CS445 Project Report

Group No: 32

Alper Cimşit 28923
Kağan Kağanoğlu 29482
Mehmet Berke Bakan 28940

1. Introduction

The increase in misinformation can be seen in the nearly 10,000% rise in the number of academic articles including the term "fake news" between 2015 and 2018 (Figueira, Álvaro, & Guimarães, 2019), which can be seen in Figure 1. Due to the increasing number of fake news on the Internet, fake news detection has become an important task in the field of NLP. In the era of postmodernism where the truth becomes ambiguous, fake news detection can contribute to differentiate fake news from true ones. The purpose of our task is to develop a model that determines the “truth rates” of given information using machine learning algorithms, which can help to identify misinformation.

After conducting a literature review and choosing the dataset we use and the paper we will implement, we first applied data analysis and preprocessing to our data. After getting familiar with the dataset we use and making the data ready for the training, we used a hybrid CNN model combining text and metadata processing, as proposed by Wang (2017). Text data is processed with pre-trained embeddings and CNNs, while metadata is encoded separately. The outputs are merged for classification, optimizing accuracy by leveraging both data types. This approach is benchmarked against simpler models like support vector machines, linear regression and naive bayes, and advanced ones like BiLSTMs.

Among all the models compared, BiLSTM with pre-trained embeddings gave better results than other simple models since it was able to handle class imbalances and subtle nuances in statements efficiently; however, accuracy of CNN was the highest among all. But CNN still struggled and was confused between labels due to data imbalance. In general, these results show the complexity of the truth prediction task and the importance of robust feature representation.

2. Methodology

2.1 Dataset

We used the LIAR dataset, introduced by Wang (2017) in the paper “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection”, which is also one of the papers we chose to implement alongside “Where is your evidence: Improving fact-checking by justification modeling” by Alhindi(2018). The dataset is composed of political statements collected from PolitiFact.com. We selected this dataset because it is widely used in the literature, providing a strong benchmark for fake news detection. Many papers have built different approaches on it, such as the BiDirectional Long Short-Term Memory (BiLSTM) approach (Alhindi, Petridis, & Muresan, 2018). The dataset includes six target categories for truthfulness, with short statements, making it a challenging task for classification. We chose the “Liar liar pants on fire” paper since it presents a clear, understandable, and

detailed approach. Moreover, since the LIAR dataset is created for this research paper, the paper includes clear explanations about the dataset too. Also, we chose the “Where is your evidence” paper since it included a clear implementation for BiLSTM.

Figure 2 displays a sample data point from the LIAR dataset. For our task, the relevant fields include “label,” “statement,” and “speaker.” The “label” is the target categorical column indicating the truthfulness of the statement, with values ranging from 0 to 5, which is also represented as pants-fire, false, barely-true, half-true, mostly-true and true respectively, where 0 represents a blatant lie and 5 represents the full truth. “statement” and “speaker” are the input fields we will use to train our model.

The LIAR dataset is pre-split into three parts: training (10.269 examples), validation (1.284 examples), and test (1.283 examples). We will maintain this split to ensure the proper benchmarking and compatibility with other studies.

2.2 Data Analysis and Preprocessing

After deciding on the dataset and paper we will use, we continued with the data analysis and preprocessing part. We utilized pandas, seaborn, matplotlib and wordcloud libraries for data analysis and visualization; and, nltk and regular expression libraries for preprocessing. After loading the data, we explored its features and conducted an analysis on it, plotting the label distribution, party distribution, top speakers, the words used most frequently, label distribution for top speakers, and the correlation between tokens and labels. We then applied preprocessing to the data, filling the empty values, converting it to lowercase, removing punctuation, numbers, and stopwords, tokenizing and finally lemmatizing. Since the dataset was already splitted, we applied preprocessing to each train, validation, and test data separately.

2.3 BiLSTM Implementation

For our approach, we first implemented a BiLSTM (Bidirectional Long Short-Term Memory) model from the paper "Where is your Evidence: Improving Fact-checking by Justification Modeling." We used pre-trained GloVe embeddings (100 dimensions) to represent words, as the paper does. The architecture consists of an embedding layer initialized with GloVe vectors, followed by a BiLSTM layer with 32 units, which captures contextual information from both past and future word sequences. The output from the BiLSTM layer is passed through a softmax layer to classify statements into six categories. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, and trained for 10 epochs with a batch size of 32. This architecture was selected due to its proven effectiveness in handling sequential text data and capturing long-term dependencies.

In addition to the BiLSTM model, we also trained alternative models for comparison, as done in the paper. These included a Logistic Regression (LR) model and a Support Vector Machine (SVM) model, both using GloVe embeddings as features, along with a Naive Bayes (NB) classifier, which was trained using a Bag of Words (BoW) representation of the text. These models’ performance was compared in terms of accuracy, F1 Score, macro-precision and recall.

