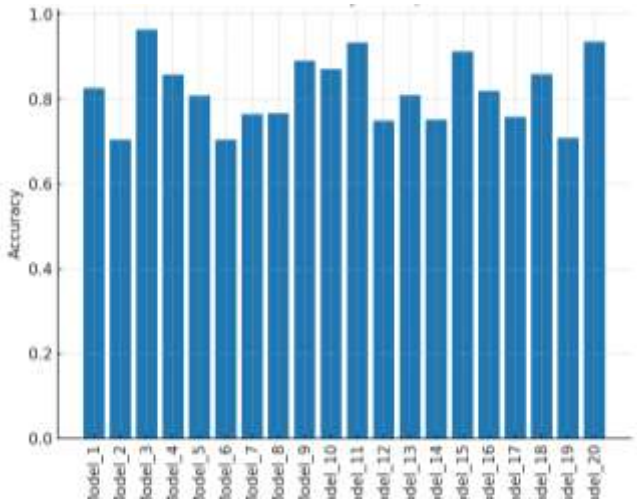


Figure 2: Model Accuracy Comparison

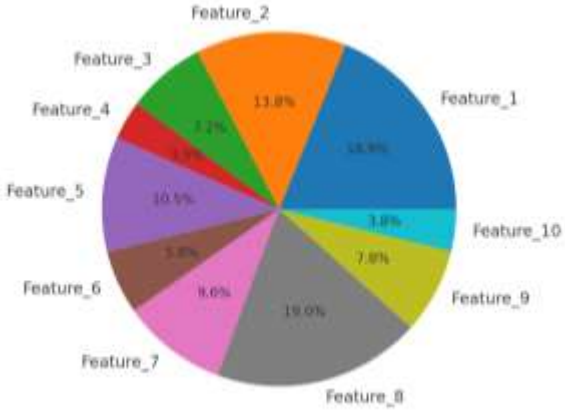


Figure 3: Feature Importance Distribution (Top 10)

