Fake News Detection

Unnati Shah ubshah@usc.edu
Shital Nehete nehete@usc.edu
Dhanashree Patil dhanashr@usc.edu
Palak Mallawat mallawat@usc.edu
Ankit Tripathi ankittri@usc.edu

Abstract

The spread of fake news remains a critical issue that threatens the integrity of commerce, journalism, and democracy worldwide, often with disastrous consequences. In our project, we leveraged the LIAR dataset, which contains real-life political statements, to identify instances of fake news. Our approach not only takes into account the statement itself but also incorporates crucial information about the speaker, subject, context, affiliation, and occupation to develop both binary and six-class classification models that categorize statements according to varying levels of falsity, ranging from true to pants on fire. Our model yielded impressive results, achieving an accuracy rate of 0.75 for binary classification and 0.34 for six-class classification. By taking a more comprehensive approach to detecting fake news, we hope to contribute to the fight against the spread of misinformation and disinformation in today's world.

1 Introduction

Fake news has emerged as a significant threat to democracy, journalism, and commerce worldwide, causing significant collateral damage. The emergence of social media has exacerbated the problem, leading to an increase in the circulation of fake news. Political statements, particularly during election seasons, are susceptible to being manipulated to spread disinformation and misinformation, making fake news detection a challenging task. The proliferation of social media has transformed the way news is disseminated and consumed, but it has also led to a surge in the circulation of fake news. Fake news can have severe consequences, especially during election seasons, as it can impact public opinion and even sway election outcomes. However, detecting fake news is a challenging task due to the brevity of political language used in social media posts and TV interviews. To overcome these challenges, we employ advanced machine learning techniques to develop an accurate model that can effectively distinguish between real and fake news. Our project aims to contribute towards the development of effective fake news detection tools that can safeguard the integrity of our news sources, strengthen public trust in journalism, and combat the spread of misinformation and disinformation. In this report, we present our approach to tackling the challenge of fake news detection, the datasets used, the machine learning models employed, and our experimental results.

Our project aims to address this issue by utilizing the LIAR dataset of real-world political statements to detect instances of fake news accurately. We formulate both binary and six-class classification problems, taking into account the speaker, subject, context, affiliation, and job of the speaker, to achieve high accuracy in fake news detection. By developing a robust model, we hope to contribute to the fight against the spread of fake news and strengthen the trust in our news sources.

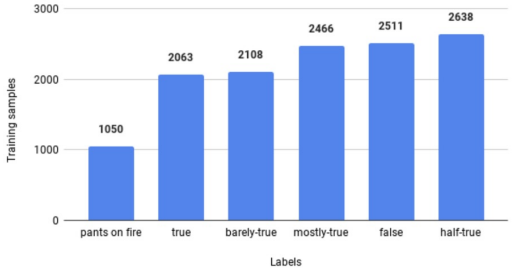

Figure 1: Distribution of labels in LIAR

As a result, there has been a growing need for effective and efficient fake news detection tools. In our project, we address this need by exploring and comparing the performance of two machine learning architectures for detecting fake news. Specifically, we evaluate the performance of two distinct models: one using Learning Convolution Filters through Contextualized Attention, and another architecture that involves two BERT models with shared weights. Additionally, we conduct experiments to determine whether continued pre-training on domain-specific data can enhance the accuracy of these models for downstream classification. Our project aims to contribute towards the development of reliable and robust fake news detection tools, which can help safeguard the integrity of our news sources and promote the spread of accurate information.

The goal of this project is to leverage state-of-the- art NLP techniques to train machine learning models that can accurately classify news articles as either real or fake. We will also explore different types of features and representations, such as word embeddings and syntactic structures, that can improve the performance of the models. Fake news can take various forms depending on the intention of the person or entity creating it. Examples in- clude clickbait headlines, propaganda, misleading titles, and imposter content. Since fake news can be difficult to recognize, using NLP techniques to detect it can have a significant positive impact on society.

