Sentiment Analysis of Tweets to Gain Insights into the 2016 US Election

by Tim Hamling, Ankur Agrawal
Department of Computer Science, Manhattan College, New York

Abstract — Social media use is at an all-time historic high for the United States, so we considered one popular social media platform, Twitter, and tried to see if we could predict how a group of people felt about an issue by only using posts from social media. For our research, we looked at tweets that focused on the 2016 United States presidential election. Using these tweets, we tried to find a correlation between tweet sentiment and the election results. We wrote a program to collect tweets that mentioned one of the two candidates, then sorted the tweets by state and developed a sentiment algorithm to see which candidate the tweet favored, or if it was neutral. After collecting the data from Twitter and comparing it to the results of the Electoral College, we found that Twitter sentiments corresponded with 66.7% of the actual outcome of the Electoral College. The overall sentiment of all tweets collected leaned more positively towards Donald Trump than it did for Hillary Clinton. Using the data that was collected, we also looked at how different geographical locations affected a candidate's popularity, analyzed what issues were most prevalent in tweets, and looked at the ratio of a state's population versus the number of tweets gathered.

I. INTRODUCTION

In United States politics, 2016 was an important year because it was an election year for the nation. The two presidential candidates were Donald Trump for the Republican Party and Hillary Clinton for the Democratic Party. As the days grew closer to November 8th, Election Day, many news outlets believed that the race between the two candidates was extremely close, with sites such as FiveThirtyEight reporting that the election could be a toss-up for either candidate [1]. The United States' presidential elections work using an Electoral College, meaning that each state has a certain number of electors who will vote for a candidate. Voters cast their vote for these electors, meaning that popular vote does not always decide the outcome [2][3]. This makes predicting the election's outcome difficult, so leading up to election day, political scientists looked to polling data as well as computer analysis programs to try and see which candidate was poised to win.

The nature of the election, as well as the two candidates themselves, was widely discussed across both the United States and other countries in the world [4]. Debates and conversations about the candidates and their policies most certainly happened verbally, but people also took to social media to voice their opinions about the election. This is no surprise, especially due to the current usage of social media sites such as Facebook, with 1.79 billion monthly active users [5], and Twitter, with around 313 million monthly active users [6]. In addition to these platforms, the rising popularity of sites such as Snapchat, with 150 million daily active users [7], and Instagram, with 500 million monthly active users [8], also contributed to social media conversations. Analyzing social media users' activity to determine what people are talking about is quite easy because many people, when posting to social media, make their posts and activity public for anyone to see.

Twitter is a social media service that allows users to post "tweets" to the site [9]. These can either be viewed publicly by anyone who wishes to see, or can be made private so that only people who have been allowed to follow a user can see that person's tweets. Regardless of the privacy, one of Twitter's features is that each tweet is limited to 140 characters, which includes whitespace characters, non-ASCII text, or links to web pages or images [10]. While this may seem like an odd restriction, it was adopted to work in conjunction with SMS messaging services, and has stuck around because it forces users to get to the point and only focus on providing important details in their tweet. Due to the quick and concise nature of tweets, some classify the social media site as a news media site [11]. Per research done by the research group Statista, as of Quarter 3 in 2016, Twitter has approximately 317 million monthly users. The monthly user base has been rising since Twitter's creation in 2006, but has slowly leveled off around the 300 million user mark since early 2015 [12].

Another key feature of Twitter is its use of hashtags, represented by a pound sign (#) preceding a word or phrase. These are used within tweets by a user to identify different topics or keywords present within the tweet. Then, other users could search other instances of that specific hashtag to see other tweets that match that topic. For example, during breaking news stories, hashtags are helpful because they allow a user to search and filter for tweets that discuss that news story. As Twitter has evolved, it is now possible for users to search by keywords that are not marked with a hashtag. This allows others to search for a work and receive a listing of all tweets that mention or use that word.

In this study, we set out to see if we could predict the results of the 2016 United States presidential election by looking at tweets that mention either candidate, and analyzing them to determine an overall sentiment. This sentiment could be leaning in favor or against either candidate. For this study, we sought to collect any tweets that were posted in the days leading up to the United States election that openly mentioned either of the two candidates by name. We then aimed to map these tweets, along with their sentiments, to locations, either within the United States or from foreign countries. For locations within the United States, we mapped each tweet to one of the states in the US, which was determined based on a user's location description tag. Finally, we collected the total number of tweets from a state and looked at the ratio of positive/negative tweets for each candidate to determine how that state, overall, feels about each candidate. These results would then be checked against the official results of the election, and we could then see if the way a state voted in the Electoral

College matched the prediction we made about it based on that state's Tweet sentiment.

Sentiment analysis of Twitter data is not a new field. Pang and Lee worked on creating algorithms to facilitate opinion mining and word analysis in their 2008 study [13]. Researchers such as Go, Huang, and Bhayani conducted a study in 2009 to train a sentiment algorithm to detect a tweet's positivity about a certain subject by using emoticons [14]. Pak and Paroubek expanded upon this in 2010 by using the subjectivity and objectivity of words in conjunction with a tweet's structure to create a classifier that could use collected data to determine a tweet's sentiment [15]. We decided to create our own sentiment algorithm that used a collection of words, each with its own sentiment value, to analyze tweets about the two candidates.

We also used the data collected to see what issues were most discussed in the week leading up to the election. Using a list of popular political issues during elections [16], we searched through our collection of tweets to see if any of these issues were being mentioned, and counted which issues were discussed most frequently. These totals were then mapped to each state, so we could then see which states talked about which issues most, and try to find correlations between the issues being discussed and the results of the election in that state.

II. METHODS

Our methodology composed of collecting tweets and associated metadata from Twitter and performing a sentiment analysis on each tweet. We developed a few algorithms and implemented them using the Java and Perl programming languages to access the tweets, clean them, perform a sentiment analysis and aggregate the results.

