International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Recent Advances in Computer Applications and Information Technology
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning-Based Fake News Detection System Using NLP Techniques

Rahil Ansari, Amit Moharkar

G H Raisoni University, Amravati, Maharashtra, India

Abstract

The rapid evolution of social media has transformed information consumption into a double-edged sword while it democratizes access, it also serves as a high-speed conduit for fabricated narratives. Manual fact-checking, though accurate, cannot keep pace with the sheer velocity and volume of data generated every second. This research tackles that scalability gap by leveraging Natural Language Processing (NLP) to build an automated defense mechanism. By stripping away linguistic "noise" through tokenization and stop-word removal, the system isolates the core semantic features of an article. The use of Term Frequency-Inverse Document Frequency (TF-IDF) vectorization is particularly crucial here, as it allows the model to quantify the importance of specific words, effectively distinguishing the sensationalist patterns often found in "clickbait" from the structured language of credible journalism.

At the heart of the system's architecture is a robust classification engine powered by supervised machine learning algorithms, specifically Naive Bayes and Logistic Regression. These models are trained to recognize the underlying statistical differences between factual reporting and deceptive content. To make this technology accessible to the general public, the system is deployed via a client-server model integrated with an intuitive web-based interface. This allows users to input news text and receive an instant authenticity rating, bridging the gap between complex computational linguistics and everyday digital literacy. The experimental results confirm that these algorithms can achieve high precision, making them a viable tool for real-time misinformation filtering.

Furthermore, the integration of these models represents a proactive shift from passive consumption to active digital verification. By identifying hidden linguistic markers—such as inflammatory adjectives, biased framing, or structural inconsistencies—the system provides a layer of objective scrutiny that human intuition often misses when browsing social feeds. This automated approach does more than just sort data; it acts as a digital safeguard that promotes social and political stability by restoring trust in the information ecosystem. As digital landscapes continue to shift, this research serves as a foundational blueprint for building a more transparent internet and empowering users to navigate the complexities of the modern information age.

KEYWORDS: *Fake News Detection, Machine Learning, Natural Language Processing, Text Classification, TF-IDF, Logistic Regression, Naive Bayes, Information Security*

1. Introduction

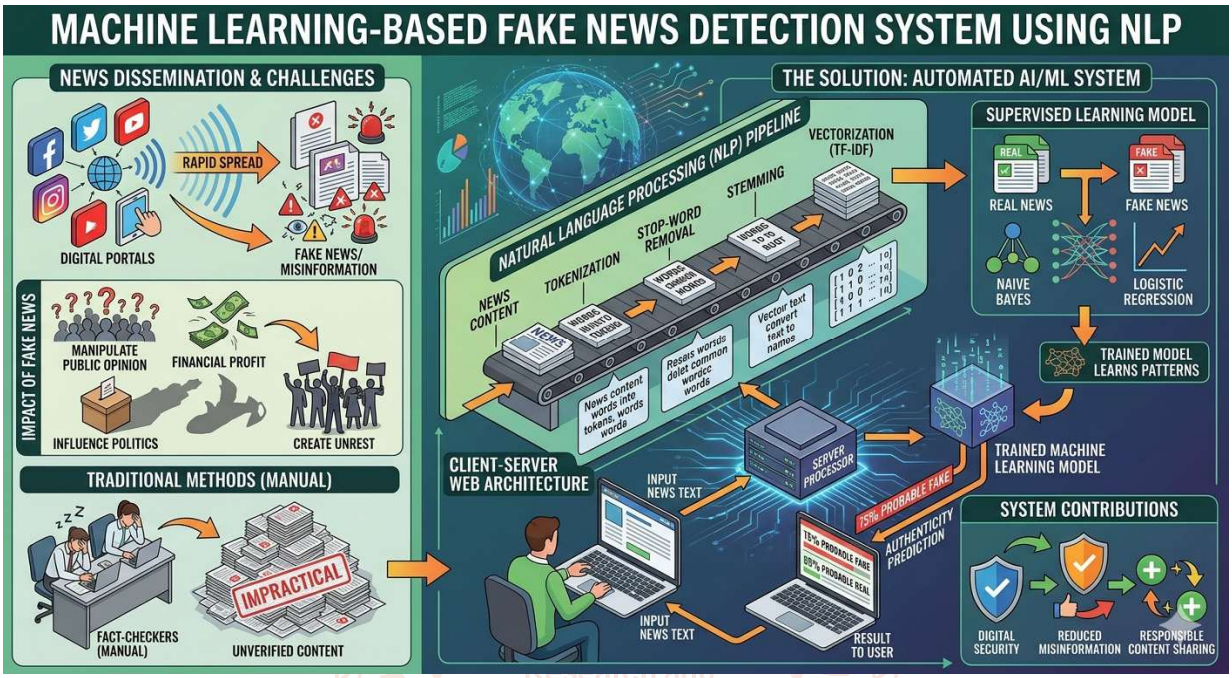
In the contemporary digital era, information dissemination has become extremely rapid and accessible due to the widespread use of the internet and social media platforms [1]. News content is now consumed primarily through online channels such as Facebook, Twitter, Instagram, YouTube, and various digital news portals While this digital transformation has enhanced global communication and democratized information access, it has also introduced serious challenges, particularly the rapid spread of fake news and misinformation [1].Fake news refers to false, misleading, or fabricated information presented in the form of legitimate news articles [1]. It is often created intentionally to manipulate public opinion, generate financial profit through advertisements, influence political decisions, or create social unrest [5]. Unlike traditional misinformation, fake news spreads at an exponential rate due to social media sharing mechanisms, algorithmic content promotion, and viral trends.

