US Election in Twitter

Compare the US Election result with the prediction from Twitter using sentiment analysis

Chenghan Song$^{1[2680951]}$, Chih-Chieh Lin$^{1[2700266]}$, and Haochen Wang$^{1[2698251]}$

Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

Abstract. Using social media data for political discourse is quite common especially around US election this time. We use sentiment analysis on twitter data ranging from 15.10.2020 to 04.11.2020 to investigate how does social media reflect the process and result of US Election 2020. According to our main research question mentioned above, we can further specify two subquestions as follows: How could we quantify the extent of support and opposition to presidential candidates from social media? What are the different attitudes among states, regions or countries? We discover that sentiment based on location and time reflects on ground public opinion to some extent.

1 Introdution

1.1 Motivation

It is always a popular issue when the US Election is going to be held. Undoubtedly, 2020 US Election became a hot topic on each social media platform during the past few months. Moreover, due to the pandemic of Corona Virus this year, the voters in the United States may be not willing to participate in the election campaign. Instead, the voters or the supporters of certain party may actively post their statements and comments on the social media in order to propagate their convincing arguments. Does this phenomenon affect the 2020 US Election? On the other hand, we would like to figure out if the data gathered from social media like Twitter provides us with some insight in 2020 US Election. Therefore, our research question for this research project is came out and shown in the next section.

1.2 Research Question

– How does social media reflect the process and result of US Election 2020, like Twitter?

According to our main research question mentioned above, we can further specify two subquestions as follows:

- How could we quantify the extent of support and opposition to presidential candidates from social media?
- What are the different attitudes among states, regions or countries?

2 Related Work

In this project, we seek to explore and answer our proposed research questions using Twitter data. There has already been done many researches on utilizing mathematics and computer science techniques to analyze social media dataset.

Laura C et al.[3] conduct an quantitative analysis of twitter posts on 2017 UK General Election during one month period running up to the election. They look at representative features such as most popular hashtags, most mentioned and retweeted accounts, and most mentioned topics by and linked to politicians, coming to the conclusion that Twitter is able to reflect spontaneous, motivated behaviour of users which means analyzing tweets contributes to help people learn that who plays an important role in setting agendas as well as shaping conversations in social media and how effective or transient their expressions are.

Social network analysis is one of the most crucial problem in data mining. The application of machine learning in the social network becomes extremely popular now such as spam content detection, recommender systems, human behavior analysis, and sentiment analysis[7]. The domains of researches relating to Twitter sentiment analyses ranging from understanding the emotional tone behind customers' reviews to transfer learning. Ussama Y et al.[10] perform a sentiment analysis of location-based twitter election data. This paper utilize two case studies including 2017 UK general election and 2016 US presidential elections with Python TextBlob library for Natural Language Processing, drawing the conclusion that location-based sentiment has a reflection on ground public attitudes.

Several methods have been proposed in research to sentiment analysis. Walaa M et al.[6] conduct a comprehensive survey on sentiment analysis algorithms and applications. The machine learning approach of sentiment analysis is divided into supervised learning and unsupervised learning. Supervised learning algorithms like Neural Network, Naive Bayes classifier, and Maximum Entropy Classifier can be applied when the labeled training data exists while unsupervised learning algorithms are used when is no target variable[7].

3 Data

3.1 Preprocessing

In order to properly take advantage of the data, we need to do preprocessing. This work is divided into several parts:

1. Pruning the data

- In this part, we first discard the data we would not use in our project. For instance, there are some column with useless information, such as user_name, user_join_date, user_description etc.
- Moreover, we only want to take the tweets posted in the few days before the election days. Therefore, the tweets posted before and after this specific period would be removed.

2. Data Cleaning
   - There might be some missing or dirty data in the dataset, such as null, non-informative data. In order to smoothly do the data analyzing in the next step, we do have to drop this missing or dirty data.
   - On the other hand, the geological information is essential for our project. Thus, the row without geo-information would be completely removed.

3.2 Analysis

Before deeply analyzing the data using sentiment analysis, we would like to have a quick and simple analysis to see through the dataset.

Figure 1 provides an overview of the tweets distribution in the world. The more spot appears in a certain region, the more people care about the 2020 US Election. It is not surprising that the US election do concern people in the United States and the Europe.

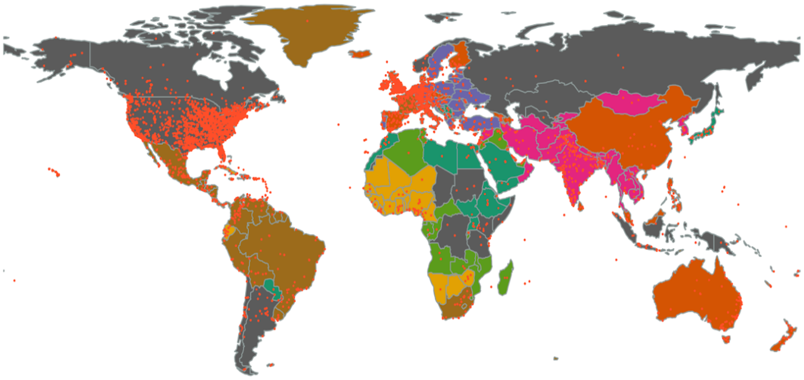

Fig. 1. The Tweets Distribution

4 Technical Methodology

We construct a bi-directional LSTM (Bi-LSTM) based deep learning model to take advantage of Word2vec word embedding for the sentiment analysis task. The frameworks of Word2vec and Bi-LSTM based model are shown respectively in Figure 2 and Figure 3. The pipeline of our work for sentiment analysis is in the following steps: splitting texts into sentences, initialization of Word2vec embedding, tokenization and padding of all tweets from the dataset, training the embeddings with padded tokens, classification of sentiment by the designed Bi-LSTM based neural networks, and the Collaborative Relation Correction (CRC) to balance the bias.