To collect data, we developed an algorithm that implemented the Twitter4J library [17]. This algorithm searched through all publicly posted tweets that mentioned, by name, either of the two presidential candidates, and stored all the Tweets that matched these criteria in text files. Our search words were "clinton", "hillary", and "trump". We did not include "donald" as a search word because it is a common name and could result in too many results that did not relate to the election. When we stored the tweets in the files, we included the tweet's text, the handle/username of the poster of the tweet, the timestamp of when the tweet was posted (using EST), the location provided by the user's Twitter bio, and a number representing which state (if any) corresponded to this location. These state numbers were determined using a database, explained in the next paragraph. Each tweet in the file, and the tweet's corresponding info, was separated from the next tweet by a series of dashes (-----). An example of how the tweet is laid out is shown in Fig. 1. Line numbers were included in this figure to aid in readability. Data from Twitter was collected for the week leading up to the US presidential election on November 8th. We separated the data by the date the tweet was posted, so we ended up with eight files containing tweets, one file for each day from November 1st to November 8th.

```
1  @Katastrophy1951 Trump only wants
       to take care of the girl and
       make sure she is not hurt
       anymore. Said he's worried
       about HER not his star.
2  Katastrophy1951
3  1 Nov 2016 05:46:02 GMT
4  Seattle ,WA.
5  46
6  -----
```

Figure 1: Example layout of a tweet in our data file with line numbers added in manually

After collecting the data, we began to analyze the tweets. First, we developed an algorithm to determine the state that the tweet corresponded with. To sort tweets, we looked at the location tag provided by the tweet's user. We parsed it for any mention of a US state, either by full name or abbreviation. If a state was found, the tweet was given a number representing that state. For example, Alabama was given a value of 00, Alaska got 01, Arizona got 02, and so on. Washington DC was also included, and was given its own value. In addition, if we detected a location origination from Mexico, Canada, or Great Britain we gave that tweet its own unique location number. These countries were included in our analyses so that we could see what people outside of the US were tweeting in regards to the election. If no US state or outside country was detected by name, then we parsed the location line for a town name. We used a database provided from SimpleMaps.com to match town names to states [18]. The database includes all Unites States towns and their matching states. If our algorithm detected multiple states associated with one town name, then we excluded that dataset. Any tweets that did not match with any state or did not have a location line were given a state value of -1 to signal that it was a tweet without a location. We still ran our algorithm on these tweets to see what results it would come up with.

The sentiment analysis of our algorithm was only run on Tweets that exclusively mentioned either candidate; tweets that mentioned both candidates were excluded from this portion of the algorithm. Sentiment was determined using the sentiment wordmap file from SentiWordNet with some modifications and additions to words associated with the election [19]. This sentiment file has a positive or negative score associated with each word. For positive words, values ranged from +0.0625 to +1.0; for negative words, values ranged from -0.0625 to -1.0. For example, a somewhat positive word like "helpful" would receive a value of +0.125 from the sentiment algorithm, or a strong negative word like "unhappy" would receive a value of -0.25. Neutral words like "a" or "the" were given neutral values of 0.0. Since these words would not affect the score, we decided to remove all neutral words from the file to improve the efficiency and speed of our algorithm. When our algorithm encountered a word in a tweet that was not in our sentiment file, it defaulted to giving it a neutral value of 0.

Another function we implemented was recognition of negation words like "not" or "don't." Our algorithm would detect these words as negations, and flip the value of the next word. So, while

"good" may return a value of +0.375, "not good" would return -0.375 because the algorithm recognized "not" as a negation word.

In addition to these functions within our algorithm, we wanted to further refine the sentiment word file. We went through the file by hand and removed or modified words that had inaccurate values. We also did sample testing on tweets from our data to see if the values of any words needed to be changed. For example, the word "investigation" was given a negative score because, in the context of the election, an investigation is never something good. Sometimes, words had to be added to the file. The word "WikiLeaks" was not included in the file originally, but we decided to add it and give it a negative value because it is normally associated with a scandal or something negative. Other words were completely removed from the sentiment file. For example, the word "trump" was taken out of the file. This was done because this verb had a positive value in the file, but since "Trump" is the name of one of the candidates, we had to remove it to prevent erroneous data.

Another aspect of our algorithm included scanning a tweet for different hashtags or phrases that would automatically indicate it as a positive or negative tweet. If a tweet included "#maga", "crooked Hillary", "lock her up", or "#trumptrain" then we knew for sure that it would be a positive tweet for Trump and a negative tweet for Hillary. On the other side, if a tweet contained "I'm with her", "Madame president", "dump trump", or "stronger together" then we could know for sure that it would be a positive tweet for Clinton, and a negative tweet for Trump. Hashtags and key phrases were collected from Top-Hashtags, which aggregated data from Twitter, Facebook, Instagram, and Tumblr [20].

To further refine our results, we looked to remove any duplicate tweets. Sometimes, multiple news organizations tweet a link to the same article, and use that article's headline as the content within the tweet. This can be seen in this tweet by @CollectedN that states "The Campaign Tour: Pop Musicians Get on the Bus (Mostly Clinton's) https://t.co/B6cpXuZ6Nf" [21] and this tweet by @thetablenz that states "The Campaign Tour: Pop Musicians Get on the Bus (Mostly Clinton's) https://t.co/SRx84KVxpA" [22]. Both are linking to an article published by the NY Times, and use the same headline as the tweet body. The only difference is the url that the tweet links to. When our algorithm collected tweets, it would save the entire tweet body, which includes any URLs. To remove duplicate tweets, such as these posts about news articles, we decided to remove any URLs from the tweets so that we could compare the raw text of each tweet. We then got rid of tweets that contained identical text, leaving us with only one copy. This improved our results by preventing repetitions within our data.

