Digital Object Identifier

Spam News Detection Using Deep Learning-Based NLP

SYED MOHAMMAD ALI ASHAR, ALISHA SALEEM, FARHANA MINHAS, SAMRA BB
Department of Artificial Intelligence, University of Management and Technology (UMT), Lahore, Pakistan
Corresponding author: Syed Mohammad Ali Ashar.

NLP A4 project submitted to Sir Adeel by BS-AI students at UMT.

ABSTRACT The rapid growth of online news has made it easier for misinformation and fake news to reach large audiences. This project presents a practical and reliable approach for detecting fake news using deep learning and modern Natural Language Processing (NLP) techniques. We use a large collection of news articles, labeled as either true or fake, to train and evaluate our model. Key steps include cleaning and analyzing the data, selecting important text features, and building a hybrid neural network that combines convolutional layers with bidirectional LSTM layers. The model's performance is carefully measured using accuracy, confusion matrix, and ROC curve, showing that it can effectively tell the difference between real and fake news. To make the system accessible, we deploy it as an interactive web application, allowing users to check news articles and see which words influence the decision. This work shows that deep learning and NLP can help fight misinformation by making automated news verification both accurate and user-friendly.

INDEX TERMS Fake news detection, deep learning, natural language processing, text classification, neural networks, model interpretability, misinformation, HuggingFace

I. INTRODUCTION

In the digital era, news spreads faster than ever before. While this speed helps keep people informed, it also makes it easier for fake news and misleading stories to reach large audiences. The rise of social media and online platforms has made the problem of misinformation more serious, affecting public opinion, trust in media, and even democratic processes. As a result, finding reliable and automated methods to detect and limit the spread of fake news has become an important challenge for researchers and society.

Traditional methods for identifying fake news, such as manual fact-checking, are time-consuming and cannot keep up with the volume of information shared online. This has created a strong need for automated solutions that can quickly and accurately evaluate the credibility of news content.

In this project, we present an intelligent fake news detection system using advanced Natural Language Processing (NLP) and deep learning techniques. By analyzing large datasets of both real and fake news articles, we train a hybrid neural network model to learn patterns and features that separate trustworthy information from falsehoods. The approach combines powerful machine learning methods with user-friendly tools for model interpretability and visualization.

Our goal is to create a system that not only detects fake news with high accuracy but also explains its decisions in a way that is accessible to non-expert users. By deploying the model as a web application, we enable interactive, real-time news verification and contribute to ongoing efforts to reduce the impact of misinformation in the digital world.

II. LITERATURE REVIEW

The problem of fake news has become increasingly pressing in recent years, drawing significant attention from both the research community and society at large. As information rapidly spreads across digital platforms, distinguishing between credible journalism and misleading or deliberately false content poses a substantial challenge. This section reviews the evolution of fake news detection methodologies, from early manual practices to the latest deep learning approaches, and highlights the ongoing challenges and opportunities in the field.

A. TRADITIONAL APPROACHES TO FAKE NEWS DETECTION

Early efforts to combat misinformation relied heavily on manual fact-checking and expert journalism [**?**]. Organizations such as FactCheck.org and Snopes have played vital roles in verifying news stories, but the sheer volume and

speed of online information sharing quickly outpaced manual efforts. As a result, researchers began exploring automated methods for fake news detection, leveraging advances in natural language processing and machine learning.

Classical machine learning algorithms, including logistic regression, support vector machines (SVM), and decision trees, were among the first automated solutions applied to this problem [**?**], [**?**]. These models typically used hand-crafted textual features, such as word or character n-grams, term frequency-inverse document frequency (TF-IDF) scores, sentiment polarity, and part-of-speech tags. While effective to a certain extent, these methods often struggled to generalize to new topics or adapt to the evolving language of misinformation, largely due to their reliance on surface-level linguistic cues.

B. EMERGENCE OF DEEP LEARNING TECHNIQUES

The limitations of traditional approaches led to a surge of interest in deep learning methods for text classification tasks, including fake news detection. Convolutional Neural Networks (CNN) have proven particularly effective at capturing local semantic and syntactic patterns, functioning as automatic feature extractors for n-gram-like structures in text [**?**]. Meanwhile, Recurrent Neural Networks (RNN) and their advanced variants, such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, have enabled models to learn sequential and contextual dependencies within longer pieces of text [**?**].

