
The Wharton School, University of Pennsylvania

A project report on

# Sentiment Analysis after US elections

*Author :*
Rahul Dhakecha

*Professor :*
Dr. Linda Zhao

Course: STAT 571, Modern Data Mining
December 23, 2016

# Summary

"Statistics and data science gets more credit than it deserves when it's correct—and more blame than it deserves when it's incorrect." - Anthony Goldbloom, CEO of Kaggle.

This statement was more than validated on November 8th 2016, witnessing mixed emotions spread throughout the country, with few people being extremely happy while others deeply disheartened. But the most serious blow was to the statisticians and data science community. We know Data Science went wrong in the 2016 US elections. Well, it was a prediction and predictions might go wrong. Should we cease to rely on such techniques? Should we stop relying on weather forecasts and other such crucial applications of learning and data science? Ofcourse, we know the answer. Data science banks on tons of assumptions, few being extremely naive. These assumptions may sometimes come out completely wrong, which eventually leads to the failure of data science. Nevertheless, data science is competent enough to give the gist of the happenings around the world.

In this report, we analyze the sentiments of people for our president elect Donald Trump. This report is mainly divided into two parts. First part analyzes the variation in sentiments of people over a period of past one month, from November 8th to December 8th 2016. Second part focuses on more recent sentiments of people, from December 10th to December 17th 2016. Second part is further broken down to analyze common sentiments over different states of the US.

# Contents

# List of Figures

# Chapter 1

# Part I

As mentioned earlier, first part of analysis focuses on the variation of sentiments across one month. For the purpose of sentiment analysis, data is downloaded from Twitter, one of the largest microblogging site in terms of users.

## 1.1 Data Acquisition

Twitter API provides an easy access for tweets to be downloaded on R platform. But this access is limited to the tweets from past one week only. It becomes a non trivial task to access a month old tweets. Packages like rvest in R are very handy to scrap through online content but owing to the flexibility and availability of Python repositories, Twitter data in this part of the report is accessed with the help of Python.

Github repository[link here] was modified to access the specific contents for this report. Tweets were filtered based on the keyword and time frame. Tweeets were encoded using utf-8 encoding which is compatible with R. Further these tweets were saved as new line separated content in a single text file. Following snapshot provides the commands used for the above mentioned steps.

```
tweetCriteria = got.manager.TweetCriteria().setQuerySearch('kudlow').setSince("2016-11-08").setUntil("2016-12-08").setMaxTweets(1000)
 tweet = got.manager.TweetManager.getTweets(tweetCriteria)[0]
tweet = got.manager.TweetManager.getTweets(tweetCriteria)
 printTweet("Hello",got.manager.TweetManager.getTweets(tweetCriteria))
for i in xrange(1000):
     tweets.append((tweet[i].text).encode("utf-8"))
f=open("/Users/rahuldhakecha/coursesfall2016/moderndatamining/finalproject/tweets_kudlow.txt","w+")
for i in tweets:
    f.write(i)
    f.write("\n")
```

Figure 1.1: Python code

In the above code snippet, 1000 tweets from November 8th 2016 to December 8th 2016 which contains the keyword "*kudlow*" is accessed and stored in a text file named

"*text_kudlow*", with new line character separating each tweet. This text file is then directly accessed through R code.

In the similar fashion, we develop five different files summarized in following table.

| File name | Description | Keyword | Since | Until | number |
|---|---|---|---|---|---|
| data_tweets_08_Nov | for sentiment analysis immediately after election | trump | 11/08/2016 | 11/16/2016 | 5000 |
| data_tweets_01_Dec | for sentiment analysis after one month of election | trump | 12/01/2016 | 12/08/2016 | 5000 |
| data_tweets_obama_speech | for sentiment analysis after Obama's speech | obama speech | 11/08/2016 | 12/08/2016 | 1000 |
| data_tweets_bannon | after nomination of Bannon | bannon | 11/08/2016 | 12/08/2016 | 1000 |
| data_tweets_kudlow | after nomination of Kudlow | kudlow | 11/08/2016 | 12/08/2016 | 1000 |

Since this sample, in any significant way, does not represent the true population of the US, limitations of this analysis is addressed in the concluding remarks. Files obatained above are raw tweets obtained from Twitter and they need to be cleaned before using it for sentiment analysis.

## 1.2 Data Cleaning

Tweets obtained in raw format has variety of contents such as punctuations, url, etc which does not contribute to any significant sentiment and thus it needs to be removed. Also, we

need to remove few essential elements like emojis, numbers, etc; which may contribute to the sentiment analysis but for simplicity we drop its usage.

Following steps are followed to clean the data and bring it into analyzable form.