Using the sentiment file in addition to our updates, our algorithm would take a tweet and break it apart into words. Next, it would give each word a value per the sentiment file. After each word had a score, the algorithm summed up all the scores to determine the overall sentiment of the tweet. We counted tweets that had a score above 0.0 as positive tweets, tweets that had a score below 0.0 to be negative tweets, and tweet with a score of 0.0 as a neutral tweets.

An algorithm was also developed to see what topics were most widely discussed throughout our collection of tweets. To do this, we looked at topics discussed during the 2016 election, using research from political science sites such as FiveThirtyEight [23] and Pew Research Center [24], as well as polls conducted by the New York Times and the Washington Post [25]. We settled on 10 different issues: Economy, Education, Election Problems, Environment, Foreign Policy, Guns, Healthcare, Immigration, Social, and Trade. The full list of words we used to cover each topic is shown in Fig. 2.

Economy: job market, minimum wage, stimulus, economy, taxpayers
Education: common core, tuition, fafsa
Election: voter fraud
Environment: greenhouse, global warming, climate, pipeline, drilling, fracking, nuclear, solar
Foreign Policy: Syria, Isis, Israel, Patriot Act, surveillance, Al Qaeda, Iran, middle east
Guns: regulation, no-fly, concealed, carry law, Bernardino, pulse
Healthcare: Obamacare, Medicare, Medicaid, health care, vaccination, affordable care
Immigration: Mexico, Muslim, boarder wall, immigrant, refugee, xenophobia, daca, dapa, h1b
Social: parenthood, abortion, lgbtq, racism, feminism, homosexual, gender, religion, gay marriage
Trade: tpp, nafta, import, export, trade, business

Figure 2: List of keywords used when searching for different topics within tweets

III. RESULTS

Table 1 shows our results from the data that we collected. Tweets from the 50 US States are listed first, followed by tweets from Washington DC, then Canada, then Great Britain, and finally Mexico. Tweets that did not include locations are listed under N/A.

Each state's electoral outcome is shown in the far-right column. These results were gathered from the New York Times [26]. There are two US states that do not give all their electoral votes to the winner of that state, and these two states are Maine and Nebraska. In these states, congressional districts decide several electoral votes while the remaining are given to whichever candidate gets more votes in that state [27]. In the 2016 election, Maine was the only state to split its votes in this manner. It awarded three electoral votes to Hillary Clinton, who won the popular vote of the state as well as a congressional district, and 1 electoral vote to Donald Trump for winning a congressional district [28]. The remaining 49 states and Washington DC cast their electoral votes entirely for either Donald Trump or Hillary Clinton. The majority of Maine's electoral votes went to Hillary Clinton, and because she won that the popular vote there as well, we marked Maine as being won by Hillary Clinton. To understand if our results matched with those of the Electoral College, we compared the percentage of positive tweets for each candidate per state to one another, as well as the percentage of negative tweets for each candidate per state. A candidate's "favorability" was found by subtracting the percentage of negative

TABLE I: Correlation between Tweet Sentiment by State and Electoral College Result

State	Clinton Tweets	% Pos (Clinton)	% Neg (Clinton)	Trump Tweets	% Pos (Trump)	% Neg (Trump)	Twitter Results	Elect. College	Correct Prediction
Total	2810051	39.6%	45.7%	4044162	43.2%	42.2%	Trump	Trump	Yes
AL	13931	37.2%	48.9%	14650	45.7%	39.4%	Trump	Trump	Yes
AK	3404	45.7%	39.7%	2924	42.8%	42.7%	Clinton	Trump	No
AZ	19111	37.3%	49.4%	20333	44.7%	42.9%	Trump	Trump	Yes
AR	6272	39.1%	46.6%	6917	44.5%	41.6%	Trump	Trump	Yes
CA	103818	40.6%	45.7%	155246	42.6%	44.0%	Trump	Clinton	No
CO	18434	39.8%	47.0%	23128	42.4%	44.1%	Trump	Clinton	No
CT	10515	41.0%	45.0%	15397	42.1%	44.0%	Trump	Clinton	Yes
DE	1613	41.7%	43.5%	2641	40.0%	45.4%	Clinton	Clinton	No
FL	84085	37.6%	48.3%	108133	44.6%	40.9%	Trump	Trump	Yes
GA	27977	39.3%	47.4%	35206	44.6%	41.5%	Trump	Trump	Yes
HI	4267	37.4%	49.5%	5298	48.4%	38.8%	Trump	Clinton	No
ID	3429	34.5%	53.0%	4454	43.2%	43.9%	Trump	Trump	Yes
IL	24821	41.5%	45.0%	35979	42.7%	44.1%	Trump	Clinton	No
IN	37409	40.2%	45.6%	56576	42.5%	43.4%	Trump	Trump	Yes
IA	5808	41.1%	45.3%	8149	43.5%	43.3%	Trump	Trump	Yes
KS	6244	39.9%	45.7%	8788	43.4%	43.9%	Trump	Trump	Yes
KY	10280	37.0%	50.1%	12817	45.5%	41.4%	Trump	Trump	Yes
LA	18136	35.5%	44.7%	25244	37.6%	40.6%	Trump	Trump	Yes
ME	7152	38.6%	45.5%	10395	41.4%	42.9%	Trump	Clinton	No
MD	19105	42.8%	44.7%	12763	44.4%	41.8%	Trump	Clinton	No
MA	29621	43.4%	43.1%	19251	42.0%	44.4%	Clinton	Clinton	Yes
MI	23349	40.4%	44.3%	33935	43.8%	41.7%	Trump	Trump	Yes
MN	10004	42.7%	43.4%	17207	44.5%	42.0%	Trump	Clinton	No
MS	5679	37.5%	47.8%	6033	44.0%	39.5%	Trump	Trump	Yes
MO	14574	38.7%	48.4%	17740	44.7%	42.4%	Trump	Trump	Yes
MT	3697	34.6%	50.5%	4050	42.9%	43.1%	Trump	Trump	Yes
NE	3992	39.5%	45.5%	5586	43.5%	42.3%	Trump	Trump	Yes
NV	13664	38.1%	48.8%	17184	44.6%	42.6%	Trump	Clinton	No
NH	4704	40.3%	45.8%	6300	44.7%	41.8%	Trump	Clinton	No
NJ	24578	39.3%	47.3%	35938	42.5%	43.8%	Trump	Clinton	No
NM	4596	39.0%	48.7%	6127	43.7%	43.4%	Trump	Clinton	No
NY	87549	42.6%	43.2%	136087	42.7%	44.1%	Clinton	Clinton	Yes
NC	28432	41.0%	45.8%	37555	44.7%	42.6%	Trump	Trump	Yes
ND	1196	38.1%	49.2%	1338	47.8%	40.7%	Trump	Trump	Yes
OH	29317	41.1%	45.2%	39320	43.7%	42.6%	Trump	Trump	Yes
OK	9847	39.9%	45.4%	10429	44.5%	41.5%	Trump	Trump	Yes
OR	14058	40.4%	47.2%	21570	43.1%	44.9%	Trump	Clinton	No
PA	32602	40.9%	45.5%	45773	44.8%	41.8%	Trump	Trump	Yes
RI	2812	40.5%	46.7%	4472	42.9%	45.0%	Trump	Clinton	No
SC	13313	38.7%	47.2%	15496	44.3%	40.3%	Trump	Trump	Yes
SD	1204	41.4%	44.3%	1512	45.2%	40.1%	Trump	Trump	Yes
TN	16916	38.8%	48.4%	19025	46.0%	40.1%	Trump	Trump	Yes
TX	80770	37.6%	48.2%	98815	44.4%	42.0%	Trump	Trump	Yes
UT	4984	42.0%	43.3%	7027	44.2%	40.7%	Trump	Trump	Yes
VT	1726	44.0%	44.2%	2653	42.5%	47.3%	Clinton	Clinton	Yes
VA	26620	38.1%	49.2%	32515	44.3%	42.8%	Trump	Clinton	No
WA	21189	41.5%	45.1%	29288	42.4%	44.7%	Trump	Clinton	No
WV	4753	38.8%	47.6%	5544	45.4%	41.6%	Trump	Trump	Yes
WI	11657	41.8%	44.2%	15866	43.4%	42.8%	Trump	Trump	Yes
WY	1254	41.1%	48.8%	1638	46.5%	41.9%	Trump	Trump	Yes
DC	34016	45.0%	39.3%	52744	43.0%	41.8%	Clinton	Clinton	Yes
CAN	43087	40.2%	44.4%	74136	41.6%	44.1%	Trump	N/A	N/A
GB	51657	44.7%	41.0%	98721	44.3%	41.5%	Clinton	N/A	N/A
MX	6751	19.7%	25.2%	11444	19.0%	28.2%	Clinton	N/A	N/A
N/A	1720072	36.5%	42.8%	2546805	38.8%	40.0%	Trump	N/A	N/A