2.4 Hybrid CNN Approach

Our hybrid CNN implementation integrates text data and metadata through a deep learning model designed for multi-modal inputs as we saw this approach from the article “Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection”. However, the text data is represented as BERT embeddings using the last hidden state of the bert-base-uncased model. Even though the article uses word2vec for text data. Using a bert model for embeddings increased the performance of our model. The metadata, consisting of numerical features, are preprocessed to enhance model performance. Metadata features are standardized using Standard Scaler. BERT embeddings are also normalized using TensorFlow's normalization utility to standardize their magnitudes across all samples.

Our model architecture processes text data through convolutional neural networks with varying kernel sizes (3, 5, and 7) to extract multi-scale features. Each layer is followed by batch normalization to stabilize learning. Outputs from these layers are concatenated to combine features extracted at different scales, and a spatial dropout is applied to prevent overfitting. A residual connection is used by reducing the concatenated features’ dimensionality using a 1D convolution and adding the result back to the original input. Global max pooling and average pooling layers are applied to the residual output, generating a fixed-size representation for text features.

Metadata is processed through a sequence of dense layers, each followed by batch normalization and ReLU activation. These layers map the metadata into a high-dimensional space, aligning its representation with the text features. The two modalities are then integrated using an attention mechanism. The text and metadata features are reshaped, and an attention layer computes their alignment, producing a combined representation that captures cross-modal interactions.

The outputs from the attention mechanism, text features, and metadata representations are concatenated to form a unified feature vector. This vector is passed through fully connected layers with L2 regularization, batch normalization, and dropout to prevent overfitting and ensure generalization. Finally, a softmax output layer is used to predict class probabilities for the given inputs.

3. Results

3.1 Preprocessing and Data Analysis Results

Firstly, the results of the data analysis and preprocessing is as follows: The severity of the fake news issue can be seen in the analysis of the dataset we use for our task, LIAR. Here in Figure 3, counts for each truth label reveal that numbers of true, mostly true, half true, barely true, and false news are quite similar to each other, leaving pants fire news behind, possibly because it is not politically beneficial to make up major lies. Also in Figure 4, it could be seen the distribution of truth labels for the top four speakers in the dataset with the highest number of statements. It reveals how frequently politicians can use partially or fully untrue information. The numbers do not seem optimistic, especially for Donald Trump, having the highest ratio for false and pants-fire labels among the four speakers. Moreover in Figure 5, the correlations between the words used in statements

the most frequently and truth labels are given. The words such as “say” or “Obama” which imply the statement mentions another person or statement, have a low truth score. The statements including the word “percent” on the other hand have a higher tendency to be true possibly because it might include statistical or numeric information from a study. Moreover, in order to have a more detailed understanding of our dataset, we analyzed the distribution of parties, the speakers with the highest number of statements, and the most frequently used words in the dataset, which can be seen in Figure 6, Figure 7, and Figure 8 respectively. And finally for this part, the resulting tokens after the preprocessing can be seen in Figure 9.

3.2 BiLSTM Results

The BiLSTM model, trained with pre-trained GloVe embeddings, outperformed the Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes (NB) models across key evaluation metrics, achieving the highest accuracy and macro F1-score. The precision-recall curve (Figure 10) highlights its robustness in managing class imbalances, while the confusion matrix (Figure 11) indicates consistent performance across most classes, with relatively lower accuracy in predicting the mostly-true and barely-true labels. Although LR and SVM, both utilizing GloVe embeddings, lagged behind, the NB model, trained with a Bag-of-Words representation, delivered comparable performance (Figure 12). These results are consistent with findings from the reference paper, reinforcing the effectiveness of sequential modeling with pre-trained embeddings for this classification task.

3.3 Hybrid CNN Results

The performance evaluation of the CNN model indicates several insights based on the training process, validation metrics, and confusion matrix. The training history demonstrates slow but steady improvements in validation accuracy over 40 epochs, with the model eventually reaching a best validation accuracy of 26.85% in the 33. epoch. The loss values show a similar trend, with a consistent decrease over epochs. However, both training and validation accuracies remain relatively low because of the difficulty of the task.

The confusion matrix (Figure 13) reveals that the model struggles to accurately classify all six categories, with considerable overlap in predictions. For example, classes such as "barely-true" and "false" are often confused with one another, indicating that the feature representations learned by the model may not be sufficiently discriminative.

Precision, recall, and F1-scores from the classification report (Figure 14) further highlight this issue. Precision values are low across most classes, indicating a high rate of false positives, while recall values vary significantly, suggesting the model's inconsistency in identifying true positives across different categories.

One notable challenge in the results is the model's tendency to underperform on classes with fewer samples, such as "pants-fire" and "true," likely due to class imbalance in the dataset. The macro average F1-score of 0.25 reflects the model's limited capability to generalize across all categories, and the weighted average F1-score of 0.25 aligns with the low overall accuracy of 26.41% on the test set.

4. Discussion

For our fake news detection system, we selected the LIAR dataset due to its widespread use, clear structure, and the challenges it presents with closely related classes and class imbalances, which influenced the model performance.

We chose two approaches: a BiLSTM model using pre-trained GloVe embeddings and a hybrid CNN model incorporating BERT embeddings for text and metadata processing. The BiLSTM model effectively captured sequential dependencies, with an advantage in handling long-term dependencies, while the hybrid CNN used multi-modal data. Both models produced results comparable to their respective reference papers but struggled with distinguishing closely related classes and showed low accuracy due to the dataset's inherent complexity and class imbalances. These limitations lowered the overall performance, especially in predicting minority classes.

