Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 21s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Deep Learning Based Fake News Detection on Social Media

Mondithoka K Prasad[1], and Prof. K. Venkata Rao[2]

[1] Department of Chemical Engineering, GMR Institute of Technology, Rajam, Andhra Pradesh 532127, India
[2] Department of Compter Science & Systems Engineering, A. U. College of Engineering, Visakhapatnam, Andhra Pradesh 530003, India

Cite this paper as: Mondithoka K Prasad, and Prof. K. Venkata Rao (2025). Deep Learning Based Fake News Detection on Social Media. *Journal of Neonatal Surgery*, 14 (21s), 246-255.

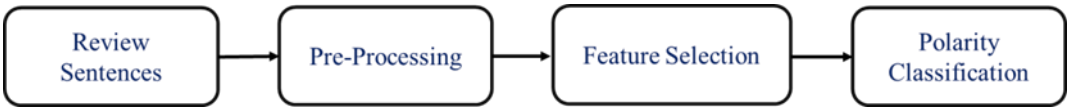
ABSTRACT

Today, social media is the primary source of news for the world. False information spread by social media has become a major global problem affecting people's lives and many aspects of society, including political, economic and social. People react to false news with disbelief, fear and disgust, often with negative emotions. We extracted some elements from the sentiment analysis of the news articles and the emotional analysis of the user reviews. The proposed two-way short-term memory and attention model for identifying false information is based on these characteristics and the element of news content. For both the training and the testing of the proposed model, we used a standard dataset called Fakeddit, which contains the headlines and comments that have been left in their wake. Using the parameters extracted, the proposed model provided high accuracy of 98 percent for the ROC curve, exceeding the accuracy of other recent studies. The findings show that the characteristics derived from emotional analysis of comments and emotional analysis of news reflect the views of the editors..

1. INTRODUCTION

Natural Language Processing (NLP) is a field of computer science and artificial intelligence that focuses on human interactions and interpreting, understanding, and generating human understandable language. It involves translating natural language sentences into numerical data for computers to interpret and understand, generating text that can be interpreted semantically by humans. NLP fills the gap in algorithms that can comprehend language, as the complexity of human language is not underestimated. Traditional NLP methods use linguistics-based mechanisms, and its ability to interpret human language is now at the core of many everyday applications, including translation software, chatbots, and voice assistants.

Sentiment analysis is a field that extracts user opinions and emotions towards entities like products, organizations, services, events, individuals, and topics. It involves modeling a system to classify reviews into labelled polarities, such as negative or positive, and adding neutral classes. Machine learning algorithms are used for efficient classification, involving several phases shown in Figure 1.



. Figure 1. Different phases involved in Sentiment Analysis

Emotion analysis classifies data based on emotions like joy, surprise, anger, fear, sadness, and disgust in texts. Emotion lexicons, developed by experts, extract emotion-based features from texts. Studies on detecting fake news use text-based features, with fake news intentionally stirring emotions to spread on social media. Emotion analysis plays a crucial role in determining user behavior towards specific topics. Vosoughi et al. found that false rumors caused fear, disgust, and surprise, while real rumors caused joy, sadness, trust, and anticipation.

Fake news is a growing issue that threatens democracy, justice, and public trust. Its popularity is due to its faster and cheaper creation and publication online. The rise of social media has also fueled this interest. Fake news detection is a challenging task, with studies in their early stages. Detection strategies focus on features like content, data propagation, and user reactions. User reactions to fake or genuine posts reflect their attitude and influence the credibility of the source.

Recent studies reveal the rapid spread of fake news, leading to widespread proliferation. Examples include anti-vaccination misinformation and rumors comparing voter numbers. This has led to global unrest and hindered the fight against COVID-19. Early detection of fake news is crucial to prevent uncertainties and prevent global unrest.

This study reviews machine learning and deep learning-based fake news detection methods, particularly in social media, and

proposes a deep learning strategy for identifying and analyzing user sentiment.

2. PREVIOUS WORKS:

G. Güler and S. Gündüz, [1] proposed automatic fake news detection for social media platforms in Turkish and English languages. It creates a real-world public dataset of Turkish fake and real news tweets using BuzzFeed and ISOT datasets. Two deep learning-based FNDSs are developed using CNN and RNN-LSTM algorithms with the Word2vec word embedding model, addressing the issue of false information spread on social media.

