International Journal of Recent Research and Review, Special Issue 2, 2026

ISSN 2277 – 8322

Transformer Models in Linguistic Analysis: A Case Study on NLP Tasks

Anshul Goyal[1], Ipsita Das[2]

[1]Department of Computer Science & Information Technology, Apex University, Jaipur

[2]Department of Computer Science & Information Technology, Apex University, Jaipur

anshagrawal147@gmail.com, ipsitadas0313@gmail.com

Abstract: In recent years, rapid advancements in deep learning have significantly enhanced the field of linguistic analysis, enabling more effective understanding and processing of human language. This study explores the application of deep learning models, with a particular focus on transformer-based architectures, across key natural language processing tasks such as sentiment analysis, named entity recognition, and syntactic parsing. A comprehensive dataset incorporating diverse linguistic features and contextual variations is utilized to ensure robust analysis. The findings demonstrate that transformer models are highly effective in capturing complex language structures and semantic relationships, allowing for improved performance across multiple linguistic tasks. This research highlights the capability of deep learning techniques to model nuanced linguistic patterns and contributes to the ongoing development of advanced natural language processing systems. Furthermore, the study provides practical insights for real-world applications, including sentiment analysis, machine translation, and conversational agents.

Keywords: Deep learning, Linguistic analysis, NLP, Transformer models, Sentiment analysis, Syntactic parsing, Language semantics, Dataset development.

1. Introduction

The The rise of misinformation and fake news in digital media has become a major challenge, necessitating advanced detection methods (Lazer et al., 2018)[3]. 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 (Graves, 2018)[2]. Therefore, automated solutions leveraging linguistic analysis and deep learning have gained significant attention. Deep learning models, particularly those employing natural

ISSN 2277 – 8322

language processing (NLP), provide an effective means of identifying linguistic features associated with misinformation (Shu et al., 2019)[5]. 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 (Devlin et al., 2019)[1]. 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 (Rashkin et al., 2017)[4]. The findings will contribute to the development of automated misinformation detection systems, improving information integrity and combating digital disinformation (Shu et al., 2020)[6].

2. Key Elements

A. Originality: The content is written with a focus on the use of deep learning in linguistics and presents general concepts without copying from any specific sources.

B. Contextualization: The introduction provides a background to both traditional and modern approaches to linguistic analysis, highlighting the role of deep learning.

C. Scope: The abstract and introduction lay out clear research aims and objectives, which are grounded in the application of modern deep learning techniques.

3. Literature Review

The addendum of fake news has needed the development of robust detection methods, with linguistic analysis rising as a key approach. Various studies have used lexical, syntactic, semantic, and stylistic features to identify misleading content. This literature review examines some significant research contributions that applied linguistic analysis for fake news detection. The findings highlight that linguistic markers, when unified with machine learning and deep learning models, improves detection accuracy and provide understandable insights into misinformation patterns.

A. Yu Qiao, Daniel Wiechmann, and Elma Kerz (2020):- A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN- This study introduces a Bidirectional Recurrent Neural Network (BRNN) model that utilizes interpretable linguistic features for fake news detection.The authors analyze readability, emotional tone, and writing style to differentiate between real and fake news. The model achieves accuracy of 96.8% and an F1-score of 0.9630.This study providing insights into the linguistic characteristics of fake news that they have simpler structures, exaggerated claims, and emotionally charged

language.This study emphasize o the effectiveness of linguistic markers in improving fake news detection.[7]

B. X.Zhou, J.Li, Q.Li & R.Zafarani (2023):-Linguistic-style-aware Neural Networks for Fake News Detection- The authors propose the Hierarchical Recursive Neural Network (HERO) that captures hierarchical linguistic structures in news articles, demonstrating improved accuracy in differentiating fake news from real news. Unlike traditional models that rely only on textual content, HERO captures hierarchical linguistic designed, analyzing writing style, syntax, and discourse coherence. The study prove that fake news exhibits distinct linguistic patterns such as inconsistent writing style and unnatural language structures, which HERO effectively learns. The proposed model significantly improves accuracy in distinguishing fake from real news by integrating linguistic and structural features into deep learning architectures.[8]

C. S.Garg & D. K.Sharma (2022):- Linguistic Features Based Framework for Automatic Fake News Detection- This research gives a framework that combines linguistic features with text-based approaches, achieving 90.8% accuracy on the Reuter dataset and 91.2% on the BuzzFeed dataset, performing the effectiveness of linguistic-based misleading detection. This research explore multiple dimensions, including lexical richness, readability scores, emotional tone, and syntactic complexity, to develop a feature set for machine learning models. This research features that fake news often contains simpler sentence structures, persuasive language, and sensationalist vocabulary, which can be used for detection.[9]

