Predicting US Presidential Elections 2020

Avinash Kumar
Student at Bennett University
E18CSE031

Aditya Singh
Student at Bennett University
E18CSE009

Nishtha Dua
Student at Bennett University
E18CSE119

Abstract

Utilizing social media to ascertain its user's opinions over an entity, more specifically Twitter to forecast Trends is a popular field of research. Employing Twitter for election campaign's ,to monitor & predict election result , capturing say of both voters & candidates in real time has escalated over the years. User's share their views , preferences on Twitter voluntarily and is publicly accessible. 'Tweets' are quick, brief real time user updates and can be extracted through Twitter API. Downloading replies on the most recent Tweets of the candidates can serve our purpose. Location parameters can help in obtaining more accurate & credible results. It is optional for a user to share its device location detail's by permitting access to it. Target tweets are derived within the USA region. Tweets obtained are in a subjective or opinionated plain text format. We would target Tweets referring to any of the two candidates namely (Donald Trump or Joe Biden) , or anything in relevance to US Presidential Elections 2020. The parameters for procuring Tweets in the interest of our analysis are CandidateName : Donald Trump or Joe Biden, Language : English , Location : US , DateOfRetrieval : (StartDate , EndDate) , #Tweets to be retrieved, and Name of output .csv comprising of @AuthorName , date of public posting of Tweet & its textual information. We intend to perform a comparative study of several ML Models to observe their performance and understand the public say for the candidates. For classification of Tweet sentiment Polarity : Positive (+1) , Neutral , or '0' ,Negative (-1) measure that indicates a user's view for a particular candidate. Neutral replies with '0' Subjectivity & Polarity does not contribute to our analysis. Hashtags, emoticons do not add any significant value to the classifier. Each Tweet is marked as Positive (+1) , Neutral , or '0' ,Negative (-1). We categorize US states as Dominantly Republican , Dominantly Democratic , to some degree Republican, to some degree Democratic , or US states for which data is not sufficient for a concrete analysis. Candidates are contrasted over the category of dominating sentiment. State-wise aggregation of Tweets after marking each Tweet as Positive (+1) , Neutral , or '0' , Negative (-1) to analyse public say about a particular applicant in a particular state.

I. Introduction

Twitter is a valuable source of user created data. Timely manipulation of the data is of tactical value to several organizations & agencies. The task of obtaining relevant data is of great complexity as it is unstructured. We opted for Twitter due to its huge expanse of relevant data and because it allows us to extract a number of geotagged tweets for locations which enables us to investigate the reasons behind the geographically segregated review. Users broadcast publicly accessible 'Tweets'. User's publicly emote opinions, issues , complaints , criticism , moods , reactions through Tweets . So, it is required to compute total sentiment and hence its association with the Tweet. Hashtags are used for hunting Tweets about a particular entity. '#' itself implicates an emotion. Employing Sentiment Analysis on Twitter voluntarily data companies' device their business tactics & understand users view towards their brand/products. It can also help in forecasting possibly risky circumstances by ascertaining a user's mood in general. We have employed a number of multi-class classifiers for evaluating views about candidates and our model is trained for each competitor singly. We intend to estimate a user's preference on a particular entity and conduct a comparative study of a number of ML models by contrasting their performances. A tweet can possibly be Positive(+1) for one candidate but Negative(-1) for another. Tweets are extracted making use of Twitter API. Classified Tweets can help us predict results. Target tweets must refer to the elections and candidates. Our research deals with checking how many negative and positive keywords are in a Tweet text. Candidates are contrasted over the category of dominating sentiment. State-wise aggregation of Tweets after marking each Tweet as Positive (+1) , Neutral , or '0' , Negative (-1) to analyse public say about a particular applicant in a particular state. Accuracy of text classification algorithms depends on the labelled training data..

II. Related Work

1. In the field of population health, Twitter data can prove useful. The candidacy of Twitter users to address personal issues, such as oral health or emotional status, suggests that this medium can be used by health care providers as a method to track the health of the public and interact with patients.
2. Twitter offers a simple way to gain a broad understanding of the behaviour's and attitudes of a community. This source of human data that is organically generated and automatically preserved enables researchers to see what people do, what they think, and how they feel

about specific problems when these behaviours and thoughts trigger.

