Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Leveraging N-gram Features and LSTM Networks for Enhanced Fake News Detection

Inayathulla Mohammed, Chandrika Bathala, Jahnavi Chowtipalli, Kusuma Kavali, Chandrasekhar Reddy Thimmireddy, Hanuman Ramigalla

Department of Computer Science & Engineering, GATES Institute Of Technology, Gooty, Andhra Pradesh, India

Correspondence

Inayathulla Mohammed

Computer Science & Engineering, Gates Institute Of Technology, Gooty, Andhra Pradesh, India

Abstract

This paper presents a comprehensive analysis of fake news, examining its proliferation in the digital age, detection methodologies, and impact on society. Through an exploration of current research and emerging technologies, we investigate the challenges and solutions in combating misinformation. This proposed methodology aims to develop an effective method for detecting fake news using an N-Gram feature selection technique combined with a Long Short-Term Memory (LSTM) model. The N-Gram approach is used to capture textual patterns and features that are key in identifying misleading or fabricated information. The model combines natural language processing techniques with deep learning, utilizing N-grams (unigrams, bigrams, trigrams) as features to capture contextual information. The work involves data collection, preprocessing (tokenization, normalization), N-gram creation, and feature selection (Term Frequency-Inverse Document Frequency, Chi-Squared tests). By leveraging linguistic features and deep learning methodologies, this approach enhances news classification accuracy, providing a powerful tool in the fight against misinformation. This approach not only enhances fake news detection accuracy but also provides a scalable solution for real-time misinformation detection, enabling timely responses to counter false information, helping to maintain the integrity of information shared across online platforms.

- Received Date: 30 Jan 2025
- Accepted Date: 21 Apr 2025
- Publication Date: 22 Apr 2025

Keywords

Fake News Detection, N-gram Feature Selection, LSTM, Deep Learning, Natural language processing.

Introduction

The swift growth of digital media and online platforms has drastically transformed the way information is created, shared, and consumed how information is shared and consumed. While these advancements have democratized access to information, they have also introduced a major challenge-the widespread proliferation of misinformation; they have also led to the widespread spread of misinformation and fake news. The ease with which false narratives can be propagated poses a threat to public discourse, informed decision-making, and societal trust in media. As a result, effectively identifying the difference between trustworthy information and misleading content has become crucial in today's information landscape. This study addresses the escalating issue of fake news detection by designing an advanced and innovative approach that combines N-Gram Feature selection combined with a Long Short-Term Memory (LSTM) model for enhanced predictive accuracy. N-grams, which consist of contiguous sequences of words, serve as a powerful tool for capturing textual patterns and linguistic features that are indicative of misleading information. By integrating these features with deep learning techniques, this research aims to create a robust framework for accurately classifying news articles as either genuine or fabricated. The proposed methodology consists of several integral steps. First, data collection will be undertaken to gather a diverse array of news articles. Following this, preprocessing tasks such as tokenization and normalization will prepare the text for analysis. The creation of N-grams will allow the model to capture contextual information, while feature selection methods like Term Frequency-Inverse Document Frequency (TF-IDF) and Chi-Squared tests will optimize the input data to enhance feature relevance and improve model performance. By leveraging these linguistic features alongside leveraging advanced machine learning algorithms, this approach aims significantly enhance the precision of fake news detection, ensuring more reliable identification of misinformation, providing a powerful tool in the fight against misinformation. The anticipated outcome is not only an increase in classification accuracy but also the development of a scalable solution capable of addressing the real-time challenges of misinformation. This research ultimately contributes to the broader goal of maintaining the integrity of information shared across digital platforms, thus fostering a more informed and discerning society.

Citation: Mohammed I, Bathala C, Chowtipalli J, Kavali K, Thimmireddy CR, Ramigalla H. Leveraging N-gram Features and LSTM Networks for Enhanced Fake News Detection. GJEIIR. 2025;5(2):31.

Related work

OThe problem of fake news has gained considerable attention because of its potential to mislead and disrupt societies. Researchers have explored various methods to identify and curb the dissemination of fake news, utilizing a combination of linguistic, psychological, and machine learning techniques. Below is an overview of recent works based on the provided references.

