Analyzing and Securing Social Networks– CS6301

Twitter Sentiment Analysis of US Presidential Elections 2016

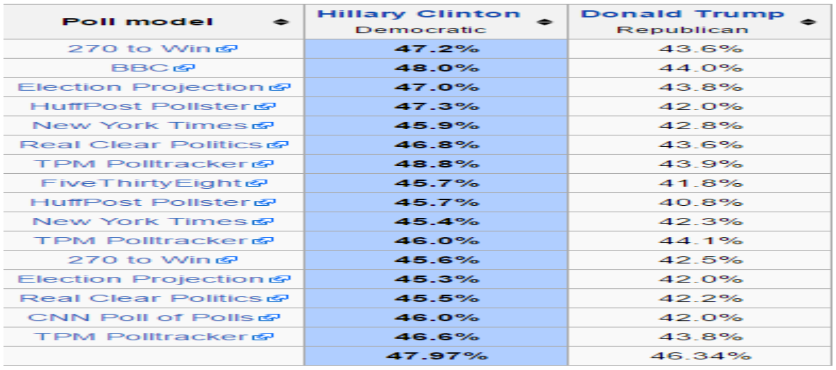
Krishna Chaitanya Dodda, Srinivas Varupula , Venkata Sai Chaitanya Vaddella

1. INTRODUCTION

The USA Presidential Elections 2016 was one of the most hotly contested elections in the history of the country. There were many opinion polls and analysis conducted by various media channels and research groups to predict who would actually win the elections. This opened up a new arena of opportunity for us to involve and participate along with the other researchers and all other people elsewhere to predict and compare the results. Before we dig into the details of the project, a brief description of the presidential election is needed. The presidential election of the United States occurs every 4 years, it is held the first Tuesday after the first Monday in November. The 2016 Presidential election was held on November 8, 2016. The tally of those votes, popular vote, does not determine the winner. Instead, Presidential elections use the Electoral College. To win the election, a candidate must receive majority of electoral votes. In the event no candidate receives the majority, the House of Representatives chooses the President and the Senate chooses the Vice President. In this context, we wish to analyze the predominant sentiment of voters with respect to a candidate and draw reasonable conclusions from it. We picked twitter as our choice of data source given its freely available API to crawl the website and also the growing volume of voters expressing their views on twitter. Our choice of source was also justified by the fact that both presidential candidates were themselves very active on twitter with a huge volume of followers. We analyzed the sentiments of potential voters to predict the outcome of the elections apart from picking suitable trends to aid the electoral process.

2. BACKGROUND

We will now look at the actual results and how the opinion pools and exit polls have predicted the results of the presidential election of 2016. The opinion polls are the surveys conducted by different organization by collecting the opinion of many different people among all the states across the United states of America. Mostly all the opinion polls have stated that the Hillary Clinton is going to win as the majority of the people from all the states have opted her, this can be seen in the following snapshot which indicates that Hillary is on the lead in most of the core states needed to win the presidential election.



