# Twitter Sentiment Analysis Of 2016 United States Presidential Election

**Authors:**
**Shaurya Shekhar (14BCE0497), Aarushi Thakral (14BCE0499)**

**Guided By:**
**Prof. Vasantha WB**

## ABSTRACT

Twitter is amongst the most popular social-networking websites today [1], with approximately 317 million monthly active users (Quarter 3 2016). Of these 67 million users are from the United States of America. Twitter being a micro-blogging platform, is widely used by people to express their opinions. Approximately 500 million tweets are posted in a day, which is around 6000 tweets per second. Assuming, even one-tenth of these tweets reflect an emotion, that results in a lot of people-generated data, which can prove to be a treasure trove of information if studied carefully. We intend to perform sentimental analysis on Twitter Data of the United States 2016 Presidential Election and then overlay our findings with respect to the two main candidates: Hillary Clinton & Donald Trump with the actual election result, to be able to categorically state whether twitter can be used as a proper indication of any election.

## INTRODUCTION

In the past few years, we have seen an exponential increase in the number of users of microblogging platforms such as Twitter. These platforms allow people to express their views and opinions on a variety of issues. These views can then be used by organizations to analyze the sentiment of the general public. Twitter has grown by leaps and bounds in the past few years, the scale of which can be guessed by the fact that in 2010, they had 30 million users whereas today they have almost 328 million users. Almost 500 million tweets are sent each day. This, allowing for Twitter data to be used as an accurate representation of the sentiments of the people.
Elections allow for citizens of democratic nations to use their power to elect the people to offices of power. A democracy allows for a government for the people to be elected by the people, and thus, every election is an important event in the timeline of every nation. It allows people to play their role in deciding whose policies would be guiding the nation and who should rule the country.
Previous attempts to do the same have been done using identification of smileys in tweets and classifying them using Bayesian networks [6].We thus, perform sentiment analysis on Twitter data to predict which of the 2 Presidential candidates: Hillary Clinton and Donald J. Trump would win in which state [2][3][7][8][9]. We have performed Twitter sentiment analysis using the R programming language [5].
Section 1 deals with how tweets have been mined and collected. Section 2 talks about the preprocessing and cleaning of tweets. Section 3 explains the processing and classification of states. Section 4 visualizes the findings in the form of bar plots and maps. Section 5 presents a brief summary and explains the result & conclusion of this project.

## 1. Data Mining

Using a registered Twitter account, we created an application for the purpose of mining data for sentiment analysis. Upon creating an application we were provided with a Consumer Key (API key), Consumer Secret (API Secret), Access Token and Access Token Secret.

The 'twitteR' package was installed in RStudio, to provide an interface to the generated Twitter web API [4].

After installing the 'twitteR' package and generating your consumer key, consumer secret, access token and access token secret, we make use of the 'setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)', which wraps the OAuth authentication handshake functions from the httr package for a twitteR session. This allows RStudio to be able to access twitter data on basis of parameters supplied to the 'searchTwitter()' function.

Data from the most populous cities of the 50 states of United States of America have been gathered. The latitudes and longitudes of these cities were found and stored.

Since, we have focused on finding tweets one day before the main Presidential election which was held on 8th November 2016, we have limited the date range to since "2016-11-07". 5 parameters were passed to the searchTwitter() function. These were:-

1. Search Phrase: which was set to either Trump or Clinton depending on the candidate whose tweets we are trying to find.
2. Number of tweets: which was set to 1000 for all the searchTwitter operations.
3. Language: which was set to "en", thus limiting the searching of tweets to only those which were in English for all searchTwitter operations.
4. Geocode: which was used to specify the latitude and longitude of the place from where tweets have to be retrieved. We also set the range (in miles) from where tweets had to retrieved from around the location specified.
5. Date: which has been to "2016-11-07" for all searchTwitter operations.