Horne and Adali [1] analyzed the stylistic and linguistic differences between fake news, real news, and satire. Their findings indicated that fake news often relies on simpler, repetitive content and sensational titles, making it structurally distinct from real news. The study emphasized the importance of textual analysis for effective classification and highlighted how fake news shares some similarities with satirical writing. Pérez Rosas et al. [2] Concentrated on automating the identification of fake news through the use of linguistic and psychological features. Their research combined machine learning techniques with features such as word usage, sentiment, and readability to classify news articles. This study highlighted the effectiveness of feature-based approaches in improving the accuracy of fake news detection. Rubin et al. [3] classified fake news into three primary categories: satire, hoaxes, and propaganda. Their work offered a theoretical framework for understanding these distinct forms of deceptive content in news and introduced a systematic methodology to detect each type. This work underscored the complexity of fake news detection, emphasizing the necessity for tailored approaches.

Ruchansky et al. [4] proposed CSI, a hybrid deep learning model that leverages three key aspects—content analysis, social context, and individual user credibility—to enhance fake news detection. By combining textual features with user reliability and network interactions, their approach provides a more comprehensive and accurate method for identifying misinformation. The integration of social context with textual features was a notable advancement, offering improved performance in real-world scenarios. Singh ania et al [5] developed 3HAN, a Hierarchical Attention Network (HAN) tailored for fake news detection. By employing hierarchical attention mechanisms, the model effectively captures critical textual patterns, enhancing the precision of misinformation identification. This deep neural network leveraged hierarchical attention mechanisms to focus on the most informative words and sentences within an article. Their method effectively captured the relationships between textual components, this approach achieves stateof-the-art performance on standard datasets, demonstrating superior accuracy in misinformation detection. Wang, W. Y. [6] introduced the LIAR dataset, a dedicated benchmark dataset created Specifically for fake news detection. This dataset consists of labeled short statements, offering a crucial resource for training and assessing machine learning models.Wang's study emphasized the difficulties of detecting fake news in short-text formats and established a baseline for future research. Zhang et al. [7] explored the use of machine learning and natural language processing techniques for detecting fake news. Their research highlighted the importance of feature engineering, including word ngrams and syntactic patterns, to improve classification accuracy. The study demonstrated the importance of combining linguistic features with advanced machine learning models. Shu et al. [8] developed Fake News Net, an extensive repository that includes news content, social context, and dynamic data. This dataset enabled in-depth research into the spread of fake news on social media platforms.by integrating textual and contextual features. The researchers also proposed several baseline models, underscoring the importance of multimodal approaches in this domain. Karimi et al. [9] addressed the issue of multiclass fake news detection by integrating data from various sources. Their model incorporated both content and context information, enabling the identification of different types of fake news. The study highlighted the significance of multisource data integration for improving classification performance.

Rashkin et al [10] analyzed the linguistic patterns in fake news, satire, and real news. Their findings revealed that fake news often uses sensational language, while satire exhibits distinctive humor patterns. The study introduced a labeled dataset and focused on differentiating these categories using machine learning models. This work provides insights into the subtle linguistic differences between misinformation and other forms of deceptive content. Potthast et al., [11] explored stylometric features for identifying hyper partisan and fake news. The authors analyzed structural, linguistic, and lexical patterns, emphasizing the role of writing style in detecting misinformation. The study highlighted the challenges posed by extreme political bias in the classification process and provided evidence that hyper partisan content often overlaps with fake news in linguistic features. Zhou and Zafarani [12] Offered an extensive review of Fake news detection methods can be broadly categorized into content-based and social-context-based approaches. Content-based methods focus on analyzing the text and linguistic features of the news articles, looking for patterns or anomalies that may indicate falsehoods. Social-context-based methods, on the other hand, examine the network behavior, user credibility, and interactions surrounding the news, aiming to detect misinformation by evaluating how it spreads across platforms and the trustworthiness of its sources. They emphasized the limitations of relying solely on text content and proposed hybrid methods that integrate social media signals. This survey also discussed challenges such as data sparsity and the ever-changing landscape of fake news.