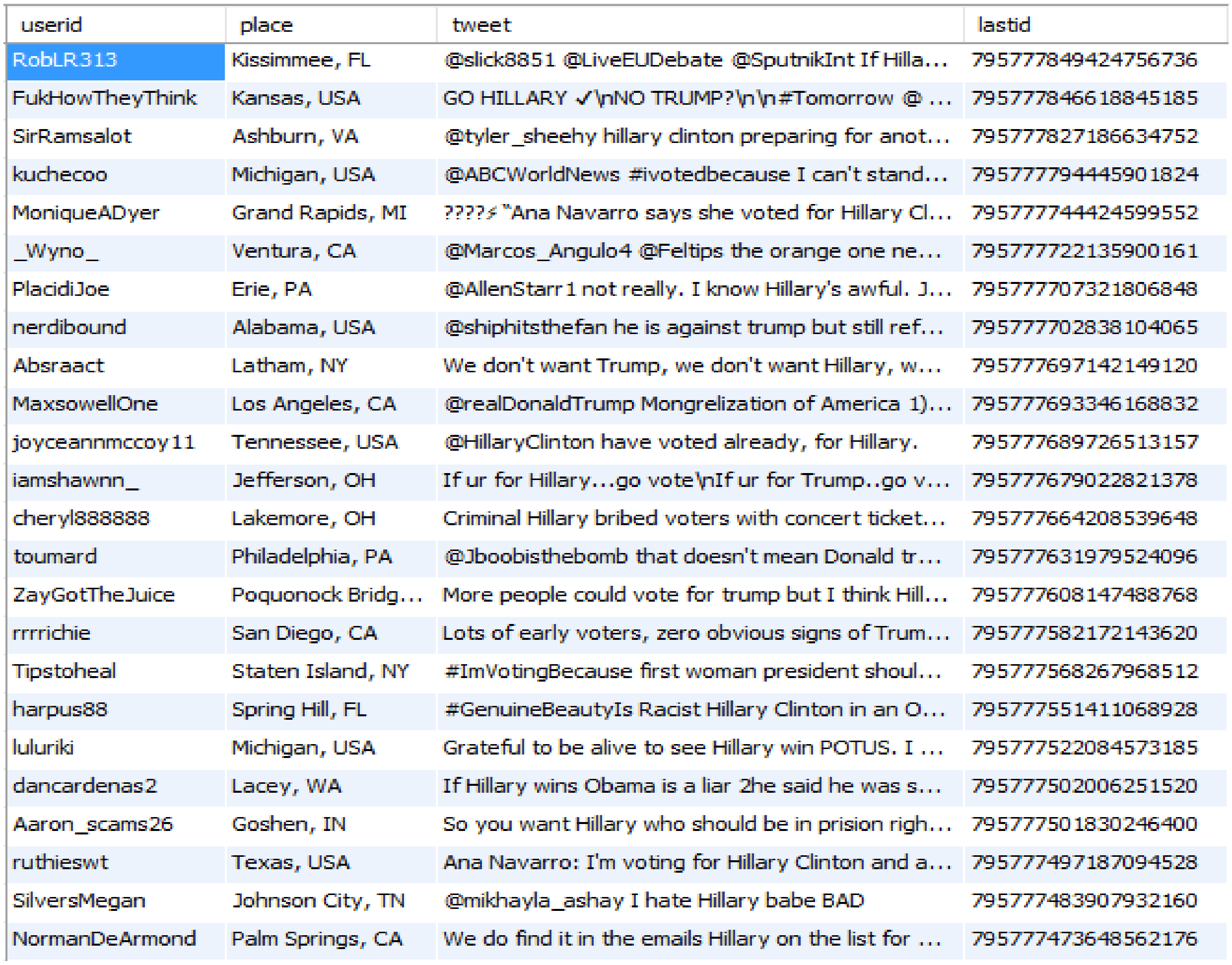
Poll model	Hillary Clinton Democratic	Donald Trump Republican
270 to Win	47.2%	43.6%
BBC	48.0%	44.0%
Election Projection	47.0%	43.8%
HuffPost Pollster	47.3%	42.0%
New York Times	45.9%	42.8%
Real Clear Politics	46.8%	43.6%
TPM Polltracker	48.8%	43.9%
FiveThirtyEight	45.7%	41.8%
HuffPost Pollster	45.7%	40.8%
New York Times	45.4%	42.3%
TPM Polltracker	46.0%	44.1%
270 to Win	45.6%	42.5%
Election Projection	45.3%	42.0%
Real Clear Politics	45.5%	42.2%
CNN Poll of Polls	46.0%	42.0%
TPM Polltracker	46.6%	43.8%
	47.97%	46.34%

Figure: source -en.wikipedia.org/wiki/Nationwide_opinion_polling_for_the_United_States_presidential_election,_2016

As seen it clearly is in contradiction with the actual results, this is where our model takes precedence as it has given more accurate results according to the dataset we have. The major drawback is that we are using the free access of the twitter streaming API and it is limited in how many tweets it provides to the users we made a little tweak around this limitation in our code to get as many tweets as possible but we got a very little tweets beyond the limitation twitter streaming API's limitation. The next section gives the description of the dataset.

