A ROBUST LSTM-BASED FRAMEWORK FOR FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING

Hamda Nadeem[1], Ms. Lubna Nausheen[2], Ms. Faiza Fatima[3]
[1]PG Scholar, Department of AI & DS, Shadan Women's College of Engineering and Technology, Hyderabad, hamd01nadeem@gmail.com
[2]Asst. Professor, Department of CSE, Shadan Women's College of Engineering and Technology lubnanausheen07@gmail.com
[3]Asst. Professor, Department of CSE, Shadan Women's College of Engineering and Technology faizafatima010@gmail.com

ABSTRACT
This study presents a deep learning-based framework for fake news detection by integrating Natural Language Processing (NLP) with a Long Short-Term Memory (LSTM) network. The rapid growth of online news platforms and social media has significantly increased the spread of misinformation, posing serious challenges to information credibility and public trust. The proposed framework utilizes NLP techniques to extract meaningful textual representations, while the LSTM model captures contextual and sequential dependencies within news articles to distinguish between genuine and fake content. By combining semantic feature extraction with deep sequential learning, the framework reduces the reliance on manual feature engineering and improves the robustness of text classification. The system is trained and evaluated using a benchmark fake news dataset to validate its effectiveness in identifying misleading content. Experimental evaluation demonstrates reliable classification performance, computational efficiency, and strong generalization capability. The proposed framework provides a scalable and practical solution for automated fake news detection and offers significant potential for real-time misinformation monitoring and intelligent news verification applications.

1. INTRODUCTION

The quick dissemination of information through social media and internet platforms has revolutionized the way that news and ideas are shared in the digital age. Although there are advantages to this democratization of information, it has also led to the pervasive problem of fake news, which is intentionally false or misleading material that is presented as reality. Fake news has far-reaching effects that impact public perception, erode institutional confidence, and even affect political results. Therefore, there has never been a greater need for effective, scalable techniques to identify and counter bogus news. Conventional techniques for identifying false news frequently depend on human fact-checkers to manually confirm the veracity of the content. Nevertheless, this method requires a lot of time and resources because to the enormous amount of content produced every day. Furthermore, the work is especially difficult due to the intrinsic intricacy and subjectivity of judging the veracity of certain statements. Due to the unpredictability of fact-checking procedures and the dynamic nature of false information, researchers are investigating automated methods that can effectively and accurately identify fake news. In order to overcome this difficulty, this project creates an automated system for detecting fake news by utilizing developments in machine learning (ML) and natural language processing (NLP). In particular, we suggest using Long Short-Term Memory (LSTM) networks, which are a kind of recurrent neural network (RNN) intended to identify long-range dependencies in sequential data. Additionally, the underlying complexity and subjectivity of determining the truth of certain claims make the endeavor particularly challenging. Researchers are looking at automated techniques that can successfully and precisely identify fake news because fact-checking processes are unpredictable and misleading information is dynamic. This research uses advances in machine learning (ML) and natural language processing (NLP) to construct an automated system for identifying bogus news in order to tackle this challenge. Specifically, we recommend the use of Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) designed to detect long-range dependencies in sequential data.

2. EXISTING SYSTEM OF PROJECT

- A sophisticated hybrid model called CNN-BiLSTM (Convolutional Neural Network - Bidirectional Long Short-Term Memory) combines the advantages of CNN and BiLSTM architectures for sequence-based tasks like sentiment analysis, text classification, and, most importantly, the identification of false news. This strategy aims to combine the contextual understanding offered by BiLSTM with the feature extraction power of CNNs.
- The CNN layer in a CNN-BiLSTM model first functions as a feature extractor, extracting pertinent features, spatial hierarchies, and local

patterns from the raw input text (often in the form of word embeddings or n-grams).