This resulted in us retrieving sets of 1000 tweets for each Clinton and Trump for each state, thus creating a dataset of 1,00,000 tweets. These tweets were in the form of a tweet list, which specified a number of parameters such as whether the tweets had been favorite or retweeted or not, whether the tweets were in reply to some user, the source link of the tweet, etc.

For further processing, these tweets were converted into a data frame using the twListToDF function, which is included in the 'twitteR' package.

## 2. Data Preprocessing

The tweets which we have extracted include emojis and links to other websites. These cannot be used to judge or gauge the sentiment of the tweets, as there exist a variety of ways an emoji can be used, such as the fact the smiling emoji could express happiness as well as sarcasm. Similarly, we are performing sentiment analysis only on the text found in the tweets and not on the text found on the pages which these tweets further link to. Thus, before we can proceed with the processing of these tweets, these tweets need to be cleaned of emojis and links.

The text columns from the data frames were selected and selectively cleared of emojis and links row by row. The emojis once downloaded into R, automatically got converted into Latin and thus had to be first converted into ASCII and substituted with “ “ (blank spaces). This made use of the sapply() function.

To clean the tweets of links, we made use of the gsub() function. We used the regular expression of hyperlinks- “(f|ht)tp(s?)://(.*)[.][a-z]” and substituted it with “ “ (blank spaces).

Lastly the cleaned text columns of these tweets are stored in a separate data frame.

## 3. Data Processing

A database of positive and negative words are loaded into RStudio.

| Table 1: Sample Words From Databases | |
|---|---|
| **Positive Words** | **Negative Words** |
| Adore, advocate, affirmative, fancy, fanfare, idolize, immense, impressive, lavish, neat, nicest, pleased, poise, qualified, revival, savior | Anxiety, antipathy, apocalypse, awful, belligerent, biased, bizarre, blunder, bogus, confront, crisis, degrading, harsh, poverty, suspicious |

A function “score.sentiment(parameters)” was defined with parameters being the negative words, positive words, and the text data frames from the tweets. The functions substituted punctuations, control words, decimal numbers and new lines with “ “ (blank spaces) to further clean the tweets, thus allowing for easier processing. The text was then converted to lower case using the tolower() function included in the “stringr” package of R. The tweets were then split word wise and stored in a list, which was further stored in a character vector. The “match()” function was used to compare every word in the character vector to the positive words and negative words database. We thus, get either the count of matches to the different databases or “NA” in cases of no matches. This, method is referred to as ‘lexical analysis’ [10]. We remove the “NA” values and sum the remaining number of matches to be able to calculate the score of the tweets using the formula

**Score = sum(pos.matches) – sum(neg.matches)**

The scores, positive scores and negative scores are then appended to each other in a single table. This leads to multiple columns of data for each of the values. This can be decomposed into a single column of data using the “melt()” function from the “reshape2” package of R. The melt function takes data in wide format and stacks a set of columns into a single column of data.

For every state, we find all the three scores (final, positive and negative), for each of the tweets, which is then summed up in final values. This is done for both Clinton & Trump.