- Data obtained from the text file is accessed in the form of table content and is needed to be converted into character strings.

- This list is then converted into corpus, which contains two parts- character string and metadata.

- Majority of tweets have URL attached to them, which are removed as they do not contribute any significant information in sentiment analysis.

- Many tweets are addressed to someone and this information is redundant in analysis; therefore all target elements are removed.

- For simplicity, effect of emoticons(emojis) in this analysis is not considered. All emojis are converted into hex codes, which are removed. This assumption is one of the naive assumption as emojis convey a lot of information about the sentiment of a tweet.

- Entire text is converted into lower case, eliminated english stopwords, punctuations and number. Removal of english stop words like ”not”,”nor” and ”no” is again a naive step as these words convey critical information about sentiment of a particular treat. For example, a sentence like ”People are not happy” is a negative statement but it will be classified as a positive sentiment after removal of english stopwords.

- When a tweet text file is created using a particular keyword, the appearance of that keyword is gauranteed in all the tweets. Such words are removed which inspite of occuring frequently does not carry any significant amount of information.

- Finally all the words are stemmed to retain their significance, and at the same time dropped of the extra content of the word.

After following the above mentioned steps we store the cleaned version of tweets as a plain text document in *corpus_elect_result*. This variable temporarily holds data for all the different text files mentioned in the section above.

## 1.3 Analysis

### 1.3.1 Wordcloud

Data rendered after data cleaning contains all significant words. To get the gist of frequently occuirng words in the tweets, a word cloud is plotted. This word cloud is obtained from *data_tweets_*08*_Nov*.txt file.

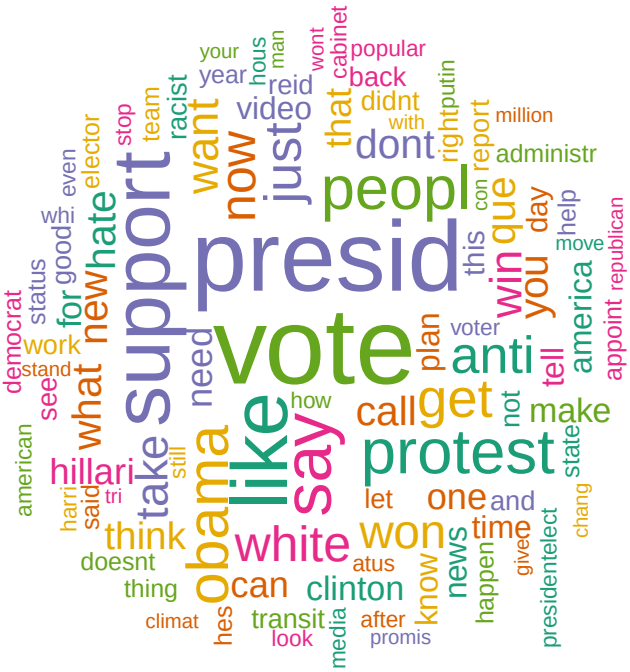



Figure 1.2: Wordcloud for

This word cloud pretty much reflects the sentiments of people during the week following the results of election. While few states were fully supporting the Republican elect Donald Trump, states like California and New York had protests for number of days. Presence of words like ”support”,”like”, ”for win”, etc clearly displays strong favor for the president elect while words like ”protest”, ”hate”, ”racist”, etc clearly shows strong disagreement of people for the results of US elections.

### 1.3.2 Sentiments immediately after US elections

To further dive into the text corpus, we try to segregate our analysis based on various sentiments like trust, anger, joy, etc. This analysis is carried out with the help of "syuzhet" package developed by Stanford NLP team. It implements Saif Mohammad's NRC Emotion lexicon. According to Mohammad, "the NRC emotion lexicon is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

For the sake of visual clearity, we represent negative sentiment by red color and positive sentiment by green color. We divide all the words in the corpus into different sentiment categories and the resultant proportion is plotted on bar graph. The plot reveals that immediately after the results, majority of population was gripped in the wave of fear. Owing to the president elect's bold statements against minorities, Muslims and women, this fear was more than obvious. Large chunks of population came out on streets to express their griefs and anger. But at the same time, there was major proportion of population which expressed their trust on the president elect. This trust owed to the promises made by Republican candidate on creating new jobs and his ability to drive profitable businesses. Overall it can be seen that negative sentiments surpassed positive sentiments during the period following immediately after election results.

### 1.3.3 Sentiments one month past US elections

It can be seen that over the period of one month, "trust" has replaced "fear" as the maximum proportionate sentiment among people. Nevertheless, this change is not significant by a huge margin. But it definitely shows that people are coming in support for new government irrespective of the differences in their thoughts and philosophy.

