International Journal of Aquatic Science
ISSN: 2008-8019
Vol 15, Issue 01, 2024

A Comprehensive Review of Fake News Detection and Classification Using Machine Learning Models with Performance Metrics

G. Revathi[1], Dr. G. K. Arun[2]

[1]Research scholar, PG & Research Department of Computer Science, Arignar Anna Government Arts College, Villupuram -605602.
[2]Assistant professor, PG & Research Department of Computer Science, Arignar Anna Government Arts College, Villupuram -605602.

Email: [1]revathigovind1987@gmail.com, [2]arunnura2370@gmail.com

Abstract: The spread of fake news threatens the reliability of information and public trust. This paper provides a detailed review of techniques for detecting and classifying fake news using machine learning models, focusing on their performance based on key evaluation metrics. It examines advanced methods, including supervised, unsupervised, and hybrid approaches, to assess their ability to identify and classify fake news. The study also highlights feature engineering strategies, such as linguistic, content-based, and network-based features, which play a crucial role in improving model performance. Moreover, it reviews commonly used datasets, explains evaluation metrics like accuracy, precision, recall, F1-score, and AUC-ROC, and identifies existing challenges and research gaps. This review serves as a valuable resource for researchers and practitioners, offering insights to develop more accurate, efficient, and scalable fake news detection systems.

Keywords: Fake News, Machine Learning Models, Classification Techniques, Feature Engineering, and Performance Metrics

1. Introduction

The exponential proliferation of social media and digital platforms has fundamentally redefined the paradigms of information consumption, concurrently facilitating the extensive propagation of deceptive content, colloquially termed as "fake news." Fake news, characterized as spurious or intentionally misleading information masquerading as credible reportage, engenders profound impediments to the veracity of information and erodes societal trust. The ramifications of this phenomenon permeate public sentiment, influence political discourse, and destabilize social equilibrium, thereby accentuating the imperative for sophisticated detection and classification methodologies.

Machine learning (ML) and deep learning (DL) frameworks have emerged as formidable instruments for the identification and categorization of fake news. These paradigms leverage large-scale datasets to autonomously extrapolate intricate patterns, often imperceptible to human cognition. This paper furnishes an exhaustive review of machine learning methodologies for fake news detection, with an emphasis on supervised, unsupervised, and hybrid models. The review elucidates salient feature engineering methodologies, encompassing linguistic, content-based, and network-oriented features, which play a pivotal role in augmenting the predictive efficacy of the models.

2. REVIEW OF LITERATURE

A myriad of scholarly inquiries has been devoted to the deployment of machine learning architectures for the detection of counterfeit news. For instance, Zhang et al. (2019) operationalized support vector machines (SVMs) to classify news content predicated on linguistic attributes, culminating in substantial enhancements in classification precision. Likewise, Gupta et al. (2020) advocated a hybridized methodology that amalgamated decision trees with ensemble learning mechanisms to bolster detection performance. Kumar et al. (2021) examined the applicability of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in the context of fake news detection, substantiating that deep learning architectures possess a pronounced capacity for apprehending contextual nuances inherent in textual data.

Feature engineering assumes a quintessential role in fortifying the prognostic proficiency of ML models. Linguistic attributes, such as lexical, syntactic, and semantic markers, have been extensively leveraged for the detection of spurious news (Shu et al., 2019). Content-oriented features, inclusive of term frequency metrics and part-of-speech annotations, aid in discerning stylistic idiosyncrasies distinctive to fake news (Pérez-Rosas et al., 2018). Network-oriented features, which scrutinize the interconnections among content, sources, and user engagement, offer a macroscopic perspective on the dissemination of misinformation (Vosoughi et al., 2018).

Benchmark datasets constitute an indispensable asset for the development and validation of fake news detection models. Prominent datasets such as LIAR (Wang, 2017) and FakeNewsNet (Shu et al., 2020) furnish standardized corpora for model training and evaluation, enabling comparative analyses of algorithmic efficacy. Evaluation indices, including accuracy, precision, recall, F1-score, and AUC-ROC, are conventionally employed to appraise model performance, thereby providing a holistic assessment of predictive robustness.

