ISSN 2723-6471

Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets

Ria Devina Endsuy A [*]
[*] STIMIK Tunas Bangsa Banjarnegara, Indonesia
[*] nezt.senna@gmail.com
* corresponding author

(Received December 21, 2020 Revised January 6, 2021, Accepted January 14, 2021, Available online January 15, 2021)

Abstract

The 2020 US Election took place on November 3, 2020, the result of the election was that Joe Biden received 51.4% of the votes, Donald Trump 46.9%, and the rest were other candidates. The period before the election was a time when people conveyed who would vote and conveyed the reasons directly or through social media, especially Twitter through keywords or tags such as #JoeBiden & #DonaldTrump. In this paper, we will compare sentiment analysis and exploratory data analysis against US election data on Twitter. The overall objective of the two case studies is to evaluate the similarity between the sentiment of location-based tweets and on-ground public opinion reflected in election results. In this paper, we find that there are more "neutral" sentiments than "negative" and "positive" sentiments. This study are focused finding sentimental tweets that people say on twitter for both presidential candidate and The dataset used is from and provided by Kaggle and has been updated on November 18, 2020, it is hoped that we hope that the academic community, computational journalists and research practitioners alike can utilize our dataset to study relevant scientific and social problems.

Keywords: US Election, Sentiment Analysis, VADER, EDA, Twitter.

1. Introduction

In this technological era, social media has become a medium for people to express their opinions. Everyone has their respective rights in expressing their opinions in the media. some of their opinions may be just satirical opinions, hoaxes, or even false opinions. however, it does not rule out the possibility that the person actually expresses his opinion, in the end it is a public medium that everyone can use and like.

Last year, to be more precise in 2020, there was a US presidential election, a sizable event that occurs every 5 years and is quite awaited by most US citizens. However, unlike previous elections there were some significant changes. Last year US citizens were quite debating about their presidential candidate, people from all parts of the US state their opinions and reasons regarding their possible candidate. unlike a few years or even decades ago the way participants express their opinions has changed tremendously thanks to social media being a platform that allows and even helps people who want to share their opinions

Almost everyone expresses their opinion on social media, there are many social media that people use to convey opinions such as Facebook, Instagram & Twitter. In this case most of the people used twitter to share their opinion about this election [1]. Twitter allows its users to post their opinions & thoughts, or commonly known tweets, on their social networks. There are approximately 330 million people around the world using Twitter and they can post approximately 150 million Twitter each day. Please note that 68 million users are US citizens, it is possible for us to collect and analyze the data in it. In this study, we will dissect and analyze data sets from opinions or opinions of Twitter users regarding the US presidential election, using sentiment analysis and comparing the two methods or models used, namely Exploratory data analysis and VADER (Valence Aware Dictionary for Sentiment Reasoning).

2. Literature Review

Interest in mining sentiment and opinion in texts has risen steadily over the last decade [2], mainly because of the increased availability of information and personal opinion messages [3]. In general, sentiment analysis is used to make predictions or measures in various fields such as the stock market, politics, and even social movements [4]. old studies of politics on social networks have ended up lacking or even lacking data because they can only take a small sample [5]. For example [6], proved that sentiment analysis on political prediction can proved accurate on the 2009 German election, On the other hand [7], Failed to predict 2011 US Presidential election ranking.

In this study we will make an analysis of how accurate the results of this study are and compare them with the selection that has been obtained. This sentiment analysis research is focused on putting together social science education with specialized quantitative methodology: our approach combines informed and correlative real-time data collection and predictive sentiment modeling, and understanding through the use of Twitter of the cultural and political practices at work.

3. Method

3.1. Exploratory data analysis

Exploratory data analysis or EDA was part of the data science process. EDA is very important before performing feature engineering and modeling because at this stage we have to understand the data first. Exploratory Data Analysis allows analysts to understand the contents of the data used [11], from distribution, frequency, correlation, and more. In practice, curiosity is very important in this process [12], understanding the context of the data is also considered because it will answer the basic problems. In general, EDA [13] is carried out in several ways:

- Univariate analysis (descriptive analysis with one variable)
- Bivariate analysis (Relationship analysis with two variables which is usually the target variable)
- Multivariate analysis (analysis using more than or equal to three variables)

In this research, we will use bivariate analysis by looking for the relationship between 2 main datasets, namely Donald Trump and Joe Biden. First of all, let's look at Figure 1 below to understand the two datasets that we will use.