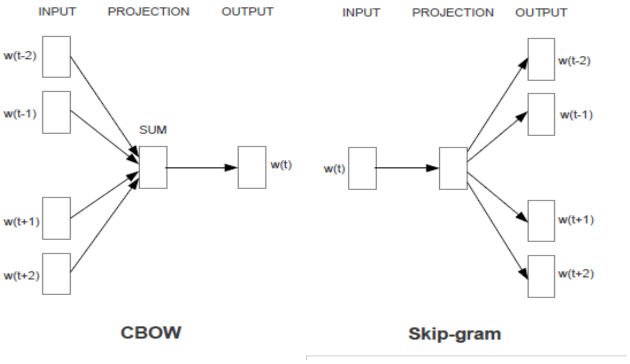

Fig. 2. The structure of CBOW and Skip-gram

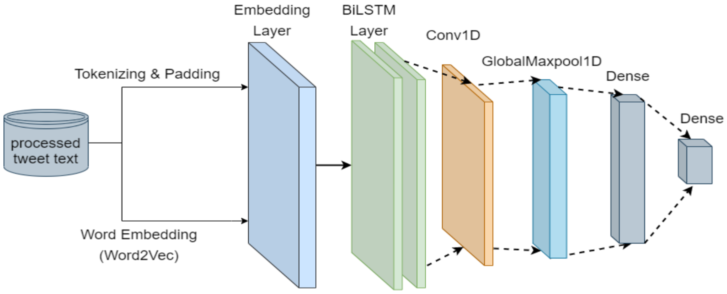

Fig. 3. The architecture of the model

4.1 Word2vec embeddings

Word embedding is an important technique in natural language processing, where words are mapped to vectors of real numbers. It can capture the meaning of a word in a document, semantic and syntactic similarity, relation with other words, which makes natural language computer readable. Further implementation of mathematical calculations on words can be used to detect their similarities. A well-trained set of word embeddings will place similar words close to each other in the vectoral space.

Word2vec is one of the most popular technique to learn word embeddings using multi-layer recurrent neural networks. We use the Gensim python library to import Word2vec model. In our project, the input is a sentence and its output are a set of vectors representing each word from the corpus. There are two main training algorithms for Word2vec, one is the Continuous Bag-of-Words (CBOW), the other is called Skip-Gram.

Continuous Bag-of-Words Model It consists of input, projection and output layers. In the input layer, previous words are encoded using 1-of- V coding, where V is the size of the vocabulary. The input layer is then projected to a projection layer P, which is shared for all words, not just the projection matrix. Therefore, all words get projected into the same position and their vectors are averaged. The order of words in the history does not influence the projection. It also takes words from the future. That is why it is called a bag-of-words model.

Moreover, it uses continuous distributed representation of the context or sentences to predict the vector of the word in the position. The architecture of CBOW model is shown in Figure 2. The weight matrix between the input and the projection layer is shared for all word positions. For the output layer, the hierarchical *Softmax* is used to represent the vocabulary as a Huffman binary tree.

Skip-gram Model The second architecture is like CBOW. Instead of predicting the current word based on the context, it optimizes to maximize classification of a word based on another word in the same sentence. Deeply, it uses each current word as an input to a log-linear classifier with continuous projection layer. Then, it will predict words within a certain range before and after the current word (mainly the latter one). The architecture of Skip-gram Model is shown in Figure 2. Some have found that increasing the range improves quality of the resulting word vectors, but it also increases the computational complexity.

The major difference between these two methods is that CBOW is using context to predict a target word while skip-gram is using a word to predict a target context. Generally, the skip-gram method can have better embeddings compared with CBOW method, for it can capture two semantics for a single word. For instance, it will have two vector representations for Apple, one is for the company and the other is for the fruit. But for the higher performance in the large scale of tweets, we choose CBOW model to obtain embeddings.

4.2 Bi-LSTM based neural networks

Bidirectional LSTM (Bi-LSTM) Long short-term memory (LSTM) units have been extensively used to encode textual sequences. The basis encoder consists of an embedding layer, LSTM layers, and dense layers for specific tasks based on the encoded features.[1] The mathematical theory is formulated in the following Equations:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$

$$g_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{4}$$

$$c_t = f_t \odot c_{c-1} + i_t \odot g_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

Where h_{t-1} represents the previous hidden state and h_t represents the current hidden state, v_t is the current input word embedded vector of LSTM layer, and W and U contribute the weight matrices.

A bidirectional LSTM, or Bi-LSTM, is a sequence processing model that consists of two layers of LSTM. The abstract structure of a Bi-LSTM is illustrated in Figure 4 [2]. The first layer of LSTM takes the input in a forward direction. The second takes in a backwards. This structure allows the networks to have both future and history information about the sequence of every time step.