- Convolutional filters allow CNN to identify crucial terms and phrases that could reveal if a news article is authentic or fraudulent. In essence, this layer finds and magnifies local patterns, which are crucial for spotting clues in text like odd word choices or deceptive claims.
- The BiLSTM layer then enters the picture, capturing the text's sequential structure. BiLSTM is a kind of Recurrent Neural Network (RNN) that may better comprehend the context and dependencies within the sequence by processing information both forward and backward throughout the text.
- BiLSTM employs two LSTMs: one that reads the text from left to right and another that reads it from right to left, in contrast to typical LSTMs that only process the sequence in one way (from left to right). This bidirectional processing enables the model to fully comprehend each word's context, accounting for both preceding and following words, which is essential for comprehending intricate linguistic patterns and context in the detection of bogus news.

Existing System Disadvantages:

- The CNN-BiLSTM model combines multiple layers and architectures, leading to increased complexity in both design and implementation.
- Due to the deep architecture and the need for processing large volumes of text data, the CNN-BiLSTM model requires extended training times, making it computationally expensive.
- The model's complexity and large number of parameters increase the risk of overfitting, especially when trained on small or imbalanced datasets, reducing its generalization ability to unseen data.

3. RELATED WORK

Scholars have used GNNs to model news documents and categorize news in order to speed up the process of identifying false news and increase the accuracy of the results. Four categories—conventional deep learning, GNNs, multi-hop GNNs, and graph transformer methods—are used in this study to classify fake news detection techniques. Below is further detail on these techniques. Traditional deep learning techniques: In Ref. [6], deep learning techniques are used to detect fake news by feeding each news sentence into an RNN. The news information is then represented by the RNN's hidden layer vector. The final classification outcome is then obtained by embedding the hidden layer vector into the classifier. The usage of CNN and the Bidirectional Long Short-Term Memory network (Bi-LSTM) to identify bogus news is then investigated in Ref. [20]. Nevertheless, the aforementioned studies overlook the ways in which various phrases interact within news articles, which are essential for identifying false information. Graph neural network method: Reference [13] models news stories using GCN. A news story is represented as a graph in Ref. [11], where sentences are nodes and sentence similarity is represented by edges. The information between the graph's nodes is then combined using GCN to create node-embedded vectors. In the meantime, semantic information is enhanced for better news modeling and classification outcomes using heterogeneous information networks [14, 15]. Ref. [22] also suggests creating a knowledge graph from both positive and bad news, and TransE is used to determine the three-factor score for identifying fake news. By comparing the entity information taken from news content with that in the knowledge graph, an external knowledge base is used in Ref. [12] to detect false news. K. K. Multi-hop graph neural network approach: In standard message-passing GNNs, node representations are updated by their immediate neighbors, or 1-hop neighbors. For instance, by sparsifying an extended type of graph diffusion, Ref. [23] combines messages from bigger (multi-hop) neighbors in each layer. The Multi-hop Attention Graph Neural network, also known as MAGNA [18], is suggested to compute attention scores in two stages: (1) determining the attention scores of neighbors who are directly connected; and (2) using an attention diffusion procedure to determine the attention scores among nodes that are not directly connected. In essence, MAGNA increases the "receptive field of the GNN at each layer" by dispersing the attention fraction throughout the network. -hop messaging is more effective than 1-hop messaging and could identify nearly all normal graphs, according to a precise mathematical formulation in Ref. [24]. Graph transformer approach: Machine translation has advanced significantly thanks to transformer architecture [27]. Transformers have been used in many other disciplines in recent studies, and the results have been outstanding. For example, preparing photos and then passing them straight into the typical transformer design yields the best results [28]. GNNs have also been utilized using transformer design [29–31]. The precise information location is represented in Ref. [33], which also takes into account the interaction between nodes and particular network features (such edges).

4. PROPOSED SYSTEM

- This project's suggested algorithm for identifying fake news makes use of Long Short-Term Memory (LSTM) networks in conjunction with Natural Language Processing (NLP) approaches. For processing sequential input, like text, where context and word associations can span large distances, LSTM, a type of recurrent neural network, works effectively.