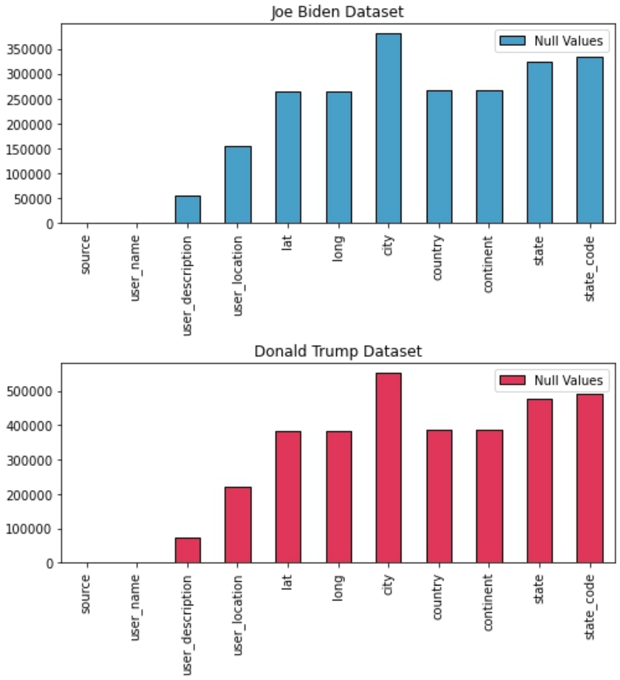

Figure. 1. Joe Biden & Donald Trump Dataset

After that, for the Geospatial features like *lat, longitude, city, country, continent, state code* spatial features are all derived from *user_location*, using the sciSpacy NER and the OpenCage API. We need to note that a number of factors exist such as the *user_location* input there are errors, our method of obtaining useful data from the wrong input of the subject, and its impact on the results returned by the OpenCage API. In order to see the more detailed relation please see Figure 2 below.

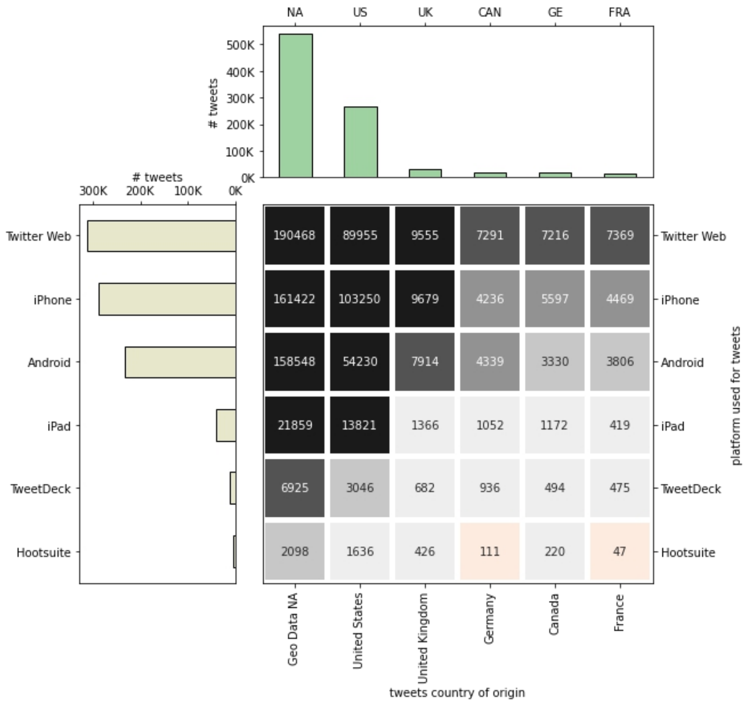

Figure. 2. Geospatial heatmap

From the heatmap above, we find that most of the tweets in the dataset, are published via the web twitter, Iphone, Android, and iPad. Which turned out to be mostly from the US and other countries. We also find that there is a relationship between the platforms used between the locations of publication, is North America tends to use Twitter on the web. We will also analyze the location where the tweets were published. We are successfully plotting all the available geo data reveals that the two presidential candidates are tweeting about many countries around the world.