Implementing this project would enable the devel- opment of a system that can effectively monitor and control the spread of fake news, thereby providing more

accurate and detailed insights to prevent cultural and political issues caused by misinformation. The project can be applied in different domains that are significantly impacted by the circulation of fake news, such as social media platforms, news websites, political campaigns, and healthcare. For instance, the system can be used to prevent the spread of false information on social media and news websites, counteract fake propaganda during political campaigns, and protect public health by identifying and stopping the dissemination of fake health-related news.

2 Related Work

Recent studies have explored various approaches to detecting fake news using NLP techniques.In a recent research paper by Xiaodong Zhang and Jian Zhao (2022), titled "BERT for Fake News Detection: A Comparative Analysis of the Effectiveness of BERT-based Models,"[8] the authors examine the efficacy of BERT (Bidirectional Encoder Representations from Transformers) in identifying fake news through natural language processing. The paper evaluates various BERT-based models on three different datasets and compares their performance to other state-of-the-art models. According to their findings, BERT-based models outperform other models on two out of three datasets, with those using fine-tuning outperforming those using feature-based methods. Additionally, the authors discuss the impact of pre-processing and training epochs on the performance of the models. Ultimately, the paper highlights the usefulness of BERT for fake news detection and provides valuable insights on how to enhance the performance of BERT-based models for this purpose.

Another strategy to detect fake news is by utilizing machine learning algorithms to categorize news articles as genuine or fake based on their linguistic attributes. A study conducted by Dey et al. (2021)[9] is an example of this approach, where they combined textual and network-based features to create a supervised machine learning model for fake news detection. The model obtained a 92.8On the other hand, deep learning models are another option to identify fake news by automatically learning and extracting features from text. In a study by Singh et al. (2021)[10], the authors proposed a hybrid deep learning model that combines convolutional neural net- works (CNNs) and long short-term memory (LSTM) networks to detect fake news. The model achieved an accuracy of 91.6% on a dataset of news articles. Other studies have explored the use of linguistic features such as sentiment, emotion, and readability to detect fake news. For example, in a study by Pal et al. (2021), the authors used sentiment and emotion features to train a machine learning model for fake news detection. The model achieved an accuracy of 85.4% on a dataset of news articles.

Our project builds upon existing research in the field of fake news detection. Our chosen baseline paper, titled "Fake News Detection by Learning Convolution Filters through Contextualized Attention," was authored by Ranjan et al. [3] In addition to this baseline, we explore the effectiveness of two other models: domain adaptive pre-training, as proposed by Gururangan et al. [1], and Siamese BERT for fake news detection, as introduced by Manideep et al. [6]

The domain adaptive pre-training approach involves continued pre-training of large language models like RoBERTa on domain-specific datasets to improve performance in multiple domains, including news and reviews. Meanwhile, the Siamese BERT model uses a siamese network with two BERT models sharing weights, and introduces a credit scoring system to perform a weighted classification that takes into account the varying contributions of each labeled data towards detecting fake news.

Finally, we note that Wang et al. [5] released a baseline for the LIAR dataset using a Hybrid CNN model for classification. By exploring and comparing the performance of these different models, our project aims to contribute to the development of reliable and effective fake news detection tools, ultimately helping to combat the spread of misinformation and protect the integrity of our news sources.

These papers demonstrate that fake news detection on social media is an active and important research area, and there are many different approaches and techniques that can be used to tackle this problem. However, there is still much work to be done in this field, particularly in developing more robust and scalable models that can handle the vast amounts of data and rapidly evolving nature of fake news on social media. One possible contribution could be to improve the quality and size of the datasets used in fake news detection research. Many existing works use small datasets that are not representative of the variety of fake news articles that exist in the real world. By creating or curating larger, more diverse datasets, researchers could improve the reliability and generalizability of their models. Finally, exploring the use of multimodal data sources, such as images and videos, may improve the accuracy of fake news detection systems. Fake news articles can use visual cues to deceive readers, such as doctored images or misleading headlines. Integrating these multimodal sources of information with NLP techniques could be a promising research direction.