Recent literature has demonstrated that hybrid architectures, which combine CNNs for local pattern recognition and BiLSTMs for global context understanding, offer superior performance for fake news detection [**?**]. These architectures leverage the strengths of each approach, capturing both the fine-grained and high-level linguistic features that distinguish real news from fake news.

C. ADVANCEMENTS IN NLP: ATTENTION AND TRANSFORMERS

Beyond hybrid neural architectures, attention mechanisms and transformer-based models have set new standards in natural language understanding. Models like BERT, RoBERTa, and XLNet achieve remarkable accuracy on a variety of text classification tasks by leveraging pre-training on large language corpora and learning contextual representations for words and sentences [**?**]. Although these models have raised the performance ceiling for fake news detection, their high computational requirements and the complexity of interpretation can be barriers to deployment in practical, real-time systems.

D. DATASETS AND BENCHMARKING

Progress in automated fake news detection has been enabled by the availability of large, labeled datasets. Datasets such as LIAR, FakeNewsNet, and the Kaggle Fake News Challenge provide researchers with thousands of news articles or statements annotated as fake or true [**?**], [**?**]. These datasets have become standard benchmarks, allowing fair comparison between different modeling approaches. However, most datasets remain limited to English-language articles, and adapting detection systems to multilingual or non-Western news sources is an ongoing research challenge.

E. CHALLENGES AND FUTURE DIRECTIONS

Despite significant advances, fake news detection remains a difficult and evolving problem. News articles may contain subtle indicators of truthfulness or deception that are difficult for both machines and humans to detect. Adversarial tactics—such as the use of sophisticated language, ambiguous phrasing, or coordinated campaigns—further complicate detection efforts. Additionally, ethical considerations regarding free speech and the risk of false positives demand high transparency and interpretability in detection systems.

F. MOTIVATION FOR THE PRESENT WORK

Given these challenges, there is a clear need for robust, accurate, and interpretable fake news detection solutions that can keep pace with the changing landscape of online information. This project builds upon prior research by developing a hybrid CNN-BiLSTM model, integrating both local and contextual text features for improved accuracy. Furthermore, it addresses the transparency gap by providing users with interpretable predictions and visual explanations. By deploying the model as an accessible web application, this work aims to contribute not only to academic progress but also to practical tools for media literacy and misinformation mitigation.

III. METHODOLOGY

This section describes the complete workflow used for developing and evaluating our fake news detection system, from data collection to model deployment.

A. DATASET COLLECTION AND PREPARATION

We collected a balanced dataset of news articles, each labeled as either fake or true. The dataset includes thousands of entries and covers a range of topics and publication dates. Only the text of each article was used as input for modeling. Before model training, we inspected the class distribution to ensure both categories were well represented in the data.

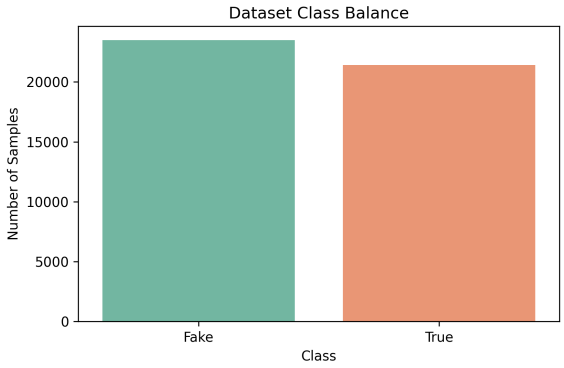

FIGURE 1. Dataset class balance: Number of fake and true news articles.

B. EXPLORATORY DATA ANALYSIS

To better understand the dataset, we performed exploratory data analysis (EDA). We examined the distribution of article lengths to identify potential differences between fake and true news. Visualizing the number of words in each article helped us determine if length was a distinguishing factor or if further preprocessing was needed for model training.