3. DATASET

The main source of the data is the tweets obtained using the twitter search API. It has 2 kinds of access to it,one is paid and is also really a lot expensive and the other is unpaid, free of cost and is limited in access to the tweets, the paid service has unlimited access to the tweets and we also get better prediction of the results. The paid service is called the Firehose access, there used to be another access level called gardenhose which gives 10% access to all of its tweets. The access which we used is called spritzer which actually gives only 1% of the access to its tweets.We have implemented the collection of the twitter data in java using a very popular library called Twitter4j, to get the tweets using this we have to first create a developer account on the twitter and then create a sample app and navigate to token management and generate the access & refresh tokens which are used to authenticate the user when he tries to get the tweets. Once this is done now we have to set the dates between which we have to actually retrieve the tweets, we have set the dates for 4 days before the original results declaration day and collected about a million tweets to analyze. This process usually takes weeks as we have limitation of tweets and once we reach that limitation we set our program to go into hibernate mode and it resumes after that, essentially it collects the tweets for about every 15 minutes. We are storing all these tweets as they come into a database which is MySQL. A snapshot of the data collected is shown in the following figure:

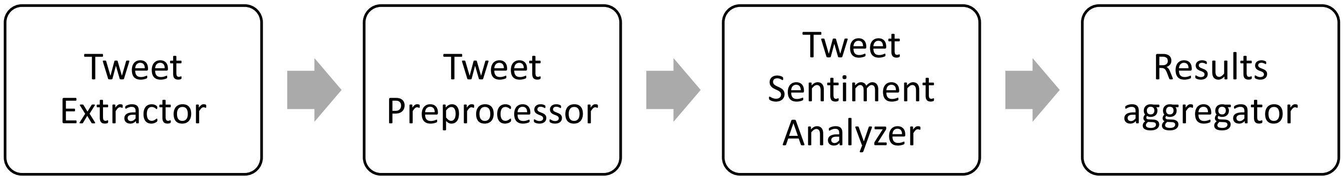


userid	place	tweet	lastid
RobLR313	Kissimmee, FL	@slick8851 @LiveEUDebate @SputnikInt If Hilla...	795777849424756736
FukHowTheyThink	Kansas, USA	GO HILLARY ✓\nNO TRUMP?\n\n#Tomorrow @ ...	795777846618845185
SirRamsalot	Ashburn, VA	@tyler_sheehy hillary clinton preparing for anot...	795777827186634752
kuchecoo	Michigan, USA	@ABCWorldNews #ivotedbecause I can't stand...	795777794445901824
MoniqueADyer	Grand Rapids, MI	????⚡ "Ana Navarro says she voted for Hillary Cl...	795777744424599552
Wyno	Ventura, CA	@Marcos_Angulo4 @Feltips the orange one ne...	795777722135900161
PlacidiJoe	Erie, PA	@AllenStarr1 not really. I know Hillary's awful. J...	795777707321806848
nerdibound	Alabama, USA	@shiphitsthefan he is against trump but still ref...	795777702838104065
Absraact	Latham, NY	We don't want Trump, we don't want Hillary, w...	795777697142149120
MaxsowellOne	Los Angeles, CA	@realDonaldTrump Mongrelization of America 1)...	795777693346168832
joyceannmccoy11	Tennessee, USA	@HillaryClinton have voted already, for Hillary.	795777689726513157
iamshawnn_	Jefferson, OH	If ur for Hillary...go vote\nIf ur for Trump..go v...	795777679022821378
cheryl888888	Lakemore, OH	Criminal Hillary bribed voters with concert ticket...	795777664208539648
toumard	Philadelphia, PA	@Jboobisthebomb that doesn't mean Donald tr...	795777631979524096
ZayGotTheJuice	Poquonock Bridg...	More people could vote for trump but I think Hill...	795777608147488768
rrrrichie	San Diego, CA	Lots of early voters, zero obvious signs of Trum...	795777582172143620
Tipstoheal	Staten Island, NY	#ImVotingBecause first woman president shoul...	795777568267968512
harpus88	Spring Hill, FL	#GenuineBeautyIs Racist Hillary Clinton in an O...	795777551411068928
luluriki	Michigan, USA	Grateful to be alive to see Hillary win POTUS. I ...	795777522084573185
dancardenas2	Lacey, WA	If Hillary wins Obama is a liar 2he said he was s...	795777502006251520
Aaron_scams26	Goshen, IN	So you want Hillary who should be in prision righ...	795777501830246400
ruthieswt	Texas, USA	Ana Navarro: I'm voting for Hillary Clinton and a...	795777497187094528
SilversMegan	Johnson City, TN	@mikhayla_ashay I hate Hillary babe BAD	795777483907932160
NormanDeArmond	Palm Springs, CA	We do find it in the emails Hillary on the list for ...	795777473648562176