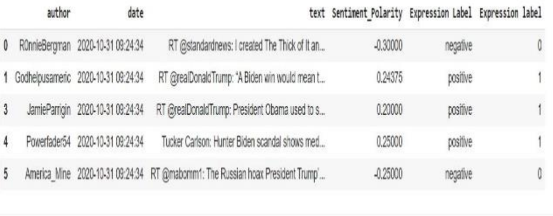
3. The easiest way to observe what people are thinking about your organisation on the Internet is brand tracking as customers voice their opinions on Twitter about brands and products.
4. Monitoring of rivals as it allows businesses to contemplate what their rivals are up to and come up with counterplans and track any number of competitors.
5. Sentiment analysis is basically the practice of describing the emotional tone of a set of terms used in an online mention to obtain an understanding of the thoughts, feelings, and attitudes conveyed.
6. Twitter is a forum where people share their views on brands, goods, or significant world events. Sentiment analysis may be conducted to classify public opinions on massive tweet data-sets.
7. Twitter has a profusion of content created by users which makes it an ideal source of data for training in machine learning.
8. The future of your business needs the foresight to keep ahead of the competition. In this fast-paced business climate, this has become all the more important.
9. Twitter is one of the most popular social media sites out there, representing all the current events and trend data important to companies.
10. To forecast stock market behaviour by evaluating people's mood on Twitter.
11. Twitter data in political analysis and election forecasting. Discover a user's political or ideological orientation, then apply it to the option. Use the chosen tweet relevant to the upcoming election and use that detail to determine the user's voting preference.
12. Various types methods can applied to identify the political leanings, such as profile of the person, user posts, Twitter-specific feature (reply / and re-tweet) and tweet content sentiment.
13. The East Java Governor Election which will will held in 2018 will be simultaneously visible on Twitter . All people will argue about their respective governor candidates, which will be both positive and negative comments.
14. A huge amount of text data is gathered with the help of surveys, comments, and reviews over the social media. All of the gathered data is used to improve products and services provided by both private organizations and governments around the world. XV. Twitter is the most used microblogging web site by the anyone to post their thoughts and opinions .

III. MOTIVATION

The recognition of social media and networking networks as critical means of communication and sharing, they have contributed significantly to the decision-making process in different domains. Here, we are interested in catching the pattern of opinion on Twitter. For this reason, the US Presidential Election 2020 enables us to access the pattern of Twitter views. We establish a supervised learning approach to evaluate the opinions of Twitter users where we do not attempt to identify tweets as voicing positive or negative feelings but as endorsing or opposing one of the top two candidates: Donald Trump and Joe Biden. In addition, Twitter is a source of user-generated data and almost 500 million tweets are posted on Twitter every day. Twitter, along with contextual and social metadata and user accessibility and interaction information, offers multimodal data containing text, images, and videos. The data created by this plays an important role in making sense of public opinions and reactions to contemporary questions. In several fields of use, Twitter data can be used for predictive analysis, from personal and social to public health and policy. Twitter data predictive analytics involves a set of techniques to derive knowledge and patterns from data, and forecast trends, future events, and historical data-based behaviour. A recurring activity is the use of the Twitter micro-blogging site as a method to forecast the effects of social phenomena. Twitter for predictive tasks: stock market forecasting, movie sales, and identification of pandemics. These emerging digital contexts include broad scale data sets with millions of users.

IV. DATASET DESCRIPTION

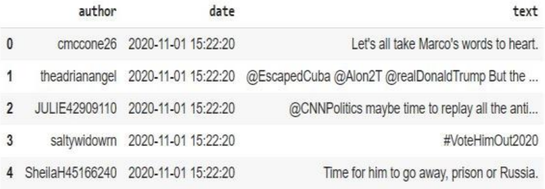
We have Four sets of datasets for our research-TrumpvsBiden, Trump, Biden,data. ‘TrumpvsBiden’ is for training the model which has columns like name, text, polarity and sentiments which is used on naïve bayes, SVM and BERT method.



	author	date	text	Sentiment_Polarity	Expression Label	Expression label
0	R0nnieBergman	2020-10-31 09:24:34	RT @standardnews: I created The Thick of It an...	-0.30000	negative	0
1	Godhelpusameric	2020-10-31 09:24:34	RT @realDonaldTrump: "A Biden win would mean t...	0.24375	positive	1
3	JamieParrigin	2020-10-31 09:24:34	RT @realDonaldTrump: President Obama used to s...	0.20000	positive	1
4	Powerfader54	2020-10-31 09:24:34	Tucker Carlson: Hunter Biden scandal shows med...	0.25000	positive	1
5	America_Mine	2020-10-31 09:24:34	RT @mabomm1: The Russian hoax President Trump'...	-0.25000	negative	0

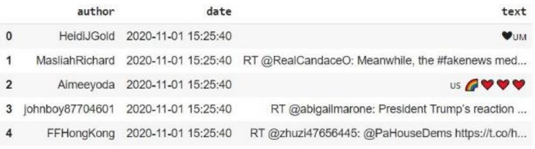
Fig: TrumpvsBiden dataset

Another Set of dataset just contain name and text which were used on the VADER and TextBlob packages to analyse the sentiments of the tweets. This dataset was generated in the real time using tweepy API. Separate dataset for Donald Trump and Joe Biden were collected at the same time to analyse them using VADER and TextBlob packages.



	author	date	text
0	cmccone26	2020-11-01 15:22:20	Let's all take Marco's words to heart.
1	theadrianangel	2020-11-01 15:22:20	@EscapedCuba @Alon2T @realDonaldTrump But the ...
2	JULIE42909110	2020-11-01 15:22:20	@CNNPolitics maybe time to replay all the anti...
3	saltywidowrn	2020-11-01 15:22:20	#VoteHimOut2020
4	SheilaH45166240	2020-11-01 15:22:20	Time for him to go away, prison or Russia.