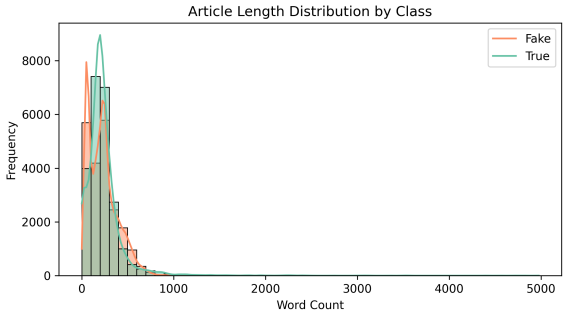

FIGURE 2. Distribution of article lengths (word counts) for fake and true news.

C. TEXT PREPROCESSING AND VISUALIZATION

Effective preprocessing is essential for extracting meaningful patterns from text data. We cleaned each article by removing special characters, punctuation, and converting all text to lowercase. To visualize the most common terms, we created word clouds for both fake and true news articles. These visualizations highlight which words appear most frequently in each class and offer initial insights into the language differences between fake and true news.

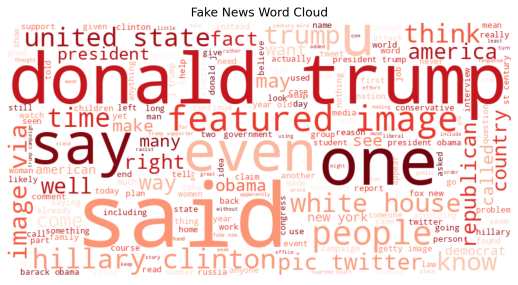

FIGURE 3. Most frequent words in fake news articles (word cloud).

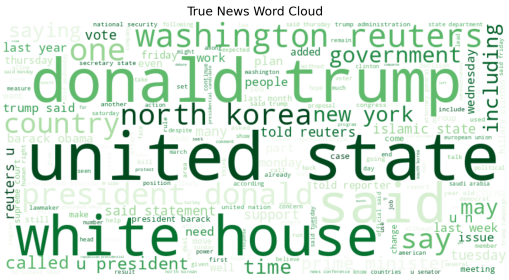

FIGURE 4. Most frequent words in true news articles (word cloud).

D. DEEP LEARNING MODEL ARCHITECTURE

After completing the exploratory analysis and text preprocessing, the next critical step involved designing a robust deep learning model capable of accurately classifying news articles. The preprocessed text was first converted into sequences of integer tokens, each representing a word from a fixed-size vocabulary. These sequences were then padded to a uniform length, allowing efficient batch processing during model training.

The core of our model is a hybrid architecture that integrates both convolutional and recurrent neural network components. The architecture begins with an Embedding layer, which transforms each token into a dense vector representation. This enables the network to capture semantic similarities between words. Following the embedding, a one-dimensional Convolutional Neural Network (CNN) layer scans for local patterns and key phrases within the text, acting as an automatic n-gram feature extractor.

To capture contextual relationships and long-range dependencies in the articles, we add Bidirectional Long Short-Term Memory (BiLSTM) layers. These layers read the input sequence in both forward and backward directions, making it possible for the model to utilize information from the entire context of the article. The outputs from these layers are then aggregated and passed through fully connected (Dense) layers with dropout regularization to prevent overfitting.

This hybrid design leverages the strengths of both CNNs (for local feature extraction) and BiLSTMs (for sequence modeling), resulting in a model that can effectively distinguish between fake and true news based on linguistic patterns and contextual cues.