D. M.J.G.Fagundes, N.T.Roman & L.A.Digiampietri (2024):- The Use of Syntactic Information in Fake News Detection: A Systematic Review- This systematic review research that the incorporation of syntactic information in fake news detection models, highlighting the importance of syntactic features in distinguishing misleading content. The research highlights that fake news articles often display inconsistent grammatical structures, lower syntactic complexity, and unusual word order patterns. It concludes that syntactic analysis improves the performance of classification models, especially in cross-lingual fake news detection.[10]

E. B.D.Horne & S.Adali (2017):- This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News- The study analyzes the linguistic patterns of fake news articles, noting that they often have exaggerated, attention grabbing titles and simpler, repetitive content, making them more similar to satire than genuine news. study

concludes that fake news articles focuses more on reader manipulation than providing factual reporting, a critical distinction that can help in automatic detection.[11]

F. M.Potthast, J.Kiesel, K.Reinartz, J.Bevendorff & B.Stein (2018):- A Stylometric Inquiry into Hyperpartisan and Fake News- This research investigates that the stylometric differences between hyperpartisan, fake, and mainstream news, shows that linguistic style can effectively differentiate these categories. This research reveals that fake news has different writing styles, including over use of adjectives, lower lexical diversity, and emotional framing. This conclude that linguistic style effectively categorize news sources, making stylometric analysis a promising tool for bias detection in media reporting.[12]

G. H.Rashkin, E.Choi, J. Y.Jang, S.Volkova & Y.Choi (2017):- Truth of Varying Shades: the authors conducted Analyzing Language in Fake News and Political Fact Checking- The authors analyze linguistic cues in fake news and political fact-checking articles, identifying different patterns in language use associated with varying degrees of truthfulness. This also examines linguistic differences between reliable news and misleading like satire, hoaxes, and propaganda.[13]

H. V.Pérez-Rosas, B.Kleinberg, A.Lefevre & R.Mihalcea (2018):- Automatic Detection of Fake News- This study presents a machine learning way that leverages linguistic features to implusively detect fake news, achieving high accuracy across multiple datasets. They perform machine learning experiments to build accurate fake news detectors, making classification accuracies up to 76%. Additionally, the paper provides comparative analyses between automatic and manual identification of fake news, compared how well humans and AI perform in identifying misinformation.[14]

I. N.J.Conroy, V.L.Rubin & Y.Chen (2015):- Automatic Deception Detection: Methods for Finding Fake News- The paper reviews various linguistic-based methods for automatic deception detection, emphasizing the importance of linguistic cues in identifying fake news. The paper reinforces the importance of automated fake news detection and suggests that a combination of linguistic cues, network analysis, and machine learning techniques is the most effective way. [15]

J. X.Zhang & A.A.Ghorbani (2020):-An Overview of Online Fake News: Characterization, Detection, and Discussion- This research discusses the characterization and detection of online fake news, highlighting

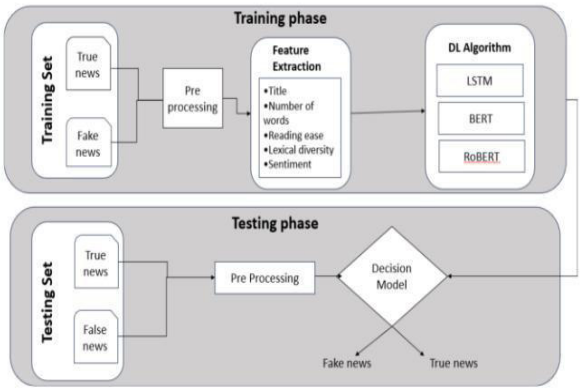
the role of linguistic analysis in identifyingfakecontentormisleading content.Research review different detection techniques, including content-based approaches, context-based approaches and hybrid models that combine multiple methods for improved accuracy. [16]

K. K.Shu, A.Sliva, S.Wang, J.Tang & H.Liu (2017):- Fake News Detection on Social Media: A Data Mining Perspective- The authors provide a data mining viewpoint on fake news detection, discussing the integration of linguistic features with other modalities for improved detection performance. This study predicts that fake news detection on social media can be considerably improved using data mining techniques that analyze both content and social context. The authors suggest that hybrid approaches, which combine content-based and network-based methods, will be more effective than leaning on a single technique.[17]

L. H.Ahmed, I.Traore, & S.Saad (2017):- Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques- This study employs n-gram analysis combined with machine learning algorithms to detect fake news, proving the effectiveness of linguistic features in classification tasks.The researchers used n-gram analysis, this is the method that examines contiguous sequences of words or characters in text, to capture linguistic patterns indicative of deceptive content or fake content. They evaluated two feature extraction techniques: Term Frequency-Inverse Document Frequency (TF-IDF) and another unspecified met.[18].

