Deep Learning Approaches for Advanced Linguistic Analysis

Kanishka Narang
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjcs103@gitjaipur.com

Karan Verma
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjcs104@gitjaipur.com

Abhilasha Sharma
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjcs006@gitjaipur.com

Vinita Sharma
Assistant Professor, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
vinita.sharma@gitjaipur.com

ABSTRACT: Recent advances in deep learning have significantly reshaped the field of linguistic analysis by enabling more accurate and context-aware processing of human language. This paper explores the application of deep learning models, with a primary focus on transformer-based architectures, across key linguistic tasks such as sentiment analysis, named entity recognition, and syntactic parsing. The study begins with the construction of a comprehensive and diverse dataset that captures a wide range of linguistic features and contextual variations, providing a strong foundation for experimental evaluation. The results demonstrate that transformer models are highly effective in modeling complex language structures and semantic relationships, allowing them to capture subtle linguistic nuances that traditional methods often overlook. The findings contribute to the expanding body of research in natural language processing and highlight the practical potential of deep learning techniques in applications including sentiment analysis, automated translation, and conversational agents.

KEYWORDS: Deep Learning, Linguistic Analysis, Natural Language Processing, Transformer Models, Sentiment Analysis, Syntactic Parsing, Language Semantics, Dataset Development, Language Structure.

1. INTRODUCTION

The rise of misinformation and fake news in digital media has become a major challenge, necessitating advanced detection methods [1]. Traditional fact-checking approaches, such as rule-based methods and human verification, are often insufficient due to the rapid spread of inaccurate content and evolving disinformation tactics [2]. Therefore, automated solutions leveraging linguistic analysis and deep learning have gained significant attention. Deep learning models, particularly those employing natural language processing (NLP), provide an effective means of identifying linguistic features associated with misinformation [3], [4]. Models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers have demonstrated strong capabilities in text classification, making them suitable for detecting tricky patterns in news articles [5]. This study applies deep learning-based linguistic analysis to a dataset of potetially misleading news articles. By verifying syntactic, semantic, and stylistic features, we aim to uncover key linguistic markers that distinguish fake news from credible sources [6], [7]. The findings will contribute to the

development of automated misinformation detection systems, improving information integrity and combating digital disinformation [8].

2. METHODOLOGY

This study uses a comparative analysis framework to evaluate the effectiveness of deep learning in linguistic analysis. The image presents a deep learning-based methodology for linguistic analysis of fake news detection. It is structured into two main phases: Training Phase and Testing Phase, Aiming on the use of LSTM, BERT, and RoBERTa as deep learning algorithms.

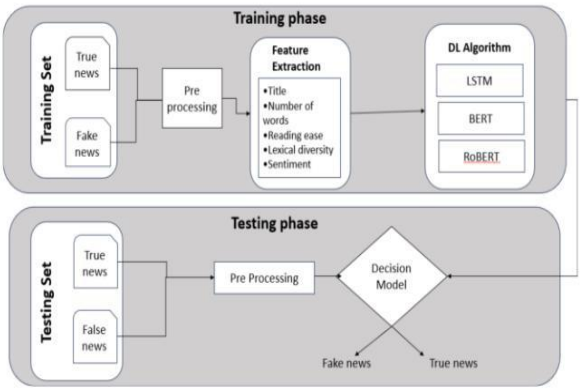

Figure 1: Training Phase

A. Dataset

In this case study we extract the dataset from the kaggle dataset name (Misinformation & Fake News text dataset). In this dataset it consist three files but for our research we use only two dataset namely:-(Dataset_Misinfo_true and DataSet_Misinfo_false) which in csv form. Both data set consist of real and fake news and the Genre of the news type in DataSet is Political news.

B. Preprocessing

Before applying Deep learning algorithms the dataset undergoes multiple preprocessing steps to ensure text consistency, correctness and improving model performance:

- **Tokenization:** The text is tokenized using the SentencePiece tokenizer. This step ensures that words are broken into meaningful subwords, allowing the model to handle unknown words more efficiently.
- **Stopword Removal:** Common English stopwords such as “the,” “is,” and “in” are removed to eliminate excess information and improve text clarity.
- **Lemmatization:** Words are reduced to their base forms (e.g., “running” → “run”) to standardize vocabulary across the dataset.
- **Noise Removal:** Special characters, punctuation, HTML tags, and non-textual data are filtered out to prevent interference in model training..
- **Named Entity Recognition (NER):** Entities such as names, locations, and organizations are identified to analyze their resemblance with misinformation.

- **Text Normalization:** Converting uppercase letters to lowercase and standardizing abbreviations to maintain regularity.