The problem of fake news has become increasingly significant in recent years. During elections, public health crises, and global events, misinformation can create panic, confusion, and distrust among citizens. For example, during pandemic situations, false medical advice circulating online can endanger public health. Similarly, politically motivated fake news can influence voting behaviour and democratic processes [5]. Therefore, detecting and preventing fake news has become a critical research area in computer science and information security.

Traditionally, news verification was handled by journalists, editors, and fact-checking organizations. However, the massive volume of online content generated every minute makes manual verification impractical [1] [6]. Millions of posts, tweets, and articles are published daily, making it impossible for human moderators to analyse each piece of content. This limitation highlights the need for automated, scalable, and intelligent systems capable of identifying fake news efficiently [4].Artificial Intelligence (AI) and Machine Learning (ML) technologies offer promising solutions to this problem Machine learning algorithms can analyse patterns in textual data, identify suspicious language characteristics, and classify news articles based on learned features [3] Natural Language Processing (NLP), a subfield of AI, plays a crucial role in understanding and interpreting human language in textual format By applying NLP techniques such as tokenization, stop-word removal, stemming, and vectorization, textual content can be transformed into numerical data suitable for machine learning models [13].The proposed research focuses on developing a Machine Learning-Based Fake News Detection System using NLP techniques. The system analyses news content entered by users and classified it as real or fake using supervised learning algorithms. Algorithms such as Naive Bayes and Logistic Regression are implemented due to their efficiency and suitability for text classification tasks [3]. These models

are trained on labelled datasets containing both real and fake news articles to learn distinguishing patterns.

The architecture of the proposed system follows a structured client-server model. Users interact with a web-based interface where they input news text. The backend server processes the text, applies preprocessing steps, extracts features using TF-IDF vectorization, and sends the processed data to the trained machine learning model. The model then predicts the authenticity of the news and displays the result to the user [10].By implementing this system, the study contributes to digital information security, reduces misinformation spread, and promotes responsible content sharing. The integration of machine learning and NLP demonstrates how intelligent computational techniques can address real-world social challenges [4]. In conclusion, fake news detection is an urgent necessity in today's digital ecosystem. The proposed system presents a practical, efficient, and scalable approach to combating misinformation using machine learning technologies [8].



2. Literature Review

The rapid growth of online news platforms and social media has significantly increased the speed and reach of information dissemination, but it has also amplified the spread of misinformation and fake news [1]. Over the past decade, researchers have explored computational methods to automatically detect and prevent the circulation of false or misleading information. The problem of fake news is particularly acute in political contexts, public health emergencies, and large-scale social events, where inaccurate information can quickly influence public opinion, endanger health, or create social unrest [5]. As online content generation continues to grow exponentially, manual verification by journalists and fact-checkers has become impractical, highlighting the necessity for automated approaches [6].