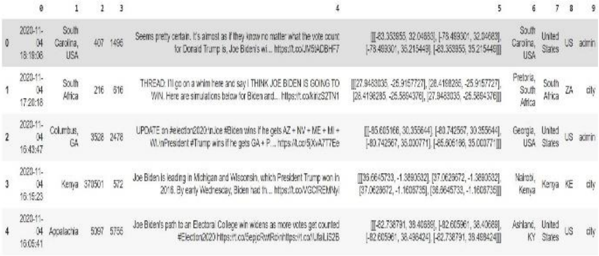
Fig: Trump dataset



	author	date	text
0	HeidiJGold	2020-11-01 15:25:40	♥UM
1	MasliahRichard	2020-11-01 15:25:40	RT @RealCandaceO: Meanwhile, the #fakenews med...
2	Aimeeyoda	2020-11-01 15:25:40	US 🌈❤❤❤
3	johnboy87704601	2020-11-01 15:25:40	RT @abigailmarone: President Trump's reaction ...
4	FFHongKong	2020-11-01 15:25:40	RT @zhuzi47656445: @PaHouseDems https://t.co/h...

Fig: Biden Dataset

We have used another dataset where we passed got longitude, latitude and location of the tweets of users along with their names and tweets.



	0	1	2	3	4	5	6	7	8	9
0	2020-11-04 18:18:06	South Carolina, USA	407	1496	Seems pretty certain. It's almost as if they know no matter what the vote count for Donald Trump is, Joe Biden's wi... https://t.co/JM5tADBHF7	[[[-83.353955, 32.04683], [-78.499301, 32.04683], [-78.499301, 35.215449], [-83.353955, 35.215449]]]	South Carolina, USA	United States	US	admin
1	2020-11-04 17:20:18	South Africa	216	616	THREAD: I'll go on a whim here and say I THINK JOE BIDEN IS GOING TO WIN. Here are simulations below for Biden and... https://t.co/kirizS2TN1	[[[27.9483035, -25.9157727], [28.4198285, -25.9157727], [28.4198285, -25.5894376], [27.9483035, -25.5894376]]]	Pretoria, South Africa	South Africa	ZA	city
2	2020-11-04 16:43:47	Columbus, GA	3528	2478	UPDATE on #election2020:\nJoe #Biden wins if he gets AZ + NV + ME + MI + WI.\nPresident #Trump wins if he gets GA + P... https://t.co/5jXvA7T7Ee	[[[-85.605166, 30.355644], [-80.742567, 30.355644], [-80.742567, 35.000771], [-85.605166, 35.000771]]]	Georgia, USA	United States	US	admin
3	2020-11-04 16:15:23	Kenya	370501	572	Joe Biden is leading in Michigan and Wisconsin, which President Trump won in 2016. By early Wednesday, Biden had th... https://t.co/VGCIREMNyi	[[[36.6645733, -1.3890532], [37.0628672, -1.3890532], [37.0628672, -1.1608735], [36.6645733, -1.1608735]]]	Nairobi, Kenya	Kenya	KE	city
4	2020-11-04 16:05:41	Appalachia	5097	5756	Joe Biden's path to an Electoral College win widens as more votes get counted #Election2020 https://t.co/5epcRwtRoi\nhttps://t.co/UfaLiS2B	[[[-82.738791, 38.40689], [-82.605961, 38.40689], [-82.605961, 38.498424], [-82.738791, 38.498424]]]	Ashland, KY	United States	US	city

Fig: Data

Since not many users keep their location on while tweeting, the sample size of this dataset is very small and it has been used to just serve our research purpose.

V. Pre-Processing steps

As we can see in the dataset the tweets are very ugly i.e. it contains RT, links, emoticons, symbols etc. Our model can't understand these symbols. So we need to clean the dataset. Our pre processing function will first remove the user handles, symbols, emoticons, brackets etc. from the tweets and convert all the text into lower case. Before feeding the tweets to our model, we also did stemming and TF-IDF vectorization. Stemming means normalising the words into its actual form by cutting suffix like 'ing','es' etc. Vectoring the word is also an important step before training any model in SVM and Naïve Bayes mehod.

TF(Term Frequency): The frequency of any word divided by the total number of words in the corpus. It is given by-

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

IDF(Inverse Data Frequency): Unlike term frequency, inverse data frequency lays emphasis on giving weights to rare words. It is given by-

$$idf(w) = log(\frac{N}{df_t})$$

N= Number of documents
df= Number of documents containing the word w.