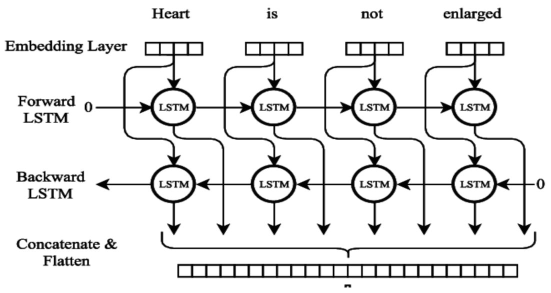

Fig. 4. The abstract structure of Bi-LSTM

Using Bidirectional, it will execute the input in two ways, one from past to future and one from future to past. The difference between this approach and unidirectional is that LSTM runs backwards you preserve information from the future and using the two hidden states combined. It is enabled in any point of sequence to preserve information from both past and future. Both *tanh* activations would be considered to calculate the output y at time t in the Equation:

$$y^{<t>} = \tanh(W_y[a^{\leftarrow<t>}, a^{\rightarrow<t>}] + b_y) \tag{7}$$

That is suitable for textual contexts in out project it is sentence. Using Bi-LSTM, the model can understand a word we need not just to the previous word, but also to the coming word.

One-Dimensional Convolution Followed by the Bidirectional LSTM, there is a one-dimensional convolution (*Conv1D*) layer. This layer creates a convolution kernel that is convoluted with the layer input over a single dimension to produce a tensor of output.

As the analogy as two-dimensional CNN, in *Conv1D*, kernel slides along only one dimension. A *Conv1D* is very effective when we expect to derive interesting features from shorter (fixed-length) segments of the overall data set and where the location of the feature within the segment is not of high relevance.[5]

One convolution layers of a *Conv1D* is presented in Figure 3. As shown in this figure, the one-dimensional filter kernels have size of 5 and we define 100 filters to train 100 different features. It first performs a sequence of convolutions, the sum of which is passed through the activation function. This is indeed the main difference between *Conv1D* and *Conv2D*, where one-dimensional arrays replace two-dimensional matrices for both kernels and feature maps.[5] As a next step, the *GlobalMaxPool1D* layer sub-samples the output of the *Conv1D* tensor of 100 dimensions and "learn to extract" such features of maximum values which will be used in the classification task performed by the Dense layers.

Consequently, both feature extraction and classification operations are fused into process that can be optimized to maximize the classification performance. This is the major advantage of *Conv1D* which can also result in a low computation complexity since the only operation with a significant cost is a sequence of one-dimensional convolution which is simply linear weighted sums of one-dimensional arrays.

4.3 Bias correction - Collaborative Relation Correction

Due to the fact that the number of tweets about Trump is much larger than those of Biden, the advocating rate of each candidates could not directly be the average of the score of sentiment. Furthermore, it is hard to say that a person who shows negative sentiment to one is equal to that he will show positive sentiment to the other. Otherwise, we would have an extreme result, where the score of Trump has huge drop with the score of Biden. Therefore, we carry out a bias correction method called Collaborative Relation Correction (*CRC*).

Here *CRC* is a name of our invented correction method, which is based on Spearman correlation.

In the first step, we can run a Spearman's correlation on a non-monotonic relationship to determine if there is a monotonic component to the association. then, we would normally pick the measure of association, that fits the pattern of the observed data. We can calculate Spearman correlation by Equation:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8}$$

Where x_i and y_i are paired scores. The Spearman correlation coefficient, r_s, can take values from +1 to -1. A r_s of +1 indicates a perfect association of ranks, a r_s of zero indicates no association between ranks and a r_s of -1 indicates a perfect negative association of ranks. The closer r_s is to zero, the weaker the association between the ranks. By this way, we can calculate the correlation indexes of four situation: *Trump positive* VS *Biden negative* and *Trump negative* VS *Biden positive.*

In the next step of CRC, we will calculate the value of natural offset between the advocating rates of two candidates. To better use the parameter of ρ and r_s, which show the correlation and coefficient respectively, we multiply them in a correlation array. Therefore, we have four values in the array represent the distance, or bias, of the advocating rates. According to the feature and structure of the definition of Spearman correlation, we decide to use harmonic mean to conclude the bias of those and set it as the value of correction. The definition of harmonic mean is as Equation:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{9}$$

Where x_i is the values of correlations and n is the number of items.

5 Visualisation and Analyses

5.1 Most concerned topic in the tweets

The citizens who care about their future would post the tweet to illustrate their idea and perspective. Therefore, the most frequent mentioned topic would give us some views in this election. Moreover, we do sentiment analysis on each tweet so that we can extract more information from the tweets.

The words in the tweets are extracted from word cloud package in python. Figure 5 shows the topics which people care about most in their tweets when they also mention Trump. Figure 6 do the same thing for Biden. We could ignore the biggest topic (Trump and Biden) since it certainly should be the name of both candidates.

We could have a look at negative part.The ”nypost” is a hot topic when the people keep negative attitude in posting tweets. Also, ”Hunter” may be seen as a negative phrase when people mention Biden. On the other hand, ”student” is a frequent mentioned issue when someone keep negative view on Trump.

There are also several phrases presented in the general part and positive. However, we are not going to in details on them. Both Figure 5 and Figure 6 can give us some view for the trend of election.