- By learning the temporal connections in the text, the LSTM model in this method is able to identify intricate patterns and contextual details that are essential for differentiating between authentic and fraudulent news. NLP approaches are used to preprocess the text data, transforming unstructured material into a format that can be utilized to train models.
- Tokenization, stopword elimination, and the use of word embeddings like Word2Vec or GloVe to convert words into vector representations are all part of this. These embeddings encode semantic information about words, allowing the LSTM to understand relationships between words in different contexts. The LSTM network processes the text sequentially, learning both short-term and long-term dependencies within the content, which is essential for understanding the overall meaning of the article. By combining NLP for feature extraction with LSTM for sequence modeling, the proposed algorithm improves the ability to detect fake news. The LSTM's capacity to capture context from both past and future words enables the model to recognize subtle linguistic patterns that may indicate misinformation.
- As a result, this approach offers a more effective and accurate method for classifying news articles, even in the face of ambiguous or evolving language used in fake news.

Proposed System Advantages:

Because LSTM is so good at processing sequential data, it is very useful for jobs like detecting fake news, where the relationships between words and phrases across the text determine the meaning of a story (increased performance on sequential data).

5. MODULES

1. Data collection

The quality and diversity of the dataset greatly affect the model's performance, making the data collection step essential. Data for this project may be obtained from several online news sources or from publicly accessible datasets.

2. Data Loading

Loading and preprocessing the dataset is the second stage in the pipeline for detecting fake news. The information may come from any custom news article dataset or from publicly accessible datasets such as LIAR and FakeNewsNet. The news text and its associated label (genuine or fraudulent) are usually included in the collection.

3. Text preprocessing

Lemmatization uses NLTK or Spacy to reduce every word to its most basic form (for example, "running" becomes "run"). After the tokens have been cleansed, text vectorization transforms them into numerical data that machine learning algorithms may use. Word2Vec, a pre-trained word embedding model, is used for this project. It converts each word into a vector of real values that captures semantic associations between words.

4. Word2Vec Embedding

Converting words into Word2Vec embeddings is a critical next step after preprocessing the text input. Mikolov et al. (2013) created the shallow neural network model Word2Vec, which learns distributed representations of words based on their context.

5. Model Building (LSTM)

One kind of Recurrent Neural Network (RNN) that works especially well with sequential data, like text, is called LSTM. LSTM can comprehend context across word sequences and learn long-range dependencies.

6. Model Training

The model must then be trained using the vectorized and preprocessed data. Using the patterns it finds in the input text, the model is trained to determine if a certain article is authentic or fraudulent.

7. Model Evaluation and Optimization

The model's ability to generalize to new data is assessed using the test dataset after training. Several methods can be used to optimize the model if it is not performing well.

8. Saving the Model (H5 Format)

The last stage is to store the model for use in deployment when it has been trained and assessed. The architecture, weights, and training configuration of models can be saved in the H5 format in Keras, a well-known deep learning system.

6. TECHNIQUES USED IN PROJECT

6.1. EXISTING TECHNIQUE: -

- **BI-LSTM**

In this hybrid method, CNN is first used for feature extraction, which involves applying filters to the text in order to identify local patterns, such as important words or word combinations, that can reveal the authenticity of a news story. CNN is particularly good at spotting these crucial local characteristics in text, which aids in the model's detection of certain language cues. However, Bi-LSTM can capture long-range relationships and word context from both past and future sequences since it processes the text in both forward and backward orientations. The model successfully captures both the fine-grained linguistic patterns and the more comprehensive contextual information required to differentiate authentic news from fraudulent news thanks to the combination of

CNN for local feature extraction and Bi-LSTM for contextual sequence learning. A strong framework that improves the precision and resilience of fake news detection systems is provided by this hybrid CNN-Bi-LSTM architecture.