This review aspires to furnish academicians and industry practitioners with an incisive comprehension of the contemporary landscape of fake news detection using machine learning paradigms. By delineating prevailing trends, spotlighting lacunae in extant methodologies, and proposing prospective avenues for research, this review endeavors to catalyze the evolution of more precise, scalable, and resilient detection frameworks.

The identification of counterfeit news has attracted considerable scholarly interest, with seminal investigations delving into linguistic and psychological indicators inherent in deceptive narratives (Rubin et al., 2015). Subsequent inquiries accentuated the role of stylometric attributes, facilitating distinctions between fabricated and hyper-partisan reporting (Potthast et al., 2018). On digital and social media platforms, data mining methodologies have been meticulously examined as tools to counteract the dissemination of misinformation (Shu et al., 2017). Zhou and Zafarani (2020) conceptualized a theoretical paradigm, elucidating detection methodologies and addressing obstacles such as dataset imbalances and the dynamic evolution of misinformation trends.

Various Machine Learning approaches with performance metrics. The present study with different experimental results and numerical illustrations using the weather dataset and its different conditions. Data mining, which serves as a valuable tool, involves the practice of examining extensive pre-existing databases to uncover previously unknown and useful information (Rajesh et al. 2017, Rajesh et al. 2019). The input data for chronic disease analysis represents specific locations as rows, while the attributes encompass topics, questions, data values, as well as low and high confidence limits. All the collected data is utilized for both training and testing purposes by employing five classification algorithms. This paper presents a detailed analysis and evaluation of the accuracy of these five different decision tree algorithms, demonstrating that the M5P decision tree approach emerges as the most effective model-building algorithm compared to the others (Rajesh et al. 2019, Rajesh et al. 2020, Rajesh et al. 2021).

Lexical intricacies and structural peculiarities in fabricated news were underscored by Horne and Adali (2017), while Pérez-Rosas et al. (2018) introduced curated datasets that enhanced the efficacy of machine learning algorithms in detecting inaccuracies. The mechanisms underlying the propagation of spurious news were scrutinized by Vosoughi et al. (2018), who juxtaposed the dissemination dynamics of veritable and fraudulent information. Visual analysis further contributes significantly, with Gupta et al. (2013) examining fabricated visual content proliferating on social platforms, and Jin et al. (2016) devising innovative visual and statistical feature sets for image authentication.

The development of benchmark datasets has profoundly advanced detection capabilities. Wang (2017) presented the LIAR dataset, which facilitated the assessment of supervised learning models for fake news categorization. Similarly, Thorne et al. (2018) introduced the FEVER dataset, concentrating on factual verification and information extraction. Hybrid models, as demonstrated by Ruchansky et al. (2017), integrate textual content with social context features, achieving notable improvements in detection precision. Trust-centric frameworks, such as those proposed by Nakamura and Levy (2021), enhance detection mechanisms by incorporating credibility metrics.

State-of-the-art progress in machine learning, particularly through transformer-based architectures like BERT, has achieved exceptional results in classifying deceptive content (Gupta et al., 2021). Convolutional Neural Networks (CNNs) have also proven effective, with Yang et al. (2019) leveraging both textual and visual elements to bolster detection accuracy. Moreover, psychological studies, such as those conducted by Pennycook and Rand (2019), explored behavioral insights, revealing how user perceptions of accuracy are modulated by cautionary labels on misleading information.