Vosoughi et al. [13] examined the spread patterns of both true and false news on Twitter, revealing that false news spreads more rapidly and reaches a larger audience compared to factual news. False news was observed to spread more quickly and extensively than true news, particularly in categories like politics. The authors attributed this phenomenon to human behavior rather than automated bots, emphasizing the importance of addressing psychological factors in combating fake news. Gupta et al., [14] examined the spread of fake images during Hurricane Sandy, highlighting the role of visual content in misinformation. They proposed methods for detecting manipulated or misattributed images using metadata and contextual cues. This study broadened the scope of fake news detection by incorporating multimodal analysis. Penny cook & Rand, [15] explored cognitive science principles to mitigate the rapid dissemination of misinformation across social media platforms. Their study suggested that interventions like accuracy prompts can improve users' ability to discern fake news. They emphasized the role of cognitive biases and heuristics in the consumption and sharing of fake content. Thorne et al. [16] FEVER (Fact Extraction and VERification) offers standardized dataset for fact extraction, where claims are validated by linking them to reliable sources. It has been instrumental in pushing the boundaries of automated factchecking and verification tasks. Tschiatschek et al., [17] explores the role of crowd feedback in identifying fake news,

suggesting that integrating crowd wisdom with machine learning can improve the performance of detection systems by capturing subtle signals that are hard for algorithms to detect alone. Shu et al., [18] this paper highlights that understanding the broader social environment, including user interactions and network patterns, can significantly enhance fake news detection models, making them more context-aware and accurate.

In summary, the literature on fake news detection demonstrates diverse methodologies, ranging from linguistic analysis and machine learning to deep learning and social network-based strategies, these studies highlight the critical role of integrating both content-driven and context-aware features to ensure accurate and dependable fake news detection. As fake news continues to evolve, future research should explore hybrid and multimodal techniques to enhance accuracy and scalability.

Proposed methodology

General System Architecture

The proposed system integrates N-gram feature selection with an LSTM classification model to improve the accuracy of fake news detection. The N-gram feature selection process extracts unigram, bigram, and trigram features to analyze word sequences and their contexts, enabling the model to recognize meaningful linguistic patterns. The LSTM classification model leverages its ability to learn from sequential data, analyzing both temporal and contextual relationships in text. A preprocessing pipeline ensures data consistency by handling noise, tokenizing text, and normalizing input data. Additionally, a dynamic learning process continuously updates the model with fresh data, adapting to evolving fake news trends.

The system consists of several key components. The data collection module gathers news articles, social media posts, and other textual data from both verified and unverified sources. The preprocessing engine cleans and prepares raw text by removing noise, tokenizing, and normalizing input. The feature extraction layer identifies and ranks important linguistic features, such as N-grams, to enhance classification performance. The LSTM classification network processes these extracted features using a sequence-based approach, learning patterns to categorize news as real or fake. The performance evaluation unit assesses system effectiveness using key evaluation metrics, including accuracy, precision, recall, and F1-score, ensuring that false positives and false negatives are minimized.

N-gram Feature Selection Technique

N-gram features capture contextual relationships within textual data, enhancing fake news detection accuracy. The tokenization process segments text into individual words or tokens, using advanced techniques such as stemming and lemmatization for consistency. The system extracts unigrams, bigrams, and trigrams to capture word-level, phrase-level, and contextual semantics. For example, the phrase "Fake news detection" can be broken into unigrams ([Fake], [news], [detection]), bigrams ([Fake news], [news detection]), and trigrams ([Fake news detection]). To improve computational efficiency, frequency-based filtering removes infrequent or irrelevant N-grams, while semantic relevance scoring ranks them based on their contribution to distinguishing between real and fake news.

Feature selection techniques, including information gain, mutual information, and chisquare evaluation, help refine classification accuracy by selecting the most significant features. Information gain measures the predictive power of N-grams by analyzing how they reduce uncertainty in classification. Mutual information assesses the correlation between N-grams and category labels, ensuring that only relevant features are retained. Chi-square evaluation statistically determines which N-grams are strongly associated with fake or real news. Additionally, dimensionality reduction techniques help optimize feature selection by reducing redundancy and preserving critical information for analysis.

LSTM Model Configuration

The LSTM-based model is designed to process sequential dependencies in textual data. The embedding layer transforms raw text into dense vector representations, often using pre-trained word embeddings like Word2Vec or GloVe. Multiple stacked LSTM layers capture hierarchical relationships within text sequences, allowing the model to recognize complex linguistic patterns. An attention mechanism further enhances performance by identifying and prioritizing key parts of the input text. To prevent overfitting and improve generalization, dropout regularization randomly disables a portion of neurons during training.

