Tweet Sentiment Analysis of the US Presidential Elections 2016

Patrick Von Angelo V. Atienza
Department of Information Systems and
Computer Science
Ateneo De Manila University
Quezon City, Philippines
patrick.atienza@obf.ateneo.edu

Paulo Jerome C. de Leon
Department of Information Systems and
Computer Science
Ateneo De Manila University
Quezon City, Philippines
paulojerome7@yahoo.com

ABSTRACT

This paper is an explorative study that aims to determine the tweet sentiments of the Twitter community based on the Twitter user's location in relevance with the Republican, Democratic during the 2016 US elections and to compare its consistency with the elections results. The study assumes that the Twitter users are classified supporters or detractors of either Republican or Democratic based on their tweets filtered from predetermined keywords and the sentiments they elicit in connection to that. Inference from the study conducted will provide substantial information with regards to the alignment of the Republican and Democratic states from their Twitter sentiments with support from objective data coming from the electoral votes their state represents.

KEYWORDS

sentiment analysis, tweets, 2016 us presidential elections

1 INTRODUCTION

The 2016 US elections event last November 8 marks as the historical landmark of Donald Trump's assumption to the presidency defeating his rival in the race Hillary Clinton. President Trump won by gathering 306 electoral votes which is 36 more than what would've been sufficient to win. His current party is Republican which is supported by various states wherein those states have supported this party historically since the olden days. On the other side of the spectrum which is the Democrat party having their own supporting states, supports presidential candidates Hillary Clinton.

On the historical day of that November 8, 2016, social media websites were trafficked with the deluge of data coming from their users. Twitter amassed 7 million tweets relevant to the election that day. The tweets were reactions from people voicing their positive, negative or neutral sentiments on the result of the presidential election. Some of the tweets from users have in them the location from where the user resides if the user sets up his account to allow location to be tracked.

From what is stated above, this paper focuses on the link between the sentiments of people from their tweets and their state location while also deriving connections to the results of the election, specifically its consistency with the electoral votes per state a given candidate received.

2 SIGNIFICANCE OF THE STUDY

Comparing the actual electoral votes result to collected sentiments of the Twitter community per state is the main focus of the study and the result would provide profound differentiation on the relationship of a user's social media sentiments versus the reality of the consensus from the electoral votes they represent. Cumulatively, this study can actively broaden the history books' data sources in the future (presumptively social media databases) and giving insights about the prevailing thought patterns of people during certain events, while also potentially giving Twitter forums weight when it comes to predicting future events from their everyday topic discussions given that innately, Twitter is where people post their day to day thoughts.

3 REVIEW OF RELATED LITERATURE

3.1 Analyzing Twitter Sentiment of the 2016 Presidential Election

Chin, Zappone and Zhao focused on tweets regarding the presidential election and its candidates. They assessed the sentiment towards each candidate based on tweets during the election daw. Using the keywords "politics" and "political candidates|, they gathered approximately 300,000 tweets. Out of the gathered tweets, 3,000 tweets contained emojis. Below are the list of emojis they used:

Figure 1: Twitter emojis that were used to determine the sentiment of tweets

They categorized all the emojis as sentiments and apply the processing of the tweet text. They used the bag--of--words model

to create the feature vector. Every distinct word from tweets is added to a feature vector - excluding URLs, usernames, and words on our list of common words that do not express sentiment. They used support vector machines (SVM), K-Nearest Neighbors (KNN), and Naive-Bayes Classification as the algorithms for the research.
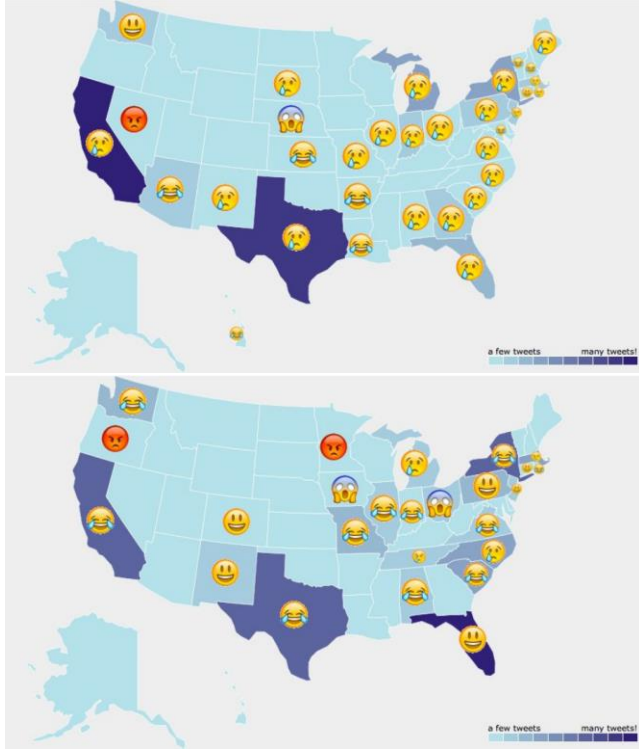