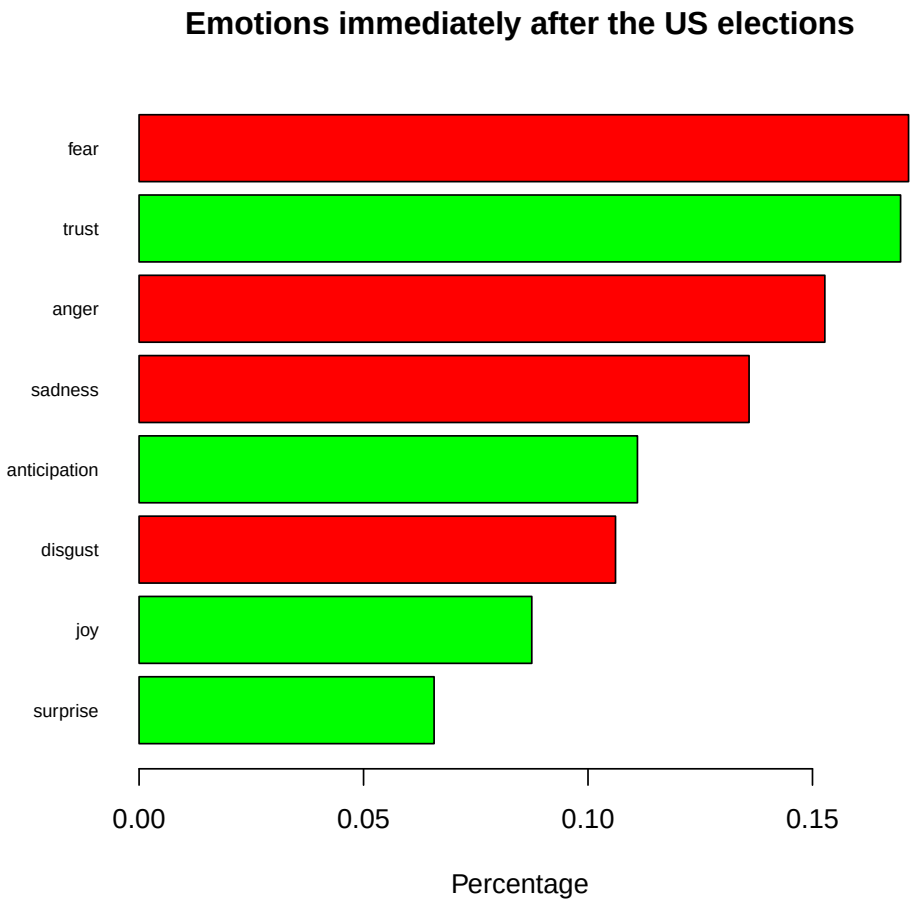
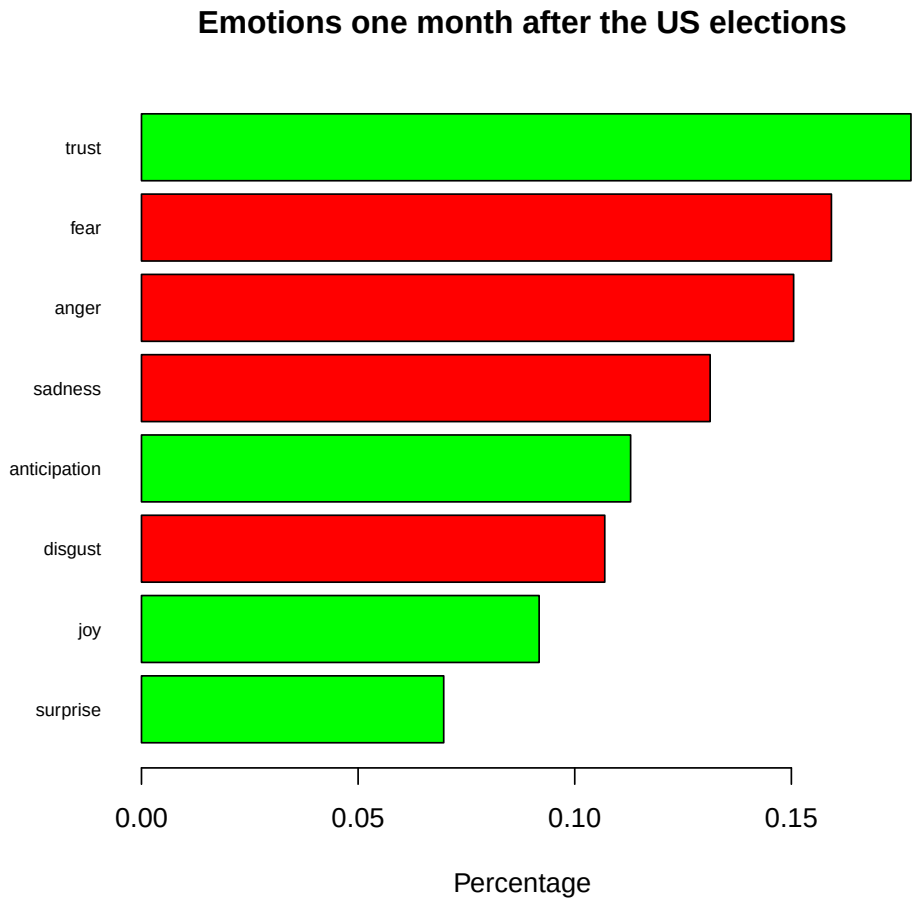