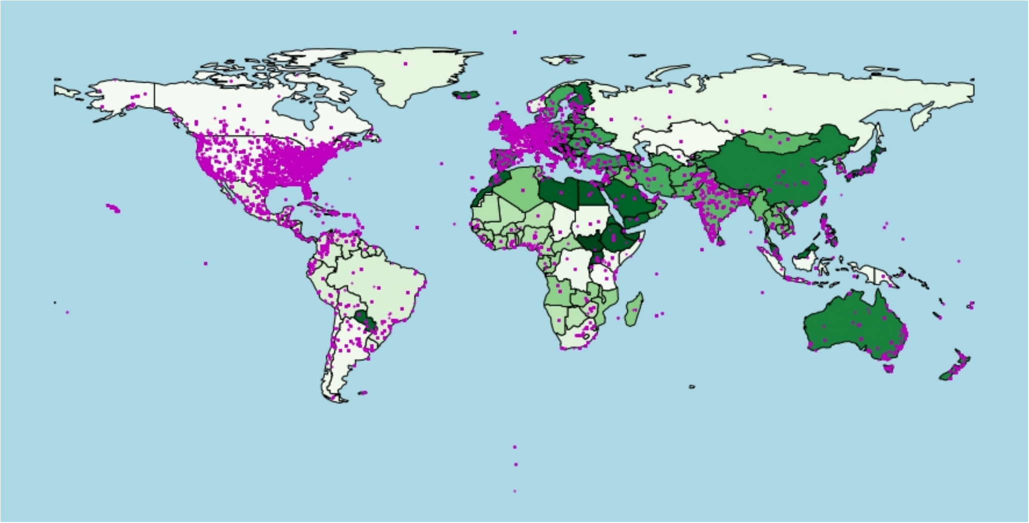

Figure. 3. Geo-data frame

After that, By combining 2 data files and removing duplicates, we can take data samples to analyze the language randomly. the results we get vary widely, there are more than 10 languages used to make these tweets, but to make it easier to do the analysis we will only take the top 5 as a reference for taking the visualization. Please pay attention to figure 4 to be able to see the results on the heat map. In order to understand the language used we use the langdetect function to sample approximately 4000 data, with the hope that it is sufficient to provide a confidence level of more than 90% with an error margin of no more than 1-2%. The heatmap below only illustrates the top 5 languages used from the top 5 countries that may make tweets, from the results we get from approximately 35 detected languages, English makes up almost 80% of all existing tweets.

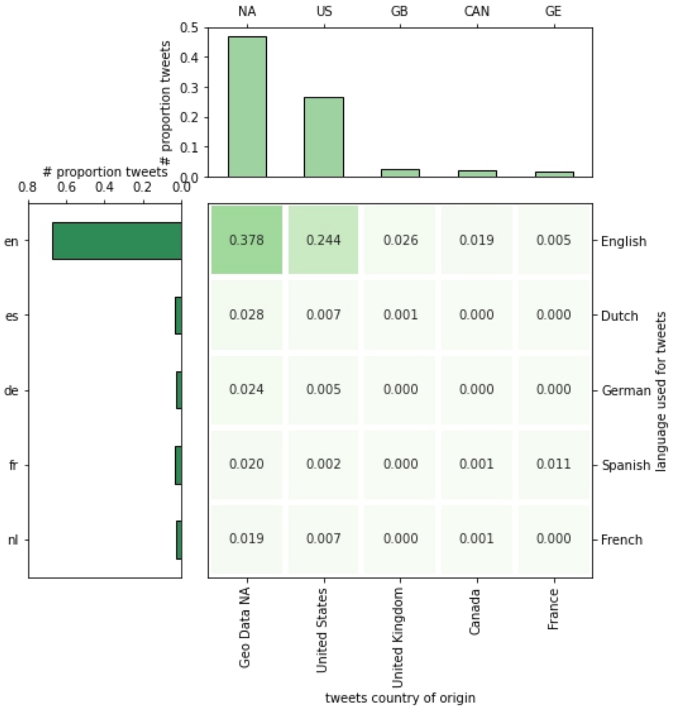

Figure. 4. The language used on tweets.

In order to create a visualization table, we need to identify the common UserId for both datasets. by creating visualization fields, narrowing down the data, and labeling for North America Geo data. We can calculate the number of tweets for each usertype and continent, to understand the results see figure 5. Most of the UserIDs posted both presidential candidates, accounting for greater than 60% of the tweets. Most of the heatmap results below are from tweets in America, at least those with Geo data.

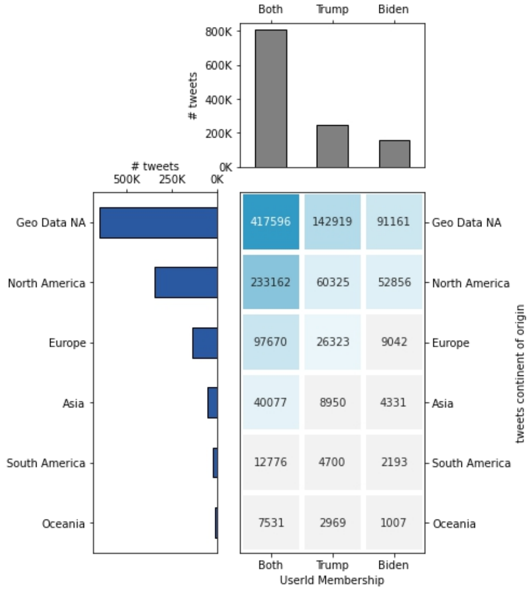

Figure. 5. UserId feed heatmap

In terms of tweets per hour and trend regularity, tweets about both presidential candidates have been reasonably consistent, with significant exceptions on the date during the last presidential debate and when we come up to and on the voting date. There are major changes in frequency consistency on both datasets on the election date itself with a steadily rising trend in tweet frequency for both presidential candidates. As we can see below (figure 6) both presidential candidates share almost the same pattern before and after election day. From the kernel distribution, we found that users in both datasets had the same frequency, but a small number of users in the Biden database posted over 1000 tweets (Figure 7).

