WDPS Assignment 2 Report

Group 03

Chih-Chieh Lin (2700266), Shuhan Pi (2689783), Sreelakshmi Muthiah Kasinathan (2668065), Xuan Zhou (2698592)

1. INTRODUCTION

This report describes our development procedure, and results of assignment 2 in Web Data Processing System course. We aim to analyze the US Election 2020 Tweets dataset using sentiment analysis on the contexts around the entities. Therefore, we may extract some useful information, such as what the people care more about, what the people be positive about, what the people keep negative view on. These information may give us some insight or observation on the US Election result this year.

2. TRAINING MODELS FOR SENTIMENT ANALYSIS

2.1 Preprocess the training dataset

First, we choose the "Sentiment140 dataset with 1.6 million tweets" as our training dataset for BiLSTM model. We then prune and clean the dataset so that the number of two labels(positive and negative) is almost the same. It makes sure that the dataset is balanced.

Afterwards, we do some preprocess on the tweet texts in order to represents the sentiment comprehensively. That is, we drop the useless information, such as url links, usernames, unmeaningful punctuation. Moreover, the emojis are replaced by some meaningful words. We then do the tokenizing, padding, and word embedding on the processed tweet texts. Afterwards, the embedding layer for the BiLSTM model is formulated.

2.2 BiLSTM model

The model structure is shown in Figure 1. We take 1 million tweets and sentiments from the dataset. However, do the training on our own laptop terribly takes lots of time. One single epoch takes us around 2 hours. We then seek for solution, since we have to train more epochs and try different models. Finally, we found that the GPU runtime type offered by Google Colab can accelerate our training procedure.

To improve the accuracy, we tried lots of ways, such as changing different types of models, using different dataset, using different preprocessing methon on raw tweet text. Finally, The metrics mentioned in this report and the BiLSTM provide us with the best performance among the others we tried. Our BiLSTM model has 83.9% accuracy on the test dataset which is better than all the others we trained.

3. NAMED ENTITY RECOGNITION

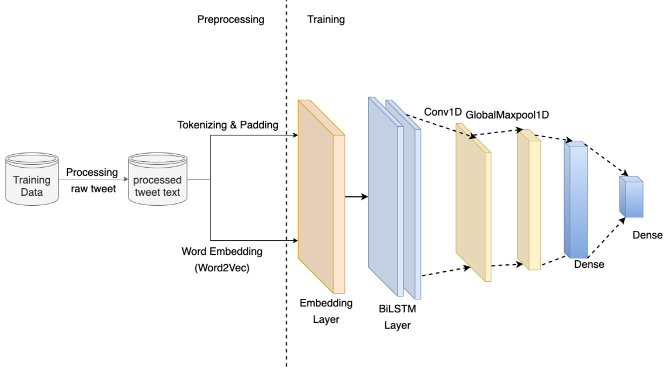


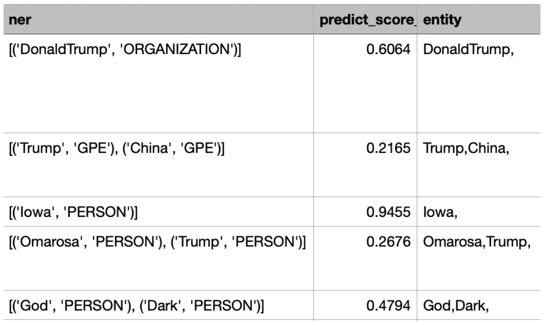
Figure 1: BiLSTM model Training Procedure

In order to analyze the sentiment around the entitiew in the tweets in the US election dataset, we need to do NER process on these tweets to extract the entities. The first thing we have to do is prepossessing these dirty data, such as deleting some useless and unmindful context, retaining only geographic information and filtering out the Twitter data for 20 days before the election. The NLTK library provides lots of packages that help us to do Named Entity Recognition, such as tokenize and POS tag. The POS tag method assigns labels to entities, but not all of the labels are accurate. This is still a problem that requires to be solved. With the labels, we can recognize entities and extract them. In the final dataset we removed the labels of the entities and only saved the entities. The Figure2 shows the schema of the final dataset.

We use the BiLSTM model mentioned in the previous section to predict the sentiment of the tweets in the processed Trump and Biden's dataset. It took 4h 9min 15s to predict the Trump data set and 2h 51min 24s to predict the Biden dataset. We stipulate that a score greater than 0.5 is a tweet with a positive view, and if the score is less than 0.5, the tweet is a negative one. These scores can help us with the analyzing work in the visualization part.