Iftikhar Ahmad et al., [2] focus on classifying fake news articles using machine learning models and ensemble techniques. Data from the World Wide Web was collected to identify patterns in text that differentiate fake articles from true news. The models were trained and parameter-tuned for optimal accuracy. Ensemble learners showed better performance metrics than individual learners. Future research should focus on identifying key elements involved in fake news spread, graph theory, and machine learning techniques. Real-time fake news identification in videos could also be explored.

Shu, K. C., Wang, S., Lee, D., Liu, H., & Narayanan, V. K. (2017) [3] provides a broad overview of the fake news detection problem, discusses various data mining techniques applied to it, and highlights the challenges and future directions in this field.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N & Polosukhin, I. (2017) [4] introduces the Transformer model and the attention mechanism, which is a key component of the approach in Document 2. It explains the benefits of attention in allowing models to focus on relevant parts of the input sequence.

Zhou, X., & Zafarani, R. (2020) [5] provides a comprehensive survey that delves into the theoretical underpinnings of fake news, explores different detection methodologies (including content-based, context-based, and propagation-based approaches), and analyzes the societal consequences of misinformation.

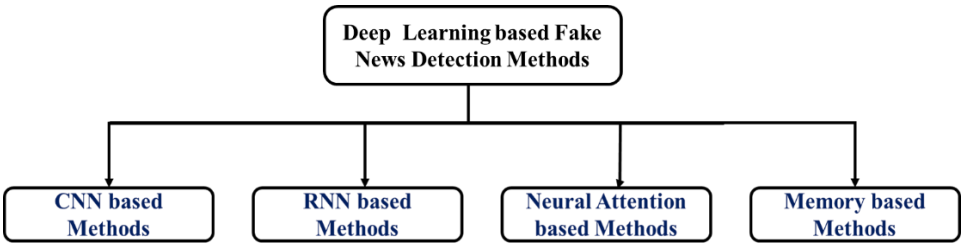
3. MATERIALS AND METHODS:

3.1. Deep Learning Networks for Fake News Detection

Lexicon-based approaches extract semantic values from words or phrases to guide classification processes. They can be classified into dictionary-based methods and corpus-based methods. Dictionary-based fake news detection uses pre-defined dictionary terms, while corpus-based methods use techniques like k-nearest neighbors, Hidden Markov Models, and Conditional Random Field.

Machine learning trains computer programs to learn from previous examples, apply the model to new samples, and monitor outcomes. Supervised learning adjusts learning data with labels, while unsupervised learning groups similar data points without labels. Reinforcement learning replaces training data with an agent learning through trial-and-error.

Traditional methods for fake news detection rely on machine learning, which relies on hand-crafted features. Recent research is focusing on developing neural network models that don't require hand-crafted features. Neural networks can be classified into Feed forward Neural Networks (FNNs) and Recurrent Neural Networks (RNNs). Deep neural networks, with more than two layers, can learn or represent input data at different levels of abstraction, generating accurate results. Deep learning strategies are used in computer vision and sequence modeling problems in Natural Language Processing (NLP). Recurrent neural networks are considered the best candidate for sentiment analysis-based fake news detection. The following Figure 2 describes the classification of deep Learning Techniques for Fake News detection.



3.2 DATASET

In this work, the dataset Fakeddit is used to test the proposed methodology. The Fakeddit dataset (https://github.com/entitize-fakeddit) is a large-scale and multi-dimensional dataset (text and images) collected by Nakamura et al. from Reddit social media platform from 19 March 2008 to 24 October 2019. They developed hybrid text-based and image-based models and performed in-depth experiments on a number of classification variants, highlighting the importance of innovative multimodal and fine-grained classification features exclusive to Fakeddit. This dataset contains over one million articles from different

topics. Several elements, such as images, comments, users, domains and other metadata, are associated with these sites, as shown in Table 1. This file contains many headlines. A single report may contain multiple reports and may not contain a reportable document. Researchers provided three labels for each site, classified into two ways, three ways of laundering, and six ways of laundering. In our study, we propose to use a two-way or binary classification approach to classify a rumor as either false or true.