Figure: Raw Data Collected

Once this is done we have to preprocess the tweets to filter based on the country the tweets come from & also we have to fill in some values for the null values present in the tweets. The life cycle of our program is shown in the following figure, which consists of four stages and each stage is explained in detail below with what that stage is and how we used that in our program to predict the outcome of the presidential election.

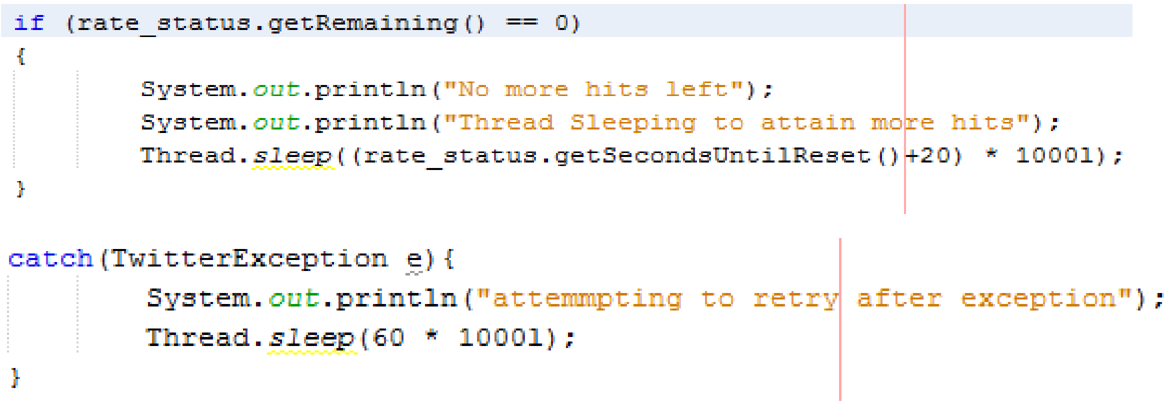
4. Process Flow



Sentiment analysis is a complex process and involves buildup across multiple stages. Broadly, the entire project is divided into 4 stages which include Tweet Extractor, Tweet Pre-Processor, Tweet Sentiment Analyzer and Results Aggregator.

4.1 Tweet Extractor

This is the most important stage of the design sentiment analysis. No sentiment analysis is good without good data and hence getting the right amount of good data was crucial to the process of sentiment analysis. However, this was not easy, twitter has rate limitations of rate of access for every 15 minutes. Also, twitter server arbitrarily shuts down the connection is it finds out heavy load by large pull out of tweets. Hence, we had to write a code to overcome these difficulties. We included the following code to overcome this situation



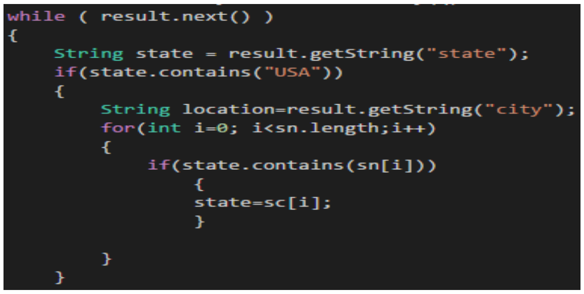
```
if (rate_status.getRemaining() == 0)
{
        System.out.println("No more hits left");
        System.out.println("Thread Sleeping to attain more hits");
        Thread.sleep((rate_status.getSecondsUntilReset()+20) * 1000l);
}

catch(TwitterException e){
        System.out.println("attemmpting to retry after exception");
        Thread.sleep(60 * 1000l);
}
```