tweets for that candidate in a state from the percentage of positive tweets for that candidate in a state. Each candidate's favorability value per state was compared, and the candidate with the higher favorability in that state was marked under the "Twitter Results" column.

From Table 1, we can see that our sentiment showed more tweets being positive towards Trump and negative towards Clinton. Overall, there were only five states plus Washington D.C. that had a higher favorability for Clinton than Trump. The five states were Alaska, Delaware, Massachusetts, New York, and Vermont. Four of these five states were in the North-Eastern region of the United States. In addition, Alaska, which was the only non-Atlantic state to support Clinton on Twitter over Trump, ended up casting its electoral votes for Donald Trump in the election.

Table 1 also shows which states tweeted most, or had the highest percentage of positive or negative tweets. For both candidates, California and New York were the two states that tweeted about each candidate the most. It makes sense for California to have the most total tweets about each candidate since it is the state with the highest population in the United States. We decided to see if the population of a state had any effect on the tweet count coming from that state, so we decided to plot the total number of tweets from a state, and then compared it to the state's population. Data was gathered using 2016 US census counts [29], and results are shown in Table 2. These results are listed in order of descending population. Fig. 3 shows that there is a correlation between state population and tweet count. The linear growth of the graph shows that there is a somewhat constant ratio between state population and tweet count.

TABLE II: Comparison of State Population to Tweet Counts

State	Population	Total Tweets
CA	39250017	184588
TX	27862596	164855
FL	20612439	171634
NY	19745289	120151
PA	12802503	57423
IL	12801539	54138
OH	11614373	57294
GA	10310371	56409
NC	10146788	51781
MI	9928301	47927
NJ	8944469	51198
VA	8411808	47809
WA	7288000	40300
AZ	6931071	48732
MA	6811779	46537
TN	6651194	54325
IN	6633053	51983
MO	6093000	33679
MD	6016447	30762
WI	5778708	30091
CO	5540545	28438
MN	5519952	23317
SC	4961119	27244
AL	4863300	32067
LA	4681666	28416
KY	4436974	24338
OR	4093465	23905
OK	3923561	20362
CT	3576452	16323
IA	3134693	10792
UT	3051217	10663
MS	2988726	11951
AR	2988248	19936
NV	2940058	19908
KS	2907289	10840
NM	2081015	8588
NE	1907116	8745
WV	1831102	8182
ID	1683140	7696
HI	1428557	8971
NH	1334795	11856
ME	1331479	9964
RI	1056426	6509
MT	1042520	5310
DE	952065	2817
SD	865454	2400
ND	757952	4600
AK	741894	37420
DC	681170	35742
VT	624594	2980
WY	585501	1254