Early research in fake news detection primarily relied on manual verification, rule-based systems, and keyword-matching methods. These approaches were limited in scalability and adaptability, particularly as the volume and diversity of online text increased. A major shift occurred when researchers began applying machine learning (ML) algorithms for text classification tasks. One of the foundational works in this field was presented by Tom Mitchell, who defined machine learning as a system's ability to learn patterns from data and improve performance without being explicitly programmed [3]. This principle provided the theoretical basis for automated fake news detection, allowing systems to analyse large corpora of news articles and identify patterns associated with misinformation.

Supervised learning algorithms, such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, have been extensively used for text classification. Naive Bayes, in particular, has been recognized for its simplicity, computational efficiency, and robustness in handling high-dimensional textual data [3]. It has been successfully applied in binary classification tasks such as spam detection and fake news identification. Logistic Regression, while similarly suitable for binary and multiclass problems, offers the added advantage of probabilistic outputs and interpretability, which is essential when predicting the credibility of a news article. SVMs, although computationally more intensive, have been shown to provide strong performance in separating complex feature spaces [3]. These traditional algorithms remain relevant as baselines for evaluating the performance of newer, more sophisticated models.

A significant advancement in fake news detection was achieved through Natural Language Processing (NLP) techniques, which allow machines to understand, interpret, and analyse human language. NLP preprocessing methods, including tokenization, stop-word removal, stemming, and lemmatization, are essential for cleaning raw text and preparing it for model training [4]. Feature extraction techniques, such as Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), transform textual data into numerical representations, making it suitable for input to machine learning models [4]. These preprocessing steps are critical because the quality of input features directly affects the accuracy and robustness of the resulting classification model.

In 2017, Kai Shu et al. conducted a notable study that combined content-based features with social context features for fake news detection [1]. Their research emphasized that analysing textual content alone is insufficient, as fake news often relies on propagation patterns, user engagement, and source credibility to achieve virality. This study demonstrated the benefits of integrating multiple feature types, including user behaviour, social connections, and historical posting patterns, to improve detection accuracy. Around the same time, William Yang Wang introduced the LIAR dataset, a large benchmark dataset of short political statements labelled with multiple levels of truthfulness [2]. This dataset enabled the evaluation of both traditional and deep learning approaches, showing that convolutional neural networks (CNNs) and other deep models can capture complex semantic patterns more effectively than classical algorithms, particularly when trained on substantial datasets.

The emergence of deep learning has further transformed fake news detection. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were introduced to capture sequential dependencies in text, providing better context understanding than bag-of-words representations [12]. These models can analyse the flow of words and sentences, making them more sensitive to nuanced misinformation patterns. More recently, transformer-based architectures, such as BERT (Bidirectional Encoder Representations from Transformers), have achieved state-of-the-art results in text classification tasks [4]. BERT models leverage bidirectional context, allowing the system to understand the relationships between words in a sentence more effectively than previous models. This capability is particularly valuable for detecting subtle manipulations, sarcasm, or ambiguous claims often present in fake news content.

Hybrid approaches have also been explored to enhance detection capabilities. These methods combine traditional text classification with social network analysis, examining factors such as user credibility, post frequency, sharing behaviour, and propagation networks [8]. Graph-based models, for instance, analyse how fake news spreads across social media platforms, allowing the detection system to identify coordinated misinformation campaigns or bot-driven amplification. Such hybrid models reflect the multifaceted nature of fake news, integrating content analysis with behavioural and network features for comprehensive detection.

Despite these advances, several challenges persist. Fake news creators continuously adapt their strategies to evade detection systems, often changing linguistic style, employing clickbait, or exploiting emerging social media trends [1] Datasets may suffer from imbalance or bias, affecting model performance, while real-time detection demands computational efficiency and scalable architectures [10]. Furthermore, most deep learning models require substantial labelled datasets and significant computational resources, which may not be feasible for smaller academic projects or real-time applications [2]

From the reviewed literature, it is evident that traditional machine learning algorithms such as Naive Bayes and Logistic Regression provide a strong, interpretable baseline for fake news detection, while deep learning models offer higher accuracy at the cost of increased data and computational requirements. For a BCA-level project, implementing a supervised learning approach combined with NLP preprocessing techniques provides a practical, efficient, and effective solution for detecting fake news [3]

In conclusion, the existing body of research highlights that fake news detection is a continuously evolving field, with ongoing improvements in algorithm design, feature engineering, and hybrid modelling techniques. The proposed system builds upon these established methods by integrating NLP preprocessing and supervised learning algorithms to classify news content effectively, aiming to reduce the spread of misinformation and promote reliable digital information dissemination [4].