6.2. PROPOSED TECHNIQUE USED OR ALGORITHM USED:

➢ **NLP(LSTM):**

This method preprocesses the text using natural language processing (NLP), which includes lemmatizing words to their root forms, eliminating stop words, and tokenizing news articles. By eliminating unnecessary information, these preprocessing techniques assist guarantee that the model concentrates on the crucial portions of the text. The LSTM network, a kind of recurrent neural network (RNN) that works well with sequential data like text, forms the basis of the suggested solution. This is especially helpful for identifying fake news, as the meaning of an article depends on comprehending the entire context rather than just certain words. The system can identify patterns and correlations in the text that reveal if a news story is authentic or fraudulent by integrating LSTM for sequence modeling with NLP for feature extraction. By combining LSTM for sequence modeling with NLP for feature extraction, the system can find patterns and correlations in the text that indicate whether a news report is genuine or fake.

7. RESULTS

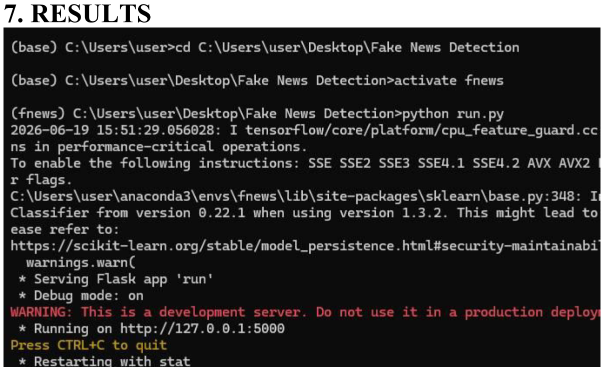

Fig 7.1 Code Launch in Terminal

Fig 7.2 Homepage

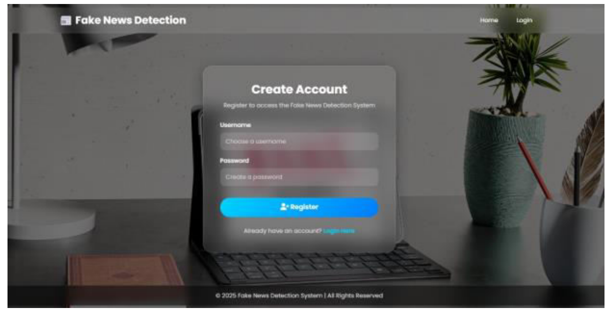

Fig 7.3 Account Creation

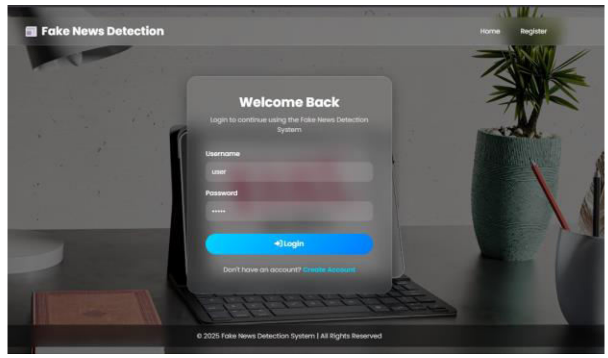

Fig 7.4 Login Page

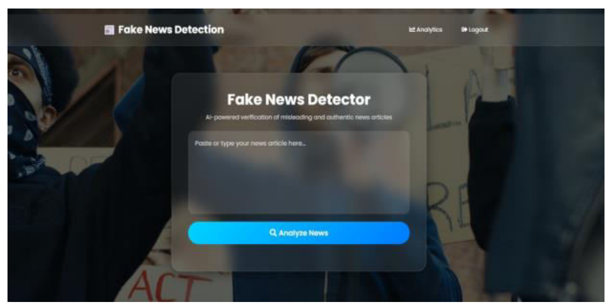

Fig 7.5 News Analysis Interface

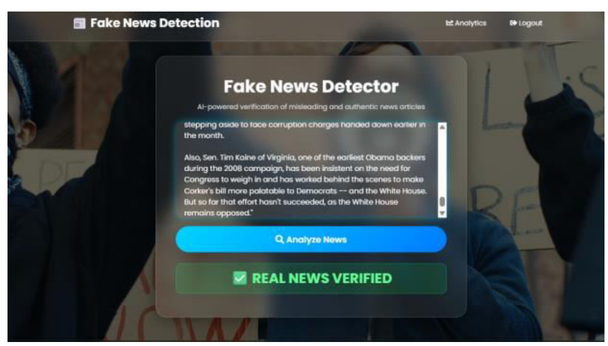

Fig 7.6: News Classification Output

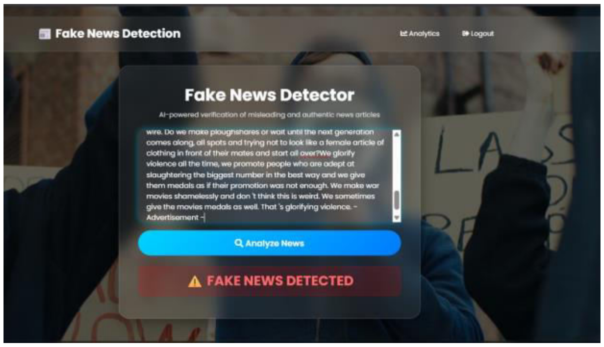

Fig 7.7: Fake News Detection Result

Fig 7.8: Statistical Analysis Of News Classification

8. CONCLUSION

In conclusion, this study addresses a crucial problem in the current digital era by showcasing the potential of utilizing NLP approaches in conjunction with sophisticated models like LSTM for fake news detection. The system can detect patterns that differentiate genuine news from fraudulent by utilizing text-based characteristics and contextual comprehension, making it a useful tool in the fight against false information. When compared to conventional techniques, the suggested system's capacity to analyze news items using both feature extraction and sequence modeling gives increased accuracy. But the study also identifies areas that could be improved in the future, like