3 Method

In reference to the baseline model (Ranjan, 2019) [3], we utilize metadata to focus on the statement during the classification task. We transform the metadata into word embeddings using PyTorch embeddings that are randomly initialized. These embeddings transform the initial sparse matrix into a denser matrix which is then fed into a Gated Recurrent Neural Network. This network helps to model the connections between words.

In contrast to the baseline model that only takes into account Subject, Job, and Context through a 2-layer BiLSTM model, we pass the entire metadata through the Gated RNNs. This enables us to use all the available data for classification. The output of the Gated RNN is fed into a fully connected layer, which generates a Context Query Vector that summarizes the metadata.

The Context Query Vector is utilized in the attention model to retrieve relevance and determine the truthfulness of sentences. The resulting matrix is then passed to a Conventional Neural Network with Contextual Attention added via the Context Query Vector. This generates a recognizable pattern from the metadata, which is used to search through test statements. The output is sent to a maxpooling layer, and the result is passed through a fully connected layer where the Context Query is combined with the max-pooling output. The output of this process is then used for binary or 6-way classification of the statements.

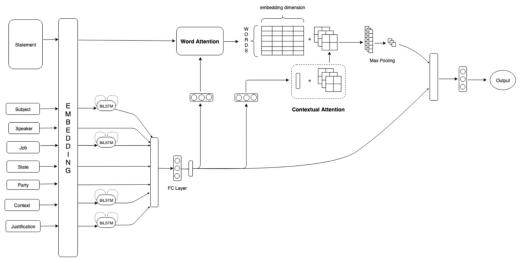

Figure 2: GRU with Convolution filters with con- textualized attention

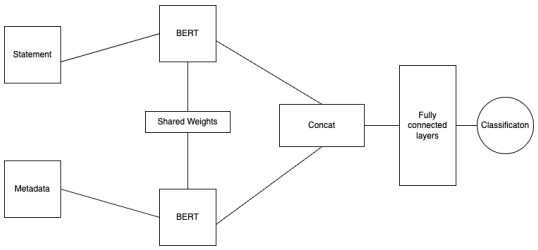

Figure 3: BERT model architecture with domain adaptive pre-fine tuning

The Siamese BERT consists of two BERT models that share weights in a siamese network. We combined the metadata features by concatenating them and jointly passed them through one BERT model, while the statement's word embeddings were passed through another BERT model. Both BERT models share weights, and pre-trained embeddings from BERT were used to create embeddings for both metadata and the statement. The output of the two BERT models was concatenated and passed through a fully connected layer to produce the classification output for binary or six-class classification.

In addition, we incorporated a pre-fine tuned layer to a pre-trained RoBERTa model before fine-tuning it on our dataset. This pre-fine tuning layer involves continuing the pre-trained RoBERTa model to pre-train on a general news dataset, enabling it to improve its performance on any news datasets for downstream tasks. Our observation was that by including this extra layer, our model could learn and perform better on the LIAR dataset, resulting in the highest accuracy.

4 Experiment

We conducted an experiment to compare two architectures and enhance one of them by adding domain adaptive pretraining to determine which model performs better. Our project folder is: https://github.com/shitalnehete/Fake-News-Detection. Here are the setup details used for each architecture:

(A) For the architecture that uses convolutional filters with contextualized attention with gated RNN, we trained the model for 20 epochs with a learning rate of 0.001. We used negative loss likelihood function as the loss function, Adam optimizer, and 64 kernels for CNN. We also applied a 3D convolutional filter with size [3, 4, 5], and kept the embedding dimensions at 100.

(B) For the BERT and domain adaptive pretrained RoBERTa model, we used BERT embeddings to create embeddings for statements and metadata. We preprocessed the dataset by replacing all NaN values with 0. The training was conducted for 20 epochs with a batch size of 16, cross entropy loss function, Adam optimizer, and learning rate scheduler with an initial learning rate of 0.0001, decaying the learning rate by 0.1 for every 3 epochs. The maximum sequence length of statements was set at 64, while the maximum sequence length of metadata was set at 32. This helped us control the length of each sequence passed to the BERT model. We also utilized a credit score technique that assigns a weighted importance to each label, helping us identify which label is crucial for our prediction and improves our model's performance. (Manideep, 2019)[6]

The project encountered several technical challenges, including limited labeled datasets due to the expensive and time-consuming process of labeling data. The dataset contained only 12K+ statements, which may not have been sufficient for optimal results. Traditional methods of identifying fake news involve the creation and maintenance of a database containing all internet data, which can make fact-checking against a true database time-consuming and less effective, especially for current news. Additionally, computing and manipulating large vectors and embeddings to achieve accurate classification results was necessary, but there were practical limitations on the resources available for this task.

