Second International Conference on Smart Computing and Communication for Sustainable Convergence (ISCCSC 2025)

Fake News Detection: A Survey on Multivariate Feature Selection and Hybrid Deep Learning

Shrikanth N G[1], Jhanavi M P[2], Rakesh G[3] and Srushti Suresh Yadahalli[4]

Department of Artificial Intelligence And Machine Learning, Alvas Institute of Engineering & Technology, Moodbidri, Mangalore, Karnataka, India

Emails: shrikanthng@aiet.org.in, jhanavimp16@gmail.com, grakesh.aiml@gmail.com, yadahallisrushti@gmail.com,

Abstract.

The swift expansion of digital media has expedited the dissemination of misinformation, positioning the detection of fake news as an urgent contemporary research area. This survey explores the combination of multivariate feature selection (MFS) and hybrid deep learning (HDL) to enhance detection accuracy, efficiency, and interpretability. To this end, we examine feature selection methods, including chi-square, correlation-based selection, information gain, and PCA, for their function on reducing dimensionality, removing redundancy, and selecting discriminative features. Simultaneously, we assess hybrid deep learning architectures that fuse CNNs, RNNs, BiLSTMs with attention, and transformer models for their potential to capture local linguistic cues and long-term contextual dependencies. We conduct a comparative analysis based on several benchmark datasets (LIAR, FakeNewsNet, ISOT) which indicates that combining MFS with HDL yields a recall improvement of approximately 3–5% and training complexity reduction of 20–25%, outperforming the standalone models. Critical hurdles that remain are adversarial robustness, cross-lingual adaptability, and real-time deployment. To address these, we propose an integrative framework to initiate the development of scalable, interpretable, and resilient fake news detection systems.

Keywords. Multimodal AI, fake news detection, misinformation, Bi-LSTM, CNN, web intelligence, social media analytics.

1. INTRODUCTION

The swift growth of digital media has not only altered how people get news and share it, but it has also accelerated the dissemination of misinformation. Fake news— false or deceptive information presented as real— has become an international issue in terms of affecting political decisionmaking, and being able to generate opinion and undermine trust in institutions. Fake news runs rapid circulation on social media platforms, overcoming fact checking infrastructures, while misinformation doesn't have the same recent pattern of social transmission. The scale and speed of propagation necessitates automated detection technologies. Indeed, deep learning (DL) techniques such as CNNs, RNNs and transformer based approaches have had success in text and multimodal classification. However, DL

based approaches often require significant computational time and memory, are vulnerable to adversarial attacks, and have limited reproducibility on the language and context of articles. Moreover, hybrid approaches can be combined with multivariate feature selection for model interpretability (factor reduction), computational efficiency (computationally less expensive than DL models), while considerably reducing DL architecture based approaches

2. LITERATURE REVIEW

[1] Brightlin et al. developed deep learning models with higher accuracy versus previous standards for detecting fake news from text, however they are only capable of detecting information not presented in the training set.

[2] Mezaris et al. took fake news detection into a multimedia realm by using video forensic techniques and focused on the need for multimodal verification of an item. Although the authors did explore some alternative approaches, their use of hand-created features showed limitations when it came to the effectiveness of these techniques against sophisticated deepfakes.

[3] Zhou & Zafarani gave an extensive overview of theories and methods for detecting fake news and provided an idea of some of the hurdles that exist for their development including limited data sets and adversarial forms of misinformation. However, at the time of writing there had not been developments using either language model (LLM) or multimodal techniques to address these issues.

[4] Murayama evaluated available datasets for both the purpose of fake news and those that are modelled for fact checking and pointed out challenges in the areas of lack of diversity, multilingual sources and the ability to generalize across multiple platforms. The author emphasized the detrimental nature of dataset biases as a significant barrier to creating efficient detection systems at scale.

[5] Caron et al. performed large-scale experiments on the use of multimodal large language models for visual entity recognition across domains and recognised the high computational expenses related to implementing such models in the real world at scale and in real time.