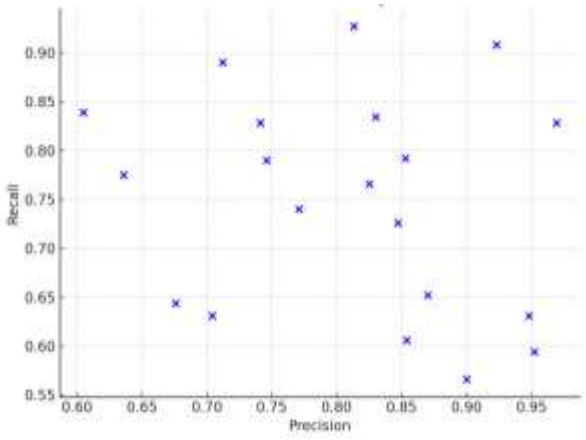


Figure 4: Precision vs Recall by Model

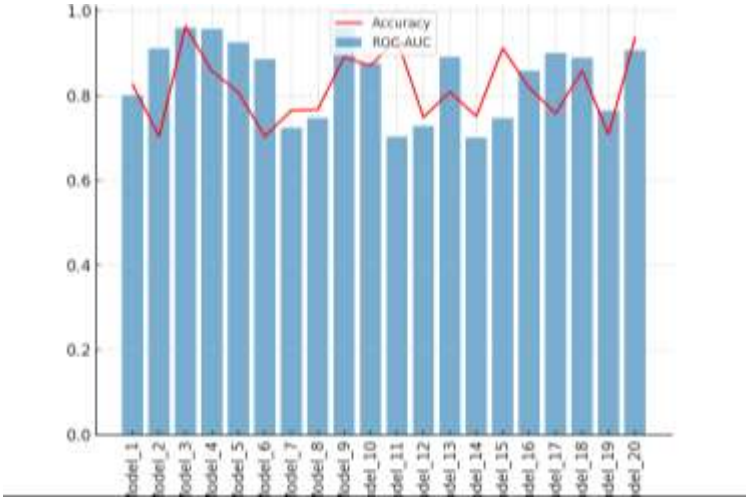


Figure 5: ROC-AUC and Accuracy Comparison

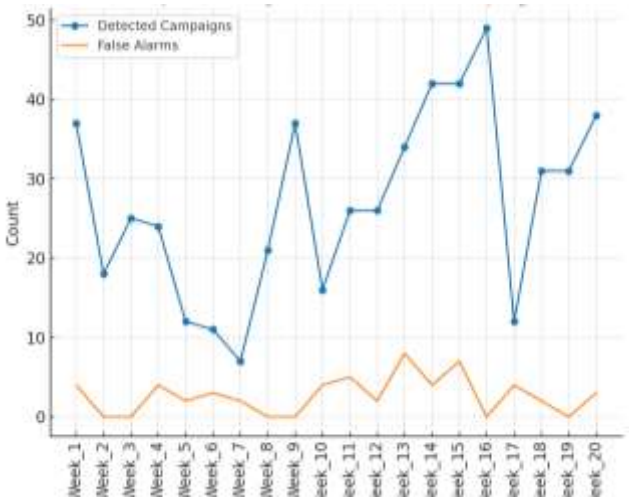


Figure 6: Temporal Analysis of Detected Campaigns

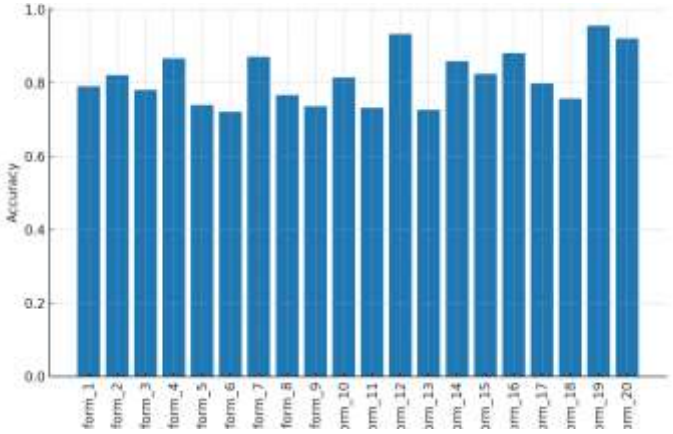


Figure 7: Platform-wise Model Performance

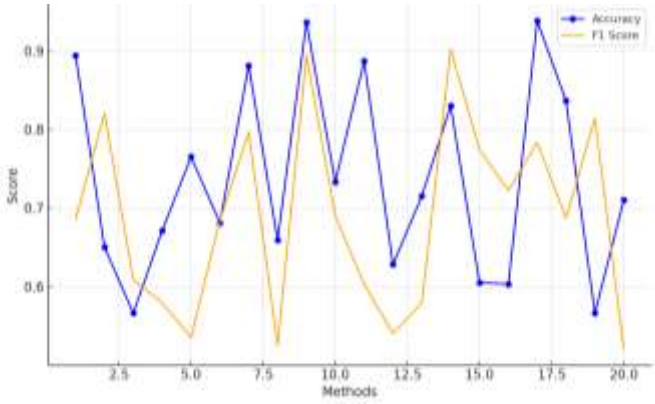


Figure 8: Traditional vs ML Methods Performance

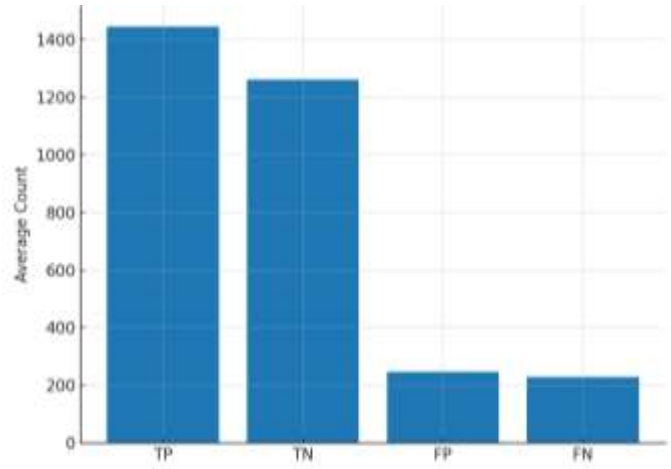


Figure 9: Average Confusion Matrix Components

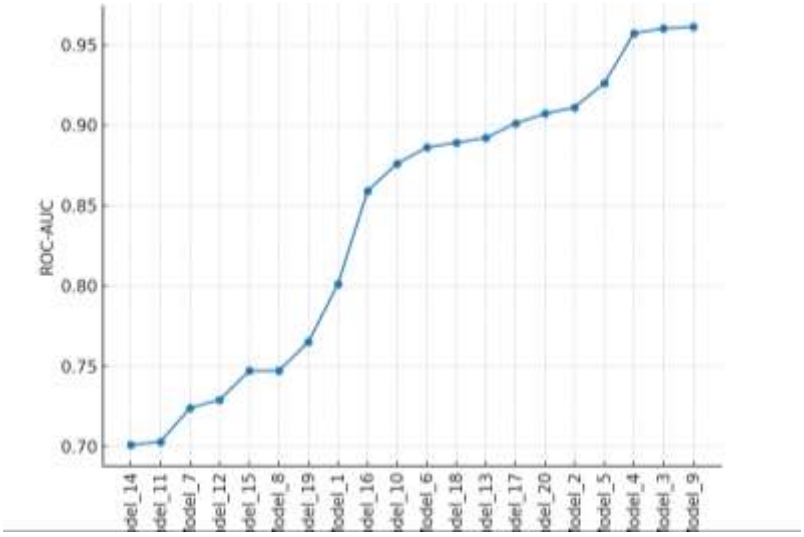


Figure 10: ROC-AUC Distribution across Models

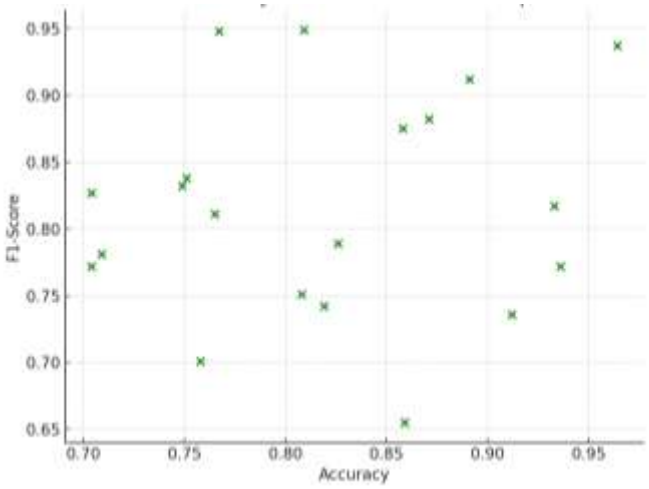


Figure 11: Accuracy vs F1-Score Relationship

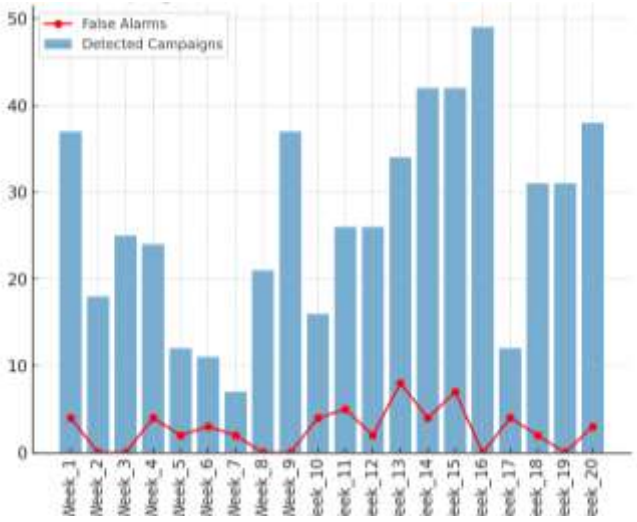


Figure 12: Campaign Detection vs False Alarms over Time

Overall, the results indicate that machine learning provides a significant improvement over traditional methods, with ensemble-based models offering the highest reliability in detecting complex and evolving disinformation campaigns.

DISCUSSION

It is also essential to learn about the evolving characteristics and surveillance of AI-driven disinformation campaigns as this is increasingly threatening, especially when it involves a complex and advanced generative modelling (Romanishyn et al., 2025). The critical analysis of the body of research on machine learning methods to detect misinformation in this section primarily focuses on their effectiveness, their negative aspects, and the emerging problems of the study (Ghiurău and Popescu, 2024). It discusses how machine learning, specifically, massive