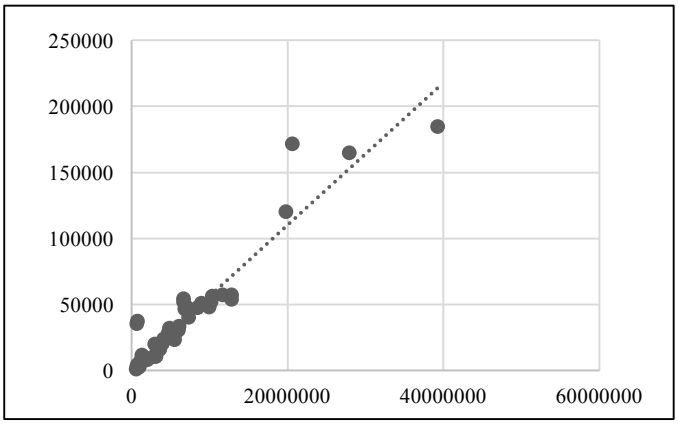

Figure 3: Graph plotting State Population vs Tweet Count, with line of best fit

In addition to looking at the 50 states and Washington D.C., we also looked at how other countries discussed the two candidates. Canada had far more tweets that discussed Trump than those that discussed Clinton, about 74,000 compared to 43,000. However, the positive and negative percentages were about the same between each candidate, with 40% to 41% of tweets being positive for Clinton and Trump respectively, and 44% of tweets being negative for both. Great Britain was similar, having about 51,000 tweets about Clinton compared to 98,000 tweets about Trump. The positive percentages for Clinton and Trump were both around 44%, while Clinton had a slightly lower negative percentage of 40% compared to Trump's 41%. In Mexico, the positive and negative percentages were both lower than anywhere else, but this is likely since many of these tweets were written in Spanish, and our algorithm used an English dictionary to operate.

States that had a calculated favor that matched their Electoral College votes were marked with a green highlight under the final column, "Correct Prediction" in Table 1. Out of the 50 states, plus Washington DC, the sentiment on Twitter for 34 states correctly aligned with the Electoral College's outcome for that state, resulting in an accuracy of about 66.7%.

We wanted to see if the geographical location of a state affected how it voted in the election. To start, we had to divide the 50 states into regions. We looked for different maps that showed various ways to section off the United States, and finally settled on a map with the following five regions: Northeast (ME, VT, NH, RI, CT, NJ, NY, PA), Southeast (MD, DE, DC, WV, VA, KY, TN, NC, SC, AR, LA, MS, AL, GA, FL), Midwest (ND, SD, MN, WI, MI, NE, IA, IL, IN, OH, KS, MO), Southwest (AZ, NM, OK, TX), and West (AK, HI,

WA, OR, ID, MT, WY, CA, NV, UT, CO). After combining state results from each state by region, the results are shown in Table 3.

In addition to the results in Table 1, we were also able to see what issues were most discussed by each state leading up to the election. In Table 4, we show the percentage of issues discussed by states that were, per the Electoral College vote, won by Donald Trump. In Table 5, we show the percentage of issues discussed by states that Hillary Clinton won. Column one lists the abbreviation for each state, and column two lists the total number of tweets from that state that mentioned one of the issues we searched for. From there, the remaining columns each list the percentage of that total that relate to each issue.

From Table 4, we can see that out of the 30 states that Donald Trump won in the election, the most frequently discussed issues were Immigration and Foreign Policy. Trade issues and Social issues were also discussed often, as well as Gun issues, Healthcare, and Environmental Issues. Finally, Economic issues and Election issues were discussed the least often, and Education was barely discussed in any tweets.

Table 5 shows that out of the 21 states that Hillary Clinton won in the election, the most frequently discussed issues were Immigration and Foreign Policy, which matches the discussion trends for states that voted for Trump. Following this, Foreign Policy, Social issues, and Trade issues were the next most discussed topics. Environment, Guns, and Healthcare were tweeted about less often than these, while Economic and Election issues were talked about even less frequently. Finally, Education issues were talked about very minimally throughout our tweets. This data shows that, overall, the candidate that a state voted for had little impact on which issues were most talked about by that state.

TABLE III: Division of Twitter Results by Geographic Location

Region	Clinton Tweets	% Pos (Clinton)	% Neg (Clinton)	Trump Tweets	% Pos (Trump)	% Neg (Trump)
Total	2810051	39.6%	45.7%	4044162	43.2%	42.2%
N.East	201259	41.7%	44.4%	276266	42.9%	43.6%
S.East	311128	39.3%	46.6%	387283	44.0%	41.3%
Midwest	169575	40.7%	45.3%	241996	43.4%	42.9%
S.West	114324	37.8%	48.2%	135704	44.4%	42.2%
West	192198	40.3%	46.2%	271807	42.9%	43.8%

TABLE IV: Issue Discussion Percentage on States that Donald Trump Won in the Electoral College