| Table 2 | | | |
|---|---|---|---|
| Clinton Scores (Final, Positive, Negative) | | | |
| **State_Column** | **Score** | **Positive** | **Negative** |
| Alabama | -134 | 498 | 632 |
| Alaska | -15 | 759 | 774 |
| Arizona | -140 | 498 | 638 |
| Arkansas | 1173 | 1446 | 273 |
| California | -345 | 370 | 715 |
| Colorado | -66 | 632 | 698 |
| Connecticut | 686 | 1290 | 604 |
| Delaware | 318 | 851 | 533 |
| Florida | -622 | 421 | 1043 |
| Georgia | -29 | 574 | 603 |
| Hawaii | 19 | 119 | 100 |
| Idaho | 46 | 589 | 543 |
| Illinois | 269 | 549 | 280 |
| Indiana | -31 | 127 | 158 |
| Iowa | -34 | 569 | 603 |
| Kansas | 213 | 676 | 463 |
| Kentucky | -54 | 536 | 590 |
| Lousianna | 82 | 576 | 494 |
| Maine | 647 | 1669 | 1022 |
| Maryland | 386 | 765 | 379 |
| Massachusetts | -64 | 539 | 603 |
| Michigan | 116 | 583 | 467 |
| Minnesota | -16 | 609 | 625 |
| Missisippi | 152 | 822 | 670 |
| Missouri | 117 | 730 | 613 |
| Montana | -438 | 451 | 889 |
| Nebraska | 118 | 708 | 590 |
| Nevada | 104 | 603 | 499 |
| New Hampshire | 313 | 865 | 552 |
| New Jersey | 323 | 875 | 552 |
| New Mexico | -465 | 434 | 899 |
| New York | 338 | 851 | 513 |
| North Carolina | -41 | 583 | 624 |
| North Dakota | 305 | 755 | 450 |
| Ohio | 229 | 826 | 597 |
| Oklahoma | 171 | 736 | 565 |
| Oregon | -651 | 419 | 1070 |
| Pennsylvania | 287 | 837 | 550 |
| Rhode Island | 297 | 841 | 544 |
| South Carolina | 192 | 801 | 609 |
| South Dakota | 214 | 773 | 559 |
| Tennessee | 189 | 805 | 616 |
| Texas | 233 | 796 | 563 |
| Utah | -369 | 537 | 906 |

| | | | |
|---|---|---|---|
| Vermont | 274 | 843 | 569 |
| Virginia | 262 | 841 | 579 |
| Washington | -531 | 457 | 988 |
| West Virginia | 203 | 827 | 624 |
| Wisconsin | 212 | 833 | 621 |
| Wyoming | 35 | 679 | 644 |

| **Table 3** | | | |
|---|---|---|---|
| **Trump Scores (Final, Positive, Negative)** | | | |
| **State_Column** | **Score** | **Positive** | **Negative** |
| Alabama | 710 | 1273 | 563 |
| Alaska | 600 | 1256 | 656 |
| Arizona | 224 | 1185 | 961 |
| Arkansas | 714 | 1385 | 671 |
| California | 1169 | 1668 | 499 |
| Colorado | 732 | 1439 | 707 |
| Connecticut | 432 | 1267 | 835 |
| Delaware | 913 | 1424 | 511 |
| Florida | 596 | 1310 | 714 |
| Georgia | 555 | 1242 | 687 |
| Hawaii | 721 | 1361 | 640 |
| Idaho | 559 | 1239 | 680 |
| Illinois | 630 | 1287 | 657 |
| Indiana | 682 | 1374 | 692 |
| Iowa | 621 | 1487 | 866 |
| Kansas | 669 | 1403 | 734 |
| Kentucky | 1034 | 1473 | 439 |
| Lousianna | 575 | 1256 | 681 |
| Maine | 659 | 1212 | 553 |
| Maryland | 453 | 1293 | 840 |
| Massachusetts | 467 | 1503 | 1036 |
| Michigan | 464 | 1205 | 741 |
| Minnesota | 481 | 1247 | 766 |
| Missisippi | 560 | 1208 | 648 |
| Missouri | 546 | 1200 | 654 |
| Montana | 760 | 1353 | 593 |
| Nebraska | 494 | 1182 | 688 |
| Nevada | 732 | 1417 | 685 |
| New Hampshire | 510 | 1175 | 665 |
| New Jersey | 612 | 1212 | 600 |
| New Mexico | 661 | 1284 | 623 |
| New York | 514 | 1178 | 664 |
| North Carolina | 672 | 1334 | 662 |
| North Dakota | 522 | 1259 | 737 |
| Ohio | 651 | 1263 | 612 |
| Oklahoma | 636 | 1277 | 641 |
| Oregon | 642 | 1378 | 736 |