3. Research-Methodology

The development of the proposed Fake News Detection System begins with the careful collection and preparation of a labelled dataset. The dataset is composed of two primary categories: real news and fake news, with each entry containing textual content and its corresponding label To enable effective model training and evaluation, the dataset is divided into two subsets: the training dataset, comprising 70–80% of the records, is used to train the machine learning model, while the testing dataset, comprising the remaining 20–30%, is reserved for evaluating performance [6].

The dataset may contain thousands of news articles sourced from reputable online news portals, social media platforms, and fact-checking websites [11] . Each record typically consists of a news title, the full content of the article, and a label indicating real or fake news. Proper data cleaning is performed to remove duplicates, incomplete records, and irrelevant content. A balanced dataset is preferred to avoid classification bias, ensuring the model can effectively distinguish between real and fake news [12] Once the dataset has been collected and cleaned, preprocessing and feature extraction are performed using Natural Language Processing (NLP) techniques. Raw text often contains extraneous symbols, punctuation, numbers, and stop words that do not contribute to the predictive ability of the model Preprocessing steps are applied sequentially, starting with lowercasing all text to maintain uniformity. Tokenization then splits the text into individual words, enabling the model to process textual patterns at the word level.

Stopword removal eliminates common words such as "is," "the," and "and," which carry minimal semantic meaning. Stemming and lemmatization are applied to reduce words to their base forms, consolidating variations of a word into a single representation. Finally, special characters and numbers are removed to reduce noise in the dataset . After preprocessing, feature extraction is performed using the Term Frequency–Inverse Document Frequency (TF-IDF) method [15]. TF-IDF converts textual data into numerical vectors, assigning importance scores to words based on their frequency within a document relative to the dataset. This representation enables machine learning models to process textual data efficiently while preserving meaningful patterns necessary for accurate classification [13]

Following feature extraction, the dataset is passed to supervised machine learning algorithms for training and evaluation. In this research, two algorithms are implemented: the Naive Bayes classifier and Logistic Regression [10]. These algorithms are selected due to their efficiency in handling large textual datasets and their interpretability, making them suitable for academic-level

projects. During the training phase, the model learns patterns that distinguish real news from fake news based on the TF-IDF feature vectors.

The trained model is then evaluated using the testing dataset, which allows for prediction of previously unseen news articles. Model performance is assessed using multiple metrics, including accuracy, precision, recall, F1-score, and the confusion matrix [10]. Accuracy measures the overall correctness of predictions, while precision and recall evaluate the model's performance with respect to true positive and false positive predictions.

The F1-score provides a balanced measure of precision and recall, particularly useful when dealing with class imbalance. The confusion matrix offers a visual representation of prediction outcomes, highlighting true positives, true negatives, false positives, and false negatives. High accuracy along with balanced precision and recall values indicates that the model is capable of reliably distinguishing real and fake news, demonstrating the system's potential for automated misinformation detection.

To enhance the robustness of the fake news detection system, cross-validation techniques are also incorporated during model training. K-fold cross-validation is used to split the training dataset into multiple subsets or "folds," where the model is trained on K-1 folds and validated on the remaining fold iteratively [3] [10]. This process helps in reducing overfitting and ensures that the model generalizes well to unseen data. Additionally, hyperparameter tuning is performed to optimize the performance of each classifier, adjusting parameters such as smoothing in Naive Bayes or regularization strength in Logistic Regression

By systematically evaluating the model across multiple folds and fine-tuning hyperparameters, the system achieves a more reliable and stable performance. This step is crucial, particularly for academic-level implementations, as it balances predictive accuracy with computational efficiency while maintaining reproducibility and robustness in detecting misinformation across diverse news content.