5.2 Geography of political participation

We create a heatmap in order to observe the state-level regional differences in political participation. Darker shades indicate higher political participation. From the figure 7 we can see that most tweets are from big states like California, New York, Florida, and Texas indicating big economically developed cities get

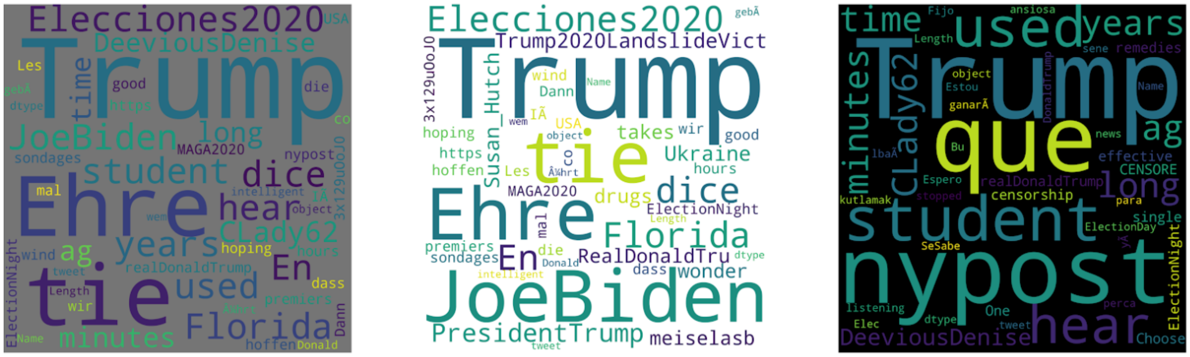

Fig. 5. Word Cloud for Trump (a) General (b) Positive (c) Negative

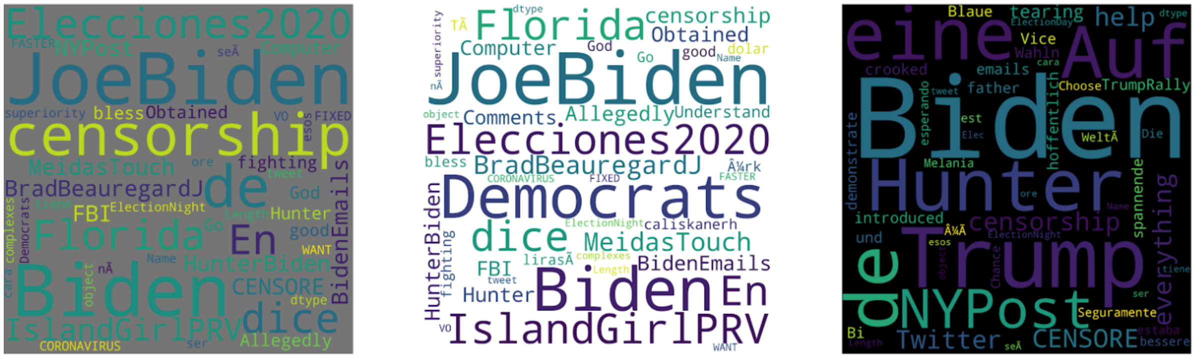

Fig. 6. Word Cloud for Biden (a) General (b) Positive (c) Negative

more political participation. In addition, most tweets mention both Biden and Trump and Trump got more referred.

From figure 8 we can see, in the city-level ranking of tweet count, people from cities like New York, Washington, Los Angeles, and Chicago are more likely to tweet about election which makes sense because these cities are cultural and political center, containing more active twitter users.

5.3 Results of Sentiment Analysis

In the first stage, we set the threshold of 0.5 to define the positive or negative tweets by predicting scores. The statistic of sentiment can be respective for two candidates. From Figure 12, we can find that the number of valid tweets about Biden is smaller than Trump's. The figure shows that the positive tweets is more than negative ones for both candidates.

Secondly, we carry out a tracker for each one by hour. As shown in Figure 10, it reveals the number of tweets posted hourly and the percentage of positive,