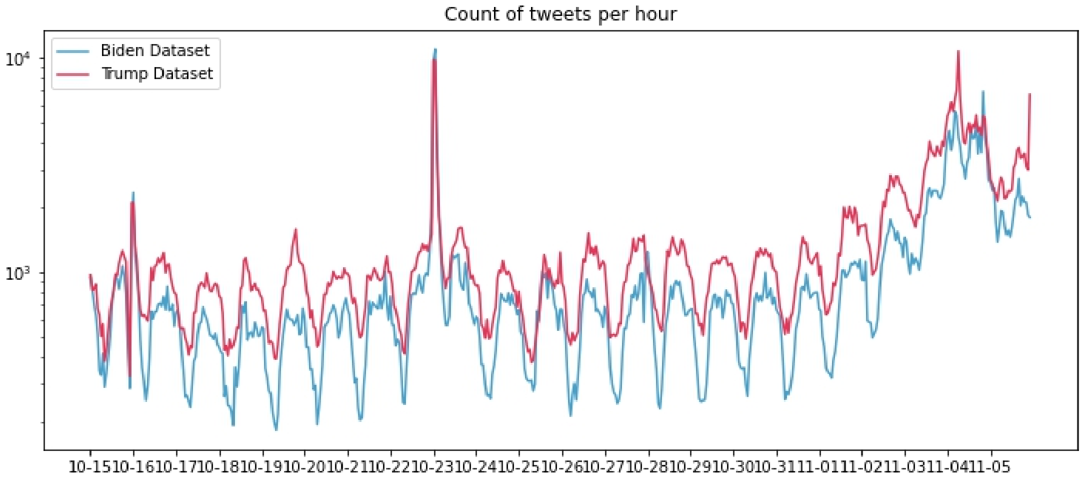

Figure. 6. Count of tweets per hour

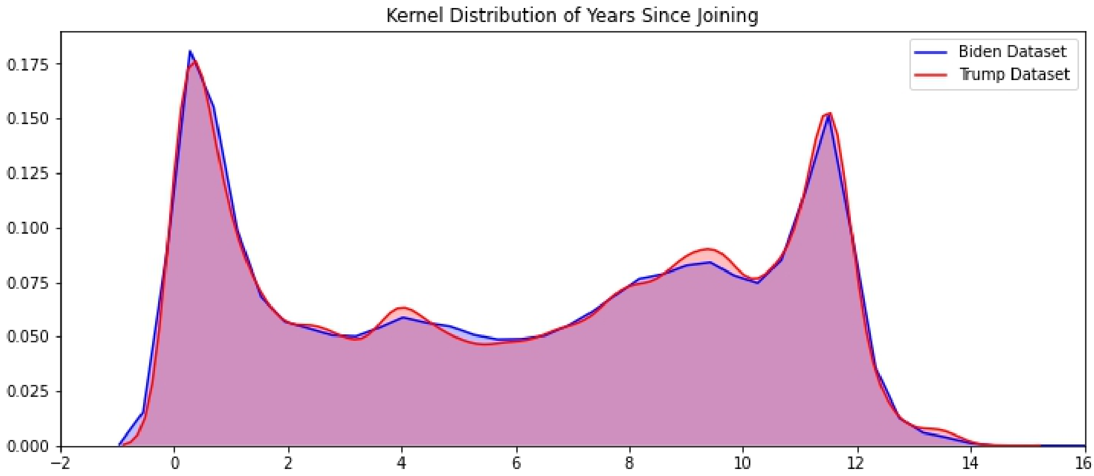

Figure. 7. Kernel distribution

The "Biden" dataset indicates a slight but constant number of posts with higher normalized "retweets" than shown in the "Trump" dataset, similar to the previous visualization. Both datasets have a comparable distribution in terms of years after joiner for users with a high at 12 years and another peak showing users entering Twitter to comment on any of the corresponding presidential candidates.

3.2. N-gram & VADER Sentiment Analysis

For the second part, we are going to perform Sentiment analysis on both datasets [8]. To perform this we will be using Valence Aware Dictionary and Sentiment Reasoner or known as VADER, which is a lexicon and sentiment analysis tool that is specifically tuned to the sentiment expressed in social media [9]. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model that is responsive both to polarity (positive/negative) and intensity (strength) of emotion used for text sentiment analysis. It is included in the NLTK kit and can be directly applied to unlabeled text data.

VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores[10]. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. The sentimental analysis of VADER is based on a dictionary mapping lexical characteristics to emotional intensities known as sentiment scores. A text's sentiment score can be obtained by summing up each word's strength in the text. For instance, Words such as love & peace all obtain a positive sentiment. VADER also is extremely intellectual to recognize the underlying meaning of these words, such as "did not love" as a negative statement. It also recognizes capitalization and punctuation emphasis, like "enjoy".

In order to do some analysis, first, we must clean the tweets to remove stopwords, Strings with “http”, “www.”, “.com”, etc which means obtaining tweets only from data that has Geo data from the US. The bar below shows us the most common Tri N-grams and Bi in each datasets, filtered from word just from the “United States of America”. The N-gram's show a clear relation to the upcoming election and each respective dataset seems to be related to each of the presidential candidates.

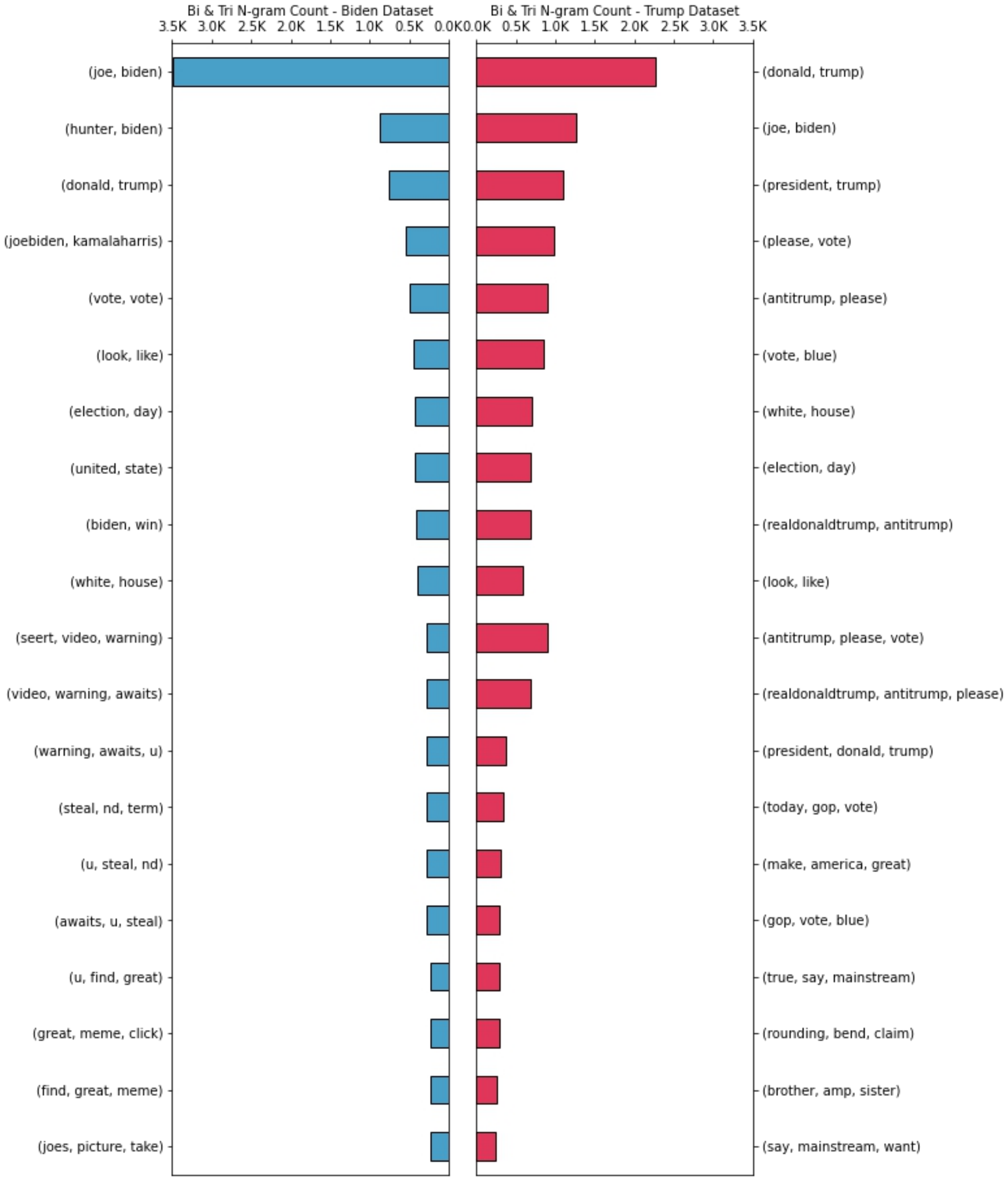

Figure. 8. Bi & Tri N-Gram count

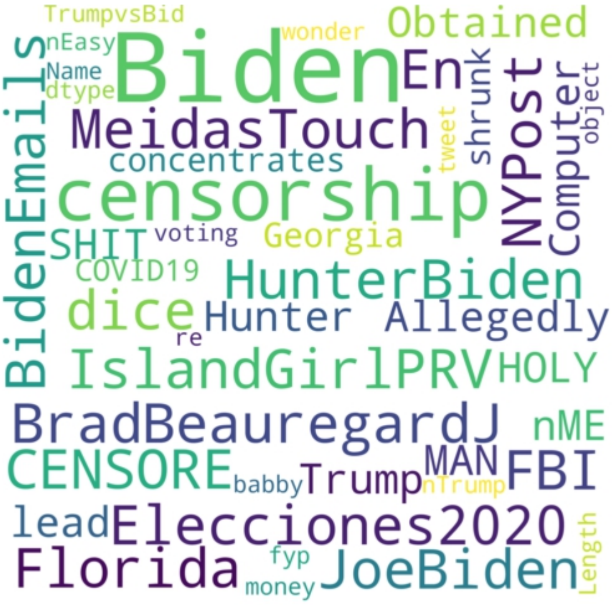

Figure. 9. Biden wordcloud

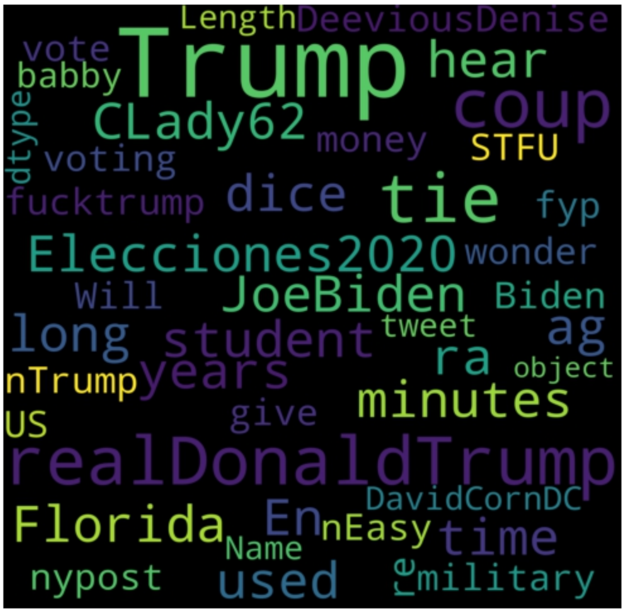

Figure. 10. Trump wordcloud

Figures 9 and 10 above were wordcloud created from both datasets, there is no much difference between both of them but also we cannot gain much information from this. however, by obtaining the sentiment score for both datasets and create a feature that allows masking data for votable states then convert it into a data frame for visualization, VADAR Analysis produces a compound sentiment score. And, also we took the mean compound score for the most recent 2 weeks and the first 2 weeks for each state. See Figure 11 below for the result.