Field Name	Description
Author	A person or bot that registers on social networking sites.
Title	The title or brief description of the post.
Domain	The source of news.
Has_image	The post whether attached with an image or not.
Id	The ID of the post.
Num of Comments	The number of comments on the post; this feature represents the level of popularity of the post.
Score	The numerical score another user gives to the post; This feature shows if another user approves or declines to share the post.
Upvote_ratio	A value representing an estimate of the approval/disapproval of other users on posts.
Body	The content of the comment

Table 2. Data Set of fake news detection

News Type	Number of Titles
Fake News	~25,000
Real News	~75,000

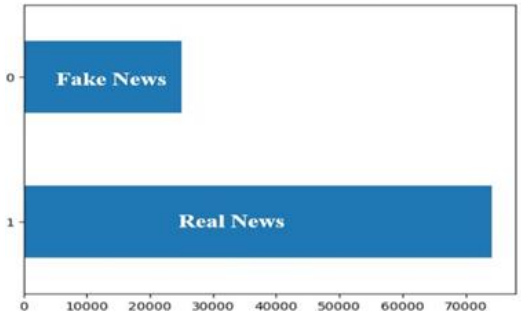


Figure 3.Proportion of Real and Fake news titles in the considered dataset

This Figure 3 visually represents the distribution of real and fake news titles within the dataset. The x-axis shows the number of titles, and the y-axis represents the news category (0 for Fake News, 1 for Real News). The graph clearly demonstrates that the dataset is imbalanced. There are significantly more real news titles than fake news titles.

Proposed Methodology

The proposed model for detecting fake news is divided into three modules, namely:

1.Emotion Analysis Module (EAM): analyses the emotional state of the comments.

2.Sentiment Analysis Module (SAM): analyses the sentiment of the news titles

3.Text Classification Module (TCM) : classifies the news based on the textual features of the titles

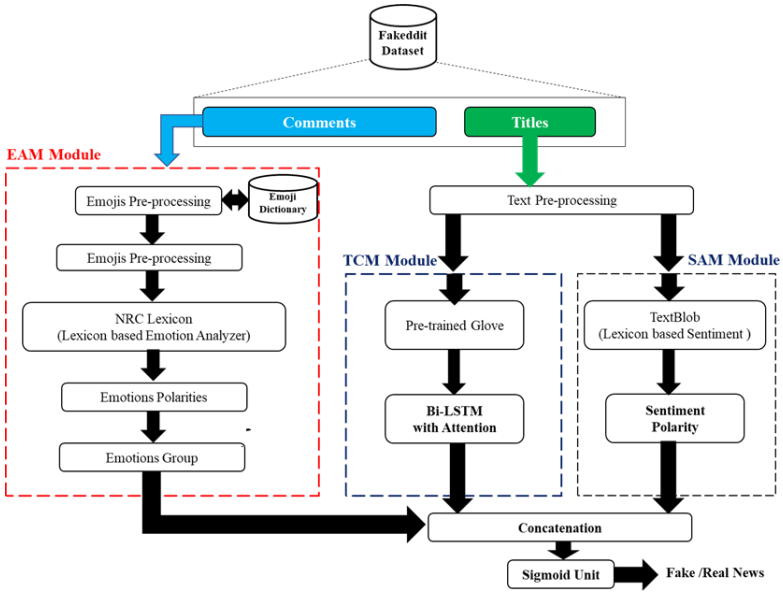


Figure 4. Modules of the Proposed Methodology

Emotion Analysis Module (EAM):

Since the comments in social media represent colloquial language because they are related to the public and contain a lot of noise, the comments are processed as per the sequence of operations shown below

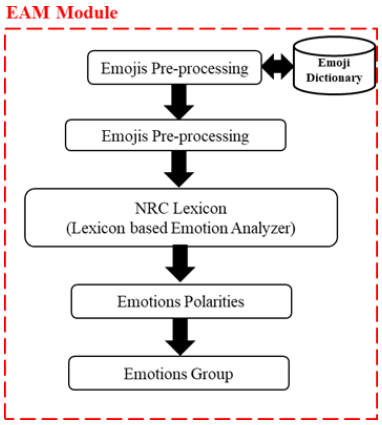


Figure 5. Flow of Operations in Emotion Analysis Module

THE STEPS INVOLVED IN THE EMOTION ANALYSIS UNIT ARE AS SHOWN BELOW.

1.Analysing each comment according to the emotions it carries based on NRCLex.

2.Categorizing each comment to the group it belongs to (novelty, expectation, and neutral) based on emotional polarity.

3.Grouping the comments for each piece of news based on ID.