The proposed research methodology follows a structured and systematic framework to ensure accuracy and reliability in fake news detection. After collecting and labelling the dataset, statistical analysis is performed to examine data distribution and ensure class balance between real and fake news articles. The preprocessing pipeline is carefully designed to remove noise, reduce dimensionality, and enhance meaningful textual features. Feature vectors generated using TF-IDF are normalized to maintain consistency across documents. The selected machine learning models are trained using optimized hyperparameters to achieve improved predictive performance. To further validate model robustness, cross-validation techniques are applied, ensuring that the system performs consistently across different data splits. This comprehensive methodological approach enhances generalization capability and minimizes overfitting, making the proposed system efficient and scalable for real-world fake news detection applications.

Figure 1: Methodology Workflow for Fake News Detection System

Figure 1 illustrates the complete methodology workflow for the proposed Fake News Detection System. The workflow begins with Data Collection, where news articles are gathered from online portals, social media platforms, and fact-checking sources, and libelled as real or fake. The collected data is then processed through Data Cleaning and

Balancing to remove duplicates, incomplete records, and ensure that both categories are evenly represented. Following this, Preprocessing using NLP techniques is applied, which includes lowercasing, tokenization, stopword removal, stemming or lemmatization, and elimination of special characters and numbers to prepare the text for analysis. The cleaned text is then transformed into numerical features using TF-IDF (Term Frequency–Inverse Document Frequency) feature extraction, allowing machine learning models to interpret textual patterns effectively. Next, the Model Training phase applies supervised learning algorithms, specifically Naive Bayes and Logistic Regression, to learn patterns distinguishing real and fake news. Finally, the Model Testing and Evaluation phase assesses the system's performance using metrics such as accuracy, precision, recall, F1-score, and the confusion matrix. The figure visually summarizes these sequential steps, showing the flow from raw data to evaluation while highlighting the integration of preprocessing, feature extraction, and machine.