TF-IDF is calculated by multiplying TF with IDF-

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

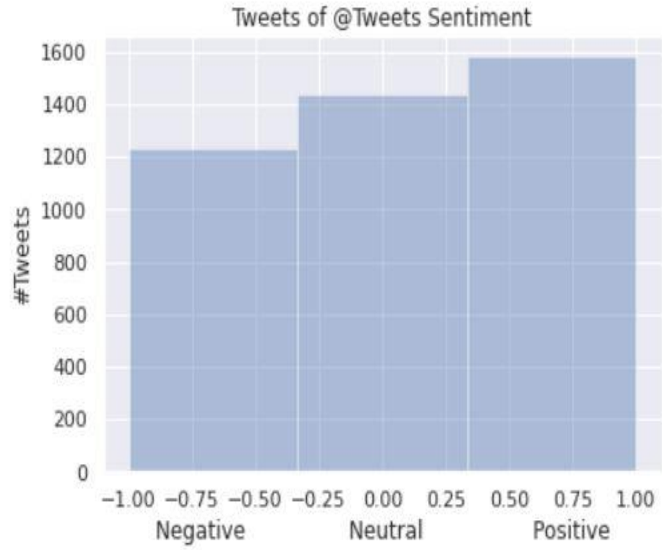
VI. Methodologies

We used various Machine Learning and Deep Learning models for predicting the sentiments of people towards the US Presidential Candidates and observed different accuracy for different methods.

A. *Vader Sentiment Analysis*

VADER(Valence Aware Dictionary for Sentiment Reasoning) is a lexicon based method to analyse sentiments based on the positive, negative and neutral metrics. The good thing about VADER is that it doesn't need training data. It can be applied directly on the streaming data and it fits perfect to be used on the social media data. When a text is passed through VADER method SentimentIntensityAnalyser(), it generates a dictionary of positive, negative, neutral and compound scores. The compound score is basically the normalisation of the positive, negative and neutral score.

When the text column containing tweets was passed through the above VADER method, compound scores were generated. To categorize them into 3 sentiments that is positive, neutral and negative , a function was written where it was defined that if the compound score is above 0.05, the tweet will be marked as positive and if it is less than -0.05, it will be classified as negative. Otherwise the tweet will be neutral. The overall dataset contained more than 4000 tweets and when those tweets were plotted against the respective sentiments, the following observations were made:



More than 1200 tweets were positive and almost 1600 tweets were positive. The next step was to plot the wordcloud of the respective sentiments. The wordcloud gives the larger size output for the most frequently occurring tokens and small size to the less occurring ones. The idea was to observe who is mentioned more in the negative tweets and who is mentioned more in the positive tweets. It was followed by extracting positive hashtag and negative hashtag from the tweets and analysing who (Trump or Biden) is mentioned in those hashtags.

B. *Textblob*

TextBlob is a python library used in text mining. It uses Natural Language ToolKit for achieving the results. TextBlob gives polarity[-1,1] and subjectivity[0,1] of texts. While polarity signifies the positive and negative sentiments, subjectivity defines about the text relevance. To analyse the sentiments for and against Donald Trump and Joe Biden, we have two separate datasets of them collected at the same time. We calculated the sentiment polarity on both the datasets and added columns for the respective sentiments based on the polarity of those tweets. To analyse it more accurately, we then deleted all the neutral tweets from both the datasets and then we balanced the datasets by picking the random 4000 columns from each dataset to analyse the sentiments against or for Donald trump and Joe Biden. We then compared the results of both i.e. plotting the corresponding positive and negative sentiments.

C. *Naïve Bayes method*

Bayes Theorem is based on the **Naive Bayes** algorithm which is a supervised machine learning algorithm. It is called 'naive' since it makes some assumptions that are naive-

- The features do not depend on one another.
- Output has equal contribution from each feature.

Using Bayes Theorem we can calculate the conditional probability of an event. It is finding probability of an event given how it is related to other events. Bayes Theorem formula is—

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

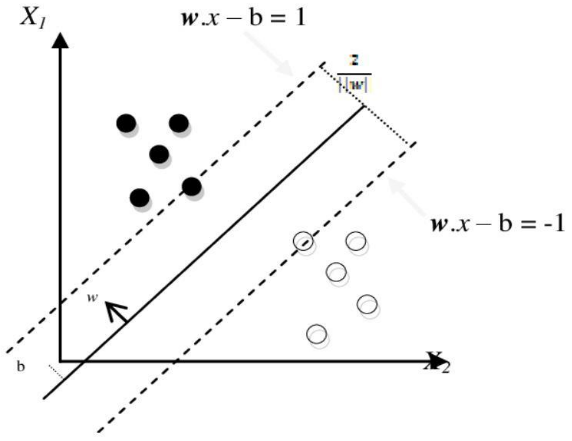
To see how we can use Bayes theorem in Naïve Bayes, we will be analysing classification problem dataset with input x1,x2,x3…….xn and output y. Using Bayes Theorem output y is-

$$P\left(y|x_1,\ldots,x_n\right) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Our dataset contain two extra columns which has sentiments(Positive and negative) and sentiment scores(+1,-1). Unlike VADER and textblob, we require to train our dataset here. To train the dataset we are using Multinomial Naïve Bayes method. For text classification of data that is distributed multinomially we use MultinomialNB. Before applying MultinomialNB we need to do feature engineering. Generally we find out how many times a word is there in the corpus Although we can also use TF-IDF for calculating the frequency as well as significance of the word. Training and testing data was split in the ratio of 80:20 and the model was trained on the sklearn's multinomial Naïve Bayes package.