[6] Setiawan et al. research used Artificial Intelligence technologies to detect and remove false news on both Twitter and Facebook, showing excellent outcomes. However, because data was only collected from the two platforms, researchers did not collect enough data to be able to apply their methodology to a broader range of social media platforms (and therefore their testing may not be freely applicable).

[7]Bai et al. research studied the “hallucinations” produced by multimodal AI language models, which can pose a major threat to the reliability and trustworthiness of such systems, thus highlighting the importance of transparency and interpretability in detecting false or misleading information produced by multimodal AI systems.

[8] Birhane et al. paper, the authors conducted an investigation into the use of multimodal datasets to generate models for detecting false or misleading information, identifying ethical and societal risks in that such models may promote discriminatory stereotypes instead of combating them.

3. METHODOLGY

3.1 Dataset Collection and Pre-Processing

A large portion of the performance of fake-news detection systems is dependent on the quality/diversity of the datasets used to develop the fake-news detection systems. Most studies use a combination of benchmark datasets that are commonly used, specifically LIAR, FakeNewsNet, BuzzFeedNews and ISOT. Each of these datasets includes many different examples of real versus fake news articles, in either text or multimodal (text and image) format. Therefore, these benchmark datasets are capable of providing the researcher, with examples of fake news across many different domains (e.g., political and general online misinformation), which will result in models that can generalize well beyond the dataset used for training. Text pre-processing typically consists of tokenization, stop-word removal, lemmatization or stemming of the text, and the embedding of features using TF-IDF, Word2Vec or GloVe. Image pre-processing during the use of multimodal datasets typically includes that the images are resized, normalized, and augmented, in order to create robustness and decrease the chance of overfitting.

3.2 Feature Selection

Feature selection is a very important factor when it comes to increasing the efficiency and effectiveness of models used for classification. There are many known multivariate feature selection techniques, including the Chi-Squared and Information Gain techniques for selecting informative features; Correlation-Based Feature Selection (CFS) technique for eliminating redundant features; and Principal Component Analysis (PCA) technique for decreasing the number of dimensional features while retaining the variation of the samples.

3.3 Architectural Types:

Hybrid deep learning architectur types have been adopted by researchers to incorporate spatial and temporal information into the predictive model. Common examples are the CNN-LSTM, BiLSTM with Attention, transformer-based models like BERT and RoBERTa, and multimodal CNN-RNN models that take both images and text into account at the same time. Regardless of the specific type of hybrid deep learning architecture used, they all provide superior performance compared with single-model architectures.

3.4 Training and Evaluation:

Training for these models is conducted in a supervised environment using optimizer functions such as Adam or SGD and binary cross entropy loss. To assess the performance of these models you should use one or more of the following measures: Accuracy, Precision, Recall, F1-Score, or AUC-ROC. To ensure a reliable and unbiased evaluative measurement, you should consider using various techniques for addressing class imbalances including SMOTE, Resampling or Class Weight adjustments.

4. CONCLUSION

This review surveyed the role of multivariate feature selection (MFS) and hybrid deep learning (HDL) in the detection of fake news. Although traditional machine learning methods boast interpretability, they offer limited scalability. Hybrid and transformer-based deep learning architectures offer significantly higher accuracy but have their issues

with efficiency and transparency. Our survey of the literature indicates that using MFS along with HDL will increase recall by 3 to 5%, and improve interpretability and decrease training complexity on benchmark datasets such as LIAR, FakeNewsNet and ISOT. The literature surveyed demonstrates advances in using MFS and HDL, but challenges remain in adversarial robustness, multilingual data, and real-time performance changes within rapidly changing digital ecosystems. Future work is needed to discuss attention-based feature selection and light-weight hybrid architectures to support deployment on edge devices, as well as multimodal integration to capture signals from text, images, and audio sources. Althought this study also extends the dialogue between MFS and HDL, the potential lies in components to design robust, efficient and accountable solutions that will provide new avenues to stem misinformation at scale.