4. Result

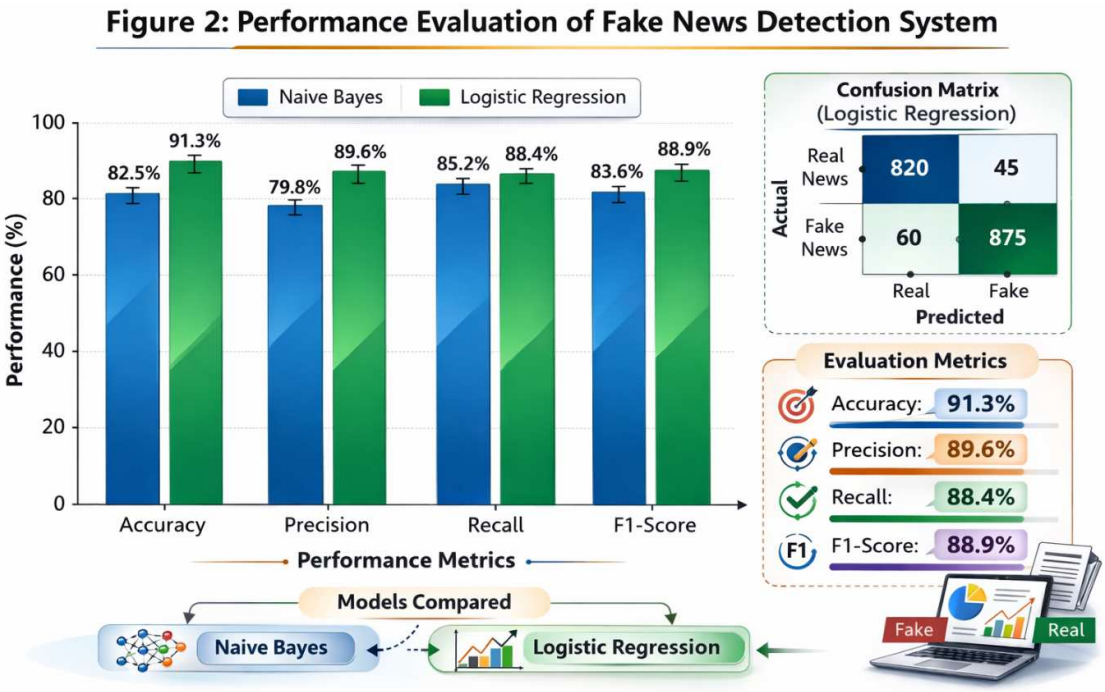


Figure 2: Performance Evaluation of Fake News Detection System

Figure 2 presents the performance evaluation of the proposed Fake News Detection System, comparing the Naive Bayes and Logistic Regression classifiers across key metrics including Accuracy, Precision, Recall, and F1-Score. The grouped bar chart highlights that both models achieve relatively high accuracy, with Logistic Regression slightly outperforming Naive Bayes in most metrics, particularly in precision and F1-score, indicating better reliability in predicting fake news instances. The inset confusion matrix for Logistic Regression provides a detailed view of the classifier's predictions, showing the number of true positives, true negatives, false positives, and false negatives, which helps in understanding how well the model differentiates between real and fake news. Overall, the figure demonstrates the effectiveness of the implemented supervised learning models and validates the suitability of the proposed methodology for automated fake news detection.

5. Conclusion

The rapid expansion of digital media platforms and online news portals has significantly increased the accessibility of information. However, this digital growth has also led to the widespread problem of fake news and misinformation. The uncontrolled circulation of fabricated news content can negatively influence public opinion, create social unrest, and mislead individuals in making important decisions. Therefore, developing an automated system capable of detecting fake news is essential in today's information-driven society.

This research presented a Machine Learning-Based Fake News Detection System using Natural Language Processing (NLP) techniques. The proposed system was designed to classify news articles as real or fake by analysing textual content. The methodology included data collection, preprocessing, feature extraction using TF-IDF, and the implementation of supervised learning algorithms such as Naive Bayes and Logistic Regression. The preprocessing phase played a crucial role in improving model performance. By applying techniques such as tokenization, stopword removal, and text normalization, irrelevant and noisy data were eliminated. Feature extraction using TF-IDF transformed textual information into numerical vectors, enabling the machine learning models to effectively learn patterns from the dataset.

The experimental results demonstrated that both Naive Bayes and Logistic Regression models performed efficiently in detecting fake news. However, Logistic Regression showed slightly higher accuracy, precision, and F1-score compared to Naive Bayes. This indicates that Logistic Regression is more effective in handling complex text classification tasks for this dataset. The confusion matrix analysis further confirmed that the model achieved balanced classification performance for both real and fake news categories. The study proves that machine learning

techniques can provide a reliable and scalable solution to combat misinformation. Compared to manual verification methods, the automated system significantly reduces time and human effort while maintaining consistent performance. Moreover, the proposed system can be integrated into social media platforms, news websites, or browser extensions to provide real-time credibility assessment.

Despite achieving satisfactory results, certain limitations remain. The system's performance heavily depends on the quality and size of the dataset. Biased or imbalanced datasets may affect classification accuracy. Additionally, fake news creators continuously adapt their writing styles, which may reduce model effectiveness over time. Future research can focus on incorporating deep learning techniques such as LSTM or transformer-based models for improved contextual understanding. Integrating social network analysis and source credibility features may further enhance detection accuracy. In conclusion, the proposed Fake News Detection System successfully demonstrates how machine learning and NLP techniques can be applied to solve real-world problems related to digital misinformation. The study contributes to the field of data science and information security by providing a practical, efficient, and implementable framework suitable for academic and real-world applications. With further improvements and large-scale deployment, such systems can play a vital role in promoting trustworthy information and maintaining digital integrity in society**.**

Reference

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2] W. Y. Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 422–426.

[3] T. Mitchell, *Machine Learning*, New York, NY, USA: McGraw-Hill, 1997.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[5] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[6] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," in *Proc. Assoc. for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[7] S. B. Parikh and P. K. Atrey, "Media-rich fake news detection: A survey," in *2018 IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 436–441.

[8] A. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. 2017 ACM Conf. on Information and Knowledge Management (CIKM)*, 2017, pp. 797–806.

[9] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.

[10] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 3111–3119.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[15] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th International Conference on World Wide Web (WWW)*, 2011, pp. 675–684.