C. Feature Extraction

In this case study we use set of features. To find most accurate feature to differentiate between real and fake news. The linguistic features extracted from the text include:

- Title: (head analysis)
- Number of words (text length-based features:- wor count of text)
- Reading ease (how complex the text is:- to make reading easy)
- Lexical diversity (variation in word usage)
- Sentiment analysis (positive, negative, or constant tone)

D. Deep Learning Algorithms

In this study, the researcher using deep learning algorithms to improve the accuracy and efficiency of fake news detection through linguistic analysis. our approach to compare LSTM, BERT, and RoBERTa deep learning algorithms to examine accuracy for textual patterns and contextual relationships in news.

LSTM (Long Short-Term Memory): LSTM is a type of Recurrent Neural Network (RNN) designed to capture long-range dependencies in serialised data, making it highly effective for text analysis and fake news detection. In our fake news detection system, LSTM processes the text sequentially, learning patterns and dependencies between words to differentiate between true and fake news. After preprocessing the text, the data is fed into an LSTM network, where it store contextual relationships about news article. Using LSTM improves our model's ability to understand pattern in news articles, making it particularly useful for detecting fake new. the accuracy by LSTM is 0.97 or 97.2%.

BERT (Bidirectional Encoder Representations from Transformers): In our fake news detection system, we use BERT to enhance contextual understanding and improve classification accuracy. Unlike traditional models, BERT processes text bidirectionally, meaning it considers both previous and next words to gain a deeper understanding of word meanings.The process begins with preprocessing. The model then extracts contextual word embeddings, capturing subtle linguistic differences between real and fake news. These features are passed through a fully connected classification layer, where a good BERT model determines whether the news is true or fake. BERT's ability to handle long texts, complex sentence structures, and semantic relationships makes it more effective than traditional models like LSTM. By integrating BERT into our system, we achieve greater accuracy and reliability in detecting misleading information, ensuring a more robust and dependable fake news detection mechanism.

RoBERTa: In this case study We incorporate RoBERTa (Robustly Optimized BERT Pretraining Approach) into our fake news detection system to enhance contextual text analysis and improve classification accuracy. As an advanced version of BERT, RoBERTa is designed to handle longer texts more effivitively by eliminating the next sentence prediction (NSP) objective and training on larger, more different datasets.

Our process begins with preprocessing, where news articles are cleaned, tokenized, and converted into word embeddings using RoBERTa's tokenizer. The processed text is then observe by RoBERTa, which generates contextual representations of words and sentences. These features are then passed through a fully connected classification process, where the model determines whether the news is true or fake.

RoBERTa overcome standard BERT by dynamically adjusting training strategies, making it more effective at detecting normal language patterns commonly found in misinformation. By using RoBERTa's deep contextual understanding, our fake news detection system achieves higher accuracy and reliability in identifying fake content.

In evaluation we presents the evaluation, evaluation metrics and performance of the model based on accuracy, recall, f1 score

Accuracy: Measures the overall correctness of the model's predictions.

ACC=(TP+TN)/(TP+TN+FP+FN)

Recall: Measures how well the model identifies all relevant instances.

REC = TP/(TP + FN)

F1-Score: Balances precision and recall for a comprehensive performance measure.

F1 = (2 × (PRE × REC))/(PRE + REC)

Where,

PRE = TP/(TP + FP)

TP : It is the number of accurately classified fake news (true positives).

FP : It is the number of incorrectly classified real news (false positives).

TN : It is the number of accurately classified real news (true negatives).

FN : It is the number of inaccurately classified fake news (false negatives).

4. RESULTS AND DISCUSSION

In this study on fake news detection using deep learning evaluated three models—LSTM, BERT, and RoBERTa—on the Misinformation Fake News Text Dataset (79k). The results show that RoBERTa performed well then both LSTM and BERT, achieving the highest accuracy (98.5%), recall (98.3%), and F1-score (98.4%). This exceptional performance is due to RoBERTa's advanced pretraining and its ability to capture deep contextual relationships in text.

BERT also delivered strong results, improving on LSTM with an accuracy of 97.8%, highlighting the effectiveness of transformer-based models in understanding linguistic nuances. Meanwhile, LSTM performed well with 97.2% accuracy, proving to be a solid baseline, though it lacks the contextual awareness of transformer models.

These findings confirm that transformer-based models like BERT and RoBERTa are more efficient than LSTM for detecting fake news. They also focus on opportunities for further improvements through fine-tuning, larger datasets, and model ensemble techniques.

Table 1: Obtained Results

DL models	Accuracy	Recall	F1-score
LSTM	97.2%	96.8%	97%
BERT	97.8%	97.5%	97.5%
RoBERTa	98.5%	98.3%	98.4%

5. CONCLUSION

This study confirms that transformer-based models are more effective than traditional deep learning architectures for detecting fake news. Among the three models analyzed LSTM, BERT, and RoBERTa RoBERTa achieved the highest accuracy (98.5%), followed by BERT (97.8%) and LSTM (97.2%). The results emphasize the ability of *BERT and RoBERTa* to better capture contextual relationships in text, making them superior to LSTMs for linguistic analysis. RoBERTa's advanced training approach and deeper contextual understanding contributed to its higher recall (98.3%) and F1-score (98.4%). These findings suggest that transformer-based architectures should be the preferred choice for fake news detection. Future research could focus on further fine-tuning, domain-specific optimizations, and ensemble techniques to enhance or improve the performance.