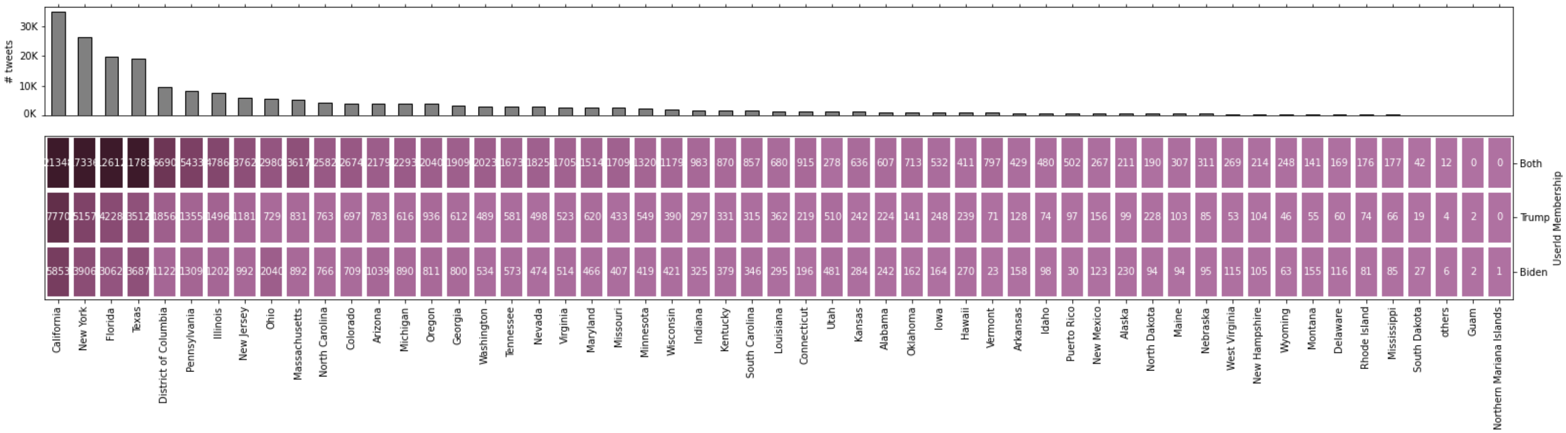

Fig. 7. Heatmap of tweet count per state

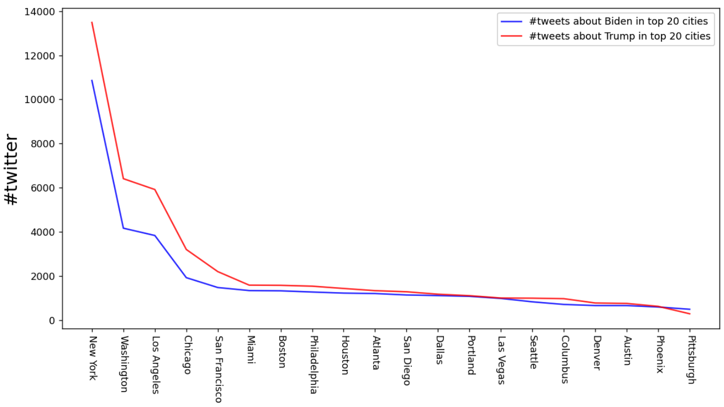

Fig. 8. tweet count of top 20 cities

negative ones. To our surprise, the tweets posted on 22:00, October 15th are missing. The general trend of volume is increased when coming to November 3rd. And for the overall percentage of positive, Biden's is almost higher than Trump's. That would indicate the main public opinion inclines to Biden. The volume of tweets about both candidates dramatically goes up on 2:00 October 23rd. We consider that it is the data retrieve problem. There are also some outstanding point. The time around 9:00 October 17th, 13:00 October 18th, 18:00 October 23rd, 12:00 October 25th, 01:00 October 31st, and the end of November 3rd in Figure 10, these points of time positive of Trump takes up more percentage than Biden.

Third, we take top 100 famous users in twitter to analyze their tweets about Trump and Biden. The dataset has the column of the number of followers, which can donate how popular and influential the user is. The user who has many followers is defined as the top user. It is meaningful to observe how these people think of two candidates. We take top 10, 100, 1000 and 10000 users to analyze, and choose top 100 uses in Figure 11 to describe. We can see from the pies that Biden takes up more positive tweets while takes less negative ones. This reveals the same result as Figure 9 and 10.

Fourthly, we visualize the score of the positive of Biden and Trump respectively in Figure 12 and Figure 13. The pictures can show how much do the US people advocate Biden and Trump. The deeper color fills in the region of states, the more advocating rate does the candidate have. In this way, we can see the public opinion various in different states, which shows different politics belief, the Democratic or the Republican.

For better comparison, we use our raised method *CRC* to correct the natural bias due to the unbalance of number of tweets about Trump and Biden. Therefore, we can compare their advocating rate in every state and predict who will win, the prediction is illustrated in Figure 14.

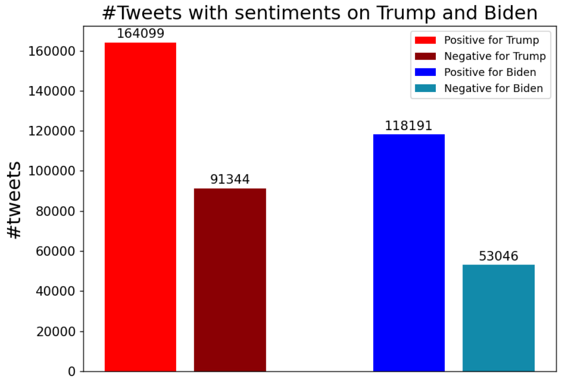

Fig. 9. The number of Tweets with sentiments on Trump and Biden

6 Conclusion

6.1 Answer to Research Question

- How does social media reflect the process and result of US Election 2020, like Twitter?

From our visualization of our prediction on the advocating rate, we can have a forecast of the winner of the Election in every state in Figure 14. Here we refer to the final result of the Election from Figure 15. We can say that in the 51 states we succeed in predicting 37 states advocating correctly, whose accuracy is 72.55%.