With more time and resources, we could improve the system by addressing these class imbalances through techniques like data augmentation or oversampling. Additionally, more fine-tuning and the use of deeper models could further improve the results.

5. Conclusion

This study addresses a critical problem in fake news detection. It applies the machine learning model on the LIAR dataset, which is a challenging benchmark database due to class imbalance and closeness of accuracy categories. In the study, a BILSTM model using pre-trained GloVe and a hybrid CNN model combining BERT embeddings with metadata are examined.

The BiLSTM model effectively captured the contextual relationships within the text. In this process, it outperformed the baseline models such as Logistic Regression, SVM, and Naive Bayes. However, it struggled to distinguish between similar classes such as "barely-true" and "false". The hybrid CNN model offers an innovative approach by combining multi-scale text feature extraction with metadata-driven attention mechanisms. However, the hybrid CNN model faced challenges such as class imbalance and low inter-class variability. This resulted in moderate accuracy and F1 scores.

A major limitation for both models was the inherent class imbalance in the dataset, which particularly affected underrepresented classes such as "pants-fire" and "true". As a result of this imbalance, the models were not able to achieve high precision and recall rates in all categories. Furthermore, the complexity of integrating multimodal data in the hybrid CNN model led to extensive tuning requirements in real-world applications.

The project could include addressing the class imbalance with techniques such as oversampling or synthetic data generation, and using more advanced embeddings such as fine-tuned BERT models to provide more robust performance. Ensemble methods that combine the strengths of both BiLSTM and CNN models could be used.

In conclusion, while our models demonstrate the potential of machine learning for fake news detection, the challenges posed by the LIAR dataset highlight the need for new research. These findings contribute to the development of reliable and scalable systems for the detection of misinformation.

6. Appendix

Figure 1: Number of hits per year in Google Scholar for the term "fake news".

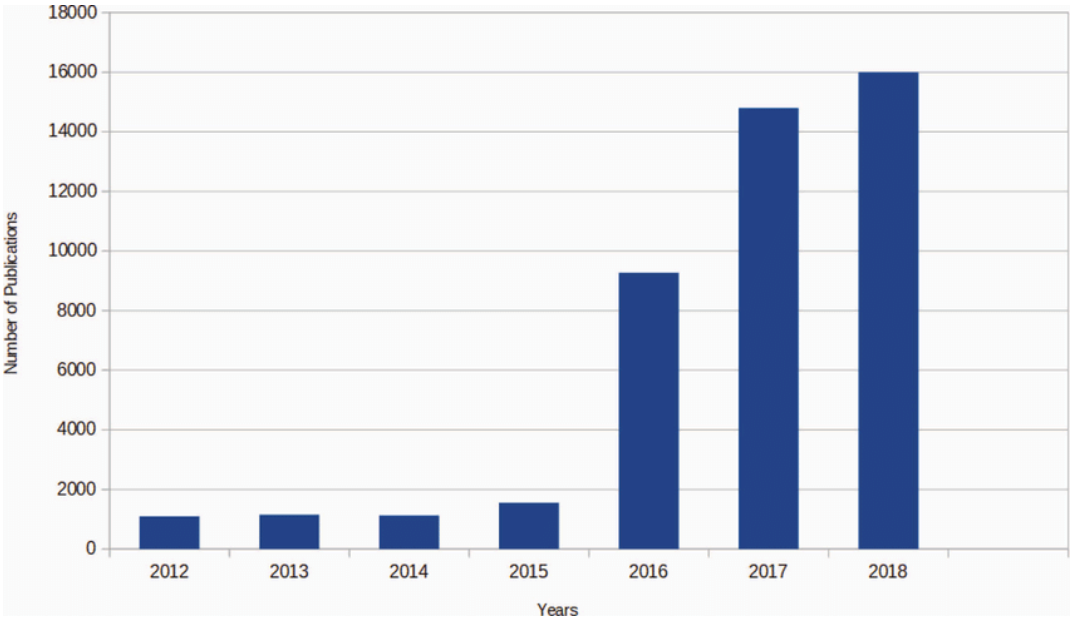

Figure 2: Example Data Point from LIAR dataset


```
{'id': '324.json',
 'label': 2,
 'statement': 'Hillary Clinton agrees with John McCain "by voting to give George Bush the benefit of the doubt on Iran."',
 'subject': 'foreign-policy',
 'speaker': 'barack-obama',
 'job_title': 'President',
 'state_info': 'Illinois',
 'party_affiliation': 'democrat',
 'barely_true_counts': 70.0,
 'false_counts': 71.0,
 'half_true_counts': 160.0,
 'mostly_true_counts': 163.0,
 'pants_on_fire_counts': 9.0,
 'context': 'Denver'}
```