Metadata-driven analysis complements textual methodologies, as evidenced by Chandra and Malaya (2020), who amalgamated natural language processing techniques with metadata-centric features. Source credibility has been another focal point, with Baly et al. (2018) evaluating the factuality and inherent biases of various media outlets. Furthermore, the employment of stylometric and sentiment analysis has proved instrumental in discerning deceptive writing styles (Ghosh & Shah, 2018). Comprehensive frameworks integrating these methodologies are increasingly prioritized for their robust scalability and heightened accuracy (Kumar & Shah, 2018). The confluence of natural language processing, machine learning, and metadata analysis has been integral to constructing resilient models for the real-time detection of fabricated news. This interdisciplinary strategy offers holistic solutions adept at navigating the complexities posed by continuously evolving misinformation. By harnessing heterogeneous datasets and cutting-edge algorithms, the research community persistently advances efforts to mitigate the widespread impact of fake news within digital ecosystems.

3. Algorithm Comparison for Fake News Detection

The detection of fake news encompasses a broad spectrum of algorithmic methodologies, with logistic regression standing out as one of the fundamental supervised learning models, delivering notable performance on datasets such as LIAR (Pérez-Rosas et al., 2018), while Support Vector Machines (SVM) have exhibited exceptional precision in discerning hyperpartisan content, particularly in applications involving the BuzzFeedNews dataset (Potthast et al., 2018). Random forest algorithms have proven adept at harnessing content-based features, achieving robust outcomes on datasets like ISOT (Zhou and Zafarani, 2020), whereas simpler models, such as Naïve Bayes and decision trees, although pivotal to early advancements, frequently lag behind when tested against more intricate datasets like BuzzFeedNews and LIAR (Rubin et al., 2015; Shu et al., 2017).

Transformer-based architectures, epitomized by BERT, have revolutionized fake news detection through the utilization of pre-trained models that enable profound contextual comprehension, achieving a remarkable accuracy exceeding 94% on the FEVER dataset (Gupta et al., 2021). Similarly, deep learning paradigms such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) have been strategically employed to integrate textual and visual features, yielding effective results on datasets like GossipCop and PolitiFact (Yang et al., 2019; Ruchansky et al., 2017). Hybrid frameworks, exemplified by the CSI model, adeptly merge social context with content features to deliver scalable and precise detection capabilities (Ruchansky et al., 2017). Furthermore, network-based approaches, including Graph Neural Networks (GNNs), are emerging as promising tools due to their proficiency in modeling the intricate relational dynamics underpinning misinformation dissemination, as demonstrated in datasets like PolitiFact and GossipCop (Monti et al., 2019).

Table 1: Algorithm Comparison for Fake News Detection

Algorithm	**Approach**	**Accuracy (%)**	**Dataset**	**Reference**
Logistic Regression	Supervised	85.6	LIAR	Pérez-Rosas et al. (2018)
Support Vector Machine (SVM)	Supervised	88.2	BuzzFeedNews	Potthast et al. (2018)
Random Forest	Supervised	86.7	ISOT	Zhou and Zafarani (2020)
Naïve Bayes	Supervised	83.5	LIAR	Rubin et al. (2015)
Decision Tree	Supervised	80.4	BuzzFeedNews	Shu et al. (2017)

BERT	Transformer-based	94.3	FEVER	Gupta et al. (2021)
Convolutional Neural Network (CNN)	Deep Learning	91.2	GossipCop	Yang et al. (2019)
Long Short-Term Memory (LSTM)	Deep Learning	89.8	PolitiFact	Ruchansky et al. (2017)
Hybrid Model (CSI)	Hybrid	92.4	LIAR	Ruchansky et al. (2017)
Graph Neural Network (GNN)	Network-based	93.0	PolitiFact, GossipCop	Monti et al. (2019)

3. RESULTS AND DISCUSSION

The comparison of fake news detection algorithms shows varying levels of effectiveness across datasets. Transformer-based models and deep learning approaches consistently outperform traditional supervised learning methods due to their advanced feature extraction and contextual analysis capabilities. Key findings include the following.

Transformer-Based Models: BERT achieves the highest accuracy (94.3%) on the FEVER dataset by leveraging contextual embeddings and pre-trained knowledge.

Deep Learning Models: CNNs and LSTMs perform well with 91.2% and 89.8% accuracy, respectively, on datasets like GossipCop and PolitiFact, excelling in processing both textual and visual features for multimodal detection.