implementing explainable AI techniques to increase model transparency, improving real-time detection skills, and integrating multimodal data. Further improvements and refinements, such multilingual support and constant learning, will allow the system adjust to new problems as fake news continues to change, guaranteeing its relevance in the ongoing battle against disinformation. In the end, our work advances the field of fake news identification by providing a basis for further study and the creation of more reliable, scalable methods.

9. FUTURE ENHANCEMENT

Combining text with multimodal data—such as pictures, videos, and social media metrics—is a crucial area for advancement. Sensational headlines or deceptive images are frequently used in fake news, thus examining these components could give more context and increase detection accuracy. Using pre-trained models like BERT or GPT, which are able to capture intricate linguistic patterns and nuances that conventional models can overlook, could be another improvement. This would enhance the system's overall performance and enable it to handle text more efficiently. Another crucial topic is dealing with data imbalance. Another crucial objective is to move the system toward real-time detection. The system would be more useful for real-world applications if the model was optimized for real-time false news identification. This would allow for instantaneous predictions when news pieces are released. Additionally, by enabling the model to explain the logic behind its predictions, explainable AI (XAI) approaches like SHAP or LIME could enhance transparency by assisting consumers in understanding why a news story was categorized as authentic or fraudulent. The technology would be more beneficial if it were expanded to cover multiple languages, enabling it to identify bogus news in various linguistic and cultural situations. Integrating social media data for user interaction analysis—such as shares, likes, and comments—may offer insightful information about the reliability of news stories.

REFERENCES

[1] D. Pogue, ''How to stamp out fake news,'' Sci. Amer., vol. 316, no. 2, p. 24, Jan. 2017.

[2] H. Allcott and M. Gentzkow, ''social media and fake news in the 2016 election,'' J. Econ. Perspect., vol. 31, no. 2, pp. 211–236, May 2017.