State	Total	Economy	Education	Election	Environment	Guns	Healthcare	Immigration	Foreign	Social	Trade
AL	1531	5.0%	0.3%	3.6%	7.1%	10.3%	8.0%	21.2%	17.1%	15.6%	11.9%
AK	352	2.8%	0.0%	1.7%	10.8%	9.9%	7.7%	20.7%	16.5%	15.1%	14.8%
AZ	2954	3.9%	0.3%	3.0%	7.8%	9.3%	8.4%	21.4%	17.7%	14.4%	13.9%
AR	787	3.0%	0.4%	3.6%	9.8%	11.4%	8.1%	18.9%	17.3%	16.9%	10.5%
FL	13178	3.8%	0.2%	3.0%	7.7%	9.1%	8.5%	20.9%	18.5%	12.9%	15.4%
GA	3719	3.7%	0.1%	2.6%	6.9%	11.5%	8.7%	20.4%	15.9%	14.6%	15.5%
ID	430	2.6%	0.0%	1.9%	7.0%	7.4%	9.3%	23.5%	14.4%	19.5%	14.4%
IN	5680	3.5%	0.3%	2.3%	7.9%	9.7%	6.0%	18.9%	22.7%	17.1%	11.6%
IA	681	3.1%	1.0%	3.2%	9.8%	9.8%	8.1%	22.9%	13.5%	16.3%	12.2%
KS	873	2.2%	0.1%	1.9%	6.8%	11.5%	5.6%	18.2%	21.0%	20.0%	12.7%
KY	1224	2.6%	0.2%	3.6%	7.0%	10.2%	8.4%	19.9%	18.3%	15.7%	14.1%
LA	2226	3.1%	0.2%	2.7%	7.7%	11.8%	6.9%	23.4%	17.8%	14.4%	11.9%
MI	3277	2.8%	0.3%	1.6%	8.2%	9.9%	7.2%	16.9%	16.1%	16.3%	20.6%
MS	567	3.0%	0.5%	3.0%	6.7%	10.9%	10.2%	18.0%	18.0%	18.0%	11.6%
MO	1871	3.4%	0.1%	2.7%	6.6%	11.5%	10.6%	19.5%	16.5%	17.0%	12.1%
MT	452	2.9%	0.0%	3.1%	6.4%	9.1%	6.6%	21.9%	22.8%	13.1%	14.2%
NE	462	4.5%	0.6%	3.2%	10.2%	10.6%	6.7%	21.2%	14.1%	16.9%	11.9%
NC	3757	4.4%	0.5%	2.4%	6.5%	13.7%	8.8%	19.1%	15.3%	16.4%	12.8%
ND	123	0.0%	0.0%	0.0%	16.3%	18.7%	3.3%	22.0%	17.1%	10.6%	12.2%
OH	3547	4.1%	0.4%	2.8%	12.2%	10.8%	7.1%	17.7%	13.5%	17.5%	14.0%
OK	1033	3.8%	0.4%	2.5%	8.5%	9.7%	7.6%	21.4%	16.4%	18.2%	11.6%
PA	4761	4.1%	0.3%	2.6%	9.2%	8.9%	9.1%	20.4%	14.7%	19.3%	11.4%
SC	1573	4.6%	0.3%	2.2%	6.3%	8.9%	9.3%	19.0%	17.9%	20.9%	10.7%
SD	158	5.1%	0.0%	1.9%	7.0%	15.8%	9.5%	21.5%	11.4%	10.1%	17.7%
TN	1983	3.8%	0.1%	2.7%	7.2%	10.8%	9.7%	20.0%	17.2%	15.3%	13.2%
TX	10846	4.3%	0.3%	3.1%	7.2%	9.6%	8.0%	22.2%	17.4%	15.3%	12.5%
UT	614	2.4%	0.3%	2.1%	10.9%	10.9%	8.5%	19.4%	15.3%	17.6%	12.5%
WV	541	4.4%	0.2%	1.8%	5.4%	13.1%	10.5%	20.0%	14.4%	16.5%	13.7%
WI	1484	4.9%	0.5%	3.7%	10.3%	12.7%	9.6%	17.4%	13.2%	15.3%	12.5%
WY	180	3.9%	0.0%	1.1%	9.4%	8.3%	8.9%	26.7%	18.3%	8.9%	14.4%

TABLE IV: Issue Discussion Percentage on States that Hillary Clinton Won in the Electoral College

State	Total	Economy	Education	Election	Environment	Guns	Healthcare	Immigration	Foreign	Social	Trade
CA	17593	4.1%	0.2%	2.4%	9.6%	9.3%	7.6%	20.4%	17.1%	15.2%	14.0%
CO	2593	2.9%	0.2%	2.5%	10.9%	9.3%	7.4%	19.5%	16.7%	16.6%	14.1%
CT	1313	4.4%	0.5%	2.6%	7.9%	10.7%	8.5%	17.6%	17.0%	15.8%	14.9%
DE	1633	2.6%	0.1%	1.0%	7.2%	6.7%	5.3%	29.3%	28.8%	11.8%	7.1%
HI	547	4.6%	0.4%	4.6%	8.8%	8.8%	8.6%	19.4%	19.7%	13.5%	11.7%
IL	3356	3.9%	0.2%	1.9%	9.7%	10.3%	9.1%	19.3%	15.6%	17.3%	12.7%
ME	900	2.1%	0.3%	1.0%	10.1%	12.0%	8.2%	19.2%	17.3%	16.6%	13.1%
MD	1823	3.9%	0.2%	2.4%	8.9%	10.3%	6.3%	22.2%	13.5%	18.7%	13.7%
MA	2894	4.5%	0.3%	2.0%	10.4%	9.8%	7.6%	21.1%	15.3%	16.7%	12.3%
MN	1561	4.2%	0.2%	2.5%	9.4%	10.5%	7.2%	21.7%	12.6%	14.3%	17.3%
NV	2180	2.9%	0.5%	1.8%	7.2%	9.7%	7.9%	20.9%	23.8%	10.9%	14.3%
NH	850	4.5%	0.1%	2.2%	10.4%	7.9%	5.8%	14.6%	10.8%	10.5%	33.3%
NJ	3768	3.1%	0.3%	2.0%	8.2%	9.5%	7.8%	21.7%	17.8%	15.8%	13.8%
NM	923	2.7%	0.0%	5.4%	8.7%	8.5%	4.4%	30.6%	14.1%	12.6%	13.1%
NY	16730	4.0%	0.2%	1.8%	9.4%	9.2%	8.3%	21.9%	16.1%	14.8%	14.3%
OR	2080	4.0%	0.4%	2.8%	10.6%	10.1%	7.4%	20.4%	13.9%	16.7%	13.7%
RI	398	6.5%	0.3%	2.0%	9.0%	13.3%	7.5%	18.6%	12.1%	14.8%	15.8%
VT	294	3.4%	0.7%	0.7%	18.4%	7.8%	11.9%	13.9%	9.9%	22.1%	11.2%
VA	3336	3.6%	0.2%	2.3%	8.2%	10.2%	7.6%	19.3%	20.3%	15.0%	13.3%
WA	2982	3.0%	0.2%	2.6%	11.0%	9.9%	6.0%	19.8%	17.8%	15.6%	14.1%
DC	6757	3.7%	0.1%	1.8%	10.9%	8.4%	11.5%	22.1%	14.3%	14.6%	12.5%

IV. DISCUSSION

The goal of this study was to determine if the sentiment on Twitter for the 2016 election matched with the electoral results of the election. Our results showed us that results from Twitter sometimes matched, but were not always accurate with the results of the Electoral College.