Approach	**Reported Accuracy**	**Strengths**	**Limitations**	**Improvements with Proposed Framework**
Traditional ML (SVM, LR, DT)	80–84% (FakeNews-Net)	High efficiency, interpretable	Poor scalability, low performance on multimodal data	Baseline reference only
Hybrid CNN+LSTM / BiLSTM+Attention	88–92% (FakeNews-Net, LIAR)	Captures local + sequential features	Computationally heavy, prone to overfitting	+2–3% recall improvement when combined with feature selection
Transformer Models (BERT, RoBERTa)	90–96% (LIAR, ISOT)	Strong contextual representation, state-of-the-art accuracy	High computational cost, black-box nature	Feature selection reduces dimensionality, improving training speed by ∼20–25%
Multimodal CNN-RNN (Text+Image)	91–94% (FakeNews-Net)	Handles both textual & visual signals	Requires large multimodal datasets; integration challenges	Integration with feature selection reduces redundancy, improving generalization
Feature Selection + Hybrid Models (Proposed)	**+3–5% recall gain across datasets**	Improves interpretability, efficiency, and robustness	Risk of discarding rare features	Outperforms standalone methods by combining dimensionality reduction with hybrid DL, yielding better scalability and resilience

TABLE I : COMPARATIVE ANALYSIS OF FAKE NEWS DETECTION APPROACHES

5. REFERENCE

[1] D. Brightlin, V. Mohanambal, N. Nagapadmavathi, and P. Vaishnavi, "Fake news detection using deep learning techniques," 2024.

[2] V. Mezaris, L. Nixon, S. Papadopoulos, and D. Teyssou, Video verification in the fake news era. Springer, 2019, vol. 4.

[3] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," ACM Computing Surveys (CSUR), vol. 53, no. 5, pp. 1–40, 2020.

[4] T. Murayama, "Dataset of fake news detection and fact verification: a survey," arXiv preprint arXiv:2111.03299, 2021.

[5] M. Caron, A. Fathi, C. Schmid, and A. Iscen, "Web-scale visual entity recognition: An llm-driven data approach," Advances in Neural Information Processing Systems, vol. 37, pp. 34 533–34 560, 2024.

[6] R. Setiawan, V. S. Ponnam, S. Sengan, M. Anam, C. Subbiah, K. Phasinam, M. Vairaven, and S. Ponnusamy, "Certain investigation of fake news detection from facebook and twitter using artificial intelligence approach," Wireless Personal Communications, vol. 127, no. 2, pp. 1737–1762, 2022

[7] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," arXiv preprint arXiv:2404.18930, 2024.

[8] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," arXiv preprint arXiv:2110.01963, 2021.

Biographies

Name: Shrikanth N G
Designation: Sr.Assistant Professor (Project Guide)
Email ID: shrikanthng@aiet.org.in
Mobile Number: +91 9880410030
Areas of Interest: As a professor in the Department of Artificial Intelligence And Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka.

Name : Jhanavi M P
USN : 4AL22AI020
Email : jhanavimp16@gmail.com
Mobile : 7996799399
Areas of Interest: She is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. She is academic focus extends to a keen interest in

Name : Rakesh G
USN : 4AL22AI038
Email :grakesh.aiml@gmail.com
Mobile : 7795362467
Areas of Interest: He is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. he is interested in research fields like Machine Learning, Data Science, Deep Learning and Artificial Intelligence. Demonstrating a commitment to these revolutionary field's knowledge advancement and

Name : Srushti Suresh Yadahalli
USN : 4AL22AI055
Email : yadahallisrushti@gmail.com
Mobile : 8618728154
Areas of Interest: She is currently pursuing a Bachelor's degree in Artificial Intelligence and Machine Learning at Alva's Institute of Engineering and Technology (AIET), affiliated with Visvesvaraya Technological University (VTU) in Karnataka. Her academic focus extends to a keen interest in areas like Web Technologies, Natural Language Processing (NLP), Generative AI, Cloud Computing, Deep Learning, Machine Learning, showcasing a dedication to advancing knowledge and fostering innovation in these transformative fields.