Figure 3: Distribution of truth labels

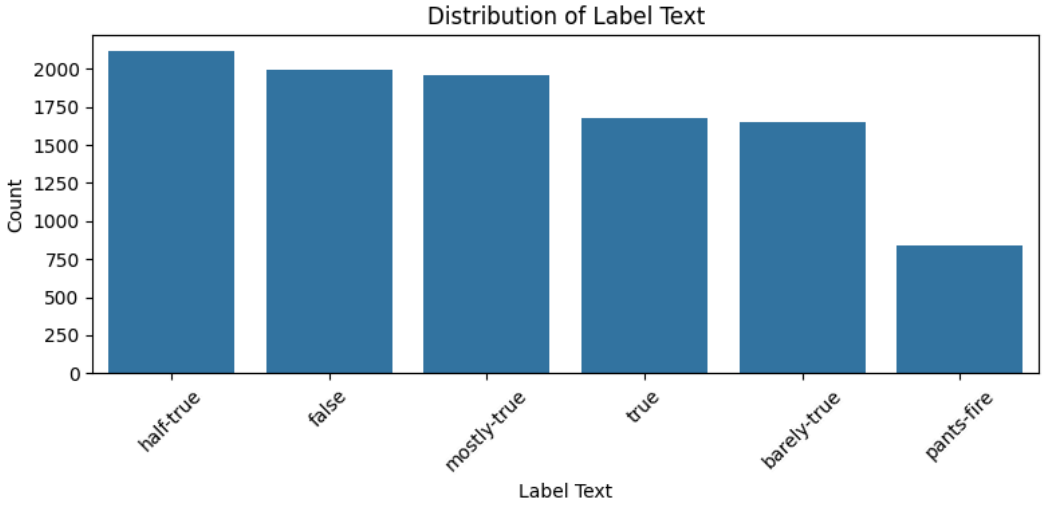

Figure 4: Label distribution for the top 4 speakers with the highest number of statements