D. *SVM(Support Vector Machine)*

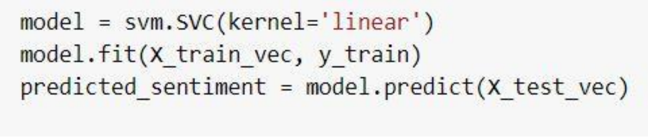
Another method where we tested the performance is Support Vector machine. This supervised machine learning algorithm was chosen because it works well for text classification and is capable enough to handle large features. In out dataset it is being used to classify the tweets into positive and negative. Support Vector Machine has a unique way of drawing hyperplane for classification and differentiating those classes.



optimal separating hyperplane between two classes

SVM uses various types of mathematical functions to convert the datapoints into hyperplanes called Kernels. Some of the widely used kernels are - linear, non-linear ,RBF, polynomial etc.

Here we are creating a linear Support Vector Machine model to train our dataset and predict the output.



```
model = svm.SVC(kernel='linear')
model.fit(X_train_vec, y_train)
predicted_sentiment = model.predict(X_test_vec)
```

E. *BERT*

Bert stands for Bidirectional Encoder Representations from Transformers. This model is pretrained on Wikipedia Corpus which contains 2,500 million words and Book Corpus which contains 800 million words that really boosts its accuracy during training. It is bidirectional meaning that it learns information from left as well as right side of each token at the time of training. It works on transformer architecture.

There are two strategies Bert uses while training:-

a. Masked Language Model

b. Next Sentence Prediction

In Masked Modelling Language we substitute 15% of words with a [MASK] token before inputting the sequence of words and then try to guess the original value of masked words, on basis of context provided by the other, non-masked, words in the sequence. Its convergence is slower than directional models as it only considers prediction of masked tokens and not of the non masked tokens.

In Next Sentence Prediction, we input sentence pairs in the model and the model learns to forecast if in the original document second sentence is the next sentence or not.

To reduce the combined loss function of these two strategies both the methods are used together during training of the BERT model.

We used the Bert model since it gives much higher accuracy than the other machine learning models like SVM, Naive Bayes etc. We first pre-processed our data and then applied the padding and vectorization on our dataset to make our data understandable by the model. We then split the dataset into 90% train and 10% test. Then we applied the Bert model on our dataset . We trained the model for 5 epochs . In each epoch the accuracy improved and we finally achieved an accuracy which is far higher than those we achieved from ML models.

VII. RESULTS

We implemented 5 methods to analyse the trends of the US Presidential election 2020- VADER, TextBlob, Naïve Bayes, Support Vector Machine and BERT. While the First two didn't require any training of the model the other three required us to train the model. Therefore we have various types of metrics as output.

VADER:

Using the VADER method, we calculated the total number of positive, negative and neutral tweets. Now by using different approaches and graphs we would see how Trump and Biden are mentioned in those positive and negative sentiments.

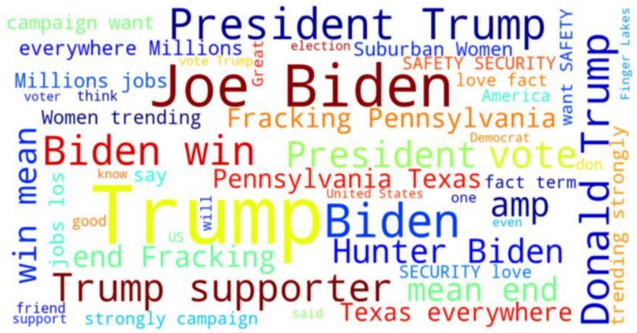


Fig: Word Cloud of Positive Tweets

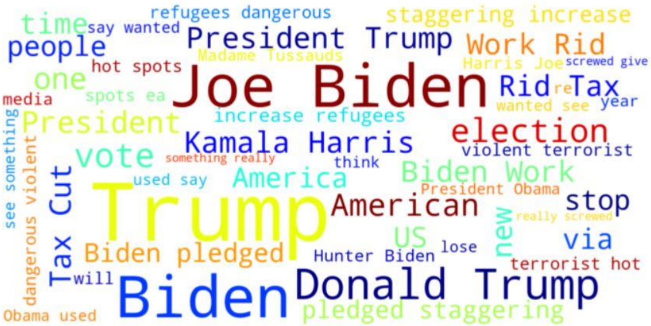


Fig: Word Cloud of negative Tweets

The word cloud gives the frequency of the most occurring word in the form of the larger text. In the positive tweets's word cloud, we cannot much differentiate between the size of words containing Trump and Biden. It shows that both are mentioned too many times in positive tweets and there is immense support for both. Similarly in the word cloud of the negative tweets it is hard to differentiate. That means hate for them is also immense against different people of different ideologies. So we'll differentiate them in some other way.