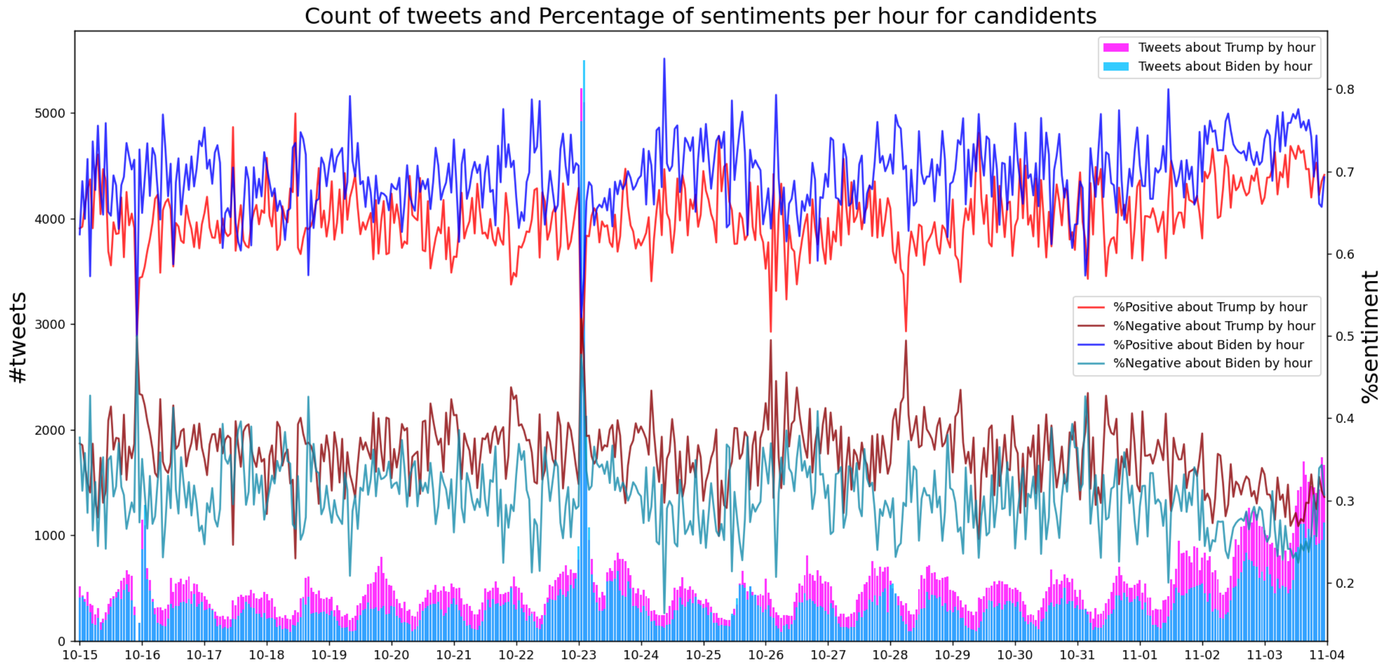

Fig. 10. Count of Tweets and percentage of sentiments per hour

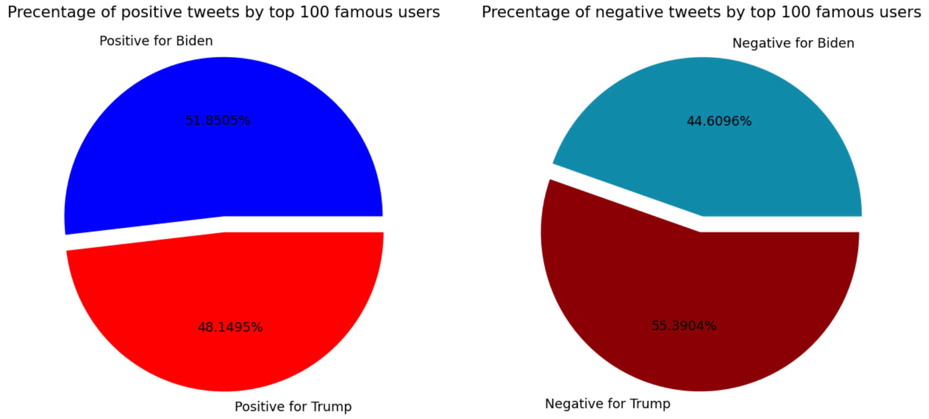

Fig. 11. Percentage of Tweets Trump and Biden by top 100 users

According to the Figure 10, we do observe the trend of change of tweets during about 20 days before the 2020 US Election. We observe that the percentage of positive tweets related to Biden is always greater than the one related to Trump; also, the percentage of negative tweets related to Trump is always greater than

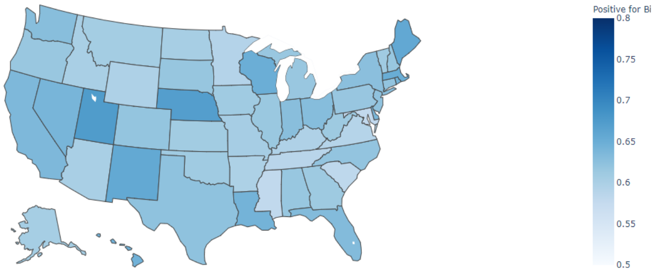

Fig. 12. Sentiment score for Biden in each State

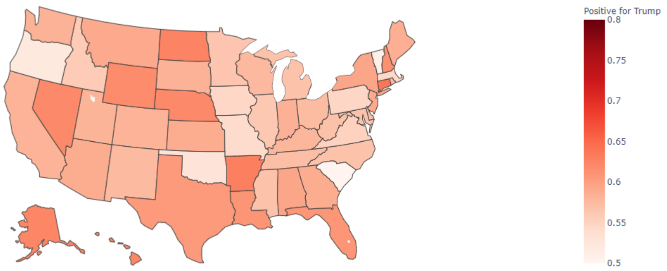

Fig. 13. Sentiment score for Trump in each State

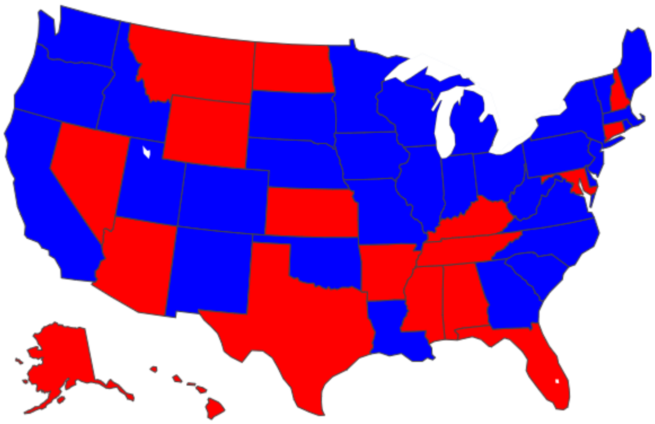

Fig. 14. More advocating rate in each State

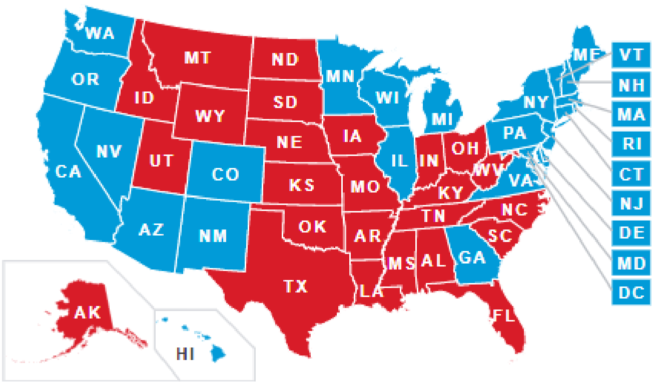

Fig. 15. The final result of Election

the one related to Biden. This observation provides us a small signal that Biden may obtain more support in this election.