REFERENCES

[1] G. Jain, M. K. Jha, "Enhancing E-Commerce Intelligence through Machine Learning-Based Sentiment Analysis and Forecasting", International Journal of Global Research in Science and Technology, vol. 10, pp. 1-7, 2025.

[2] M. K. Sain and N. Sharma, "A study of research issues and challenges of big data analytics," Journal of Advances and Scholarly Researches in Allied Education, vol. 16, no. 5, pp. 1699–1707, 2019.

[3] N. Sharma, "An analytical study of distributed data store using big data analysis technique," Research Methods, Imparc, 2019.

[4] A. Bohra, K. Paliwal, S. Soni, "Online code editor: A cloud-based platform for real-time web development," International Journal of Global Research in Science and Technology, vol. 9, pp. 52–76, 2024.

[5] R. Joshi, M. Farhan, U. Sharma, S. Bhatt, "Unlocking Human Communication: A Journey through Natural Language Processing", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 3, pp. 245-250, 2024.

[6] S. A. Saiyed, N. Sharma, H. Kaushik, P. Jain, G. K. Soni and R. Joshi, "Transforming portfolio management with AI and ML: shaping investor perceptions and the future of the Indian investment sector," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1108-1114, 2025.

[7] N. Soni, N. Nigam, "Recent Advances in Artificial Intelligence and Machine Learning: Trends, Challenges, and Future Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 1, pp. 9-12, 2025.

[8] Y. Sharma, N. Mulani, M. K. Jha, "Artificial Intelligence-Driven Cybersecurity for Modern Digital Ecosystems", International Journal of Global Research in Science and Technology, vol. 10, pp. 34-39, 2025.

[9] K. Gautam, M. Dubey, N. Jain, "Face Detection and Recognition for Patient", International Journal of Biomedical Engineering, Vol. 8, Issue. 2, pp. 1-7, 2022.

[10] A. Gautam, R. Ajmera, D. K. Dharamdasani, S. Srivastava, and A. Johari, "Improving climate change predictions using time series analysis and deep learning," Global and Stochastic Analysis, vol. 12, no. 4, Jul. 2025.

[11] G. Sharma, N. Hemrajani, S. Sharma, A. Upadhyay, Y. Bhardwaj, and A. Kumar, "Data management framework for IoT edge-cloud architecture for resource-constrained IoT application," Journal of Discrete Mathematical Sciences and Cryptography, vol. 25, no. 4, pp. 1093–1103, 2022.

[12] R. Misra, Dr. R. Sahay, "A Review on Student Performance Predication Using Data Mining Approach", International Journal of Recent Research and Review, Vol. 10, Issue. 4, pp. 45-47, 2017.

[13] S. Tiwari, K. Gautam, R. Kumar, "A Survey on Deep Learning", National Conference on Renewable Energy & Digitalization Resources for the Development of Rural Areas, 2020.

[14] H. Kaushik, "Artificial Intelligence in Healthcare: A Review", International Journal of Engineering Trends and Applications (IJETA), Vol. 11, Issue. 6, pp. 58-61, 2024.

[15] S. Pathak, S. Tiwari, K. Gautam, J. Joshi, "A Review on Democratization of Machine Learning In Cloud", International Journal of Engineering Research and Generic Science, Vol. 4, Issue. 6, pp. 62-67, 2018.

[16] H. Kaushik, "Artificial Intelligence: Recent Advances, Challenges, and Future Directions", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 2, 2025.

[17] A. Maheshwari and R. Ajmera, "A comprehensive guide to natural language processing in Sanskrit with named entity recognition," in Proc. ACM Int. Conf. on Information Management & Machine Intelligence, 2023.

[18] A. Sharma and K. Gautam, "Flood prediction using machine learning technique," 2nd International Conference on Pervasive Computing Advances and Applications (PerCAA 2024), pp. 319-327, 2024.

[19] M. K. Jha, "Recent Trends and Emerging Applications of the Internet of Things: Transforming the Way We Live and Work", International Journal of Engineering Trends and Applications (IJETA), Vol. 12, Issue. 4, pp. 239-244, 2025.

[20] A. Maheshwari, R. Ajmera and D. K. Dharamdasani, "Unmasking Embedded Text: A Deep Dive into Scene Image Analysis," 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), pp. 1403-1408, 2023.

[21] P. Upadhyay, K. K. Sharma, R. Dwivedi and P. Jha, "A Statistical Machine Learning Approach to Optimize Workload in Cloud Data Centre," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 276-280, 2023.

[22] R. Joshi, A. Maritammanavar, "Deep Learning Architectures and Applications: A Comprehensive Survey", International Conference on Recent Trends in Engineering & Technology (ICRTET 2023), pp. 1-5, 2023.