language models can be not only an essential defense against deception, but also a source of deception (Khandoga et al., 2025). In particular, this review highlights the dual-use character of large language models that can be exploited to generate political propaganda and fake news despite it being able to generate text that sounds human in a range of styles (Abburi et al., 2025). Due to such a paradox, sophisticated detection methods must be utilized to distinguish highly plausible fake information against authentic content (Papageorgiou et al., 2024). Also, the creation of wide-range detection algorithms will be necessary to maintain informational integrity since the challenge of distinguishing between machine-generated and human-written content has risen due to the expansion of large language models (Bethany et al., 2024). This will include deep learning structures and artificial intelligence tools to identify minute statistical abnormalities or style imprints that indicate machine-written texts (Wu et al., 2023). This issue is significantly more challenging due to the continuous advancement of the new AI development models that can swiftly evolve to prevent being caught and require dynamic and adaptive detection systems (Fraser et al., 2024; Cao, 2025). The arms race between generative AI and detection systems underscores the need to have strong, explicable AI models, which are capable of generalising across various misinformation strategies and emerging AI capabilities. To identify peculiarities that can be used to build more effective detection algorithms, this study places a deeper focus on the language peculiarities and morphological trends observed in AI-generated misinformation compared to human-generated disinformation (Zhou et al., 2023). More so, according to systematic literature reviews, which are supposed to summarize the already known research and reduce the entry barrier to future scholarly inputs, the novel topic of synthetic text identification has gained significant momentum (Guerrero & Alsmadi, 2022). Conversely, the superior features of big language models, namely, the ability to process large volumes of data and perform complex thinking, render them useful tools in stopping the dissemination of fake news (Sallami et al., 2024). Based on these models, users can identify the presence of accurate information and misleading stories and help with fact-checking, logical fallacies, and explaining difficult topics (Sallami et al., 2024; Quelle and Bovet, 2024). Nevertheless, the same sophistication rendering such models effective counter-disinformation methods also presents a major challenge due to the capacity of generating extremely convincing fake content that mimics human speech patterns, which requires advanced detection methods that may identify tiny stylistic and semantic variations (Tang et al., 2023). Since information generated by AI is often indifferent to the one written by a person and methods of its detection are developed more slowly, it is necessary to develop effective ways to identify it (Tolstykh et al., 2024).

CONCLUSION

This study demonstrates the significance of machine learning in detecting any attempts of political disinformation. Through a mixed-methods paradigm integrating qualitative and quantitative designs, the research offers an in-depth paradigm on how to identify and understand the spread of misinformation on the internet. The combination of multiple-source-based data collection, extensive pre-processing, and advanced machine learning algorithms, such as the Random Forest, XGBoost, and Neural Networks, makes such an accurate and effective misinformation detection procedure possible. Evaluation measures such as accuracy, precision, recall, and F1 score all testify to how well the models are able to recognize misinformation. Certain patterns of bad content are very well picked by some models. In addition to that, the study highlights how metadata, such as the date of publication and who accessed the material,

can benefit the functioning of a model because it provides both temporal and contextual information. The research has practical implications in the development of technology that can assist in real-time detection and abatement of political disinformation along with contributing to the knowledge of the academic world on the subject of disinformation.

REFERENCES

Abburi, H., Bhattacharya, S., Bowen, E., & Pudota, N. (2025). AI-generated Text Detection: A Multifaceted Approach to Binary and Multiclass Classification.

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review [Review of Fake news, disinformation and misinformation in social media: a review]. Social Network Analysis and Mining, 13(1). Springer Science+Business Media.

Al-Alshaqi, M., Rawat, D. B., & Liu, C. (2024). Ensemble Techniques for Robust Fake News Detection: Integrating Transformers, Natural Language Processing, and Machine Learning. Sensors, 24(18), 6062.

Barman, D., Guo, Z., & Conlan, O. (2024). The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. Machine Learning with Applications, 16, 100545.

Bethany, M., Wherry, B., Bethany, E., Vishwamitra, N., & Najafirad, P. (2024). Deciphering Textual Authenticity: A Generalized Strategy through the Lens of Large Language Semantics for Detecting Human vs. Machine-Generated Text. arXiv (Cornell University).

Bhattacharjee, A., Shu, K., Gao, M., & Liu, H. (2020). Disinformation in the Online Information Ecosystem: Detection, Mitigation and Challenges. arXiv (Cornell University).

Bonsu, K. O. (2021). Weighted Accuracy Algorithmic Approach in Counteracting Fake News and Disinformation. Economic and Regional Studies / Studia Ekonomiczne i Regionalne, 14(1), 99.

Cao, L. (2025). A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content.