| | | | |
|---|---|---|---|
| Pennsylvania | 651 | 1243 | 592 |
| Rhode Island | 711 | 1286 | 575 |
| South Carolina | 678 | 1304 | 626 |
| South Dakota | 676 | 1345 | 669 |
| Tennessee | 570 | 1219 | 649 |
| Texas | 464 | 1183 | 719 |
| Utah | 607 | 1342 | 735 |
| Vermont | 551 | 1220 | 669 |
| Virginia | 545 | 1235 | 690 |
| Washington | 681 | 1419 | 738 |
| West Virginia | -607 | 807 | 1414 |
| Wisconsin | -2853 | 49 | 2902 |
| Wyoming | 598 | 1307 | 709 |

Depending on the basis of who has greater final score in the state, we then used used classification to predict who would win in which state.

| **Table 4** | |
|---|---|
| **Predictions** | |
| **State_Column** | **Predicted Winner** |
| Alabama | Trump |
| Alaska | Trump |
| Arizona | Trump |
| Arkansas | Clinton |
| California | Trump |
| Colorado | Trump |
| Connecticut | Clinton |
| Delaware | Trump |
| Florida | Trump |
| Georgia | Trump |
| Hawaii | Trump |
| Idaho | Trump |
| Illinois | Trump |
| Indiana | Trump |
| Iowa | Trump |
| Kansas | Trump |
| Kentucky | Trump |
| Lousianna | Trump |
| Maine | Trump |
| Maryland | Trump |
| Massachusetts | Trump |
| Michigan | Trump |
| Minnesota | Trump |
| Missisippi | Trump |
| Missouri | Trump |
| Montana | Trump |

| | |
|---|---|
| Nebraska | Trump |
| Nevada | Trump |
| New Hampshire | Trump |
| New Jersey | Trump |
| New Mexico | Trump |
| New York | Trump |
| North Carolina | Trump |
| North Dakota | Trump |
| Ohio | Trump |
| Oklahoma | Trump |
| Oregon | Trump |
| Pennsylvania | Trump |
| Rhode Island | Trump |
| South Carolina | Trump |
| South Dakota | Trump |
| Tennessee | Trump |
| Texas | Trump |
| Utah | Trump |
| Vermont | Trump |
| Virginia | Trump |
| Washington | Trump |
| West Virginia | Clinton |
| Wisconsin | Clinton |
| Wyoming | Trump |

## 4. Data Visualization

The positive and negative scores can be represented using a bar plot for both Clinton & Trump. The barplot was generated using the "barplot()" function of the "ggplot2" package of R.