Assign the largest group (most frequent group) of all comments for each piece of news. In the case of equality between the two groups (novelty, expectation), such comments are assigned to neutral group/sentiment. As the model can handle with numeric data, it is required to normalize the emotion groups and convert them to numeric values between 0 and 1, (0 = expectation, 0.5 = neutral, 1 = novelty). The generated column denotes the emotion-based feature and this feature will be concatenated with other columns in the concatenate layer

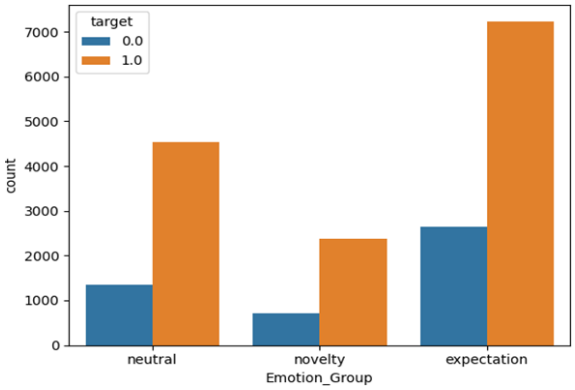


Figure 6. Proportion of Emotion Groups

3.2.2. SENTIMENT ANALYSIS MODULE (SAM)

The SAM unit analyzes the sentiment of news titles. This unit uses a lexicon- based sentiment analyser, which is TextBlob (It is a python-based library that is designed based on NLTK Toolkit), to analyse embedded sentiments in each title and assigns a polarity score ranges from -1(Negative Polarity: disgusting", "terrible", and "pathetic") to +1(Positive Polarity: great, best) to each news title. Each title has been pre-processed the steps which are listed in the previous section. The output obtained from this pre-processing stage will be supplied to both SAM and TCM modules as shown in figure 7. TextBlob processes each input sentence (a ordered collection of words), and separate score for each word will be calculated. Finally, the overall sentiment of the sentence will be calculated by pooling procedure.

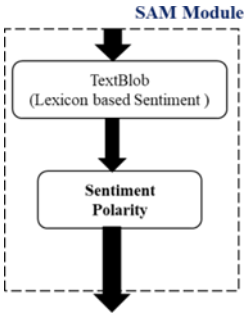


Figure 7. Flow of Operations in Sentiment Analysis Module

3.2.3. TEXT CLASSIFICATION MODULE (TCM)

The main function of this module is to extracts text-based features of labelled samples of both fake and real news. Here, Bidirectional-LSTM (Bi-LSTM) and Bi- LSTM with Attention are employed to classify a new title.

The main steps involved in this module are

4. Division of pre-processed samples into Training Set and Test Set in the ration of 80:20 respectively.

5. Tokenization and padding

6. Generation of Text Embedding using Glove Language Model, a pretrained language models which captures the context of every sentence. This model Embedding converts each word in the text string into vectors of n dimensions, where n is the embedding dimension, which represents 300d in this proposed model. The distance between the two embedding vectors indicates the proximity of the words in terms of their relationship to each other