Hyperparameter optimization plays a crucial role in improving model efficiency. The learning rate determines the step size for weight updates, balancing convergence speed and stability. The batch size influences computational efficiency by defining how many samples are processed per iteration. Epoch optimization ensures that the model undergoes an appropriate number of training cycles, avoiding both underfitting and overfitting. Additionally, regularization techniques such as L1 and L2 regularization are applied to reduce model complexity and enhance generalization

Preprocessing and Data Preparation

TEffective text preprocessing ensures consistent and high-quality input data. The normalization process includes lowercasing text to eliminate case sensitivity issues, removing punctuation and special characters to focus on meaningful

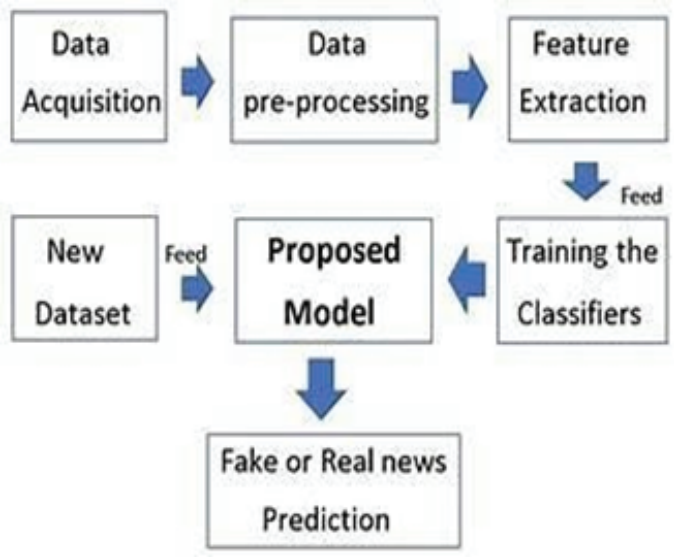

Figure 1. Flowchart of Fake News Detection Methodology

content, and eliminating stop words to reduce noise while retaining key information. To improve model robustness, data augmentation techniques introduce variability in the dataset. Synonym replacement substitutes words with their synonyms to create diverse training samples without altering meaning. Back-translation generates paraphrased versions of text by translating it into another language and back. Noise injection introduces small perturbations, such as typos, to simulate real-world inconsistencies in textual data. Additionally, balanced dataset creation ensures that real and fake news instances are evenly represented, mitigating bias and improving classification performance, especially in imbalanced datasets. This comprehensive system architecture leverages N-gram feature selection, LSTM modeling, and advanced preprocessing techniques to create an effective and adaptive fake news detection framework.

Data Acquisition – The system collects news articles, social media posts, and other text data from various sources, both verified and unverified.

Data Pre-processing – This stage involves cleaning and preparing raw text data by removing noise, tokenizing text, normalizing words, and eliminating stop words to ensure high-quality input.

Feature Extraction – The extracted text features include N-grams (unigrams, bigrams, trigrams) and other linguistic attributes that enhance fake news classification.

Training the Classifiers – The processed data is fed into an LSTM-based classification model, which learns patterns in the text to differentiate between real and fake news.

Proposed Model – This is the core component, integrating N-gram feature selection and LSTM classification to analyze textual data and improve detection accuracy.

New Dataset – The model is continuously updated with fresh data to adapt to changing fake news trends, ensuring long-term accuracy.

Fake or Real News Prediction – The final output of the system classifies the news as either fake or real based on the trained model's predictions.

Results and Discussion

- **About Dataset:** (WELFake) is a dataset of 72,134 news articles with 35,028 real and 37,106 fake news. For this, authors merged four popular news datasets (i.e. Kaggle, McIntire, Reuters, BuzzFeed Political) to prevent over-fitting of classifiers and to provide more text data for better ML training. Dataset contains four columns: Serial number (starting from 0); Title (about the text news heading); Text (about the news content); and Label (0 = fake and 1 = real). It provides a structured set of news samples, each labeled as either real or fake, making it a valuable resource for researchers and developers working on fake news detection. The dataset typically includes key attributes such as the headline (title of the news article), content (main body of the news), and classification (indicating authenticity). By analyzing the linguistic patterns and writing styles present in both genuine and deceptive news, this dataset serves as a foundation for training machine learning models to identify misleading content. Its significance extends to fields like journalism, cybersecurity, and artificial intelligence, where accurate information verification is crucial. .