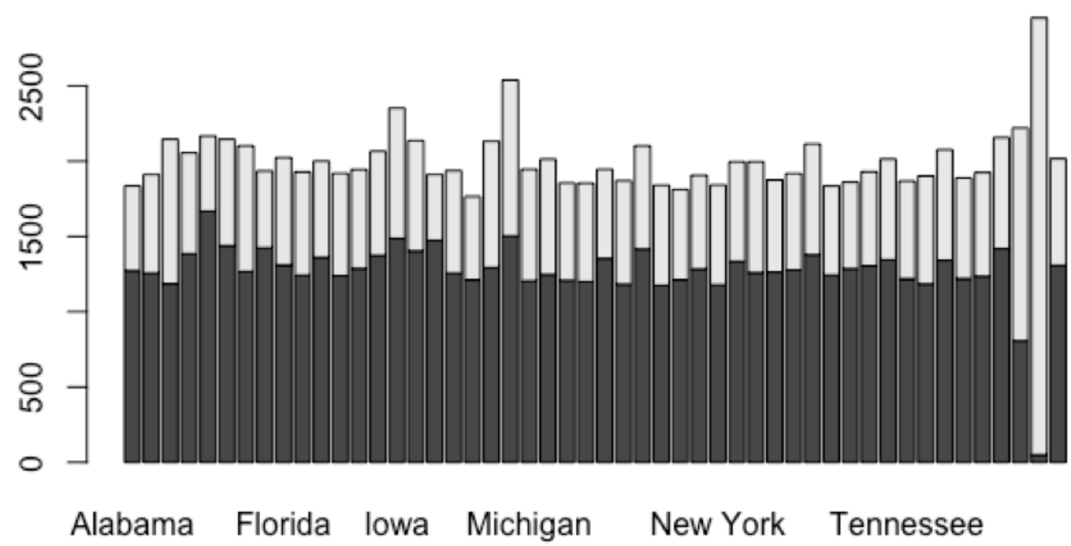


Figure 1: Trump Barplot

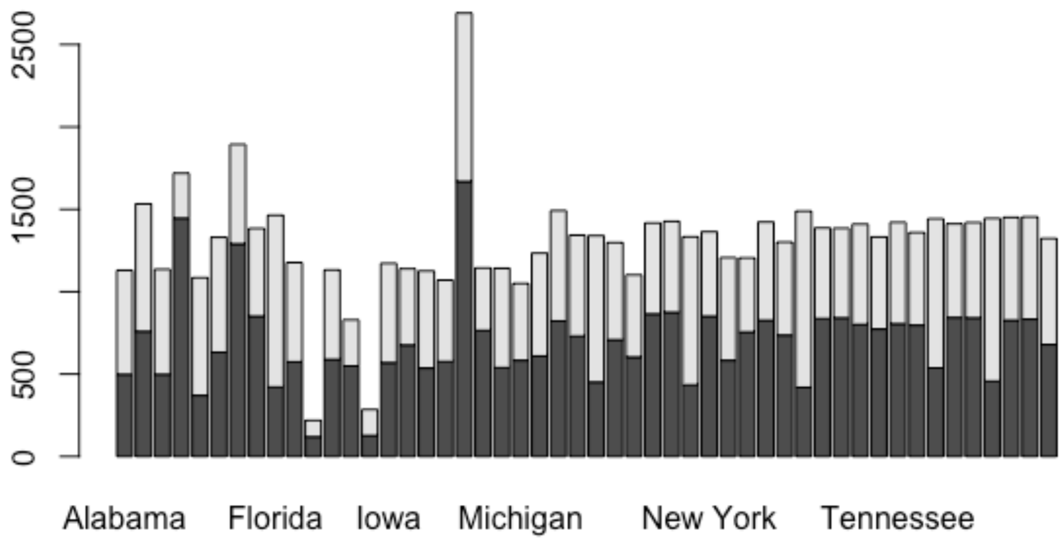



Figure 2: Clinton Barplot

**Negative Score** **Positive Score**

The overall scores of Trump & Clinton with respect to the different states can also be visualized using the “barplot()” function of the “ggplot2” library of R.