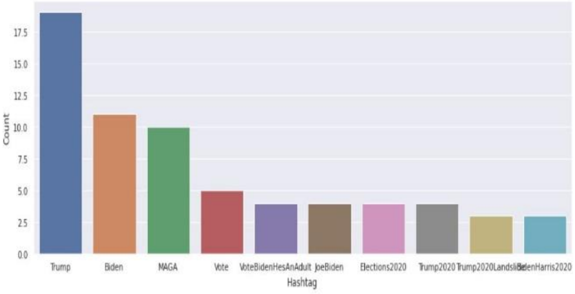


Fig: Positive Hashtag

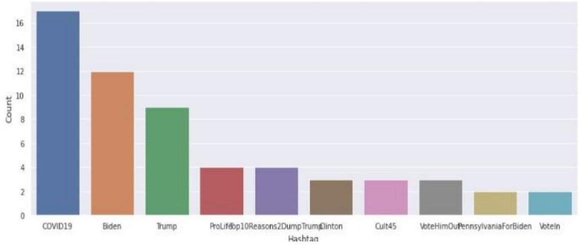


Fig: Negative Hashtag

We can see that Trump is mentioned more in the positive hashtag tweets and the negative hashtag graph conforms our basis because it shows that Biden is mentioned more in those. Although the difference between Trump and Biden is not much.

TextBlob

Using the textBlob package, we got quite interesting results. The positive, negative and neutral tweets were separately plotted for Trump and Biden.
Green-positive
Red- Negative
Blue- Neutral

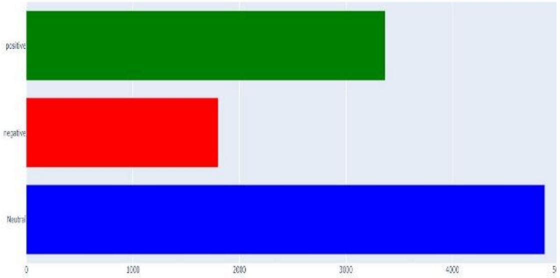


Fig: Trump dataset

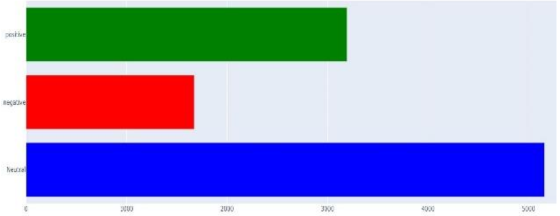
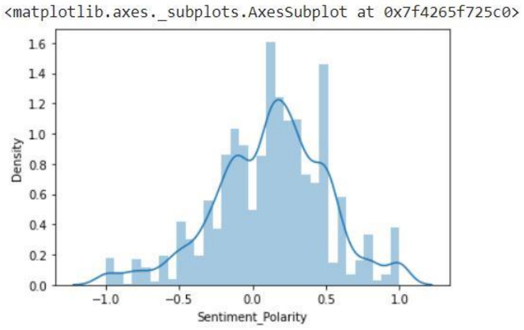


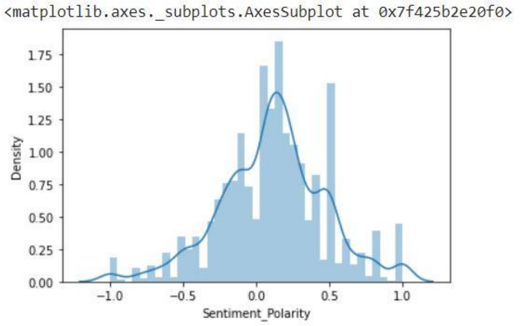
Fig: Biden dataset

As we can see from the graph, neutral tweets in both the datasets are approximately 5000 and our research doesn’t require those neutral tweets. So we dropped all the neutral tweets from both the datasets and scaled the datasets to 4000 rows each to analyse. When we plot the polarity of those tweets against the density function, we got the following graphs-



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4265f725c0>
```

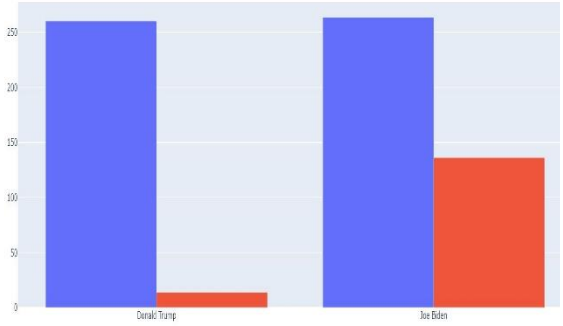
Fig: For Trump



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f425b2e20f0>
```

Fig: For Biden

From the above two graphs, we conclude that most of the sentiments polarity lies between -0.5 to +0.5.
After scaling both the datasets to 4000 rows, we then compare the positive sentiments for Trump and Biden –



We found that Biden has got slightly more positive sentiment than Donald Trump but interestingly negative sentiments for Biden is way more than Donald Trump.