4. VISUALIZATION AND RESULT

Visualisation was carried out on the results obtained. In the graphs, positive and negative tweets are taken to be the good and bad ones; and this applies to the entities as well. According to the our project, the sentiment around the entities were found to an extent and the graphs were



ner	predict_score_	entity
[('DonaldTrump', 'ORGANIZATION')]	0.6064	DonaldTrump,
[('Trump', 'GPE'), ('China', 'GPE')]	0.2165	Trump,China,
[('Iowa', 'PERSON')]	0.9455	Iowa,
[('Omarosa', 'PERSON'), ('Trump', 'PERSON')]	0.2676	Omarosa,Trump,
[('God', 'PERSON'), ('Dark', 'PERSON')]	0.4794	God,Dark,

Figure 2: Named Entity Recognition

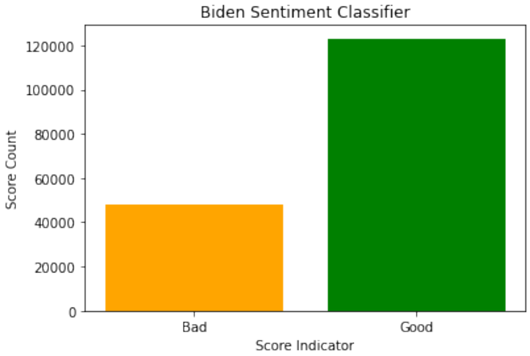


Figure 3: Sentiment Classifier Biden

plotted based on those results. We will describe the findings using the results obtained by the previous section.

The first graph (see figure 3) over the major classification of tweets between the candidates. It is observed that both the candidates received a good number of ”Good Tweets” but on a comparison it is also observed Biden got less percentage of ”Bad Tweets” which could have been a good point in his victory.

A view of the top 20 entities of Biden (see figure 4) and Trump (see figure 5). For this purpose, a list of entities with their corresponding count was grouped together and from that we picked the top 20 for Biden and Trump individually. From the graphs depicted, there is a conclusion assumed on the appearance of the entities. For example, the entity ”COVID19” [1] appeared as the 20th entity in Biden with very less number of counts whereas in the Trump list, it appeared twice with different names like ”COVID19”and ”COVID”. Also, it can be noted that the position of the entity ”COVID19” came up in the top 10 of the trump list. On the whole, both the entities mean the same issue which came up as the important issue during the election times. Based on this analysis, we found another reason for the victory of Biden.

The figure 6 and depict the good entity occurrence count for Biden. The entity ”JoeBiden” occurred many times in good occurrence whereas it occurred few times with few counts in bad occurrence category and also the position in the graph says in a bad context the entity did not occur many times which once again concludes that Biden had less hate tweets.

Similarly the figure **??** gives the bad entity occurrence count for Trump. Here, the entity ”COVID19” and ”COVID”