Hybrid Approaches: The CSI model, which integrates social context and content features, achieves 92.4% accuracy on the LIAR dataset, demonstrating the strength of hybrid frameworks in combining different data sources for better detection.

Traditional Supervised Learning Models: Algorithms like Logistic Regression (85.6%) and SVM (88.2%) perform well on simpler datasets, but their effectiveness diminishes with complex data. Naïve Bayes (83.5%) and Decision Tree (80.4%) are less effective due to their inability to manage the complexity and high dimensionality of fake news data.

Network-Based Models: Graph Neural Networks (GNNs) achieve 93.0% accuracy on datasets like PolitiFact and GossipCop, showcasing their ability to analyze relational and network-based features in misinformation spread.

4. CONCLUSIONS

Dataset Complexity: Simpler models like Decision Trees perform better on less complex datasets but fail with nuanced, high-dimensional datasets like FEVER. Transformer-**Based Models:** BERT's superior performance highlights the value of contextual word embeddings and transfer learning for detecting linguistic subtleties.

Multimodal Approaches: CNNs and LSTMs are effective for scenarios involving both text and images, often used in fake news dissemination. **Hybrid Methods:** The success of the CSI model emphasizes the advantage of integrating content and social context to enhance detection accuracy.

GNNs' Potential: GNNs offer promise in understanding misinformation propagation through network analysis, paving the way for innovative research. Limitations of Traditional Models: While simpler algorithms are easier to interpret, they are less capable of capturing the complexity required for modern detection tasks.

Future Directions

Develop robust hybrid models combining multimodal and network-based features. Expand datasets to include diverse languages and cultural contexts. Explore unsupervised and semi-supervised approaches to address data imbalance and limited labeled samples. Enhance explainability in models like BERT and GNNs to improve transparency in decision-making. This analysis provides insights into the strengths and weaknesses of different algorithms, guiding the development of more accurate and scalable fake news detection systems.

5. REFERENCES

[1] Alam, F., Dalvi, F., & Sajjad, H. (2021). Feature selection for fake news detection: A comparative study. Knowledge-Based Systems, 211, 106-122.

[2] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting factuality of reporting and bias of news media sources. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3528-3539.

[3] Chandra, A., & Malaya, D. (2020). A hybrid approach for fake news detection using machine learning and NLP. Journal of Information Technology Research, 13(2), 1-14.

[4] Chandra, A., & Sharma, D. (2019). Comparative study of traditional and deep learning methods for fake news detection. Applied Computing and Informatics, 7(2), 89-102.

[5] Chen, W., & Wu, Z. (2021). A hybrid deep learning approach for fake news detection. Machine Learning Research, 17(2), 111-125.

[6] Faris, R., Roberts, H., & Zuckerman, E. (2018). Fake news detection using content and network-based features. Harvard Misinformation Review, 4(2), 25-36.

[7] Gao, L., Zhang, X., & Lin, J. (2021). Leveraging transfer learning for fake news detection. Journal of Artificial Intelligence Research, 45(1), 78-90.

[8] Ghosh, S., & Shah, C. (2018). Toward automatic fake news classification. Proceedings of the 2018 ACM SIGIR Conference on Research and Development in Information Retrieval, 321-324.

[9] Gupta, A., Lamba, H., & Kumaraguru, P. (2013). Faking sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. Proceedings of the 22nd International Conference on World Wide Web, 729-736.

[10] Gupta, H., Sahu, T., & Shrivastava, D. (2021). Detection of fake news using a transformer-based model. Applied Intelligence, 51(2), 984-998.

[11] Gupta, R., Yadav, P., & Sharma, K. (2020). A hybrid approach to fake news detection using decision trees and ensemble methods. Information Systems Research, 26(3), 112-130.

[12] Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 217-226.

[13] Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2016). Novel visual and statistical image features for microblogs news verification. IEEE Transactions on Multimedia, 19(3), 598-608.

[14] Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. Social Media Analytics, 1-34.