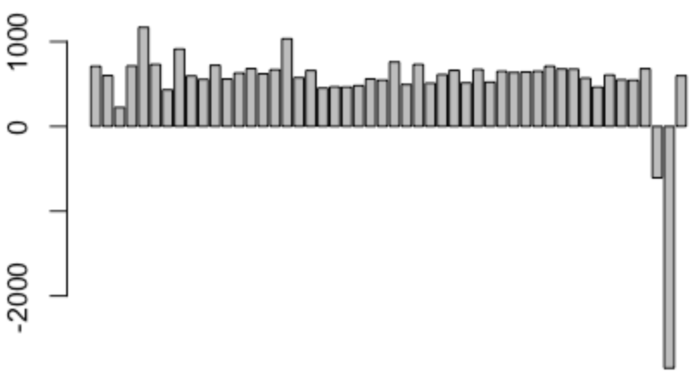



Figure 3: Trump Overall Score Barplot

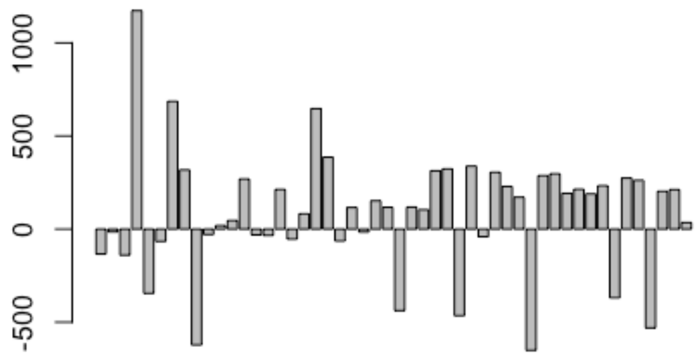
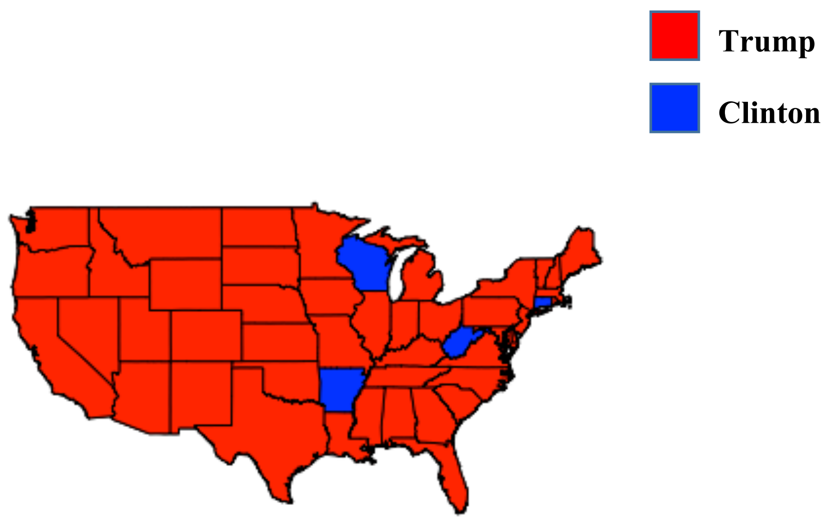


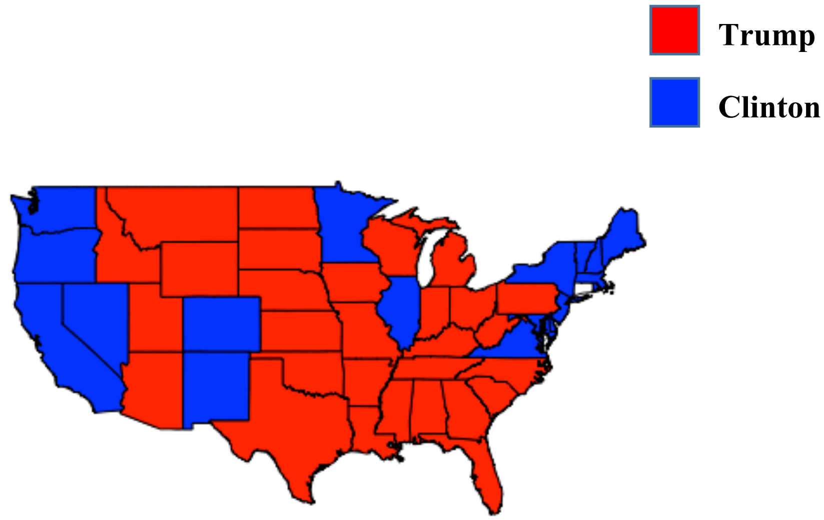
Figure 4: Clinton Overall Score Barplot

We then proceeded further to represent our findings on the map using the “maps” package of R.