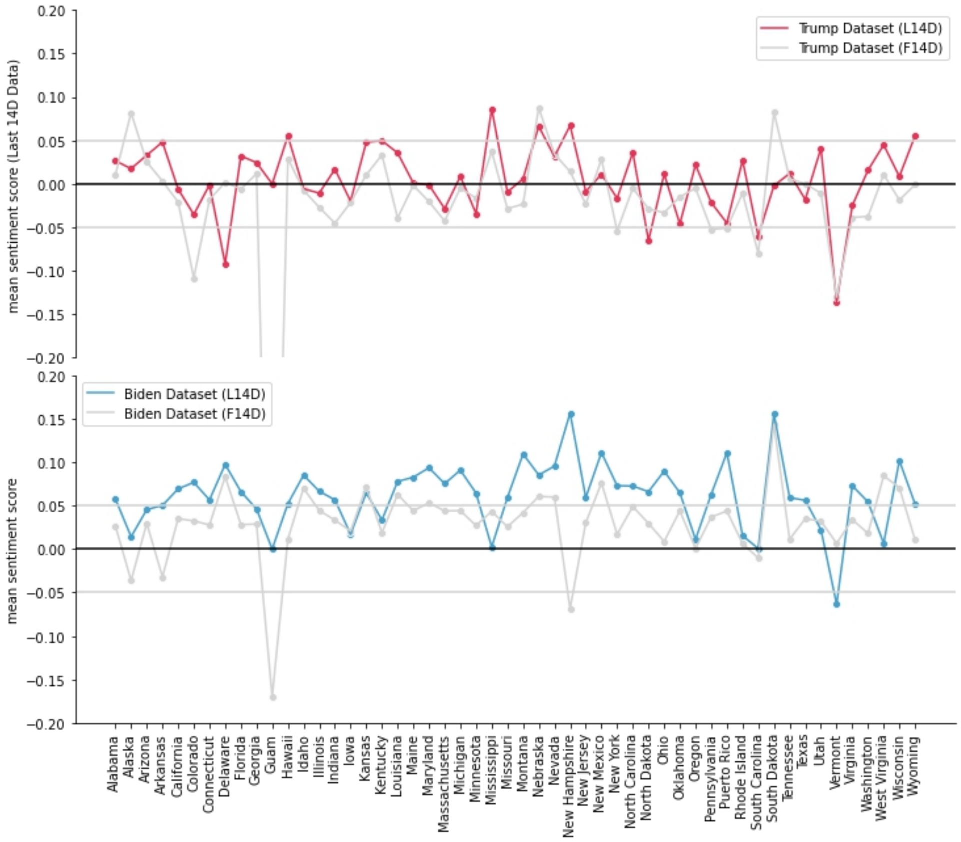

Figure. 11. Compound sentiment score

First of all its should be noted that 0.05 and -0.05 is considered “Neutral”. The results appear to suggest that a significant number of states are switching to a "positive" sentiment score from the previous more "neutral" sentiment for the democratic candidate. Although most countries with the Republican candidate are now essentially "Neutral". By calculating sentiment proportions and feed into dataframes for visualization the visualization below was generated by first assigning each tweet a positive, neutral or negative sentiment then assume those for each day and calculating the proportion for each sentiment group.

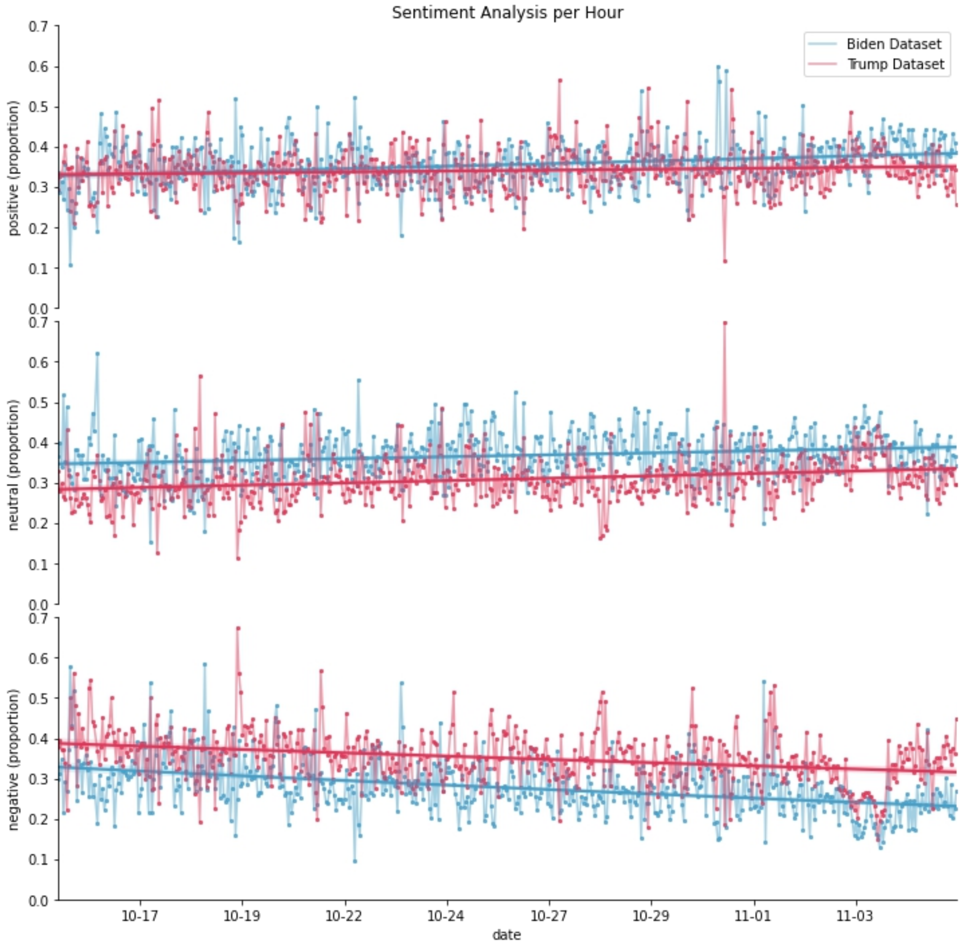

Figure. 12. Sentiment analysis per hour

Then by using logistic regression to find the best fit line we found that, there is an increasing "Positive" and "Neutral" sentiment, with decreasing negative sentiment, which is the pattern over the entire timeframe of the dataset for both presidential candidates. Nevertheless, fascinating changes occurred when we approached but after election day, where we see a faster rise in "Positive sentiment for the democratic candidate over the Republican, where we see a significant difference in the lines of logistic regression developing. The gaps between two candidates remain relatively constant until there is a slight fluke on election day where the gap temporarily drops before returning to the previous steady difference, moving onto the "Neutral" sentiment. We see an increase in "Negative" sentiment for the republican candidate on the post-election day.