6.2 Privacy & ethical considerations

With the exponentially development of online social platforms, the easier availability of personal information has increased the risk of user privacy leakage which prompts a question that whether sentiment is personal data? Issues like privacy, balance of power, and security evoked by these technologies[8]. After some research, we think that sentiment is personal data if they can be associated to an individual. Our project indicates that tweets can reveal personal political leanings and may cause more serious social divisions.

Another ethical consideration is balance of power and unfair competition. Twitter accounts with massive followers can be concentration of opinion power which can yield great social influence. Politicians may use those accounts to spread misleading information about opponent to try to manipulate the political atmosphere of social networks.

6.3 Limitation

John Scott raised concern of abuses of techniques on sociology. Mathematical and computer science measurements make sense only when there are good theoretical or empirical reasons for making them[9]. Through this project, we have a better understanding that we should focus on answering research questions themselves instead of fancy methods.

We found that the amount of tweets mentioned Biden is much less than that of Trump, which could cause bias during analysis. We use CRC algorithm to correct that. In addition, even though precision of the current model is not bad, still sentiment analysis may get incorrect understandings of sentiment because human communication is complicated especially when there are sarcasm, word ambiguity, negation, and multipolarity[4].

Another limitation is that due to privacy, we are not sure about who will actually vote. A significant portion of active users may not be old enough to vote. The ideal dataset should be able to identify age range and voting eligibility. In addition, a real random sample of the likely voters twitter dataset is still unattainable because only those who tend to actively tweet about election can be observed.

6.4 Future Work

There are several parts we could do further work, such as improving our performance of model, comprehensively illustrating our results of sentiment analysis, presenting our project results in a better way.

First part could be improved is that increasing the accuracy of sentiment analysis model. To increase the accuracy, we could adjust the model structure, tune the parameters, or refine the training dataset to enhance the performance of model. The accuracy of sentiment analysis significantly affects our work on analyzing since it would make our results be closer to the reality.

Second, to present our project result in a better way, there are lots of frameworks and tools could be utilized. For instance, some figures in this report are human-interactive. However, we cannot present this kind of function in a paper report format. Therefore, creating a statistic website may be a better option for presenting our results, and also makes user realize our work easily.

Finally, There might be some blind spots or mistaken idea in our works since the authors of this report come from similar backgrounds. We are all studying in the field of Computer Science, so there are definitely something we may lost or we cannot figure out from our own perspective. To solve this limitation, we will ask people from different background especially the one from sociology.

References

1. Chen, Y., Yuan, J., You, Q., Luo, J.: Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 117–125 (2018)
2. Cornegruta, S., Bakewell, R., Withey, S., Montana, G.: Modelling radiological language with bidirectional long short-term memory networks. arXiv preprint arXiv:1609.08409 (2016)
3. Cram, L., Llewellyn, C., Hill, R., Magdy, W.: Uk general election 2017: a twitter analysis (2017)
4. Eremyan, R.: Four pitfalls of sentiment analysis accuracy (Oct 2018), https://www.toptal.com/deep-learning/4-sentiment-analysis-accuracy-traps
5. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J.: 1d convolutional neural networks and applications: A survey. arXiv preprint arXiv:1905.03554 (2019)
6. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal **5**, 1093–1113 (2014)
7. R V, B., Kanaga, E., Kundu, S.: Application of Machine Learning in the Social Network, pp. 61–83 (06 2020). https://doi.org/10.1002/9781119551621.ch4
8. Royakkers, L., T.J.K.L.e.a.: Societal and ethical issues of digitization. Ethics Inf Technol **20**, 127–142 (2018)
9. Scott, J.: Social network analysis. Sociology **22**, 109–127 (1988)
10. Yaqub, U., Sharma, N., Pabreja, R., Chun, S.A., Atluri, V., Vaidya, J.: Location-based sentiment analyses and visualization of twitter election data. Digit. Gov.: Res. Pract. **1**(2) (Apr 2020). https://doi.org/10.1145/3339909, https://doi-org.vu-nl.idm.oclc.org/10.1145/3339909

A Chenghan Song

- Jupyter notebook assignment 1: For the first assignment, Chih-Chieh, Haocheng, and me meet at Chih-Chieh's house to work together. We did a lot of research on how to retrieve social web data and applied for Twitter API. This assignment is quite new to me and I learned how to use Twitter API to retrieve trends and search results from Twitter

- Jupyter notebook assignment 2: I did the Exercise 2 in this assignment, Chih-Chieh did the first exercise 1, and Haochen did the third exercise. After finishing our own part, we held a group meeting to go through everything and improve the final notework. In this assignment, In addition, I learned how to parse data and gain the knowledge of microformat.