7. Classification by Bi-LSTM Model (Bi-LSTM) with attention model is trained and tested)

8. Figure 8 and 9 show the basic architecture of Bidirectional LSTM and the Bi-LSTM with attention models respectively

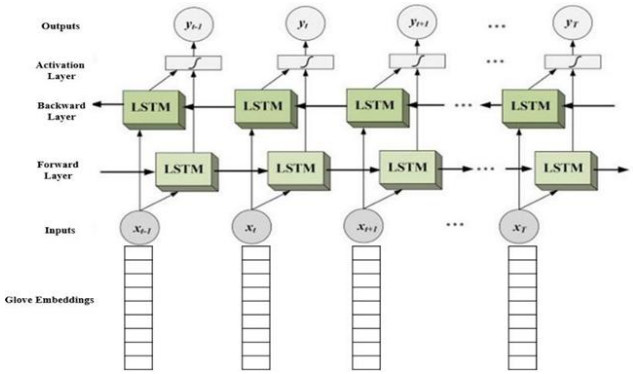


Figure 8: Basic Structure of Bi-LSTM

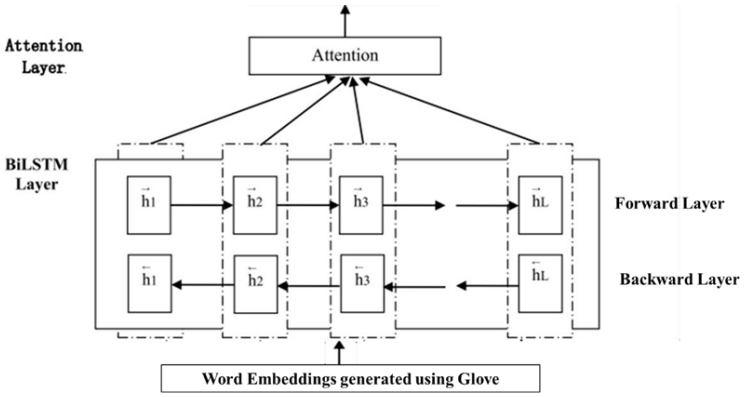


Figure 9: Structure of Bi-LSTM with Attention

During the training process, tuning the hyper parameters of is a crucial step. The proposed Bi-LSTM model's hyper parameters are tuned and selected based on the highest results after test and comparison processes where the outputs of the classifier with different learning paradigms (different optimal hyper parameters and architectures). The output of the Bi-LSTM/Bi-LSTM with attention model will be concatenated with the other text features extracted from other modules and class of news items (identified by its ID) will be decided based on the output of the sigmoid function as shown in figure 9.

3.2.4 CONCATENATION LAYER

In this layer, the emotions feature and the sentiment feature generated from the previous two units will be combined with the output of the Bi-LSTM with Attention layer in the text classification unit. The composite vector will be passed to sigmoid layer as shown in Figure 10, where the NLP input represents the textual features of the news title and the meta input represents the two features of emotions and sentiment.

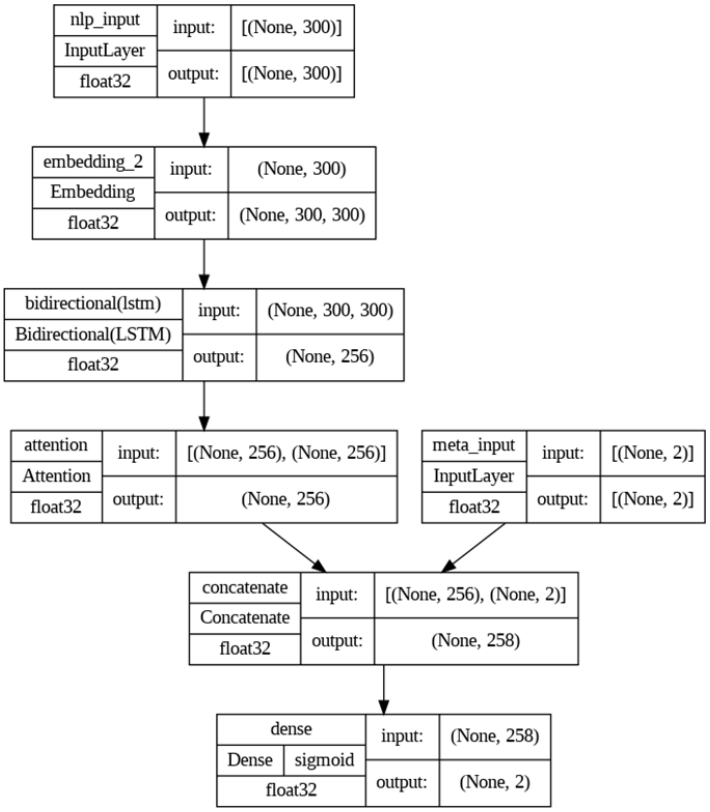


Figure 10: Concatenation Layer

4. EXPERIMENTAL RESULTS:

Evaluation of fake new classification model is an important aspect in developing Fake news detection systems. As discussed in the previous sections, the polarity of a news title along with emotion and sentiment polarity can be any of the two polarities: fake or real. Fake news polarity classification methods are evaluated based on their classification ability in a set of test samples (20% of the total data). Researchers utilize different evaluation measures based on context.

The following Table 3 shows how the accuracy of the model improves over training epochs. Both training and validation accuracy increase, indicating that the model is learning to classify news titles. The validation accuracy being close to the training accuracy suggests that the model is generalizing well to unseen data and not overfitting significantly.