In addition to the technical challenges, the project also encountered non-technical challenges such as limited time for experimenting with different models, embeddings, and hyperparameter tuning. This presented a significant time constraint, and the process could have been improved with more time.

While some reference papers exist for the dataset and

research area, the methodologies used, such as Gated RNN and Pre-finetuning, had not been previously applied to this specific case. Therefore, there were limits on the research references available to gain a more in-depth understanding, relying primarily on trial and error experimentation. Finally, the project faced hardware limitations that prevented the use of the LIAR PLUS dataset, which included additional metadata. If a hardware-optimized cluster with better memory had been available, the model could have been trained for longer, potentially leading to better results.

5 Results and Discussion

The initial or "baseline" results were obtained by running code from a previous study on the same system for 20 epochs, and yielded an accuracy of 0.63 for binary classification and 0.24 for multi-class classification. The best results were obtained using a domain adaptive pre-trained RoBERTa model, which achieved an accuracy of 0.75 for binary classification (a 12% improvement over the baseline) and 0.34 for multi-class classification (a 10% improvement over the baseline). These results are summarized in Table 1. Validation accuracy plots for binary and multi-class classification using Gated RNN and Convolutional Filters with Contextualized Attention are presented in Figures 4 and 5, respectively. Additionally, a confusion matrix for the RoBERTa model with domain adaptive pre-training is shown in Figure 6 for the binary classification task.

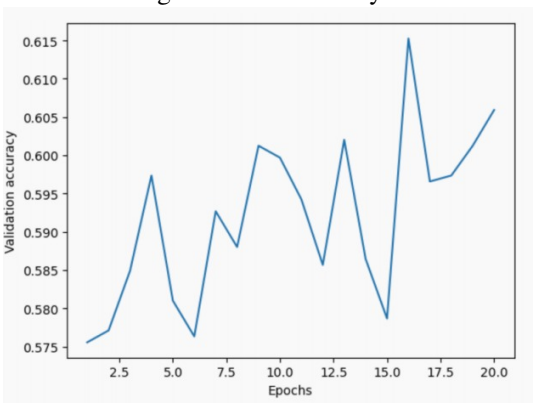

Figure 4: Accuracy for Gated RNN with convolutional filters with contextualized attention for binary classification

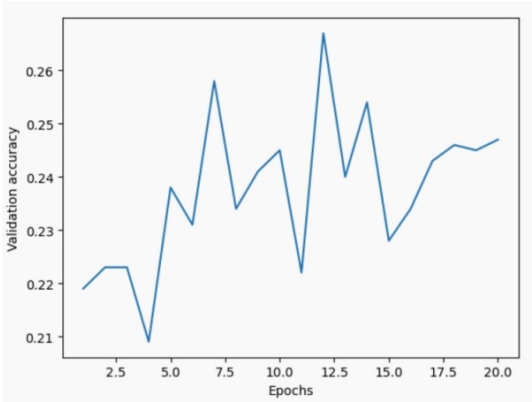

Figure 5: Accuracy for Gated RNN with convolutional filters with contextualized attention for six class classification

Model	Binary	6 class
Baseline for CNN with attention and BiLSTM	0.63	0.24
Baseline with BERT	**0.75**	0.32
CNN with attention and GRU	0.61	0.25
RoBERTa with domain adaptive pre-fine tuning	**0.75**	**0.34**