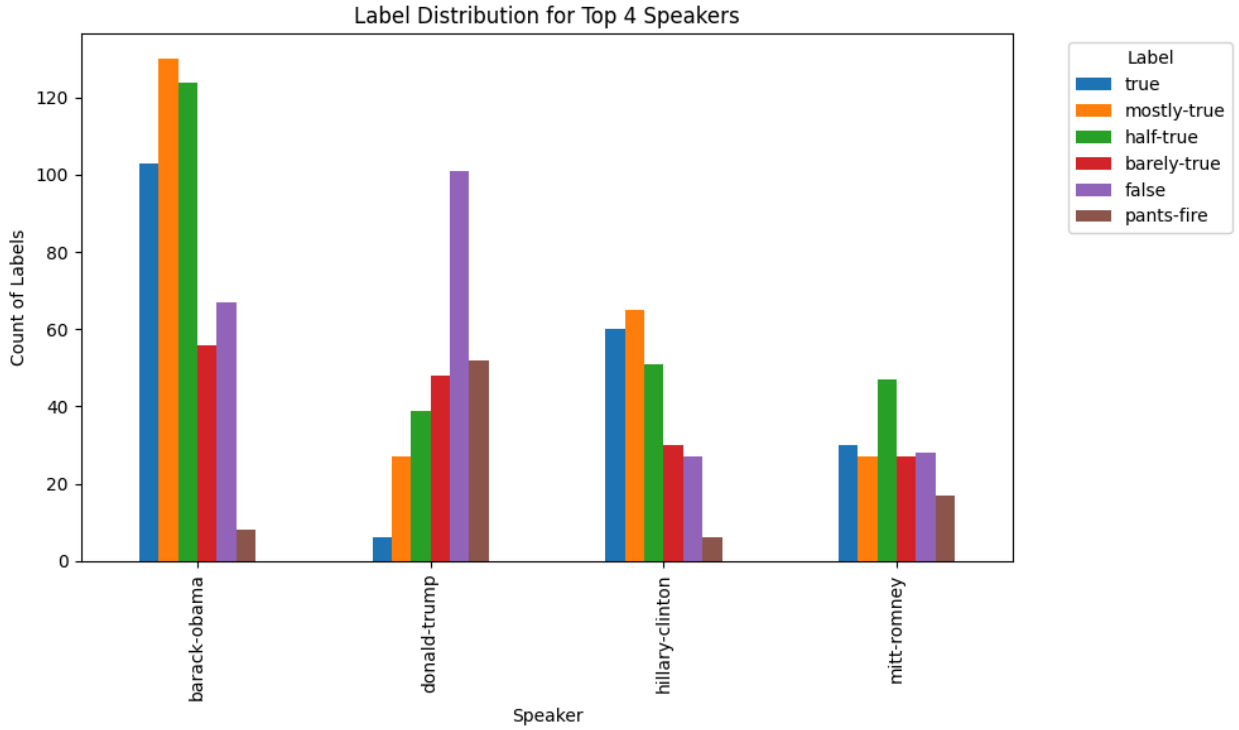

Figure 5: Correlation between labels and words used the most frequently

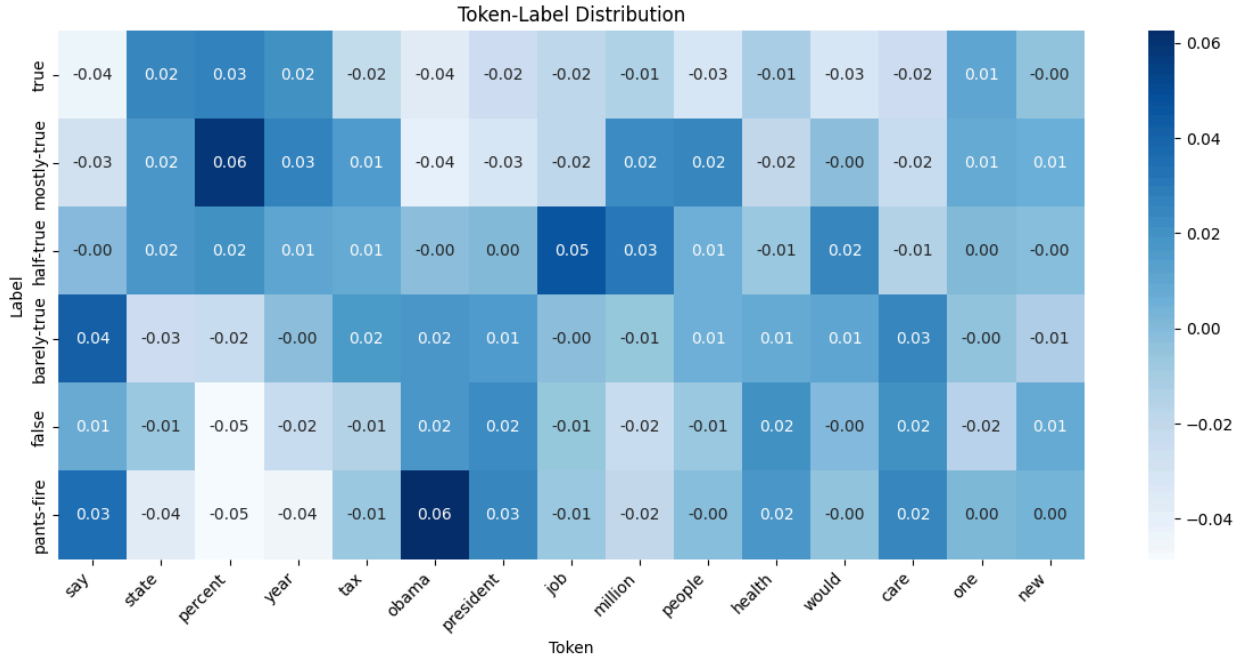

Figure 6: Distribution of parties

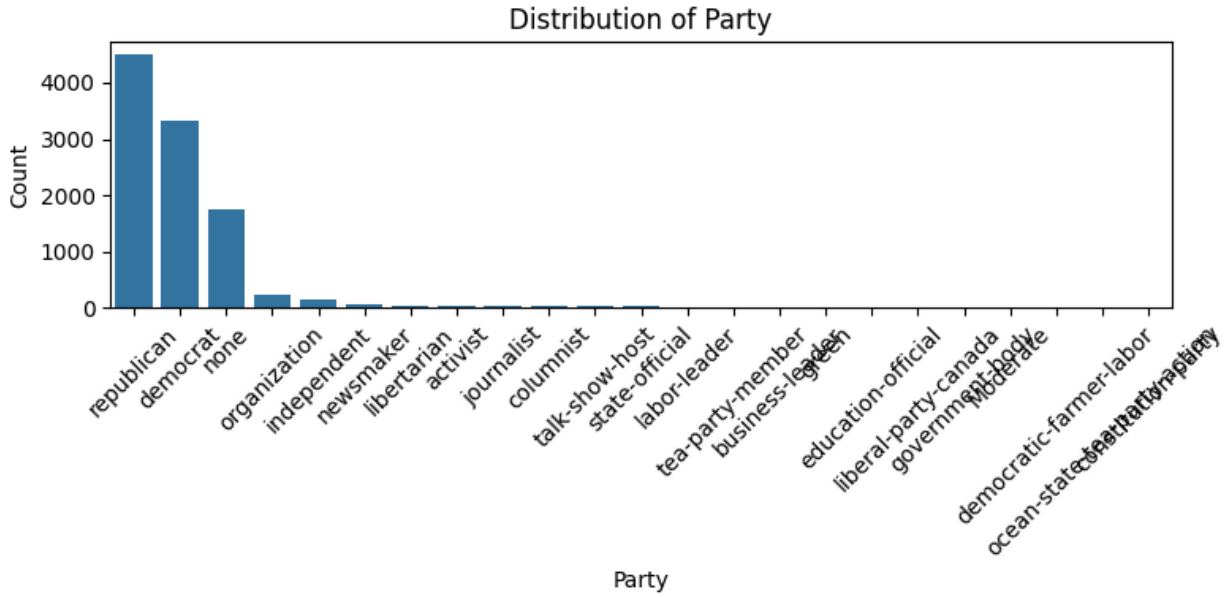

Figure 7: Top speakers by frequency

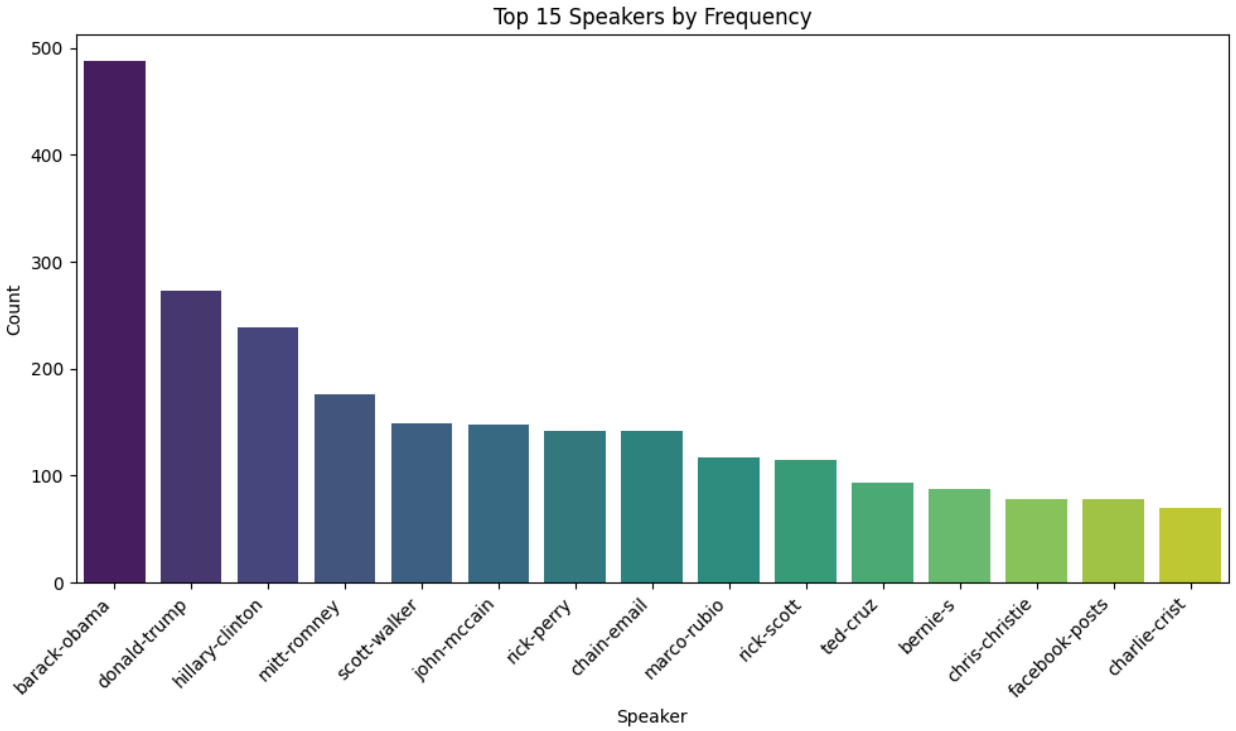

Figure 8: The most frequently used words

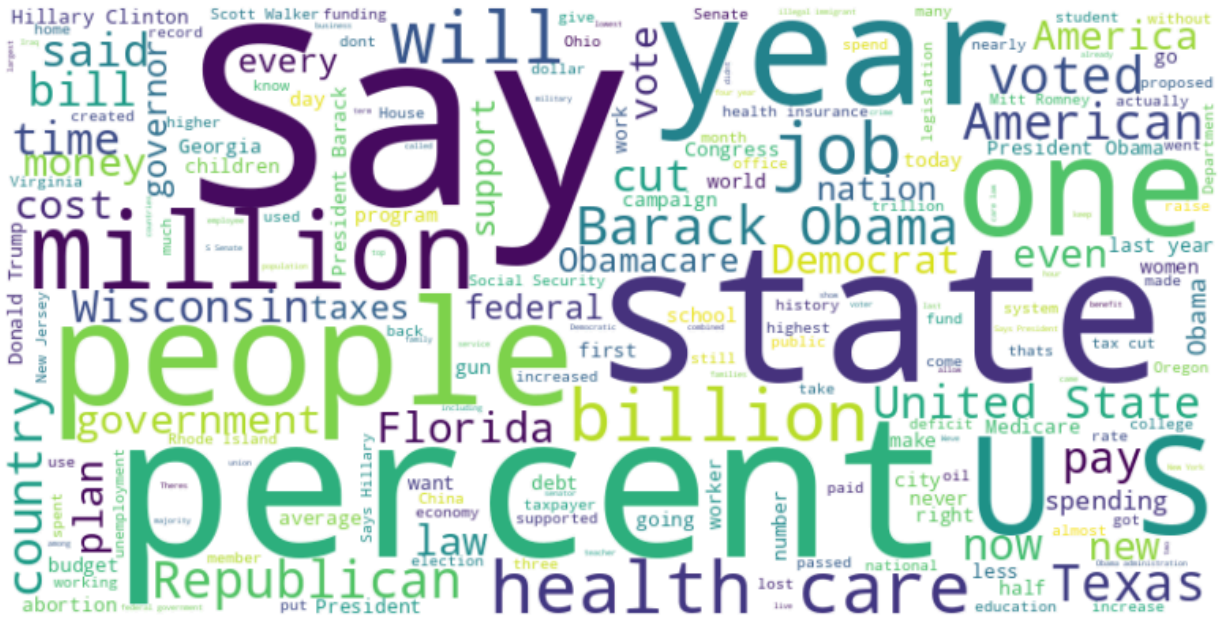

Figure 9: Preprocessed data

	processed_statement
0	[say, annies, list, political, group, support,...
1	[decline, coal, start, started, natural, gas, ...
2	[hillary, clinton, agrees, john, mccain, votin...
3	[health, care, reform, legislation, likely, ma...
4	[economic, turnaround, started, end, term]
...	...
10235	[larger, number, shark, attack, florida, case,...
10236	[democrat, become, party, atlanta, metro, area...
10237	[say, alternative, social, security, operates,...
10238	[lifting, cuban, embargo, allowing, travel, cuba]
10239	[department, veteran, affair, manual, telling,...

10238 rows × 1 columns

Figure 10: Precision-Recall Curve for each class

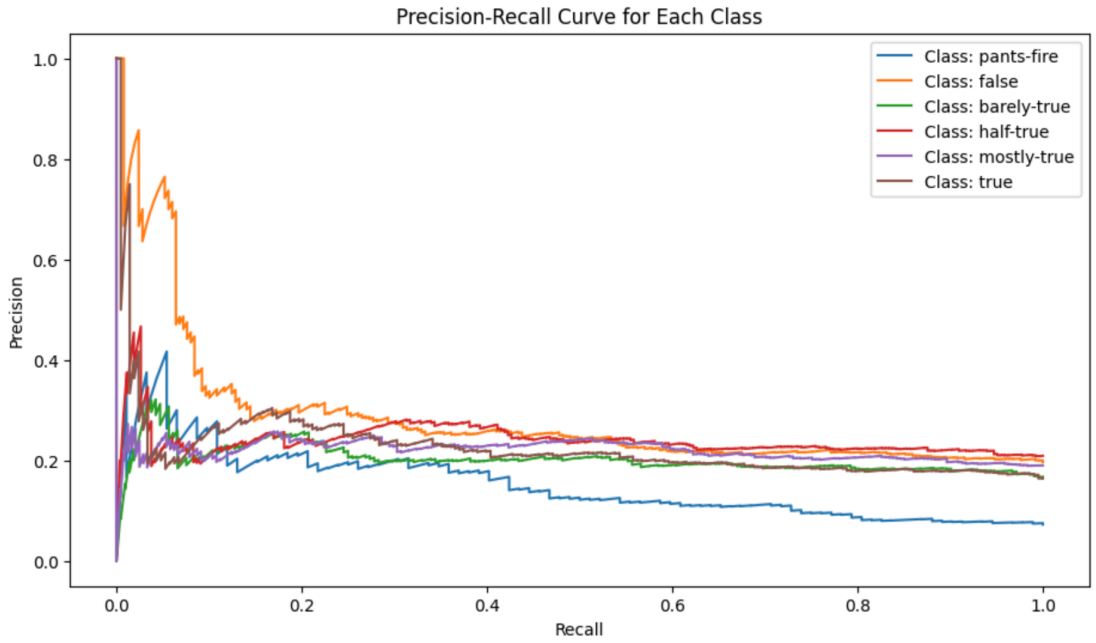

Figure 11: Confusion matrix for BiLSTM

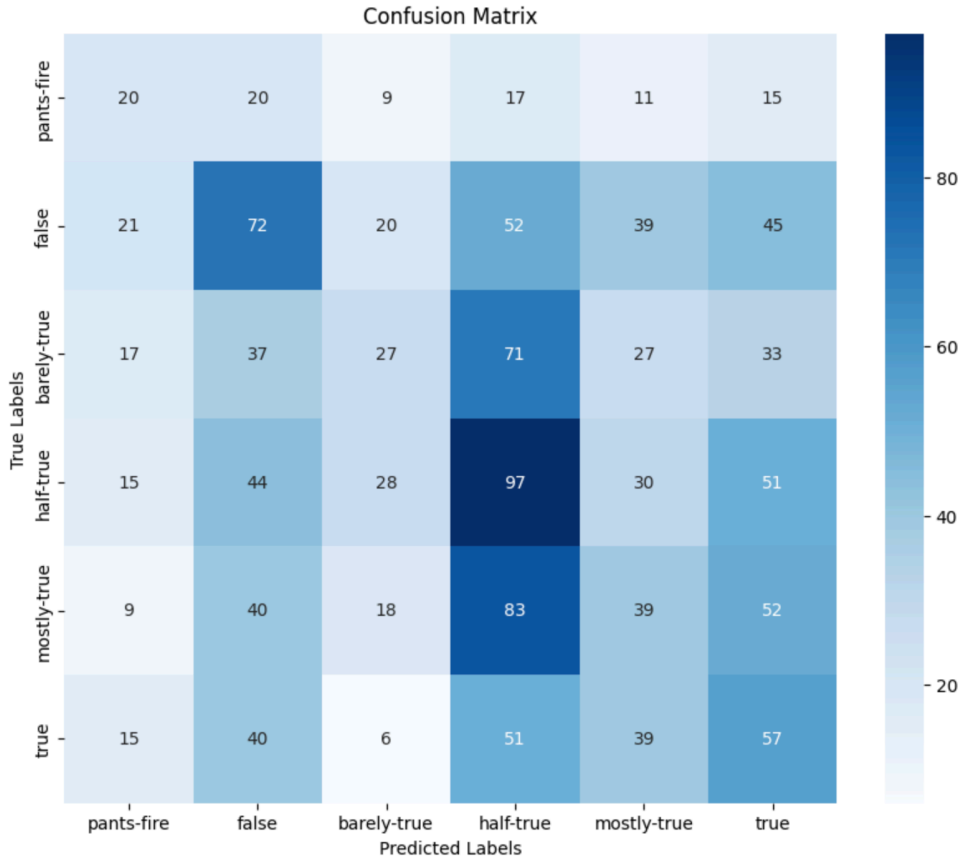

Figure 12: Comparing Results of LR, SVM, Naive Bayes and BiLSTM

MODEL (NO METADATA)	ACCURACY ON TEST	F1 SCORE ON TEST
Logistic Regression	0.192	0.110
SVM	0.202	0.144
Naive Bayes with BoW	0.240	0.230
BiLSTM	0.246	0.234

Figure 13: Confusion Matrix for Hybrid CNN

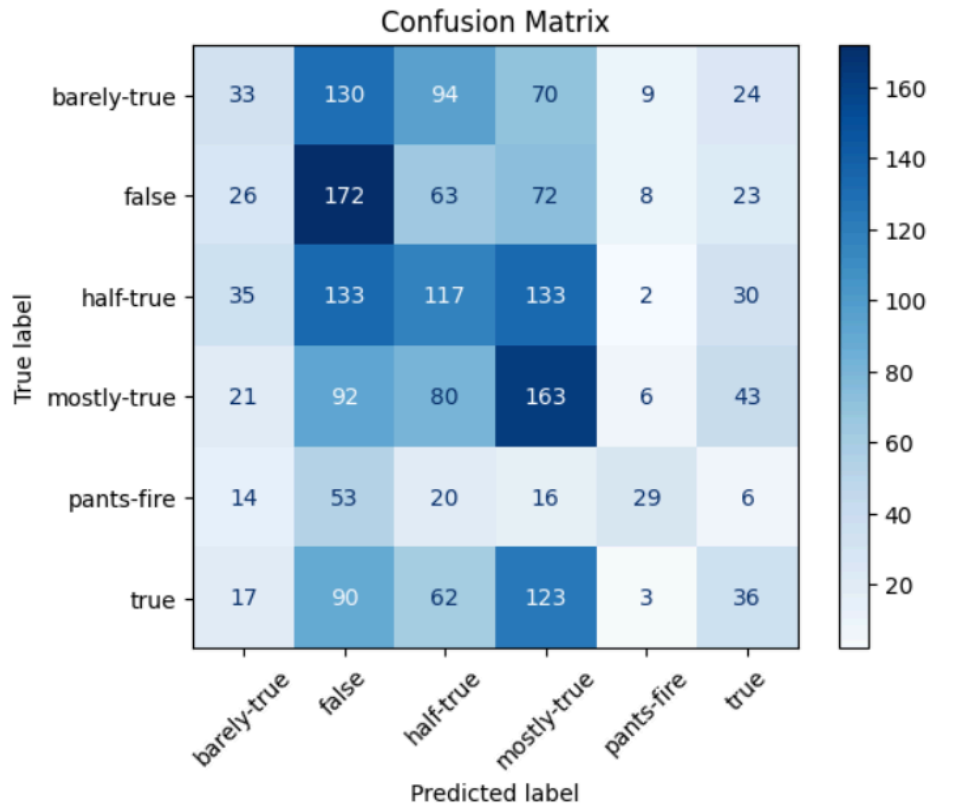

Figure 14: Classification Report for Cnn

Classification Report:

	precision	recall	f1-score	support
0	0.23	0.09	0.13	360
1	0.26	0.47	0.33	364
2	0.27	0.26	0.26	450
3	0.28	0.40	0.33	405
4	0.51	0.21	0.30	138
5	0.22	0.11	0.15	331
accuracy			0.27	2048
macro avg	0.29	0.26	0.25	2048
weighted avg	0.27	0.27	0.25	2048

7. References

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modeling. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER). https://doi.org/10.18653/v1/w18-5513

Figueira, Á., Guimarães, N., Torgo, L. (2019). A Brief Overview on the Strategies to Fight Back the Spread of False Information. Journal of Web Engineering (JWE). 18. 319-352. 10.13052/jwe1540-9589.18463.

Wang, W. Y. (2017). “Liar, Liar Pants On Fire”: A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). https://doi.org/10.18653/v1/p17-2067