Figure :Code Snippet

4.2 Tweet Preprocessor

This is the stage where we actually filter the tweets based on various factors. For example all the tweets have to be filtered out based on the country of the user that is posting the tweet, this is done as we are considering only the people who are citizens or residents of the united states of America as most of them have the right to vote and these tweets are the ones that are going to be useful in determining the winner and performing the sentiment analysis. The next preprocessing that needs to be done is fill out all the missing values with the null values. The following picture shows the snapshot of how we filtered the tweets based on the country.



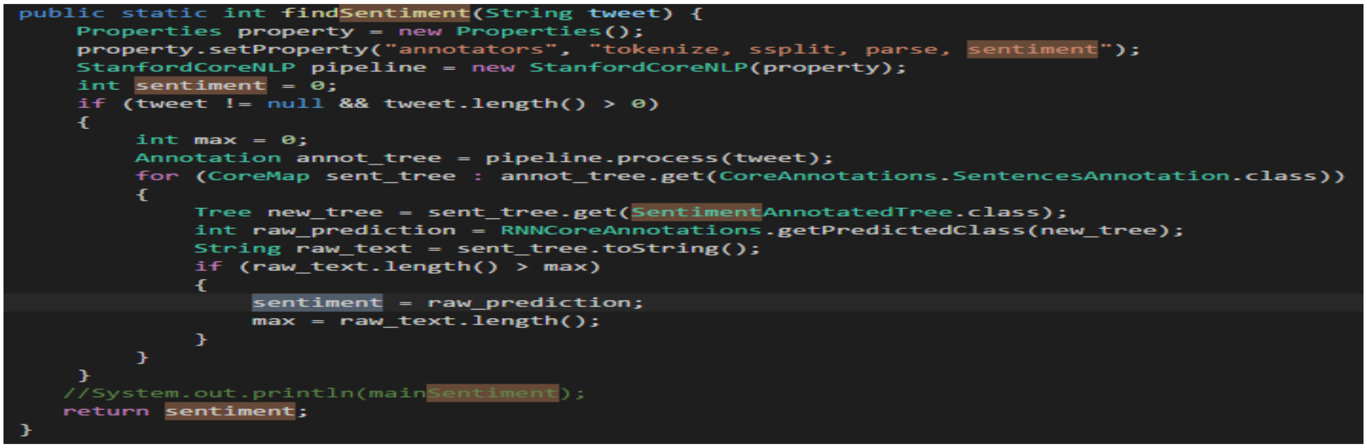
```
while ( result.next() )
{
    String state = result.getString("state");
    if(state.contains("USA"))
    {
        String location=result.getString("city");
        for(int i=0; i<sn.length;i++)
        {
            if(state.contains(sn[i]))
                {
                state=sc[i];
                }

        }
    }
```