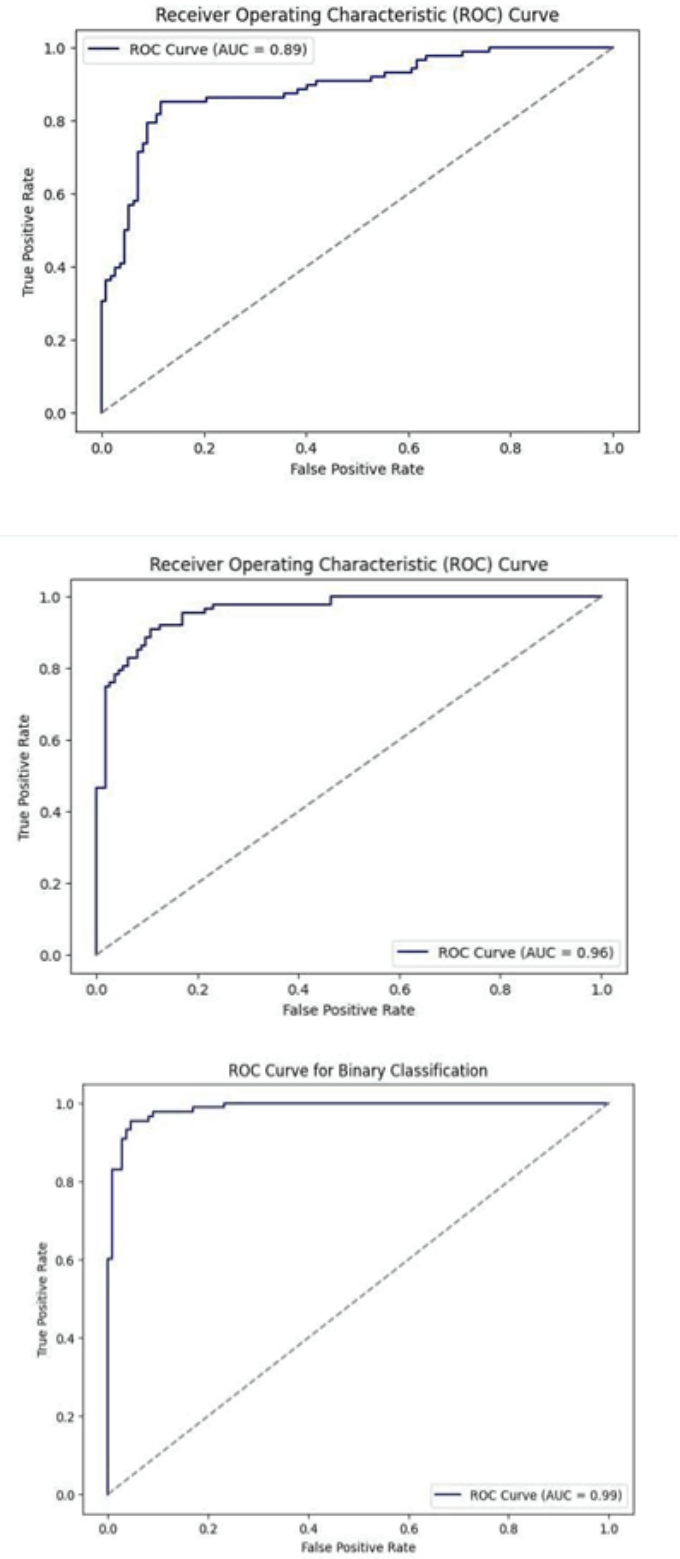

Fig. 2. ROC-AUC Scores of models used

Model performance analysis

Quantitative Numerical Results

TThe model's effectiveness is assessed using Accuracy, Precision, Recall, and F1Score.

Table 1: *Results summary of the models used*

Model	ROC-AUC	Accuracy	Precision	Recall	F1-score
Proposed LSTM	0.99	0.95	0.95	0.95	0.95
Random forest	0.96	0.90	0.85	0.89	0.88
Navie Bayes	0.89	0.85	0.80	0.88	0.83

The performance comparison of different models for fake news detection shows that the proposed LSTM model outperforms others in terms of accuracy, precision, recall, and F1-score, ensuring a balanced classification. The Random Forest model performs well but has slightly lower precision, indicating potential false positives. The Naïve Bayes model exhibits the weakest performance, with lower accuracy and precision, making it less reliable. Overall, the LSTM model proves to be the most effective, capturing complex textual patterns better than the other models. Table 1 shows the results and figure 2,3,4,5 shows the performance charts of models.