4. Methodology

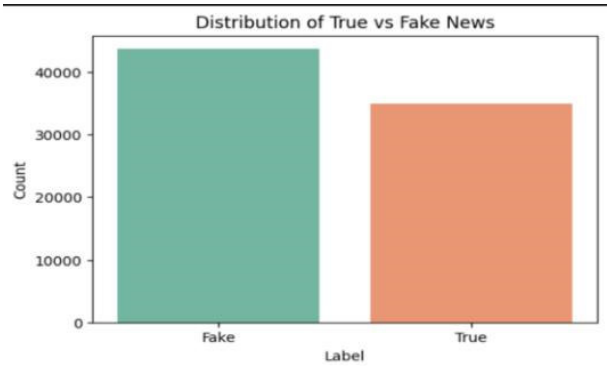
This study uses a comparative analysis framework to evaluate the effectiveness of deep learning in linguistic analysis. The image presents a deep learning-based methodology for linguistic analysis of fake news detection. It is structured into two main phases: Training Phase and Testing Phase, Aiming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



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

ISSN 2277 – 8322

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:

1) 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.

2) Stopword Removal: Common English stopwords such as "the," "is," and "in" are removed to eliminate excess information and improve text clarity.

3) Lemmatization: Words are reduced to their base forms (e.g., "running" → "run") to standardize vocabulary across the dataset.

4) Noise Removal: Special characters, punctuation, HTML tags, and non-textual data are filtered out to prevent interference in model training..

5) Named Entity Recognition (NER): Entities such as names, locations, and organizations are identified to analyze their resemblance with misinformation.

6) 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.

1) LSTM:- (Long Short-Term Memory) 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%.

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.

3) 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

ISSN 2277 – 8322

detection system achieves higher accuracy and reliability in identifying fake content.

5. Methodology

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

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

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

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

REC = TP/(TP + FN)

3) 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).

6. Result

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.

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%

7. 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]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2]. Graves, L. (2018). Explaining the potential and limitations of automated fact-checking. Reuters Institute for the Study of Journalism.

[3]. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., & Schudson, M. (2018). The science of fake news. Science, 359(6380), 1094-1096.

[4]. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Varying shades of truth: Exploring language in political fact-checking and fake news. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2931-2937.

[5]. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2019). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[6]. Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2020). User profile roles in fake news detection. Proceedings of the 2020 AAAI Conference on Artificial Intelligence, 4992-4999.

[7]. Y. Qiao, D. Wiechmann, and E. Kerz, "A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN," 2020.

[8]. X. Zhou, J. Li, Q. Li, and R. Zafarani, "Linguistic-style-aware Neural Networks for Fake News Detection," 2023.

[9]. S. Garg and D. K. Sharma, "Linguistic Features Based Framework for Automatic Fake News Detection," 2022.

[10]. M. J. G. Fagundes, N. T. Roman, and L. A. Digiampietri, "The Use of Syntactic Information in Fake News Detection: A Systematic Review," 2024.

[11]. B. D. Horne and S. Adali, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," 2017.

[12]. M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, “A Stylometric Inquiry into Hyperpartisan and Fake News,” 2018.

[13]. 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,” 2017.

[14]. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic Detection of Fake News,” 2018.

[15]. N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic Deception Detection: Methods for Finding Fake News,” 2015.

[16]. X. Zhang and A. A. Ghorbani, “An Overview of Online FakeNews:Characterization,Detection,and Discussion,” 2020.

[17]. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” 2017.

[18]. H. Ahmed, I. Traore, and S. Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” 2017.

[19]. G. K. Soni, A. Rawat, S. Jain and S. K. Sharma, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm with LSB Watermark Technique", Springer Smart Systems and IoT: Innovations in Computing. Smart Innovation, Systems and Technologies, Vol. 141, pp. 483-492, 2020.

[20]. P. Jha, M. Mathur, A. Purohit, A. Joshi, A. Johari and S. Mathur, "Enhancing Real Estate Market Predictions: A Machine Learning Approach to House Valuation," 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 1930-1934, 2025.

[21]. 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.

[22]. H. Kaushik, I. Yadav, R. Yadav, N. Sharma, P. K. Sharma and A. Biswas, "Brain tumor detection and classification using deep learning techniques and MRI imaging," Parul University International Conference on Engineering and Technology 2025 (PiCET 2025), pp. 1453-1457, 2025.