Figure :Code Snippet

4.3 Tweet Sentiment Analyzer

Sentiment analysis is also known as the opinion mining and it refers to the use of the technologies mainly like the natural language processing and Text processing mainly to extract the useful information in the dataset we are interested in. Sentiment Analysis is mainly used in social media and also mainly in the areas like marketing the products etc, to match the user tastes based on the location of the user. The sentiment is a score of 0 to 4 with 0 being a negative sentiment, 4 being a positive sentiment and a neutral sentiment having a score of 2.We used the library from the Stanford educational community which is free of cost to perform our sentiment analysis on the data we collected. It includes many NLP techniques to and libraries which are very useful. The following code snippet provides a snapshot of how we performed the sentiment analysis.



```
public static int findSentiment(String tweet) {
    Properties property = new Properties();
    property.setProperty("annotators", "tokenize, ssplit, parse, sentiment");
    StanfordCoreNLP pipeline = new StanfordCoreNLP(property);
    int sentiment = 0;
    if (tweet != null && tweet.length() > 0)
    {
        int max = 0;
        Annotation annot_tree = pipeline.process(tweet);
        for (CoreMap sent_tree : annot_tree.get(CoreAnnotations.SentencesAnnotation.class))
        {
            Tree new_tree = sent_tree.get(SentimentAnnotatedTree.class);
            int raw_prediction = RNNCoreAnnotations.getPredictedClass(new_tree);
            String raw_text = sent_tree.toString();
            if (raw_text.length() > max)
            {
                sentiment = raw_prediction;
                max = raw_text.length();
            }
        }
    }
   //System.out.println(mainSentiment);
   return sentiment;
}
```

Figure : code snippet

4.4 Results Aggregator

This is the final stage where we collect all the sentiments that we generated for each tweet and categorizations based on the sentiment for each tweet and actually compute the percentage share of the tweets of the presidential candidates Trump & Hillary using the java code- get_results.java .

5. Analysis And Results

After aggregating our results we found the following prediction on the popular vote tally of the presidential elections as shown in the table below.

	Predicted	**Actual**
Trump	**0.304**	**0.475**
Hillary	**0.493**	**0.477**
Neutral	**0.203**	**0.048**

Table 1: Popular Vote

From the table, at the first instance the immediate temptation is to predict Hillary Clinton as the clear winner with a 49% predicted vote share. But this is not the case in actuality. Also, the Trump vote share of 30.4 % is way below the actual values. Hence, one would also be tempted to draw the conclusion that the model failed to predict the actual outcome. However, we believe that the model correctly captured the trend of the presidential elections and the reasons why it couldn't capture the exact outcome include for reasons like the size of the Dataset, the nature of US presidential polity, Social setup of voters and more importantly the role of Swing Voters. We will be analyzing each one of the reasons as follows.

5.1 Dataset

Firstly, we think any exact prediction of the true outcome is not possible because of fundamental technological and dataset limitations. Twitter doesn't make it readily available of all residents of the USAs tweets. Only users who declare in their privacy settings that they would not bother sharing their location settings can have their location parsed. So there were a significant number of users who were potential voters but whose tweets we did not consider because we had no evidence they were US residents. Also, after processing over a million tweets we shortlisted approximately 50,000 tweets and analyzed a sizeable portion of that. This by no means the complete representation of the 120 million voters in the USA and most of them would not be vociferous on twitter let alone the 4 days window we chose for getting our data.

5.2 Social Setup of the US voters

We noticed after reading countless articles on the analysis of the USA presidential elections and various expert views that the hinterlands and the countryside played a crucial role in the outcome of the presidential elections. These voters are most likely not active on twitter unlike the urban agglomerative youth voters who are predominant on twitter. This pocket of vote bank included industry workers, laborers, small farm owners, small businessmen in the hinterlands. Its been widely acknowledged that while Hillary won in the urban places, she was outperformed in the non-urban areas and eventually lost the swing states. Wisconsin and Michigan are one such cases. Hence it was no wonder that our twitter data could not pick up this trend.

5.3 Winner takes all Policy

The very nature of electoral process in the USA is fundamentally different for doing a sentiment analysis on the popular vote. The outcome of election is decided by an electoral college consisting of a winner take all policy. So even though Hillary might have come close to beating trump or was close behind, she would still lose all the electoral college seats for that particular state. Hence even though Hillary won the popular vote as we predicted she would, she would still not be guaranteed winning the presidency.

5.4 Role of Swing voters

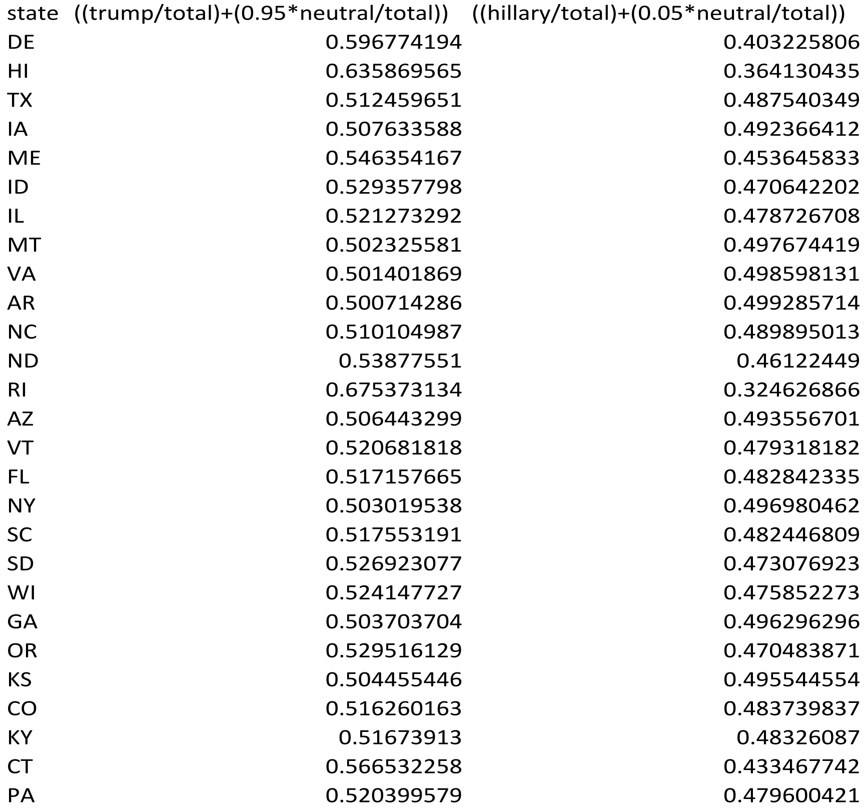
In our analysis we believe that the role of the swing voters played the most crucial role in shifting the course of the elections. If we observe closely, we predicted a neutral vote share of 20% which was different from the actual vote share of 4.8%. This is consistent to typical voter behavior. While many voters might be indifferent to a candidate while expressing their view on twitter, they would most likely choose a candidate to vote at the time of voting. We believe that Trump received the lion’s share of this swing votes across all the electoral polling stations. There is collaborating evidence from the fact that Hillary lost most of the swing states. Moreover, if we normalize our neutral votes prediction to the actual prediction and allocate that share to Trump, our prediction comes close to the actual prediction.

	Predicted	**Actual**
Trump	**0.459**	**0.475**
Hillary	**0.493**	**0.477**
Neutral	**0.048**	**0.048**

Table 2: Prediction adjusted to neutral voters assigned to trump.

5.5 Statewise Prediction

Our prediction of state-wise election gave a much deeper insight. We could clearly see Hillary winning in states like California, Washington, New Jersey, also confirming our earlier analysis that a twitter sentiment analysis could more accurately predict the outcome in states dominated by urban youth who are most likely users of twitter. Also , it gave us insight into some of the potential swing states. Even Republican stronghold Texas did not show a wide margin for trump over Hillary as was seen in the actual elections too. We could see states like PA ,FL , WI , NC , IL being potential swing states. Moreover , we could draw an analysis that if trump gets 95% of the neutral vote share then , he would win in 27 states which is close to the actual prediction too thereby further confirming our earlier analysis of the role played by swing voters.