4. Results and Discussion

With tweets from 40 different languages, but with a large proportion of tweets in English and originating from the US, there is interest in US elections from many different countries in the world. The sentiment analysis was conducted only on data from the "United States of America" that had geo-data to try to establish the sentiment in each specific dataset and thus each presidential candidate. A significant majority of states were trending to a "Positive" sentiment score for the democratic nominee from the previously more "Neutral"

ISSN 2723-6471

sentiment while analyzing sentiment at the state level as we approached the election date. Whereas several states with the Republican candidate are now relatively Neutral. When viewing the sentiment analysis from a data perspective, this pattern is correctable.

5. Conclusion

From the research above, it is evident that the use of sentiment analysis in analyzing the results of the 2020 US presidential election has high accuracy, the use of sentiment analysis is also very perfect in analyzing any data that has data in the form of text or sentences which of course requires an analysis in it, in this research data We get it from Twitter, maybe for further research it can be done using data in the form of surveys or other social media platforms. The results we get, of course, can still be improved again, compared to neutral, positive, and negative sentiments, it is possible for further research to provide a more specific grouping of sentiments.

The use of EDA and VADER in this research can be said to be quite accurate and can produce good visualization, although of course some further improvements are needed. However, as a conclusion, this Sentiment analysis research is expected to provide sufficient encouragement to readers to carry out similar research, to help election participants independently analyze their election results, and perhaps both candidates to find out what comments or opinions from their voters.

References

[1] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, “A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle,” *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 115–120, 2012, doi: 10.1145/1935826.1935854.

[2] A. G. Greenwald, C. T. Smith, N. Sriram, Y. Bar-Anan, and B. A. Nosek, “Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election,” *Anal. Soc. Issues Public Policy*, vol. 9, no. 1, pp. 241–253, 2009, doi: 10.1111/j.1530-2415.2009.01195.x.

[3] J. Shin, L. Jian, K. Driscoll, and F. Bar, “Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction,” *New Media Soc.*, vol. 19, no. 8, pp. 1214–1235, 2017, doi: 10.1177/1461444816634054.

[4] D. Karol and E. Miguel, “The electoral cost of war: Iraq casualties and the 2004 U.S. presidential election,” *J. Polit.*, vol. 69, no. 3, pp. 633–648, 2007, doi: 10.1111/j.1468-2508.2007.00564.x.

[5] B. S. Kiousis and T. N. York, “New York Times Issue Coverage 2000 U.S. Presidential.pdf,” pp. 71–87, 2004.

[6] M. Cerezo, “Las nociones de Sachverhalt, Tatsache y Sachlage en el Tractatus de Wittgenstein,” *Anu. Filos.*, vol. 37, no. 2, pp. 455–479, 2004, doi: 10.15581/009.37.2.455-479.

[7] E. S. Han and A. goleman, daniel; boyatzis, Richard; Mckee, “US Presidential Election 2012 Prediction using Census CorrectedTwitter Model,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.

[8] J. Wilson and C. Hernández-Hall, “Octava Conferencia Internacional AAAI sobre Weblogs y Redes Sociales,” Eighth Int. AAAI Conf. Weblogs Soc. Media, p. 18, 2014, [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

[9] S. Elbagir and J. Yang, “Twitter sentiment analysis using natural language toolkit and Vader sentiment,” Lect. Notes Eng. Comput. Sci., vol. 2239, pp. 12–16, 2019.

[10] C. W. Park and D. R. Seo, “Sentiment analysis of Twitter corpus related to artificial intelligence assistants,” 2018 5th Int. Conf. Ind. Eng. Appl. ICIEA 2018, pp. 495–498, 2018, doi: 10.1109/IEA.2018.8387151.

[11] D. R. Brillinger, H. K. Preisler, A. A. Ager, and J. G. Kie, “An exploratory data analysis (EDA) of the paths of moving animals,” J. Stat. Plan. Inference, vol. 122, no. 1–2, pp. 43–63, 2004, doi: 10.1016/j.jspi.2003.06.016.

[12] J. T. Behrens, “Principles and Procedures of Exploratory Data Analysis,” Psychol. Methods, vol. 2, no. 2, pp. 131–160, 1997, doi: 10.1037/1082-989X.2.2.131.

[13] T. Astuti and I. Pratika, “Product Review Sentiment Analysis by Artificial Neural Network Algorithm,” IJIIS Int. J. Informatics Inf. Syst., vol. 2, no. 2, pp. 61–66, 2019, doi: 10.47738/ijiis.v2i2.15.