Naïve Bayes, SVM and BERT

In these 3 methods we have trained our model and did hyperparameter tuning to obtain better results.
The evaluation criteria will be accuracy, f1-score, recall and precision.
These four metrics will be used for evaluating our sentiment analysis model.
Accuracy: It is the measure of correct predictions by our model divided by the total number of all predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: It is the measure of the correctly marked positive cases divided by the total number of predicted positive cases.

$$precision = \frac{TP}{TP + FP}$$

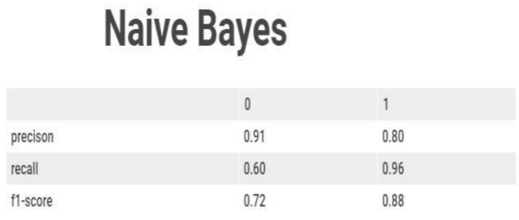
Recall: It is the measure of the correctly marked positive cases divided by the total positive cases.

$$recall = \frac{TP}{TP + FN}$$

F1-score:

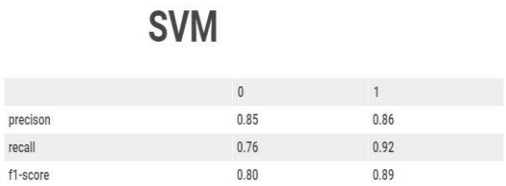
$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

When we evaluated the classification report of each model, we got some interesting table regarding precision, recall and f1-score-



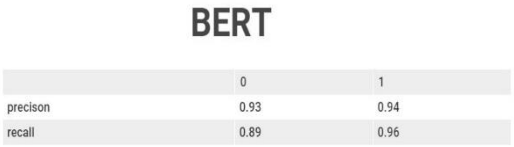
Naive Bayes

	0	1
precison	0.91	0.80
recall	0.60	0.96
f1-score	0.72	0.88



SVM

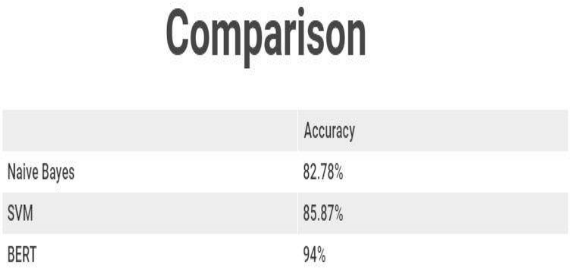
	0	1
precison	0.85	0.86
recall	0.76	0.92
f1-score	0.80	0.89



BERT

	0	1
precison	0.93	0.94
recall	0.89	0.96
f1-score	0.91	0.95

We can see that as we are moving towards more complex and the more reliable model, our metrics's values are increasing. Now lets have a look at the accuracies of our different model:

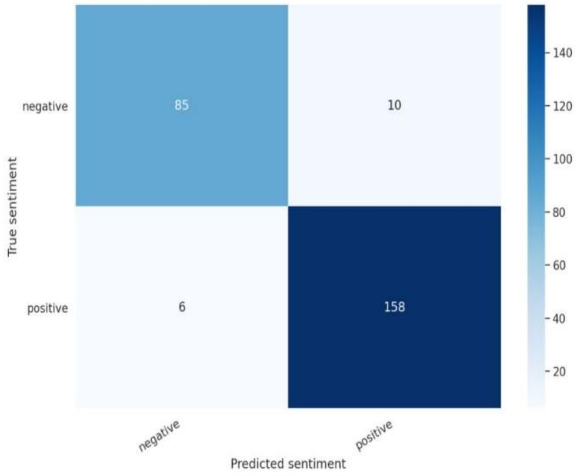


Comparison

	Accuracy
Naive Bayes	82.78%
SVM	85.87%
BERT	94%

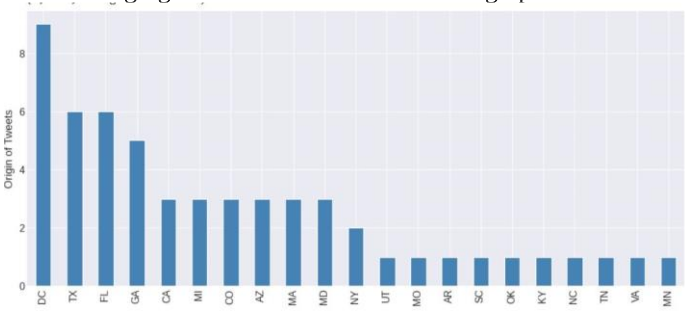
We can see that our naïve bayes model almost gives the accuracy of 83% and Support vector Machine almost gives the accuracy of 86% but BERT performs exceptionally well and gives the accuracy of 94%. This was bound to happen also because BERT also considers the after and before words to predict the sentiment of the current word and hence is used widely because of its good accuracy.