Table 3. Training and Validation accuracy of Bi-LSTM with Attention

Epoch	Training Accuracy (%)	Validation Accuracy (%)
0	~94-95	~94-95
5	~98	~97
10	~99	~98

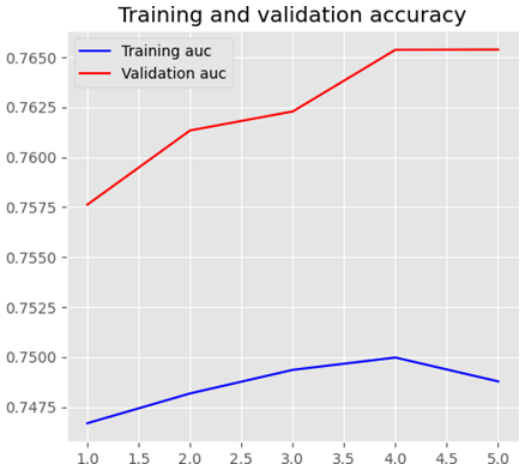


Figure 11. Training and Validation Accuracy of Bi-LSTM with Attention

This Figure 11 plots the training and validation accuracy of the Bi-LSTM with Attention model over training epochs. The x-axis represents the epochs (iterations of training), and the y-axis represents the accuracy (percentage of correctly classified titles). Both training and validation accuracy increase as the model trains, indicating that it's learning to classify news titles more effectively. The validation accuracy is close to the training accuracy, which suggests that the model is generalizing well and not over fitting. Overfitting occurs when a model performs very well on the training data but poorly on unseen data. The model achieves high accuracy (around 98%) by the end of training.

The following Table 4 illustrates the decrease in loss (error) during training. Both training and validation loss decrease, which is desirable as it means the model is becoming more accurate. A small difference between training and validation loss is good, as it indicates the model's ability to generalize. This also provided graphically in Figure

Table 4. Training and Validation Loss of Bi-LSTM with Attention

Epoch	Training Loss	Validation Loss
0	~0.1-0.2	~0.1-0.2
5	~0.05	~0.05
10	~0.02	~0.04

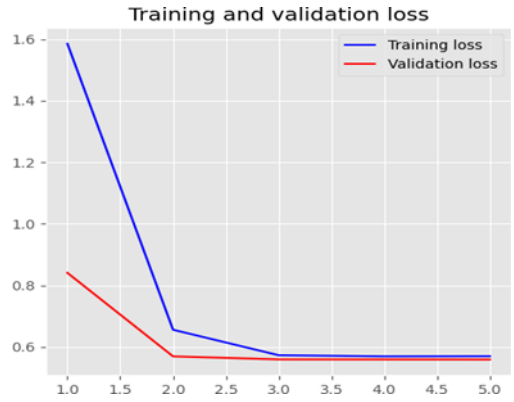


Figure 12. Training and Validation Loss of Bi-LSTM with Attention

This Figure 12 plots the training and validation loss of the Bi-LSTM with Attention model over training epochs. The x-axis represents the epochs, and the y-axis represents the loss (a measure of error). Both training and validation loss decrease as the model trains, indicating that it's making fewer errors in its predictions. The validation loss is relatively close to the training loss, further supporting the idea that the model is generalizing well. The overall loss is low, which corresponds to the high accuracy observed in the previous graph.

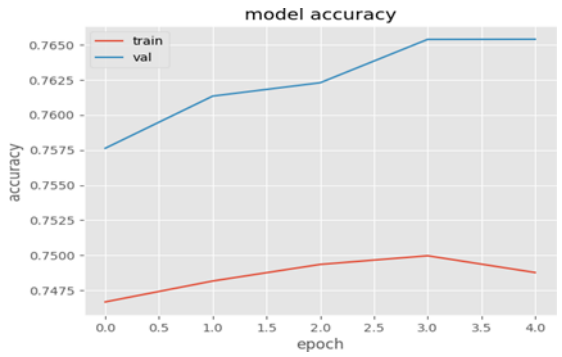


Figure 13. Final Accuracy of Bi-LSTM with Attention

This Figure 13 provides the final accuracy 98 % achieved by the Bi-LSTM with Attention model. This provides a single, clear metric of the model's overall performance

5. COMPARISON BETWEEN THE MODELS

The Table 5 compares final accuracy of the Bi-LSTM model and the Bi-LSTM with Attention model and also represented visually in Figure 14. The Bi-LSTM with Attention model slightly outperforms the base Bi-LSTM model, indicating that the attention mechanism helps improve classification accuracy.