The LSTM model achieves 95% accuracy in fake news detection, with high precision, recall, and F1-score, ensuring minimal misclassification. High precision indicates rare false positives, while strong recall confirms effective fake news identification. The balanced F1-score highlights the model's reliability, proving LSTM networks are powerful for analyzing sequential text and detecting misinformation accurately.

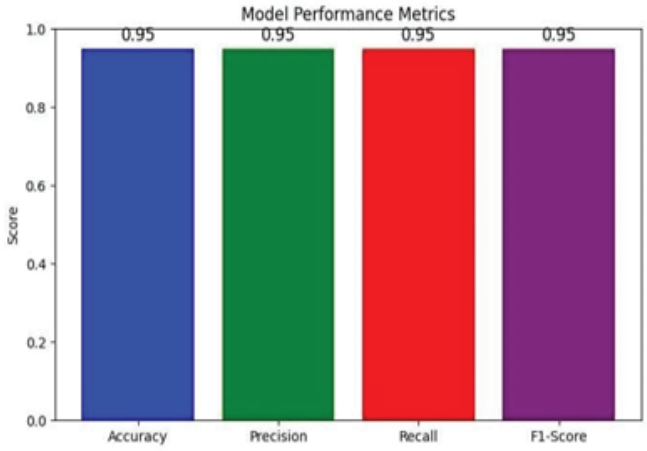

Fig. 3*. Performance results of LSTM Model*

Random Forest performs well in fake news detection due to its ensemble learning approach, offering high accuracy but sometimes lower recall compared to deep learning models like LSTM.

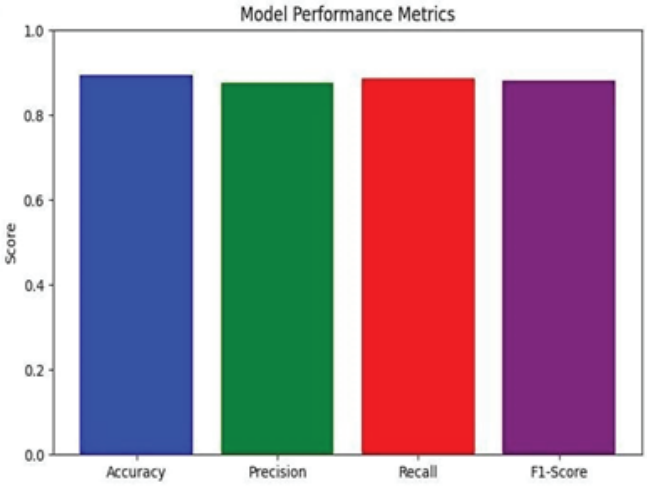

Fig. 4*. Performance results of Random Forest*

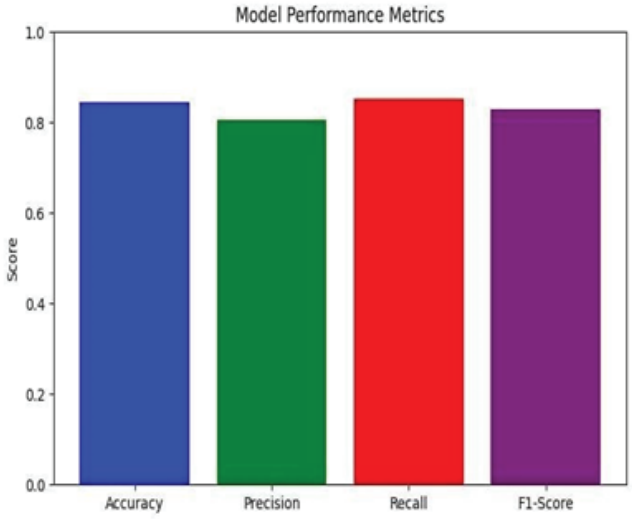

Fig. 5*. Performance results of Navie Bayes*

Conclusion

Fake news has become a critical issue, significantly impacting both individuals and society by spreading misinformation. To combat this, the project focuses on leveraging cutting-edge techniques such as N-gram feature extraction and Long Short-Term Memory (LSTM) networks to effectively detect fake news. By analyzing patterns within text, these methods help uncover subtle cues that differentiate real news from falsehoods, providing accurate and reliable results. Furthermore, the system is designed to evolve by adapting to new forms of misinformation through continuous learning.