Of the 50 states plus Washington DC, 34 states had results from Twitter that matched their Electoral College results. Of the states that Twitter correctly predicted, the only state that supported Hillary on Twitter but voted for Trump was Alaska. There were 29 states, Alabama, Arizona, Arkansas, Florida, Georgia, Idaho, Indiana, Iowa, Kansas, Kentucky, Louisiana, Michigan, Mississippi, Missouri, Montana, Nebraska, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, South Carolina, South Dakota, Tennessee, Texas, Utah, West Virginia, Wisconsin, and Wyoming, all voted for Trump in the election and showed more support on Twitter for Trump as well. There were five states, including Washington DC, that backed Hillary Clinton in our finding as well as in the election. These states were Delaware, Massachusetts, New York, Vermont, and Washington DC. This leaves 17 states that voted for Clinton yet supported Trump more on Twitter: California, Colorado, Connecticut, Hawaii, Illinois, Maine, Maryland, Minnesota, Nevada, New Hampshire, New Mexico, Oregon, Rhode Island, Virginia, and Washington state.

When looking at states that had differing results on Twitter, there were more states that voted for Clinton yet supported Trump on Twitter than states that voted for Trump yet supported Clinton on Twitter. Of the 30 states that Trump won in the election, Twitter sentiments matched for 96.7% of them. For Clinton's 20 states plus D.C., Twitter sentiments matched 28.6% of the state results. Comparing these percentages shows that, while overall the Twitter sentiments lined up with state results 66.7% of the time, most of this percentage came from states that supported Donald Trump on Twitter as well as in the election.

These results point to Twitter being either more positive towards Donald Trump, more negative towards Hillary Clinton, or both. In fact, Donald Trump was discussed in Tweets far more often than Hillary Clinton was; there was not a single state that tweeted about Hillary Clinton more than Donald Trump. Although this does not correlate to an increase in positive tweets for Trump, it could certainly be a factor. A final piece of evidence is that, per our tweet analysis, Donald Trump had a higher total favorability rating than Hillary Clinton. The percentage of positive tweets for Trump was 43.47% while Clinton had a lower percent, with 40.03% of tweets about her being marked as positive. Hillary Clinton's negative tweet percentage was 46.03%, which is a bit higher than Donald Trump's 42.71%.

By looking at the most discussed issues, we can see that issues relating to the two candidates was by far the most discussed out of any of the topics. This is likely because, if someone is going to be tweeting about a candidate, there is a high chance that they will be talking about something related to that candidate, such as an issue or scandal revolving around the candidate. Following this, immigration was the next most discussed topic, which again makes sense because it was a very heavily discussed topic during the election. News sources such as USA Today claimed "Immigration at front of 2016 presidential race" and our data shows this is likely true per Twitter [31].

There are many reasons for why the correlation is not fully accurate and leaned heavily towards Trump as opposed to Clinton. Firstly, our sentiment algorithm is not perfect when used to analyze the word structure. It works by summing up values of individual words, but cannot detect the sentiment from complex word clauses or phrases consisting of multiple words. One example of this is the following tweet by Twitter user @buttonlol who tweeted "another example of above the law… Hitlerary [sic] Clinton…" [32]. Although this tweet is understood to be negative towards Clinton by the usage of the phrase "above the law," the three words that

make up that phrase carry little to no connotation on their own, and thus our sentiment algorithm would be unable to classify it as negative.

In addition, we only dealt with tweets that exclusively mentioned a candidate. If a tweet were to mention both Trump and Clinton, our algorithm could not identify who the subject was or who to give the sentiment score to, so we discarded all these tweets. If we could include these tweets and isolate which candidate the tweet was focusing on to attach a sentiment, the numbers for both candidates would certainly be changed. Whether this would balance out the results or further increase the divide is unknown, but it would certainly make the results much more accurate.

Sarcasm is another weakness of our algorithm, and when testing, we saw that it was impossible for our algorithm to detect sarcasm at this stage. One example of sarcasm is a section from a tweet by Twitter user @followingleads who tweeted "…nice unbiased piece on Trump today…" [33]. Out of context, it appears this tweet is supporting an article posted about Trump. However, this tweet was posted as a sarcastic response to an anti-Trump journalist. This context can be determined by looking at who the tweet was directed at, as well as reading into the rest of the tweet, where @followingleads writes that this is sarcastic praise. A computer would not be able to determine this sarcasm.

In addition, since tweets can be posted by anyone and have no quality or content assurance, spelling and grammar is not guaranteed to be fully accurate, so if key words were misspelt, they could be ignored or incorrectly analyzed by our algorithm. One example is the following tweet by user @komptonmusic who tweeted "I'm suddenly voting Hilary [sic]" [34]. In this tweet, the user misspelt Hillary Clinton's first name, which means that our searching algorithm did not pick it up at all. If the algorithm had seen the spelling error and collected the tweet, then it would have most likely been counted as an additional positive tweet for Clinton.

Looking at related studies shows similar trends. One study by Kunal Jagtap [35] showed that, per tweets collected for 5 days, Donald Trump was tweeted about negatively more than Hillary Clinton was. However, other algorithms and systems that used a wider field of data came up with different results. An AI system named "MoglA", created by Sanjiv Rai [36], took in data from numerous social media platforms, such as Google, Facebook, Twitter, and YouTube, and used this information to come up with a conclusion. Rai's system found that people were engaging more with content related to Donald Trump than they were with content from Hillary Clinton, and in the past three elections, the candidate with more engagement ended up winning the election. From this, Rai's AI determined that Donald Trump would win the election over Hillary Clinton. In our study, we found that there was a total of 4,044,162 tweets originating from the United States that exclusively mentioned Donald Trump, and a total of 2,810,051 tweets originating from the United States that exclusively mentioned Hillary Clinton.

Though most predictions showed Clinton winning the election, Trump ended up with the victory. One of the main reasons for why Twitter was not able to accurately predict these election results is because it is not a good sample of the population. Results from Twitter are only showing the sentiments of those who are actively using the platform. Not every American uses Twitter, and in addition, discussions about the candidates on Twitter do not always correlate to how people vote.

V. CONCLUSION

Twitter is used globally by millions of users, and discussions on the social media platform can sometimes be used to see what the public's opinion is regarding a certain issue. After collecting tweets about the US presidential election and analyzing them to determine how people viewed the two candidates, we saw that the sentiment according to Twitter was somewhat accurate. If sentiment analysis algorithms are improved and further developed, they could be used to predict election results in the future.

REFERENCES

[1] N. Silver, "Election Update: The How-Full-Is-This-Glass Election," Last modified November 2, 2016, http://fivethirtyeight.com/features/election-update-the-how-full-is-this-glass-election/

[2] "What happens if the US presidential election is close?" Reuters, last modified November 4, 2016, http://www.hindustantimes.com/world-news/factbox-what-happens-if-the-u-s-election-is-close/story-ZpB0gvmqTVRkAkqtCYmZiK.html

[3] J. Uscinski, "The real presidential election is in December when the Electoral College votes," Last modified November 7, 2016, http://www.miamiherald.com/opinion/op-ed/article113171073.html

[4] S. Reich, "What will the US presidential election mean for Europe?" http://blogs.lse.ac.uk/europpblog/2016/11/01/trump-clinton-nato-europe/

[5] "Number of monthly active Facebook users worldwide as of 3rd quarter 2016 (in millions)" https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[6] "Twitter Usage / Company Facts" Twitter Inc., https://about.twitter.com/company

[7] "Number of daily active Snapchat users from March 2014 to June 2016 (in millions)" https://www.statista.com/statistics/545967/snapchat-app-dau/

[8] "Number of monthly active Instagram users from January 2013 to June 2016 (in millions)" https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

[9] "Getting Started with Twitter," Twitter Inc., https://support.twitter.com/articles/215585

[10] "Posting a Tweet," Twitter Inc., https://support.twitter.com/articles/15367

[11] H. Kwak, C. Lee, H. Park, S. Moon, "What is Twitter, a social network or a news media?" ACM pp. 591-600, 2010.

[12] "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2016 (in millions)" https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis." Foundations and Trends in Information Retrieval, vol 2, pp. 1-135. 2008.

[14] A. Go, L. Huang, and R. Bhayani, "Twitter Sentiment Classification using Distant Supervision." The Stanford Natural Language Processing Group, 2009, https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[15] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc, vol. 10, pp. 1320-1326. 2010.

[16] "Big {Political} Data," https://www.isidewith.com/polls

[17] "Introduction" Twitter 4J. http://twitter4j.org/en/index.html

[18] "US Zip Codes Database," SimpleMaps, http://simplemaps.com/data/us-zips

[19] SentiWordNet, http://sentiwordnet.isti.cnr.it/

[20] "Hashtags for #election2016 in Instagram, Twitter, Facebook, Tumblr, ello," Top-Hashtags. https://top-hashtags.com/hashtag/election2016/

[21] @CollectedN. Twitter Post. November 3, 2016, 6:45 PM. https://twitter.com/CollectedN/status/794309570759946241

[22] @thetablenz. Twitter Post. November 3, 2016, 11:24 PM. https://twitter.com/thetablenz/status/794379871342313472

[23] "The Big Issues of the 2016 Campaign," Five Thirty-Eight, last modified November 19, 2015. https://fivethirtyeight.com/features/year-ahead-project/

[24] "Top voting issues in 2016 election," Pew Research Center, last modified July 7, 2016. http://www.people-press.org/2016/07/07/4-top-voting-issues-in-2016-election/

[25] "Problems and Priorities," Polling Report. Polls conducted September 5-8 and October 28-November 1. http://pollingreport.com/prioriti.htm

[26] "Presidential Election Results," The New York Times, http://www.nytimes.com/elections/results/president

[27] "What is the difference between the winner-takes-all rule and proportional voting, and which states follow which rule?" Frequently Asked Questions, National Archives and Records Administration, https://www.archives.gov/federal-register/electoral-college/faq.html

[28] "Maine Results," The New York Times, http://www.nytimes.com/elections/results/maine

[29] Unites States Census Bureau, https://www.census.gov/quickfacts/table

[30] "Immigration at front of 2016 presidential race," USA Today, http://www.usatoday.com/story/news/politics/elections/2015/05/15/immigration-2016-presidential-race/27360717/

[31] @buttonlol. Twitter Post. October 23, 2016, 7:42 PM. https://twitter.com/buttonlol/status/790337640222560256

[32] @followingleads. Twitter Post. November 3, 2016, 7:58 PM. https://twitter.com/followingleads/status/794328035071705088

[33] @komptonmusic. Twitter Post. November 3, 2016, 7:51 PM. https://twitter.com/KomptonMusic/status/794326138742972416

[34] K. Jagtap, "2016 USA Presidential Election Twitter Sentiment Analysis & Topic Modelling." LinkedIn Pulse, April 5, 2016. Accessed December 5, 2016, https://www.linkedin.com/pulse/2016-usa-presidential-election-twitter-sentiment-analysis-jagtap

[35] A. Kharpal, "Trump will win the election and is more popular than Obama in 2008, AI system finds." CNBC, October 28, 2016. Accessed October 28, 2016. http://www.cnbc.com/2016/10/28/donald-trump-will-win-the-election-and-is-more-popular-than-obama-in-2008-ai-system-finds.html