We can also see the confusion matrix of our BERT model to see how it evaluated and performed on our test data-

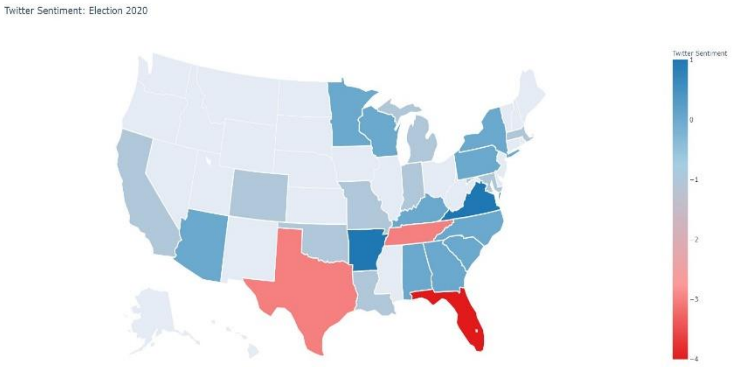


Results on Maps

We also have a fourth dataset whose sample size is small but it contains location , latitude, longitude of the user tweeting. We then parsed the tweets and plotted on the tweets of those users belonging to US states. One look at the graph-



We can see the places in US from where the maximum number of users tweeted. We then parsed the total sentiments for Trump based on the location. Then using folium library we added the latitude and longitude of US states json file and plot those sentiments on the US map. Our final map looks like-



CONCLUSION

We tried different methods to determine people's sentiments regarding US Presidential Candidates based on their tweets. We also used different types of datasets and evaluated them on different models like SVM, Naïve Bayes, Textblob and BERT and we found out that BERT provided the highest accuracy since it is a bidirectional model so maintains the context on both side on word(right and left) so it could understand different meaning of same word and give proper classification of text . However our goal was not just to find the highest accuracy method but to see how different models performed on our dataset and why they performed in such manner.

ACKNOWLEDGMENT

We would like to thank the Computer Science department of Bennett University to give us the opportunity to do this project. We would specially like to thank Professor Vipul Mishra who guided us in our project. We would also like to thank Mohit Bhat who gave us his timely advise on how to approach our project.

REFERENCES

1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135. Retrieved from https://www.scirp.org/(S(czeh2tfqyw2orz553k1w0r45))/reference/ReferencesPapers.aspx?ReferenceID=2289762

[2] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proceedings of the tenth ACM international conference on Knowledge discovery and data mining, Seattle, 2004, pp. 168-177. Retreived from https://www.scirp.org/(S(czeh2tfqyw2orz553k1w0r45))/reference/ReferencesPapers.aspx?ReferenceID=2289762

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 2002, pp. 79-86. Retrieved from http://www2.denizyuret.com/bibtex.php?field=booktitle&fn=select&value=Proceedings+of+the+ACL-02+conference+on+Empirical+methods+in+natural+language+processing-Volume+10

[4]Ali Hasan1, Sana Moin1,"Machine Learning-Based Sentiment Analysis forTwitter Accounts ", Retrieved from https://res.mdpi.com/d_attachment/mca/mca-23-00011/article_deploy/mca-23-00011-v4.pdf

[5]Kia Dashtipour Scotland, United Kingdom "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques", Springer, 2016

[6]Kigon Lyu Korea University, Korea "Sentiment Analysis Using Word Polarity of Social Media", Springer, 2016

[7]Monu Kumar Thapar University, Patiala "Analyzing Twitter sentiments through big data", IEEE, 2016

[8]V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1345-1350, doi: 10.1109/SCOPES.2016.7955659.

[9]Jorge A Balazs University of Chile "Opinion Mining and Information Fusion- A survey", 2015

[10]Donglin Cao Xiamen University, China "A cross-media public sentiment analysis system for microblog", Springer, 2014

[11]Min Chen Huazhong University of Science, China "BigData: A Survey", Springer, 2014

[12]Rafeeque Pandarachalil Govt. College of Engineering, Kannur "Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach", Springer, 2014

[13]Seven Rill Goethe University Frankfurt, Germany "PoliTwi- Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis", Elsevier, 2014

[14] Gurshobit Singh Brar ," Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques",International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12788-12791 © Research India Publications. http://www.ripublication.com

[15]Giuseppe Di Fabbrizio A&T Research Labs, USA "Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling", Digital Object Identifier IEEE, 2013