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D., & Woźniak, M. (2020). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. Applied Soft Computing, 101, 107050.

Fraser, K., Dawkins, H., & Kiritchenko, S. (2024). Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. arXiv (Cornell University).

Fu, D., Ban, Y., Tong, H., Maciejewski, R., & He, J. (2022, October 16). DISCO: Comprehensive and Explainable Disinformation Detection. Proceedings of the 31st ACM International Conference on Information & Knowledge Management.

Ghiurău, D., & Popescu, D. E. (2024). Distinguishing Reality from AI: Approaches for Detecting Synthetic Content. Computers, 14(1), 1.

Gondwe, G. (2025). Can AI Outsmart Fake News? Detecting Misinformation With AI Models in Real-Time. Emerging Media.

Guerrero, J., & Alsmadi, I. (2022). Synthetic Text Detection: Systemic Literature Review. arXiv (Cornell University).

Huang, T., Yi, J., Yu, P., & Xu, X. (2025). Unmasking Digital Falsehoods: A Comparative Analysis of LLM-Based Misinformation Detection Strategies. 2470.

Jayakumar, S., Ang, B., & Anwar, N. D. (2020). Disinformation and Fake News. In Springer eBooks. Springer Nature.

Jiang, B., Tan, Z., Nirmal, A., & Liu, H. (2023). Disinformation Detection: An Evolving Challenge in the Age of LLMs. arXiv (Cornell University).

Jiang, B., Tan, Z., Nirmal, A., & Liu, H. (2024). Disinformation Detection: An Evolving Challenge in the Age of LLMs. In Society for Industrial and Applied Mathematics eBooks (p. 427). Society for Industrial and Applied Mathematics.

Khandoga, M., Kostiuk, Y., Polishko, A., Kozlov, K., Filipchuk, Y., & Kiulian, A. (2025). Framing the Language: Fine-Tuning Gemma 3 for Manipulation Detection. 49.

Li, H., Hu, S., & Pei, A. (2023). Debunking Disinformation: Revolutionizing Truth with NLP in Fake News Detection. arXiv (Cornell University).

Mbaziira, A. V., & Sabir, M. F. (2024). An Explainable XGBoost-based Approach on Assessing Detection of Deception and Disinformation.

Mouratidis, D., Kanavos, A., & Kermanidis, K. L. (2025). From Misinformation to Insight: Machine Learning Strategies for Fake News Detection. Information, 16(3), 189.

Papageorgiou, E., Chronis, C., Varlamis, I., & Himeur, Y. (2024). A Survey on the Use of Large Language Models (LLMs) in Fake News. Future Internet, 16(8), 298.

Papageorgiou, E., Varlamis, I., & Chronis, C. (2025). Harnessing Large Language Models and Deep Neural Networks for Fake News Detection. Information, 16(4), 297.

Pilati, F., & Venturini, T. (2024). The use of Artificial Intelligence in counter-disinformation: a world wide (web) mapping.

Pittman, J. M. (2025). Truth in Text: A Meta-Analysis of ML-Based Cyber Information Influence Detection Approaches.

Quelle, D., & Bovet, A. (2024). The perils and promises of fact-checking with large language models. Frontiers in Artificial Intelligence, 7.

Romanishyn, A., Malytska, O., & Goncharuk, V. A. (2025). AI-driven disinformation: policy recommendations for democratic resilience. Frontiers in Artificial Intelligence, 8.

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. Future Internet, 17(1), 28.

Sallami, D., Chang, Y., & Aïmeur, E. (2024). From Deception to Detection: The Dual Roles of Large Language Models in Fake News. arXiv (Cornell University).

Tang, R., Chuang, Y.-N., & Hu, X. (2023). The Science of Detecting LLM-Generated Texts. arXiv (Cornell University).

Tolstykh, I., Tsybina, A., Yakubson, S., Gordeev, A. A., Dokholyan, V., & Kuprashevich, M. (2024). GigaCheck: Detecting LLM-generated Content. arXiv (Cornell University).

Trivedi, A., Suhm, A., Mahankal, P., Mukuntharaj, S., Parab, M. D., Mohan, M., Berger, M., Sethumadhavan, A., Jaiman, A., & Dodhia, R. (2021). Defending Democracy: Using Deep Learning to Identify and Prevent Misinformation. arXiv (Cornell University).

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. arXiv (Cornell University).

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions.