Fake News Detection using LIAR

Amulya Sethurao
University of Southern California
sethurao@usc.edu

Harshith Acharya
University of Southern California
hacharya@usc.edu

Mohammed Zubair Khan
University of Southern California
mkhan217@usc.edu

Purva Makarand Kulkarni
University of Southern California
purvamak@usc.edu

Vinya Somayajula
University of Southern California
vsomayaj@usc.edu

Abstract

Fake news is considered to be one of the greatest threats to commerce, journalism and democracy all over the world and with huge collateral damages. In our project, we use the LIAR dataset of real world political statements for our task of detecting fake news. Our project not only focuses on the statement but also on the speaker, subject, context, affiliation and job of the speaker to formulate both binary and 6 class classification problem with labels as different levels of fakeness ranging from true, mostly true to false and pants on fire. We achieved an accuracy of 0.75 for binary classification and 0.34 for six class classification.

1 INTRODUCTION

Traditionally, news sources have been trusted but due to extensive use of social media, fake news has been increasingly circulated, especially around election season which can be observed in the case of 2016 United States Presidential elections. The problem of fake news detection is that it is more challenging than detecting fake reviews, since the political language on TV interviews as well as posts on Facebook and Twitter is mostly short statements (Wang, 2017).

The primary goal of our project is to identify fake news in political statements. The dataset used for our project is a pre-existing dataset named ”LIAR” (Wang, 2017) that was made by William Yang Wang. It has 12,836 rows and 14 columns of data which were manually labeled and retrieved from POLITIFACT.com and this data is gathered from a variety of places like conferences and articles etc which accounts for the context in the dataset. The table gives us an overview of the columns in the dataset.

Column No.	**Column Name**
Column 1	Statement ID
Column 2	Label
Column 3	Statement
Column 4	Subjects
Column 5	Speaker
Column 6	Job title of speaker
Column 7	The State
Column 8	Party
Column 9	Count of barely-truths
Column 10	Count of false
Column 11	Count of half-truths
Column 12	Count of mostly-truths
Column 13	Count of pants-fire
Column 14	The Context

Table 1: Overview of columns in dataset

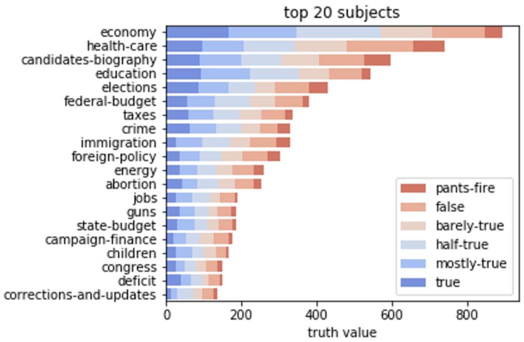

Figure 1: Mapping the frequency of class labels to subject (3)

The information is classified as True, Mostly-True, Half-True, Mostly False and Pants on Fire. We performed both binary classification, i.e True or False and a 6 class classification on the dataset.

In this project, we have compared performance of two different architectures, detection by Learning Convolution Filters through Contextualized Attention (Ranjan, 2019) and another model architecture involving two BERT models (Manideep, 2019) with shared weights. Further, we performed experiments to check if continued pre-training on domain specific data would improve the performance on downstream classification.

2 RELATED WORK

The chosen baseline paper performs Fake News Detection by Learning Convolution Filters through Contextualized Attention (Ranjan, 2019). Domain adaptive pre-training (Gururangan, 2020) shows that continued pre-training large language models like RoBERTa on domain specific datasets improved performance in multiple domains including biomedical, news and reviews. Siamese BERT has two BERT models in a siamese network with shared weights between both models. (Manideep, 2019) Siamese BERT for fake news detection introduces a credit scoring system to perform a weighted classification as it has been identified that each label data contributes differently towards detection fake news. (Wang, 2017) released a baseline for the LIAR dataset in the original paper by using Hybrid CNN to perform classification.

3 METHOD

Keeping the baseline model (Ranjan, 2019) as a reference, we use metadata to attend over the statement in the classification task. The metadata is converted into word embeddings using PyTorch embeddings that are randomly initialized. The embeddings convert the initial sparse matrix into a dense matrix. It is then passed to Gated Recurrent Neural Network which helps in modelling the dependencies between words. Unlike the baseline paper, which only considers Subject, Job, Context to be passed through a 2 layer BiLSTM model, we have passed the entire metadata through the Gated RNNs. By doing so, we are making use of all the available data to be classified. The data attained as the output of the Gated RNN is passed through a fully connected layer. The output of the fully connected layer is a Context Query Vector which is essentially a summarized version of the metadata.

The Context Query Vector is passed into the attention model which is used to retrieve the relevance. This will help us find the truthfulness of the sentences. The new matrix is passed to a Conventional Neural Network where Contextual Attention is being added using the Context Query Vector. This allows a recognizable pattern to be generated from the metadata which will used to search through test statements. The entire result is sent to a maxpooling layer and the resultant is then passed through a fully connected layer where the Context Query is concatenated with the maxpooling result. The output is generated through this which will classify the statements as either a binary classification or a 6-way classification.

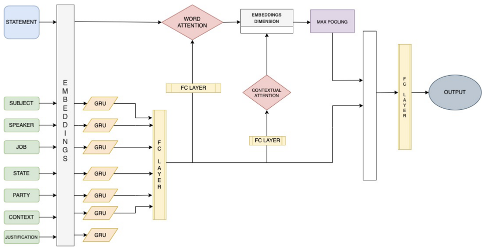

Figure 2: GRU with Convolution filters with contextualized attention

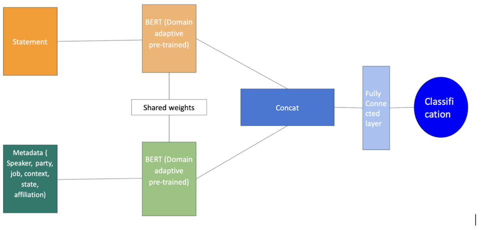
Statement
BERT (Domain adaptive pre-trained)
Shared weights
Metadata (Speaker, party, job, context, state, affiliation)
BERT (Domain adaptive pre-trained)
Concat
Fully Connected layer
Classification

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

Siamese BERT has two BERT models in a siamese network with shared weights between both the models. We combined the metadata by concatenating the metadata features and jointly pass the metadata word embeddings through a BERT model and passed word embedding of statement through another BERT model and both these BERT models share weights. We used BERT pre-trained embeddings to create embeddings for metadata and the statement. Further, we concatenated the outputs of the two BERT models and passed it through a fully connected layer before getting the classification output for binary or six class classification.

We also added a pre-fine tuned layer to a pre-trained RoBERTa before finetuning on our dataset. This pre-fine tuning layer is where the pre-trained RoBERTa model is continued to pre-train on a general news dataset where it can learn to perform better on any news datasets for the downstream task. We observed that by adding this additional layer, our model is able to learn and perform better on LIAR dataset giving us the best accuracy.

4 EXPERIMENT

Our experiment was to compare two architectures and add an additional domain adaptive pretraining to see which model performs better. Our experiments are recorded at `https://github.com/Mohammedzubairkhan/Fake_news_detection` The setup used for the different architecture are:-

- For the Convolutional filters with contextualized attention with Gated RNN architecture, we ran the model for 20 epcohs to train the model along with a learning rate of 0.001. Negative Loss Likelihood funciton was used as loss function with Adam optimizer. The number of kernels for CNN used was 64 and a 3D convolutional filter with size as [3,4,5] was used. The embeddings dimensions were kept at 100.

- For BERT model and domain adaptive pretrained RoBERTa model, we used BERT embeddings for creating embeddings for statements and the metadata. We replaced all NaN values with 0 as part of preprocessing of the dataset. The training was done for 20 epochs with batch size 16, loss function as Cross Entropy loss function, Adam optimizer and learning rate scheduler with initial learning rate as 0.0001 to decay learning rate by 0.1 for every 3 epochs. Max sequence length of statement was taken as 64 while maximum sequence length of metadata as 32 as the metadata has shorter sentence lengths. This helped us to control the length of each sequence passed to the BERT model. We also incorporated something called a credit score (Manideep, 2019) that gives weighted importance to each label and this helps in identifying which label is important for our prediction and we also see that this improves our performance.

Some of the technical challenges we encountered while working on the project are:-

1. **Labeling** : Labeling is an expensive task and hence there are limited labelled datasets available in this domain. Our dataset consists of only 12K+ statements and we expect to yield better results given a large enough dataset.

2. **Updating the database** : Traditional systems of identifying the fake news involves a maintaining and updating a system with all the data from the internet. This makes fact checking for recent news against a true database less efficient.

3. **Resources** : We want to be able to compute the embeddings and manipulate vectors of very large sizes to obtain accurate classification results but practically, there is always a limit on the resources available for us to use.

Other than the technical challenges, we faced challenges like -

1. **Time constraint**: We had limited time to experiment different embeddings, different models and hyperparameter tuning to understand which combination works the most efficient all the while continuing with our other coursework as well. This posed a time challenge and we could've improved this process.

2. **References**: Although this dataset and this area of research has been experimented and has some reference papers, the methodologies we are trying to use to achieve the results (Gated RNN, Pre-finetuning, etc) has never been tried out before for this particular case and hence there is a limit on the research references which we can use to learn about this in depth. Hence, we are majorly relying on the trial and hit process of experimentation.

3. **Hardware**: We could not run LIAR PLUS dataset which had additional metadata due to the hardware limitations, but if we had a hardware optimized cluster with better memory, we could have trained the model longer and achieved better results.

5 RESULTS AND DISCUSSION

The baseline results of the dataset for binary classification is 0.63 accuracy and 6-way classification is 0.24 accuracy. We got this baseline results by running the code from our baseline paper on our system for 20 epochs for a fair comparison. We observe the best accuracy results for the domain adaptive pre-trained RoBERTa model. This model yields an accuracy of 0.75 for binary which is 12%

better than binary classification accuracy for baseline and 0.34 for multi class classification which is 10% better than multi class accuracy for baseline. All the results are seen in the table 2. Validation accuracy plots for binary and multi class classification for Gated RNN and Convolutional Filters with Contextualized Attention are given in Figure 4 and Figure 5 and the Confusion Matrix on test set for RoBERTa model with domain adaptive pre training for binary classification is shown in fig 6.

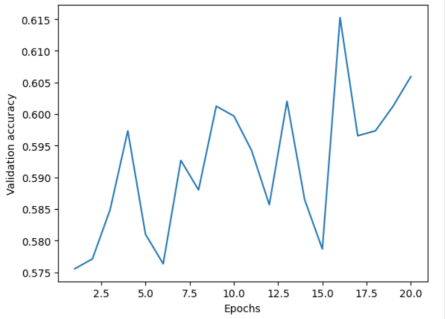

Figure 4: Accuracy for Gated RNN with convolutional filters with contextualized attention for binary classification

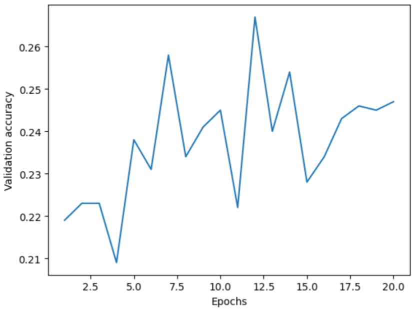

Figure 5: Accuracy for Gated RNN with convolutional filters with contextualized attention for six class classification