- Jupyter notebook assignment 3: In this assignment, I was responsible for Task 6 and Task 7 while Chih-Chieh did Task 1,2,3 and Haochen did Task 4, 5. After finishing the work, we met on zoom meeting with TA to discuss and go through the whole notebook. In this assignment, I learned about centrality measures like betweenness centrality, eigenvector centrality, and closeness centrality as well as how to use networkx library to construct and plot social network.

- Jupyter notebook assignment 4: I did Exercise 3,4 in this assignment while Chih-Chieh did first two exercises and Haochen did the last two exercises(the last one was cancelled later). After finishing the assignment, we meet with TA to discuss the problem we encounter about the task of building a recommender. In this assignment, I learned similarity measures which provide recommendations and explore the textual similarity using NLTK libirary.

- Final assignment: In the final group assignment, we first implement different solutions respectively. I implemented the Naive Bayesian method, Chih-Chieh implemented a Bidirectional LSTM model, and Haochen trained a LSTM model. We compared the results of three models and select Bidirectional LSTM as out final solution because it has better performance. Then Haochen and Chih-Chieh did the prediction part. After we have the final data, we divided the visualization and analysis task into three parts and assign them to everyone. For the report, I did the abstract, related work, my part of visualization and analysis, privacy & ethical considerations, and limitation part. After finishing the draft, I gather together with Chih-Chieh and Haochen to polish the final version of the report. I think Chih-Chieh and Haochen are both good teammates and I had a good time working with them.

B
Chih-Chieh Lin

- Jupyter notebook assignment 1: Three of us work with this assignment together. We meet one another at my home and discuss this assignment. I still remember that we are sort of confused about applying the Twitter API key.

- Jupyter notebook assignment 2: I do the first exercise, which let us practice the bs4 package and then convert the output into KML. Moreover, Chenghan do the second exercise, Haochen do the third exercise. We do this separately and upload our own works on my github repository. Also, we checked and discussed our answer with TA.

- Jupyter notebook assignment 3: Task 1,2,3 in this assignment are done by me. I use the graph concept(edges, nodes) to represent the social networks. I plot plenty of graphs in this assignment. Moreover, Haochen do the task 4,5; Chenghan do the task 6,7. We did discuss and share our own results with one another after finishing our own parts. Also, we did check the our answer and discuss some points of view with TA.

- Jupyter notebook assignment 4: Exercise 1 and 2 are done by me; exercise 3 and 4 are done by Chenghan. Haochen focused on Exercise 5; found some blind spots in it and reflected our view to the TA. During this assignment, I implement the similarity between some users using Euclidian distance, pearson correlation. Also, I use this similarity features to create a simple recommendation function.

- Final assignment: First, Three of us try to implement different method for sentiment analysis. I tried to train a Bidirectional LSTM model; Haochen trained a LSTM model; Chenghan implement the Naive Bayesian method. We choose the model with the highest accuracy, which is BiLSTM model. Second, Haochen prune and clean the 2020 US Election dataset. Haochen and I use the model to predict each tweet's sentiment in the 2020 US election. Third, We divided the analyzing and visualization into three parts. These three parts of work could be illustrated by three notebooks we submitted. Finally, in paper work, I finished the section 1 (Introduction), section 3 (Data), section 5.1 (Most concerned topic in the tweets), section 6.1 (Answer to Research Question), section 6.4 (Future Work). Last but not least, we did some discuss, modify each other's work and give one another some comments. It is really nice to do this project with them.

C
Haochen Wang

- Jupyter notebook assignment 1: I went Chih-Chieh's home in one day's afternoon. We discuss the first notebook assignment about Twitter API. I did the notebook from the head to the end together with Chih-Chieh and Chenghan. I really enjoyed the moment of discussion.

- Jupyter notebook assignment 2: I did Exercise 3 about rdf, like Schema.org, and answer the questions in the notebook. Then I look into the microformats to find the difference among them. The first task is done by Chih-Chieh, while Chenghan finished the second part. After each work, we had a heated discussion and then turned to TA Nihat to solve some tough problems.

- Jupyter notebook assignment 3: I did the part of task 4 and 5 in the notebook namely Time for Data. I learned networkx library to consturct social networks for the public Facebook dataset. And answering some corresponding questions in the third notebook assignment. Chih-Chieh chose the first three tasks and Chenghan did the last two. As usual, we discussed the assignment together before meeting with TA. Then we checked our answer with Nihat.

- Jupyter notebook assignment 4: I did the last task of this assignment, which is Building a Reddit Recommender. It ran with the former results of task in this notebook. Therefore, I checked and ran the codes before to better understand some algorithms. I tried to use different topics as objects of recommendation and got pretty good results. In this assignment, Chih-Chieh did the first two tasks and Chenghan answered some questions in the notebooks. After several discussion in our chatting group and Zoom meeting. We found it hard to understand the last work. Therefore, we directly turned to Nihat and reported our tough problem. Unfortunately, we might not get the feedback from him.

- Final assignment: For the part of sentiment analysis, I tried LSTM, while Chih-Chieh used Bi-directional LSTM and Chenghan used Naive Bayes. After discussing and comparison, we found that Bi-directional LSTM shown the best accuracy and performance. I did the data preprocessing, like cleaning and regular expression. Chih-Chieh and I did the prediction work together. I took charge of some visualization. To specific, I visualized all of the output of sentiment analysis, in the paper from Figure 9 to 14. For the reporting part I wrote Methodology part and Results of Sentiment Analysis part. For the presentation, I showed the methods and data we used about the 2020 US Election. Finally, when we finished our work, we together check the article and improve the details then fit into Latex format using Overleaf. I really enjoy the final project. Chih-Chieh and Chenghan are both excellent groupmates.