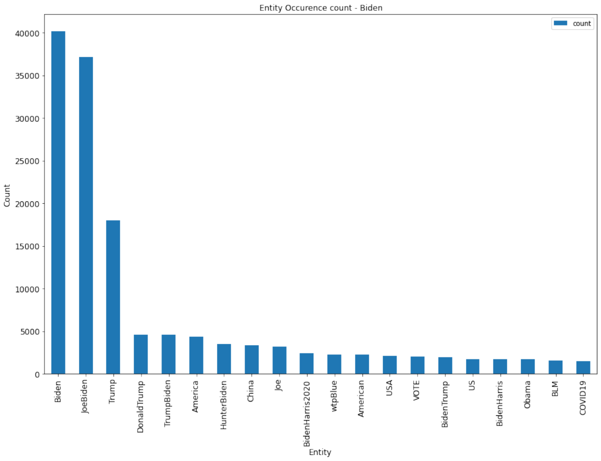


Figure 4: Biden general entity occurrence count

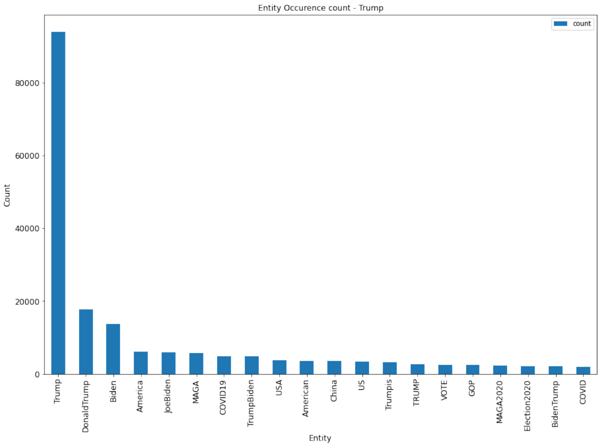


Figure 5: Trump general entity occurrence count

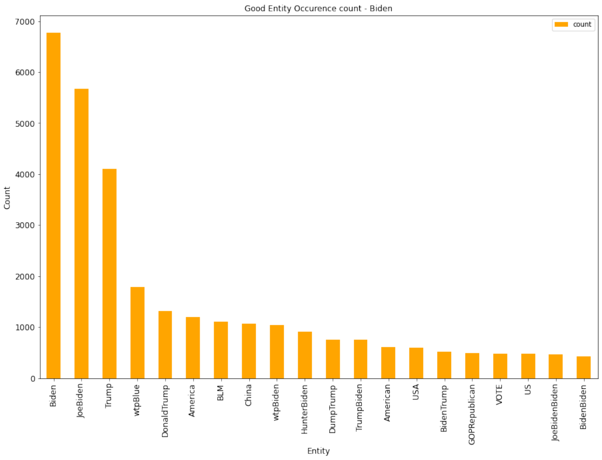


Figure 6: Good entity occurrence count - Biden

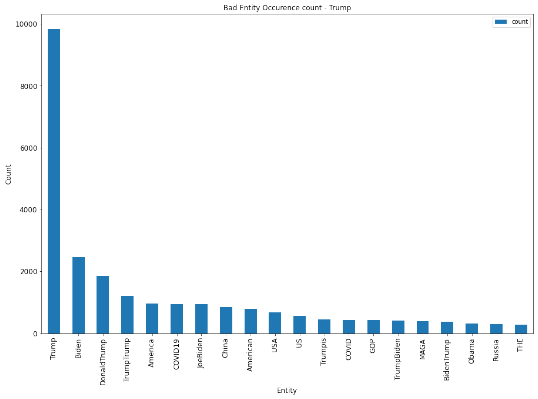


Figure 7: Bad entity occurrence count - Trump

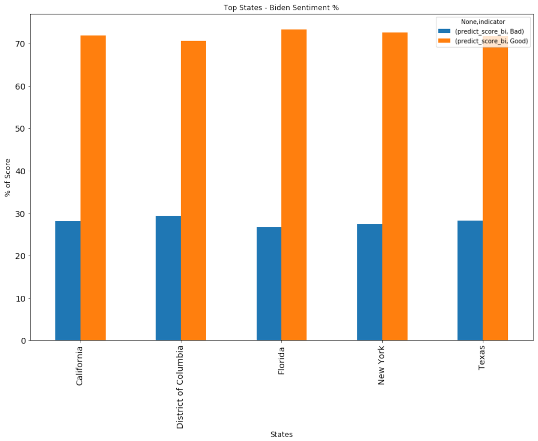


Figure 8: Percentage for Biden

occurred in both the lists but the position of the occurrence varied. In the bad occurrence count graph, it is observed that these entities have moved few position up in the top 20. This again brings a conclusion, that this Covid issue happened to be one of the negative influences for the defeat of Trump in the US elections 2020.

Finally, the figures 8 and 9 depict the percentage of good and bad tweets both the candidates received in the states of United States of America and the top five was chosen. In the percentage graph of Biden (see figure 8) it is witnessed that almost more than 70% of tweets were positive(good) and only the range of 20% to 30% occurred to be negative (bad). In the percentage graph of trump (see figure 9), it could be visualised that the positive (good) tweets of Trump ranged only between 60% to 70% and not more than that whereas the negative(bad) tweets ranged between 30% to 40% which is more than the range of Biden's negative(bad) tweets. Thus, we also conclude here that Biden had less number of hate tweets from people which turned the US elections 2020 in favour of Joe Biden.

5. CONCLUSION

From the results we analysed, the paper concludes that Joe Biden received less percentage of negative tweets(bad). Also, Biden received more percentage of positive tweets(good). The most importantly, the Covid pandemic played a vital role in the results since the entities related to Covid was found more in the Trump set and the positions varied in the bad occurrence count list which helped us predict that this could have been a major drawback in the defeat of Trump. Finally, its Biden who won the elections and that result was predicted based on the various analysis that was performed and mentioned in this paper.

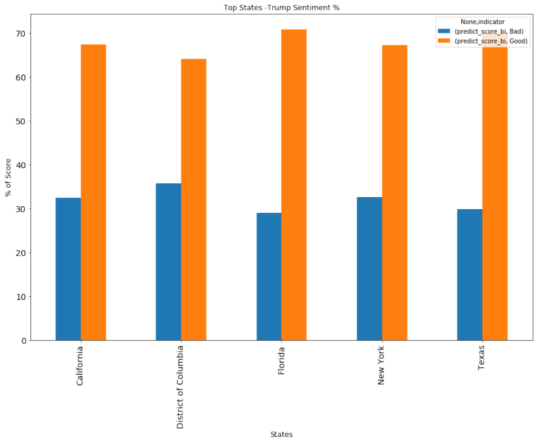


Figure 9: Percentage for Trump

6. REFERENCES

[1] L. Doggett. Timeline of trump's coronavirus responses.

APPENDIX

A. CONTRIBUTION OVERVIEW

The contribution of each part of our assignment procedure is shown in Table 1.

Person	Parts
Chih-Chieh Lin	Train the BiLSTM model. Data preprocessing
Shuhan Pi	Named Entity Recognition
Xuan Zhou	Named Entity Recognition
Sreelakshmi Muthiah Kasinathan	Visualization

Table 1: Contribution Overview