Figure 2: Sentiment maps for Donald trump (above) and Hillary Clinton (below)

Based on their results, Donald Trump's map showed huge majority of the sad sentiment. This maybe caused by relatively sad tweets that were then retweeted many times. If retweets were disregarded, laughter was the most common emoji from the unique tweets for every candidate.

3.2 Social Bots Distort the 2016 US Presidential Election Online Discussion

Bessi and Ferrara described the investigation that brought them to unveil the pervasive presence and activity of social bots involved in the 2016 US Presidential Election on Twitter. They collected Twitter data from September 16 to October 21, 2016 that included all three Presidential debates. They gathered a total of 20 million tweets from 2.8 distinct users.

Figure 3: List of hashtags monitored via Twitter Search API during September 16 to October 21 2016

Based on their results, it was estimated that 400,000 users out of 2.8 million users are bots that are engaged in the political discussion about the Presidential election. The bots were responsible for 3.8 million bot tweets.
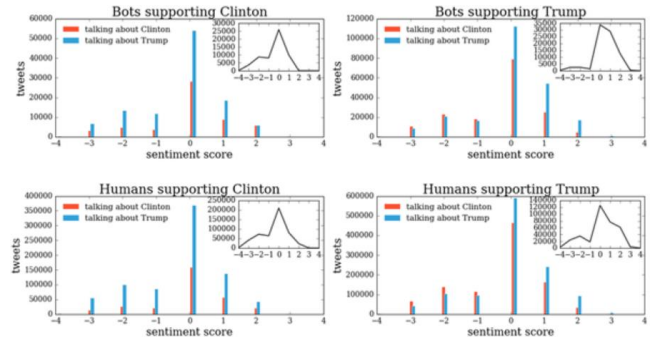

Figure 4: Distributions of the sentiment of bots and humans supporting Trump and Clinton

They found out that on average, the tweets produced by Trump's supporters are significantly more positive than that of Clinton's supporters, regardless of whether the source is human or bot. This generated a stream of support that is at staggering odds with respect to the overall negative tone that characterizes the 2016 Presidential election campaigns. The fact that bots produce systematically more positive content in support of a candidate can bias the perception of the individuals exposed to it, suggesting that there exists an organic, grassroots support for a given candidate, while in reality it's all artificially generated.

4 SCOPE AND LIMITATIONS

The study will be focusing solely with the 2016 US presidential elections and tries to avoid events not associated with the elections. This is done with the narrowing down of focus with the keywords used in performing the Tweets search.

Table 1: Keywords used to filter the tweets

Related to Democrat	Related to Republican

hillary	trump
clinton	yourefired
democrat	republican
madampresident	gop
kaine	makeamericagreatagain
	pence

Also, the study assumes that potential conflicting sentiments of retweets from the same user are resolved with the result of the internal tallying of their retweets.

Having discussed the limitations of the study, it is also at the same time reflecting the potential development of this research. Studies of a similar nature from which this research was conceived made this study even more worth delving into as it can attract the attention of several researchers in the future. Interested on this topic can range from politicians and historians to data scientists and sociologists.

The actual electoral votes for Maine was 3 for the Democrat and 1 for the Republican. However, since we cannot gather sufficient data for each county in that state, we assume that the party with the most votes will get all the electoral votes. In this case, the total electoral votes for the democrat and the republican parties are 233 and 305 respectively instead of the actual 232 and 306.

5 METHODOLOGY

5.1 Dataset Gathering

The dataset used in the study is titled 'Tweets from election day 2016' by Github user 'chrisalbon' and is from a Github repository found public in the internet [11] The data in the repository contained only Twitter data relevant to the 2016 presidential elections and only contained tweet IDs.

Another dataset formatted in a .CSV file containing electoral vote counts per state was also gathered for the mid-portion of the study.

5.2 Data Extraction from Tweet IDs

Through the use of a tweet collecting program written in the Python language, the group gathered historical data in Twitter using the keywords as found in **Table 1**. These tweet IDs were then extracted for their tweet message, user information, location, and the like with the command line tool "twarc" . After the tweets were collected by the program, they were compiled in a .CSV file and were saved in the group's Github repository . Initially, the tweets included had some states that were unincorporated territories but were removed from the database during the cleanup. These included the states: AS, FM, GU, MH, MP, PW, PR, VI.

5.3 MongoDB and MapReduce

Installation of the MongoDB was done inside of virtual machine with 64-Bit Ubuntu Mini OS installed in it. Multiple VMs were configured in order to assess the sharding and replication part of the experiment.

In addition for confirming that MongoDB server can be run, initiating the 'mongod' command inside the bin folder of the dumped zip would activate the MongoDB server. Errors such as 'mongo.lock' and 'wiredTiger' may occur during this setup if a previous mongo session was shut down incorrectly or a 32-bit limitation of the OS happens respectively, though it can be easily fixed by resources to be found online specifically stackoverflow.com. Finally, to check if the running MongoDB server works, issuing the 'mongo' command of the bin folder must bring the user to the mongo shell.

Upon extraction of the desired tweet information from the tweet_id's provided by our data set, mapping and reducing the relevant data for our study becomes the next objective. To start the process, it can be done via copy-pasting all the insert statements inside the chosen database and specified collection or if the dataset is too big, to make use of the mongoimport function. After this, projection of the desired fields to be outputted via 'map' and doing a grouping of fields via 'reduce' was performed. Next, 'mapreduce' was made with the parameters 'map', 'reduce' and the parameter 'query' to count occurrences of the study's desired reduced fields. The result yields the number of people who rooted for Republican or Democrat per state. This became the group's first data yielding useful results and is recorded permanently despite being in the drafting of codes as it does reveal a significant study finding when paired with the other refined versions of this mapreduced functions.

On that note, the upgraded mapreduced function aimed to eliminate retweets from a same user by classifying their sentiments in a clustered concise manner. This includes capturing all the users tweets and retweets and processing them to a single output based on the recorded sentiments of the tweets. The modification of the group's mapreduce code was made by isolating the user's id attribute from the map function to elicit a single value for the reduce function later on.

Below are the list mapreduce methods done to obtain the desired collections:

1. Get tweets by state and party rooted for
2. Get party rooted for by state
3. Merge the actual election results and electoral votes per state
4. Get tweets by user id and determine the party he/she rooted for
5. Repeat step 1 for unique users
6. Repeat step 2 for unique users
7. Repeat step 3 for unique users

Following the mapreduced collection of the tally party made, another mapreduce was necessitated to compare the electoral votes' actual winner with the inclinations of the users to either Democrat or Republican per state based from their tweets. This double mapreduce is achieved by the two map functions, the actual vote and the tweets' vote by being joined together to elicit a comparison. The end result was a side by side comparison which showed if whether the sentiments of the people in who they're rooting for were consistent with the electoral votes that actually transpired, per state; Same parties means consistency and differing parties meant otherwise.

5.4 Sharding and Replication

For the final portion of the study, replication and sharding was performed on the database to establish data security and ease of access of data chunks across the entirety of the data set. To show the sharding and replication process of the experiment, we have split the setup into two topologies. First is replication-only topology. This includes a replica set consists of 3 replication servers.
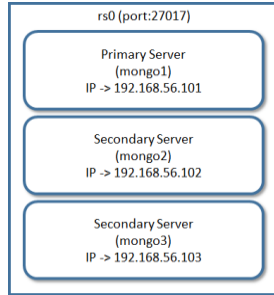

Figure 5: Replication-only topology

Second is the sharding-only topology. This topology includes 1 app server, 3 config servers and 2 independent shard servers.
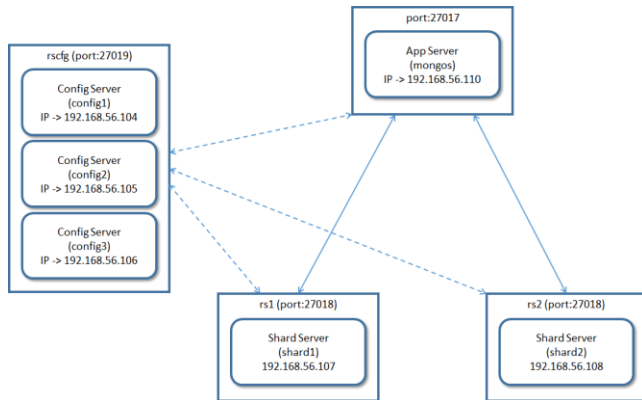

Figure 6: Sharding-only topology

Each server is running on a separate virtual machine (VM) that equipped with host-only network configuration. Each virtual machine had Ubuntu Mini 16.04 64-bit OS installed in it and had 512MB ram, 10GB hard disk space.

6 RESULTS AND DISCUSSION

6.1 Visualization of Results

For the figures below, the red indicates Republican while the blue for Democrat. We used Vincent visualization library in Python to visualize our results and display the results using Jupyter Notebook.
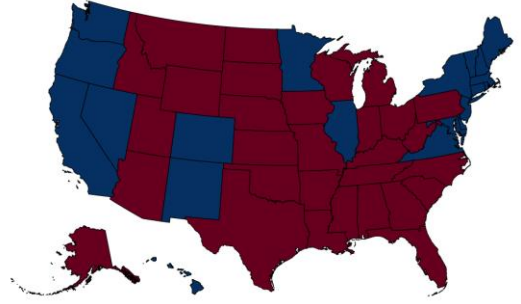

Figure 7: Actual Electoral Votes Winner per State Map

As mentioned in the scope and limitation, we assume that the party with the most votes in the state of Maine would get all the electoral votes. The electoral vote count for the Republicans is 305 while the Democrats have 233 electoral votes. The state of Nebraska would also have the assumption, but since the Democrats got all the electoral votes, it did not matter for our analysis.
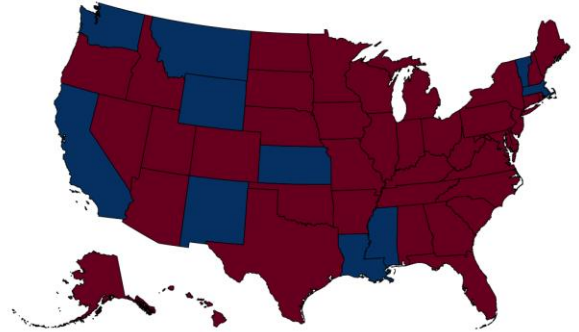

Figure 8: Tweet Votes Count and Winner per State Map

The Republicans and the Democrats gained a total of 426 and 112 electoral votes respectively based from the gathered tweets. As shown in **Figure 8**, most states on the northeast part of the map flipped to states who rooted for Republican.
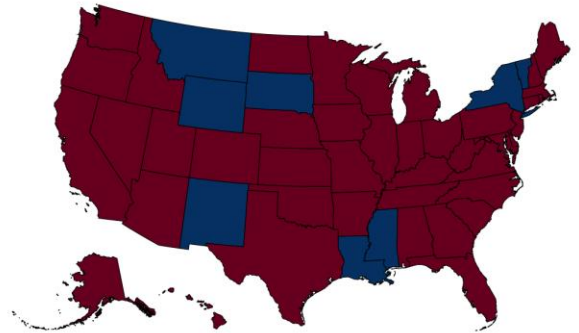

Figure 9: Tweet Votes with Unique Users Count and Winner per State Map

As shown in **Figure 9**, the Republican almost wiped out the whole map. Vote-rich states like California and Massachusetts made a change of heart to Republican that made the election results a much more landslide than before.

6.2   Data Analysis of Popular Votes

Based on the actual presidential election results, Hillary Clinton received a total of 65,853,625 votes or 48.0% of the total votes compared to Donald Trump that only have 62,985,106 votes or 45.9% of the total votes. An article from The Independent stated that Donald Trump had lost the greatest margin of popular votes in the history of US Presidential Elections. The democrat lost by 2,868,519 votes or 4.1% of the winner of the popular votes. If we only count total votes from the Democrat and the Republican, Clinton would have 51.11% of the votes while Trump would have 48.89%

Table 2: Popularity Sentiment Votes from Tweets and its difference from the Actual Vote %

Party	Vote Count	% From Total	Difference from Actual %
Republican	87856	50.80%	+1.91%
Democrat	85075	49.20%	-1.91%

As shown from **Table 2**, the Republican had a boost of 1.91% (from 48.89% to 50.80%) from the actual popularity vote percentage. Take note that those results are only based from Tweets that have retweets and duplicate users. The Democrat sentiment votes decreased from 51.11% to 49.20%.

Table 3: Popularity Sentiment Votes from Unique User Tweets and its difference from the Actual Vote %

Party	Vote Count	% From Total	Difference from Actual %
Republican	51670	51.41%	+2.52%
Democrat	48828	48.59%	-2.52%

As shown from **Table 3**, the Republican widened vote percentage margin by having a boost of 2.52% (from 48.89% to 51.41%) from

the actual popularity vote percentage. The Democrat sentiment votes plummeted to 48.59%. This shows that there are more people supported the Republican regardless of removing the unique users and retweets or not.

6.3Data Analysis of Electoral Votes

The Republicans won the electoral votes in the Presidential Elections despite losing to the popularity votes. The Republicans gathered 306 electoral votes while the Democrats only had 232 electoral votes. As stated, we will assume that the Republicans had 305 actual electoral votes and Democrats had 233 actual electoral votes.

Table 4: Electoral Votes from Tweets and its difference from the Actual Electoral Votes

Party	Electoral Votes	% From Total	Difference from Actual
Republican	426	79.18%	+121
Democrat	112	20.82%	-121

As shown from **Table 4**, the Republican gained a whopping 121 electoral from its actual electoral votes (from 305 to 426 electoral votes) while the Democrat fell from 233 to 112 electoral votes. Take note that those results are only based from Tweets that have retweets and duplicate users.

Table 5: Electoral Votes from Unique User Tweets and its difference from the Actual Electoral Votes

Party	Electoral Votes	% From Total	Difference from Actual
Republican	478	88.84%	+173
Democrat	60	11.16%	-173

As shown from **Table 5**, the Republican increased the lead even more by gaining a total of 173 electoral votes from the actual electoral votes (from 305 to 478) while the Democrat plunges from 233 to 60 electoral votes. Based from the tweets, the Republicans won with landslide 418 electoral vote lead over the Democrats.

6.4 Discussion of Results

On a winner per state accuracy of the tweet votes based from the sentiment analysis proved to be at 60.8% (31/51) for the tweet votes that didn't take into account the retweets of unique users versus the actual electoral votes count derived from **Figure 7** and **Table 2** and at 52.9% (27/51) for the unique user tweet votes versus the actual electoral votes count derived from **Figure 8** and **Table 3**. Additionally, the comparison between the tweet votes and the tweet votes with unique users makes a significant difference for the tallying of the Republican or Democratic party classification as seen with the 88.2% (45/51) accuracy the tweet votes have with the tweet votes with unique users with each other and vice-versa, based on the difference between **Table 4** and **Table 5**.

7 CONCLUSIONS

Based on the weighted value of electoral votes per state ratios and the winner per state accuracy, it can be inferred that the study was still able to predict the candidate winner on a holistic level, to the point of an output clairvoyance from the ratios and a near to midway depiction of electoral state winners from the winner per state map accuracy. This can be seen on the seeming conundrum that despite the tweets' votes higher accuracy with the tweets' votes of unique users (the presumed more refined data) when compared to the actual electoral votes result on the candidate winner per state scale, consistency is still seen with what the definitive 'landslide' answer of the tweets' votes with unique users elicits when it comes to the their electoral votes conversion based on tweet sentiments. On that same token, it can be inferred from the study that those users who continually spam retweets and spambots are not to be underestimated as they do have the capability of turning the tides for either candidates to a mild scale.

From the entirety of what the study has gathered, Twitter to a certain extent provides a means of objectifying subjective data from the tweet sentiments to support a factual claim. Based from the results, the data determined potentially raises the chances of debunking conspiracy theories related to the election.

ACKNOWLEDGMENTS

REFERENCES

[1] D. Chin, A. Zappone, and J. Zhao. 2016. Analyzing Twitter Sentiment of the 2016 Presidential Candidates. Retrieved from https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf.

[2] A. Bessi and E. Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/7090/5653.

[3] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo. 2015. Approaches, Tools and Applications for Sentiment Analysis Implementation. In *International Journal of Computer Applications*, 125, 3 (September 2015), 26-33.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passoneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics (June 2011), 30-38.

[5] A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, 10, 2010 (2010), 1320-1326.

[6] E. Kouloumpis, T. Wilson, and J. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!. In *Proceedings f the Fifth International AAAI Conference on Weblogs and Social Media*, 11, 164 (2011), 538-541.

[7] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. 2011. Election Forecasts With Twiiter: How 140 Characters Reflect the Political Landscape. In *Social science computer review*, 29, 4 (November 2011), 402-418.

[8] P. Metaxas, E. Mustafaraj, and D. Gayo-Avello. 2011. How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialComm)*, IEEE, 167-171.

[9] J. Chung and E. Mustafaraj. 2011. Can collective sentiment expressed on twitter predict political elections?. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 11 (August 2011), 1770-1771.

[10] A. Bermingham and A. Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2-10.

[11] M. Thelwall, K. Buckley, and G. Paltoglou. 2011. Sentiment in Twitter events. In *Journal of the Association for Information Science and Technology*, 62, 2 (February 2011), 406-418.

[12] Chris Albon. 2016. Tweet Election Day 2016. Retrieved from https://github.com/chrisalbon/election_day_2016_twitter