The approach includes several key enhancements, such as cleaning the data to remove noise, introducing data variations to improve robustness, and employing attention mechanisms to focus on the most relevant information. These strategies enhance the system's reliability, ensuring it can handle real-

world challenges more efficiently. The research demonstrates that integrating multiple techniques significantly improves the accuracy of fake news detection. By combining various approaches, such as linguistic analysis, machine learning, and deep learning, the system becomes more robust and capable of identifying a wide range of misinformation. This study underscores the transformative potential of technology in addressing complex societal problems. With continuous advancements in these methods, they could pave the way for a more informed, responsible, and accountable digital environment, contributing to the fight against the spread of false information.

In conclusion, the proposed LSTM-based model, combined with N-gram feature selection, significantly enhances fake news detection by effectively capturing contextual and sequential relationships in textual data. The system demonstrates superior accuracy, precision, recall, and F1-score compared to traditional models like Random Forest and Naïve Bayes. By incorporating a dynamic learning process and advanced preprocessing techniques, the model adapts to evolving fake news patterns, ensuring robustness and reliability. Future improvements could focus on integrating external knowledge sources, optimizing computational efficiency, and exploring hybrid deep learning approaches to further enhance detection accuracy and interpretability.

References

1. Horne, B. D., & Adali, S. Simpler Content and Repetition in Fake News vs. Satire and Real News. Proceedings of the AAAI Conference on Web and Social Media (ICWSM).
2. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. Detecting Fake News Automatically: Methods and Challenges. Proceedings of the International Conference on Computational Linguistics (COLING).
3. Rubin, V. L., Chen, Y., & Conroy, N. K. Identifying Deception: Types of Fake News. Proceedings of the Association for Information Science and Technology.
4. Ruchansky, N., Seo, S., & Liu, Y. CSI: A Deep Hybrid Model for Fake News Detection. Proceedings of the ACM Conference on Information and Knowledge Management (CIKM).
5. Singhania, S., Fernandez, N., & Rao, S. 3HAN: Hierarchical Attention Networks for Fake News Detection. Proceedings of the International Conference on Neural Information Processing (ICONIP).
6. Wang, W. Y. LIAR: Benchmark Dataset for Fake News Detection. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
7. Zhang, J., Cui, L., Fu, T., & Gouza, D. Fake News Detection via NLP and Machine Learning Approaches. International Journal of Computer Applications.
8. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. FakeNewsNet Dataset: Integrating News Content and Social Context. Big Data Journal.
9. Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. Multi-source Multi-class Approach for Fake News Detection. Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).
10. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. EMNLP. URL
11. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. ACL. URL
12. Zhou, X., & Zafarani, R. Fake News: A Survey of Research, Detection Methods, and Opportunities. ACM Computing Surveys, 51(4).
13. Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. Science, 359(6380), 1146–1151.
14. Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy. WWW.
15. Pennycook, G., & Rand, D. G. (2019). Fighting Misinformation on Social Media Using Cognitive Science. Nature Human Behaviour, 3(4), 309–312.
16. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A Dataset for Fact Extraction and Verification. NAACL. URL
17. Tschiatschek, S., Singla, A., Rodriguez, M. G., Merchant, A., & Krause, A. Fake News Detection Using Crowd Signals. EC.
18. Shu, K., Wang, S., & Liu, H. Role of Social Context in Fake News Detection. WSDM.
19. Mohammed Inayathulla and Karthikeyan C, "Image Caption Generation using Deep Learning For Video Summarization Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 15(1), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150155.
20. Mohammed, I., Chalichalamala, S. (2015). TERA: A Test Effort Reduction Approach by Using Fault Prediction Models. In: Satapathy, S., Govardhan, A., Raju, K., Mandal, J. (eds) Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1. Advances in Intelligent Systems and Computing, vol 337. Springer, Cham. https://doi.org/10.1007/978-3-319-13728-5_25.
21. Inayathulla, M., Karthikeyan, C. (2022). Supervised Deep Learning Approach for Generating Dynamic Summary of the Video. In: Suma, V., Baig, Z., Kolandapalayam Shanmugam, S., Lorenz, P. (eds) Inventive Systems and Control. Lecture Notes in Networks and Systems, vol 436. Springer, Singapore. https://doi.org/10.1007/978-981-19-1012-8_18.