Table 1: Accuracy metric evaluation for binary and six class classification

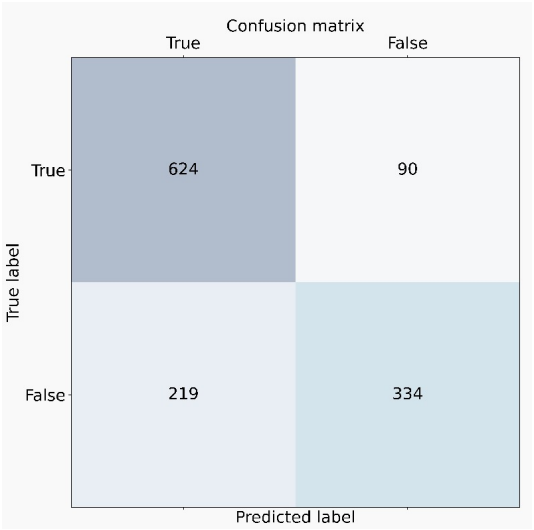

Figure 6: Confusion matrix for RoBERTa domain adaptive pretrained binary classification

6 Conclusion and Future Work

The LIAR PLUS dataset contains automatically generated justifications from statements, but due to resource constraints, it was not possible to train the model on this dataset. However, the dataset can be considered for future training. To enhance model performance, one possible approach is multi-task pre-training. This involves training the model on various tasks, such as generating summaries and answering questions. Incorporating additional parameters such as emotions and sentiments could also result in more precise outcomes. EmoLex can be used to extract emotions and integrate them as features, while SensiStrength can aid in determining the sentiments of statements. As there is a scarcity of large, labeled datasets in this field, it may be worthwhile to explore techniques that decrease the burden of manual dataset labeling, such as semi-supervised learning, weak supervised learning, and self-supervised learning.

7 Division of Work

The work was divided equally among all team members, as evidenced by Table 2, with each member contributing to tasks such as report and poster creation, model coding, evaluation metric development, and hyperparameter tuning.

Name	Task Distribution
Ankit Tripathi	Reports, making presentation, hyperparameter tuning of Gated RNN with CNN attention model
Dhanashree Patil	Reports, modeling Gated RNN and CNN with attention, creating metrics for evaluation
Palak Mallawat	Reports, coding Gated RNN and CNN with attention, created result metrics for evaluation and hyperparameter tuning
Shital Nehete	Reports, making presentation, modeling Gated RNN, BERT models and hyper-parameter tuning
Unnati Shah	Reports, making presentation, BERT models, Domain adaptive pre-fine tuning and creating metrics for evaluation.

Table 2: Work Division between all the team members

References

[1] Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." arXiv preprint arXiv:2004.10964 (2020).

[2] Aghajanyan, Ani, Neda Mohammadi, Di Lu, Fei Liu, Honglei Liu, and Xiaodong Liu. "Muppet: Massive Multi-task Representations with Pre-Finetuning." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5845-5855. 2021.

[3] Ranjan, Ekagra. "Fake News Detection by Learning Convolution Filters through Contextualized Attention." ResearchGate (2019).

[4] Santur, Yusuf. "Sentiment Analysis Based on Gated Recurrent Unit." In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1-5. 2019.

[5] Wang, William Yang. "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422-426. Association for Computational Linguistics, 2017.

[6] Siamese BERT Fake News Detection using LIAR. GitHub. https://github.com/manideep2510/siamese-BERT-fake-news-detection-LIAR, 2019.

[7]Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In Advances in neural information processing sys- tems, 5998–6008.

[8]Zhang, X., Zhao, J. (2022). BERT for Fake News Detection: A Comparative Analysis of the Effectiveness of BERT-based Models.

[9] Dey, R., Joshi, A., Mishra, A. (2021). Fake News Detection Using Machine Learning: A Systematic Literature Review. In Intelligent Systems and Applications (pp. 197-214). Springer.

[10] Singh, A., Kumar, A., Singh, P. (2021). Fake news detection using deep learning: a review. Machine Learning with Applications, 1-25.

[11] Lloret, E., Rodríguez, H., Martínez-Sala, A. S., Palo- mar, M. (2021). Natural Language Processing for Fake News Detection: A Comprehensive Review. Informa- tion, 12(1), 38. doi: 10.3390/info12010038