[15] Kumar, V., Singh, A., & Mishra, S. (2021). Deep learning for fake news detection: A review of CNN and RNN-based approaches. Neural Computing & Applications, 32(2), 177-195.

[16] Li, H., & Yu, Z. (2020). Role of deep learning models in automated fake news detection. Expert Systems with Applications, 42(1), 561-574.

[17] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using graph neural networks. arXiv preprint arXiv:1902.06673.

[18] Nakamura, K., & Levy, K. (2021). Trust-based approaches for fake news detection. Information Processing & Management, 58(4), 102650.

[19] Nasir, J., Hussain, A., & Usman, M. (2020). Identifying fake news using BERT and transfer learning. IEEE Access, 8, 18132-18141.

[20] Patwa, P., Sharma, S., & Pykl, S. (2021). Fighting COVID-19 misinformation on social media: A case study of fake news detection. Social Network Analysis and Mining, 11(1), 44-62.

[21] Pennycook, G., & Rand, D. G. (2019). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Management Science, 66(11), 4944-4957.

[22] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. Proceedings of the 27th International Conference on Computational Linguistics, 5(2), 3391-3401.

[23] Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 231-240.

[24] Rajesh, P., & Karthikeyan, M. (2017). A comparative study of data mining algorithms for decision tree approaches using weka tool. Advances in Natural and Applied Sciences, 11(9), 230-243.

[25] Rajesh, P., Karthikeyan, M., Santhosh Kumar, B., & Mohamed Parvees, M. Y. (2019). Comparative study of decision tree approaches in data mining using chronic disease indicators (CDI) data. Journal of Computational and Theoretical Nanoscience, 16(4), 1472-1477.

[26] Rajesh, P., & Kumar, B. S. (2020). Comparative studies on Sustainable Development Goals (SDG) in India using Data Mining approach. J. Sci, 14(2), 91-93.

[27] Rajesh, P., & Govindarasu, V. (2021). Analyzing and Predicting Covid-19 Dataset in India using Data Mining with Regression Analysis. International Research Journal on Advanced Science Hub, 3(7S), 91-95.

[28] Rajesh, P., & Karthikeyan, M. (2019). Data assimilation of gross domestic product (GDP) in India using stochastic data mining approach. Journal of Computational and Theoretical Nanoscience, 16(4), 1478-1484.

[29] Reddy, Y., & Reddy, C. (2022). Evaluation of ML models for fake news detection. Artificial Intelligence Review, 45(3), 75-94.

[30] Roy, R., Sinha, K., & Choudhary, A. (2022). Performance analysis of fake news detection using neural networks. Neural Networks, 54(3), 124-136.

[31] Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: An exploratory study. Decision Support Systems, 57, 254-264.

[32] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797-806.

[33] Shu, K., Mahudeswaran, D., & Liu, H. (2020). FakeNewsNet: A data repository with news content, social context, and spatiotemporal information. Data Mining and Knowledge Discovery, 34(1), 1-6.

[34] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[35] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2019). Detecting fake news on social media: Features, benchmarks, and models. ACM Transactions on Information Systems (TOIS), 38(1), 1-27.

[36] Singh, R., Garg, N., & Thakur, N. (2021). Comparative analysis of machine learning algorithms for fake news detection. International Journal of Data Science and Analytics, 6(3), 78-95.

[37] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 809-819.

[38] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

[39] Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 3(1), 422-426.

[40] Yang, Y., Zheng, L., Zhang, J., Cui, Q., & Li, Z. (2019). TI-CNN: Convolutional neural networks for fake news detection. Information Retrieval Journal, 22(3-4), 259-288.

[41] Zhang, X., Wu, Y., & Zhao, Y. (2019). Fake news detection using linguistic features and support vector machines. Journal of Computational Social Science, 12(4), 45-60.

[42] Zhao, Y., Wang, J., & Wu, Y. (2020). Content and sentiment analysis for fake news classification. Pattern Recognition Letters, 144, 88-97.

[43] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and challenges. ACM Computing Surveys (CSUR), 53(5), 1-40.