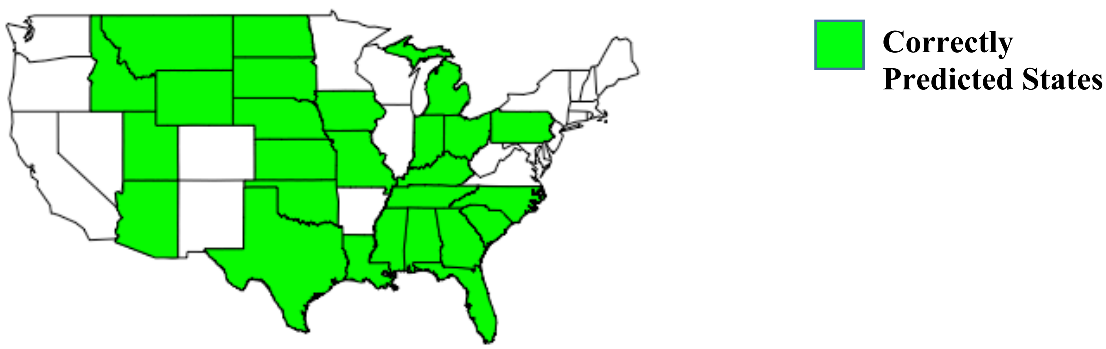

Map 1: Predicted Wins

We then compared it to the actual results map of the United States Presidential Election 2016.




Map 2: Actual Wins

We have then compared the above two maps (Map 1 & Map 2) and displayed the correctly predicted states in Map 3 in green.




Map 3: Correct Predictions (in green)

**Results**

We have performed data retrieval using the Twitter API and imported them into RStudio. We then cleaned the tweets of emojis, links, punctuations, control words, decimal words and new lines to prepare the tweet texts for further processing.

Upon performing lexical analysis with respect to the pre-defined positive and negative word databases, we then calculate positive scores and negative scores of every tweet. Scores are then calculated by subtracting the negative score from the positive scores.

On the basis of scores, states are classified, we have thus, been able to classify 31 of the 50 states correctly, and have achieved 62% accuracy.

As can be seen from Figure 1 & Figure 2, the number of positive scores (darker region) is much greater than the number of negative scores (lighter region) in the case of Trump, than with

Clinton. The only anomaly in the case of Trump would be Wisconcin, which presented a stark contrast to what the other states, by giving a major negative score to Trump.
Similarly in Figure 3 & Figure 4, while Trump managed an overall positive score across almost all the states, Clinton had a more varied score across the country. Trump had an overall negative score only in West Virginia and Wisconsin.
Map 1, 2 and 3, are efficient representations of our findings on the United States map state-by-state. While 1 represents our predictions, 2 represents the actual results and 3 compares map 1 and map 2 and colors the correctly predicted states in green.
Twitter sentiment analysis has many wide-ranging applications and elections are just one of them. There exist many ways to perform sentiment analysis and the one we have performed is called "lexical analysis" and has been followed up with classification. There exists tremendous scope for us to expand these methods so as to allow them to factor-in greater number of factors while classifying tweets as 'positive' or 'negative'.

## References


1. Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior. In *ICWSM* (Vol. 30, pp. 5-314).
2. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*(1), 178-185.
3. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social science computer review*, *29*(4), 402-418.
4. Makice, K. (2009). *Twitter API: Up and running: Learn how to build applications with the Twitter API*. " O'Reilly Media, Inc.".
5. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media* (pp. 30-38). Association for Computational Linguistics.
6. Chin, D., Zappone, A., & Zhao, J. (2016). Analyzing Twitter Sentiment of the 2016 Presidential Candidates.
7. Wang, Y., Li, Y., & Luo, J. (2016). Deciphering the 2016 US Presidential campaign in the Twitter sphere: A comparison of the Trumpists and Clintonists. *arXiv preprint arXiv:1603.03097*.
8. Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, *9*(3), 372-378.
9. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*(1), 178-185.
10. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.