state	((trump/total)+(0.95*neutral/total))	((hillary/total)+(0.05*neutral/total))
DE	0.596774194	0.403225806
HI	0.635869565	0.364130435
TX	0.512459651	0.487540349
IA	0.507633588	0.492366412
ME	0.546354167	0.453645833
ID	0.529357798	0.470642202
IL	0.521273292	0.478726708
MT	0.502325581	0.497674419
VA	0.501401869	0.498598131
AR	0.500714286	0.499285714
NC	0.510104987	0.489895013
ND	0.53877551	0.46122449
RI	0.675373134	0.324626866
AZ	0.506443299	0.493556701
VT	0.520681818	0.479318182
FL	0.517157665	0.482842335
NY	0.503019538	0.496980462
SC	0.517553191	0.482446809
SD	0.526923077	0.473076923
WI	0.524147727	0.475852273
GA	0.503703704	0.496296296
OR	0.529516129	0.470483871
KS	0.504455446	0.495544554
CO	0.516260163	0.483739837
KY	0.51673913	0.48326087
CT	0.566532258	0.433467742
PA	0.520399579	0.479600421

Figure : Statewise Predictions after adjusting the neutral votes . Trump Leads in 28 states

6. Conclusion

This project was a great opportunity for us to test our social analytic skills on real world real time problem statements. In this case it was the USA presidential elections. We believe that we gained a lot of knowledge in this process. Broadly, our prediction model was fairly accurate in analyzing the users twitter sentiment. We think that this election was one of its kind and it will not be possible to accurately predict the outcome of an election which is unconventional in many ways in the history of US elections. Most of the opinion polls conducted by various agencies failed to predict the outcome of the elections. However, the sentiment analysis can be a barometer for trend picking. In this particular case, it served as a barometer for trends like voter swings, role of hinterland , impact of the USA electoral process on the potential outcomes and impact of the dataset. As said by a Professor Pedro Domingos, “Correlation does not imply causation”. What a twitter sentiment analysis or any machine learning based predictive model gives is more a guide to decisions and policy making and need not be the ultimate or final say. And in this function twitter sentiment analysis can play a great role in trend picking and analyzing situations like the US presidential elections.

7. Challenges

We encountered many challenges while working on the project. The first significant challenge was collecting the data itself. Twitter has restrictions on extracting the tweets and we had to create multiple accounts and run our code across multiple systems to download nearly a million tweets and process those. We faced problems of sudden code exception errors because the twitter server timed out our requests. Also, storage and processing of the tweets was hugely time taking. It took days to download the tweets alone and equally for deriving the sentiment for these tweets. The last challenge was also trying to analyze and draw meaning out of them especially after the recent outcome. We faced the same difficulties as other opinion poll predictors in analyzing the results and comparing it with the reality. This meant us researching into the political as well as the social setup of the US polity. Hence , in a way , this project helped us to broaden our scope and expand our knowledge base at the same time getting a grip on the technicalities of doing a sentiment analysis on twitter data.

8. Future Work

In the future we plan to develop an interactive UI using Tableau to construct a representation of our results with a cool visualization of the USA map. We also plan to expand our research by analyzing other data sources like Facebook (using its graph API). We plan to use this approach in doing sentiment analysis on other hot topics like customers response to a new release of an iPhone or the analysis on a recent disaster or extraordinary event.

9. Contribution of Team Members

We are a group of three – Venkata Sai Chaitanya, Krishna chaitanya and Varapula Srinivas . The project provided us a great opportunity to test our skills at a both technical and not technical level. At a non-technical level, it helped us in efficient time management and perform as a team and not as a group of isolated individuals. We formed whatsApp groups and scheduled meetups at the Library as well as the UTD open lab. At a technical level, we congregated our efforts and technical skills in a properly coordinated fashion. We brainstormed which libraries and coding language and also the appropriate way to approach the project. Sai chaitaya's knowledge in Tweet Extraction, Krishna's experience in working in Natural Language Processing in Enterprises and Srinivas's knowledge in MySQL helped in efficient planning and completion of project. We sat together and discussed the algorithm, Sai chaitanya helped with preprocessing of data, Krishna tuned the model and analyzed the results and Srinivas sat through the SQL Connections & did all the Backend work. Overall we had an equal contribution in the entire project and also the report**.**

10. References

1. http://twitter4j.org/oldjavadocs/4.0.4/
2. Course Lecture Slides
3. Stanford NLP toolkit guide
4. en.wikipedia.org/wiki/Nationwide_opinion_polling_for_the_United_States_presidential_election,_2016