Table 5. Accuracy comparison table

Model	Accuracy (%)
Bi-LSTM	~96
Bi-LSTM with Attention	~98

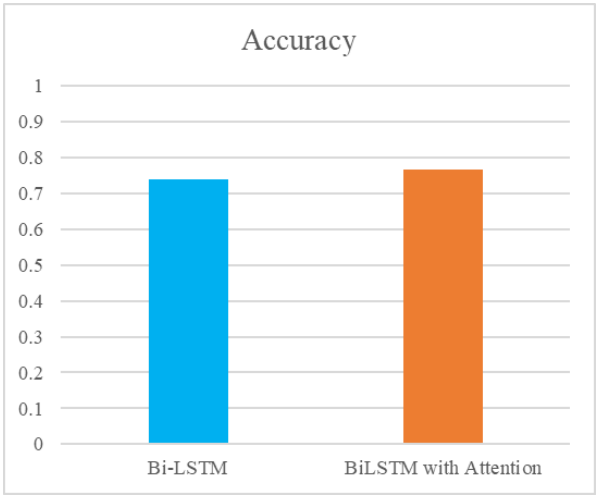


Figure 14. Accuracy (Bi-LSTM vs Bi-LSTM with Attention)

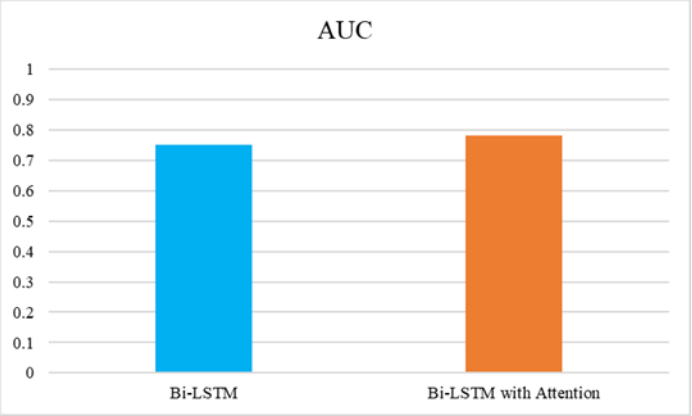


Figure 15. AUC (Bi-LSTM vs Bi-LSTM with Attention)

The Figure 15 compares the Area Under the Curve (AUC) for both models. AUC is a measure of a classifier's ability to distinguish between positive and negative classes. Similar to the accuracy comparison, the Bi-LSTM with Attention model shows a slightly better AUC score than the base Bi-LSTM model

CONCLUSION:

This paper presented a deep learning approach for fake news detection on social media, leveraging a Bi-directional Long Short-Term Memory (Bi-LSTM) network enhanced with an attention mechanism. The proposed model effectively captured crucial features from news titles and user comments, demonstrating a high accuracy of 98% on the Fakeddit dataset, as indicated by the ROC curve analysis. Furthermore, the Bi-LSTM with attention outperformed the standard Bi-LSTM model, highlighting the effectiveness of the attention mechanism in focusing on the most relevant information for classification. The findings suggest that incorporating sentiment and emotional analysis, as reflected in the user comments and news content, provides valuable cues for distinguishing between real and fake news. This research contributes to the growing body of work aimed at mitigating the spread of misinformation on social media platforms.

REFERENCES

1. G. Güler and S. Gündüz, “Deep Learning Based Fake News Detection on Social Media”, IJISS, vol. 12, no. 2, pp. 1–21, 2023, doi: 10.55859/ijiss.1231423.

2. Iftikhar Ahmad & Muhammad Yousaf & Suhail Yousaf & Muhammad Ovais Ahmad, 2020. "Fake News Detection Using Machine Learning Ensemble Methods," Complexity, Hindawi, vol. 2020, pages 1-11, October.

3. Shu, K. C., Wang, S., Lee, D., Liu, H., & Narayanan, V. K. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

5. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and social impacts. ACM Computing Surveys (CSUR), 53(5), 1-36.

6. Ahmad, I., Taboada, M., & Brooke, J. (2020). Detecting opinion spam using ensemble methods. Information Processing & Management, 57(1), 102144.

7. Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

8. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

9. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

10. Liu, W., Gao, S., & Rui, Y. (2015). Content-based spam detection using recurrent neural networks. In Proceedings of the 2015 ACM on Conference on Information and Knowledge Management (CIKM '15) (pp. 1683-1692).