[3] R. Zafarani, X. Zhou, K. Shu, and H. Liu, ''Fake news research: Theories, detection strategies, and open problems,'' in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, Jul. 2019, pp. 3207–3208.

[4] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, ''The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review,'' J. Public Health, vol. 31, no. 7, pp. 1007–1016, Jul. 2023.

[5] S. Vosoughi, D. Roy, and S. Aral, ''The spread of true and false news online,'' Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[6] C. Silverman, This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook. New York, NY, USA: BuzzFeed News, 2016.

[7] C. Xu and N. Yan, ‘‘AROT-COV23: A dataset of 500k original Arabic tweets on COVID-19,’’ in Proc. 4th Workshop Afr. Natural Lang. Process., 2023, pp. 1–9.
[8] C. Colomina, H. S. Margalef, R. Youngs, and K. Jones, The Impact of Disinformation on Democratic Processes and Human Rights in the World. Brussels, Belgium: European Parliament, 2021.
[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, ‘‘Fake news detection on social media: A data mining perspective,’’ ACM SIGKDD Explor. Newslett., vol. 19, no. 1, pp. 22–36, 2017.
[10] X. Zhang and A. A. Ghorbani, ‘‘An overview of online fake news: Characterization, detection, and discussion,’’ Inf. Process. Manage., vol. 57, no. 2, Mar. 2020, Art. no. 102025.
[11] X. Zhou and R. Zafarani, ‘‘A survey of fake news: Fundamental theories, detection methods, and opportunities,’’ ACM Comput. Surveys, vol. 53, no. 5, pp. 1–40, Sep. 2020.
[12] J. Shang, J. Shen, T. Sun, X. Liu, A. Gruenheid, F. Korn, A. D. Lelkes, C. Yu, and J. Han, ‘‘Investigating rumor news using agreement-aware search,’’ in Proc. 27th ACM Int. Conf. Inf. Knowl. Manage., Oct. 2018, pp. 2117–2125.
[13] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, ‘‘Defending against neural fake news,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 9054–9065.
[14] W. Wang, ‘‘‘Liar, liar pants on fire’: A new benchmark dataset for fake news detection,’’ in Proc. 55th Annu. Meeting ACL (Short Papers), vol. 2. Vancouver, BC, Canada, Jul. 2017, pp. 422–426.
[15] N. Vo and K. Lee, ‘‘Where are the facts? Searching for fact-checked information to alleviate the spread of fake news,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 7717–7731.
[16] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M . Akhtar, A. Ekbal, A. Das, and T. Chakraborty, ‘‘Fighting an infodemic: COVID-19 fake news dataset,’’ in Proc. Int. Workshop Combating Line Hostile Posts Regional Lang. During Emergency Situation. Cham, Switzerland: Springer, 2021, pp. 21–29.
[17] L. Zadeh, ‘‘Fuzzy sets,’’ Inf. Control, vol. 8, no. 3, pp. 338–353, 1965.
[18] L. Zadeh, Fuzzy Logic. New York, NY, USA: Springer, 2023, pp. 19–49.
[19] J.-S. R. Jang, ‘‘ANFIS: Adaptive-network-based fuzzy inference system,’’ IEEE Trans. Syst. Man, Cybern., vol. 23, no. 3, pp. 665–685, Jun. 1993.
[20] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, ‘‘A hierarchical fused fuzzy deep neural network for data classification,’’ IEEE Trans. Fuzzy Syst., vol. 25, no. 4, pp. 1006–1012, Aug. 2017.
[21] R. Das, S. Sen, and U. Maulik, ‘‘A survey on fuzzy deep neural networks,’’ ACM Comput. Surv., vol. 53, no. 3, pp. 1–25, May 2020.
[22] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, ‘‘Fake news on social media: The impact on society,’’ Inf. Syst. Frontiers, vol. 26, pp. 443–458, Jan. 2022.