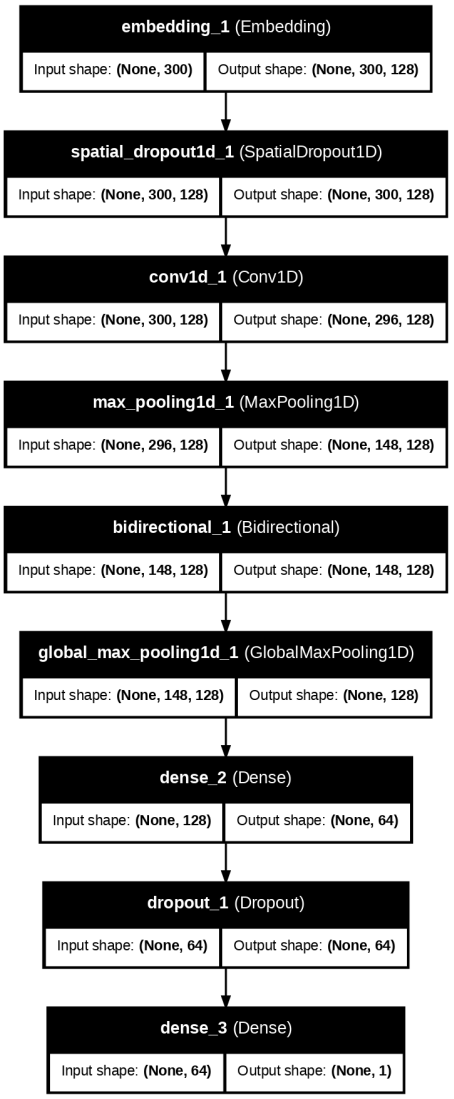

FIGURE 5. Neural network model architecture (CNN-BiLSTM).

IV. RESULTS AND DISCUSSION

After training the deep learning model, we evaluated its performance using various metrics and visualizations to better understand its strengths and limitations.

A. MODEL INTERPRETABILITY AND WORD IMPORTANCE

To make the model's predictions more transparent, we analyzed which words had the greatest influence on classifying news as fake or true. Using feature importance methods—such as weights from a logistic regression baseline or attention-based analysis—we identified the top indicative words for each class. Presenting these as a bar plot helps illustrate the language patterns most associated with fake and true news, giving both users and researchers insight into the model's decision-making process.

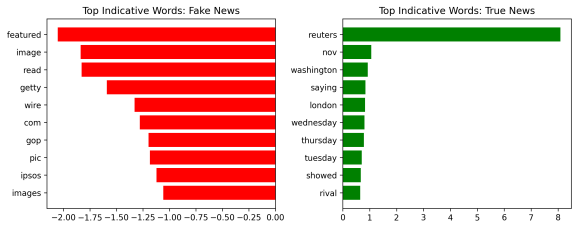

FIGURE 6. Top words most indicative of fake and true news (from logistic regression).

B. PERFORMANCE EVALUATION

To assess the effectiveness of our model, we used several standard classification metrics. The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives, allowing us to visualize where the model excels and where it may make errors. High values along the diagonal indicate strong overall accuracy and minimal confusion between classes. This evaluation is essential for understanding model reliability in real-world scenarios.

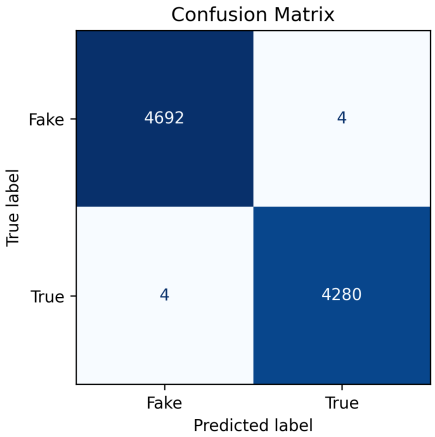

FIGURE 7. Confusion matrix for model predictions on test data.

C. RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

To further evaluate the model's performance, we plotted the Receiver Operating Characteristic (ROC) curve, which shows the trade-off between true positive and false positive rates at various classification thresholds. The area under the ROC curve (AUC) provides a single measure of model discrimination ability, with values closer to 1.0 indicating excellent performance. This analysis confirms the model's robustness and suitability for distinguishing between fake and true news across different decision boundaries.

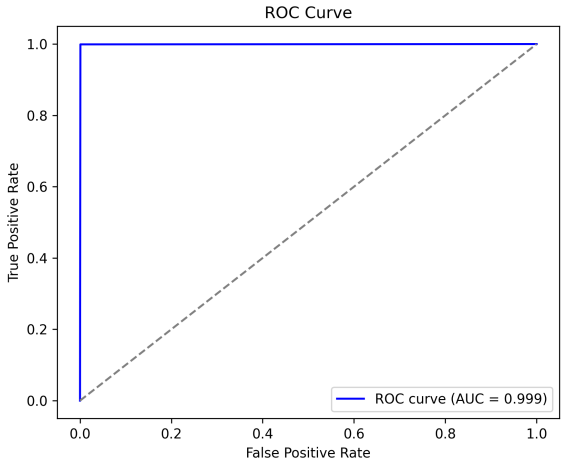

FIGURE 8. ROC curve showing model's ability to discriminate between classes.