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 2: Accuracy metric evaluation for binary and six class classification

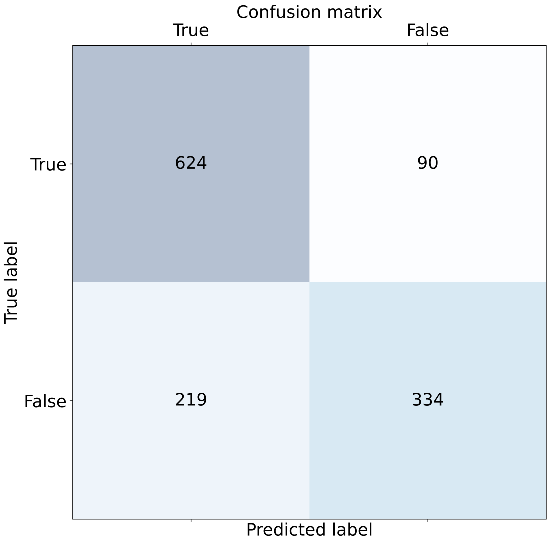

Figure 6: Confusion matrix for RoBERTa domain adaptive pretrained binary classification

6 FUTURE DIRECTION

- LIAR PLUS dataset has justification which is automatically generated from statements which we could not train on our model due to our limited computational capability. Hence, we can extend our model training to it.

- Multi task pre-training is one task we wanted to incorporate in our training but due to the time constraint were not able to. Pre-training a model on multiple tasks ranging from summary extraction to question-answer generation has said to improve the performance on downstream tasks.

- Adding extra parameters like sentiments and emotions will give a more accurate results. EmoLex will help deriving emotions and adding them as features whereas SensiStrength will help in getting the sentiments of the statements.

- Since there are less large, labeled datasets available in this domain, an approach that reduces this complexity of manually labeling datasets would be a good direction such as Semi supervised learning, weak supervised learning and self-supervised learning.

7 Division of Work

Work was equally divided between all team members including making reports, posters, working on code for the models, creating metrics for evaluation and hyperparameter tuning as shown in table 3.

Member	Work Division
Amulya Sethurao	reports, making poster, hyperparameter tuning of Gated RNN with CNN attention model
Harshith Acharya	reports, modeling Gated RNN and CNN with attention, creating metrics for evaluation
Mohammed Zubair Khan	reports, coding Gated RNN and CNN with attention, created result metrics for evaluation and hyperparameter tuning
Purva Makarand Kulkarni	reports, making poster, modeling Gated RNN, BERT models and hyperparameter tuning
Vinya Somayajula	reports, making poster, BERT models, Domain adaptive pre-fine tuning and creating metrics for evaluation

Table 3: Work Division between all the team members

References

William Yang Wang. 2017. *"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), pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Ranjan, Ekagra. 2019. *Fake News Detection by Learning Convolution Filters through Contextualized Attention*. 10.13140/RG.2.2.20829.84968.

Aghajanyan et al. 2021. *Muppet: Massive Multi-task Representations with Pre-Finetuning*. https://aclanthology.org/2021.emnlp-main.468, EMNLP 2021

Y. Santur. 2017. *Sentiment Analysis Based on Gated Recurrent Unit*. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-5, doi: 10.1109/IDAP.2019.8875985.

Gururangan, Suchin and Marasović, Ana and Swayamdipta, Swabha and Lo, Kyle and Beltagy, Iz and Downey, Doug and Smith, Noah A. 2020. *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. Computation and Language (cs.CL), Machine Learning (cs.LG), FOS: Computer and information sciences, FOS: Computer and information sciences, 2020, pp. 1-5, doi: 10.48550/ARXIV.2004.10964.

2019. *Siamese BERT Fake News Detection using LIAR*.

https://github.com/manideep2510/siamese-BERT-fake-news-detection-LIAR