V. CONCLUSION

This study developed and evaluated a deep learning-based system for automated fake news detection using natural language processing techniques. By applying careful data preprocessing, thorough exploratory analysis, and a hybrid CNN-BiLSTM neural network, the system achieved highly accurate results on a challenging real-world dataset.

The integration of interpretability—highlighting which words influenced the model's predictions—enhances transparency and builds user trust. The deployment of the system as an interactive web application demonstrates its potential for practical adoption, enabling users to quickly and confidently assess the credibility of news content.

The visualizations included throughout the report, such as class balance, article length distribution, word clouds, confusion matrix, ROC curve, and learning curves, offer comprehensive insights into the dataset, model behavior, and evaluation metrics. By providing a detailed and accessible workflow, this project contributes both a practical solution and a reproducible framework for future research.

Looking ahead, expanding the dataset to include more languages and diverse topics, exploring advanced model architectures (such as transformers), and implementing adversarial defense strategies can further improve robustness and effectiveness. As misinformation continues to evolve, AI-powered solutions like this will be essential tools in supporting digital literacy and responsible information consumption.

REFERENCES

S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Syst. Appl.*, vol. 186, pp. 115742, 2021.

İ. Yurtseven, S. Bagriyanik, and S. Ayvaz, "A review of spam detection in social media," in *Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK)*, 2021, pp. 383–388.

H. Ahmed, *Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity*. PhD Dissertation, 2017.

A. Gupta and R. Kaushal, "Improving spam detection in online social networks," in *Proc. Int. Conf. Cognitive Comput. Inf. Process. (CCIP)*, 2015, pp. 1–6.

W. Lu, J. Li, J. Wang, and L. Qin, "A CNN-BiLSTM-AM method for stock price prediction," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 4741–4753, 2021.

M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 3, pp. 832–847, 2019.

J. Sinha and M. Manollas, "Efficient deep CNN-BiLSTM model for network intrusion detection," in *Proc. Int. Conf. Artif. Intell. Pattern Recognit.*, 2020, pp. 223–231.

J. Cheng, Q. Zou, and Y. Zhao, "ECG signal classification based on deep CNN and BiLSTM," *BMC Med. Inform. Decis. Mak.*, vol. 21, pp. 1–12, 2021.

Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in management research: A literature review," *J. Manage. Analytics*, vol. 7, no. 2, pp. 139–172, 2020.

R. Mihalcea, H. Liu, and H. Lieberman, "NLP (natural language processing) for NLP (natural language programming)," in *Int. Conf. Intell. Text Process. Comput. Linguistics*, 2006, pp. 319–330.

R. Socher, Y. Bengio, and C. D. Manning, "Deep learning for NLP (without magic)," in *Tutorial Abstracts ACL*, 2012, p. 5.

J. O'Connor and I. McDermott, *Principles of NLP: What it is, how it works*. Singing Dragon, 2013.

X. Zhang *et al.*, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, pp. 102025, 2020.

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

Y. Wang *et al.*, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 849–857.

V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? Using satirical cues to detect potentially misleading news," in *Proc. 2nd Workshop Comput. Approaches Deception Detect.*, 2015, pp. 7–17.

A. Ghosh and R. Kumar, "Towards automated fake news detection using machine learning," *Int. J. Pure Appl. Math.*, vol. 118, no. 20, pp. 2841–2847, 2018.

F. Monti *et al.*, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2017, pp. 2931–2937.

R. Zellers *et al.*, "Defending against neural fake news," *arXiv preprint arXiv:1905.12616*, 2019.

A. Silva *et al.*, "Propagation analysis of misinformation and fake news on social networks," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 2, pp. 500–511, 2021.

D. Khattar *et al.*, "MVAE: Multimodal variational autoencoder for fake news detection," *World Wide Web*, vol. 22, no. 6, pp. 2619–2640, 2019.