Figure 1.3: Loss functions

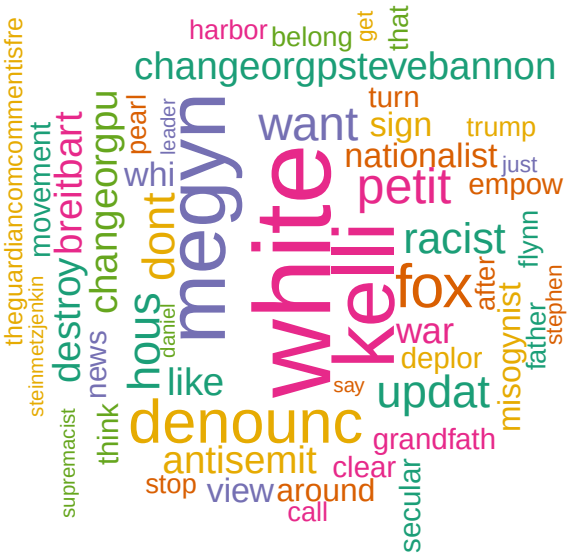



Also, it can be seen that there is still around 30 percent of people expressing sadness and anger in some or the other way. One month after the results has witnessed several major and minor events which has swayed people's emotions. We take a look at few of the prominent ones and try to reflect the sentiments based on the word cloud.

### 1.3.4 Sentiments after Obama's speech over President elect


Despite strong differences in opinions on matter of sheer importance, current president Barack Obama gave congratulatory speech on Novemeber 9th at White House. This speech addressed the positivies of the new government and was in the direction of soothing the angry mob. He assured smooth transition of government and assured befitting atmosphere for the coming generations. Enhancing the spirit of patriotism, this speech definitely wooed the mass.

### 1.3.5 Sentiments after appointment of Steve Bannon as chief White House strategist




Steve Bannon was CEO of Donald Trump's presidential campaign and was later appointed as Chief Strategist at White House on November 13th 2016. This appointment saw severe opposition from many groups especially from Anti Defamation League and Council on American Islamic relations. Bannon is often considered racist, white nationalist and anti Semitic. Words like "white", "antisemit", "racist", etc portray this polarity in Bannon's view. Words like "stop","don't", etc displays negative public opinion towards this appointment.

### 1.3.6 Sentiments after appointment of Larry Kudlow as Chairman of Council of Economic Advisor


Another major decision which raised eyes was president elect's nomination of Larry Kudlow as Chairman of Council of Economic Advisor. While Kudlow does not have controversies following him as Steve Bannon, but his academic inexperience leaves many loop holes to be filled. Position which was once chaired by great economists such as Janet Yellen, Ben Bernanke and Alan Greenspan demands a lot of intellect and decisive power. Words like "idiot", "danger", etc reflects a sense of doubt in public.

# Chapter 2

# Part II

In this part we directly access the data from Twitter API without involving Python. We set authorization with twitter and save the accessed data.

## 2.1 Data Acquisition

We download the twitter data for 10 prominent states of the US in terms of population and economy. We also pay special attention to the swing states Michigan, Pennsylvania, Florida, Ohio, Iowa, North Carolina and Wisconsin which were the deciding states in 2016 US elections. For each state, we take its capital, feed its geographical information in twitter and get 1000 tweets within a radius of 100 miles.

```
[1] "Using direct authentication"
```

## 2.2 Analysis

### 2.2.1 New York sentiment

In this part, we list all positive words along with their corresponding counts. For this analysis we use "tidytext" package which gives us further flexibility to handle text data. Inherent structure of this package allows us to create different groups of sentiments. This package handles sentiments with the help of three lexicons:

- AFFIN - This lexicon rates different words on the scale of -5 to +5, where more negative score representing negative sentiments and vice versa.
- Bing - This lexicon divides all the words into either positive or negative.

- nrc - This package divides all the words into 8 different sentiments, similar to that of ”syuzhet” package.

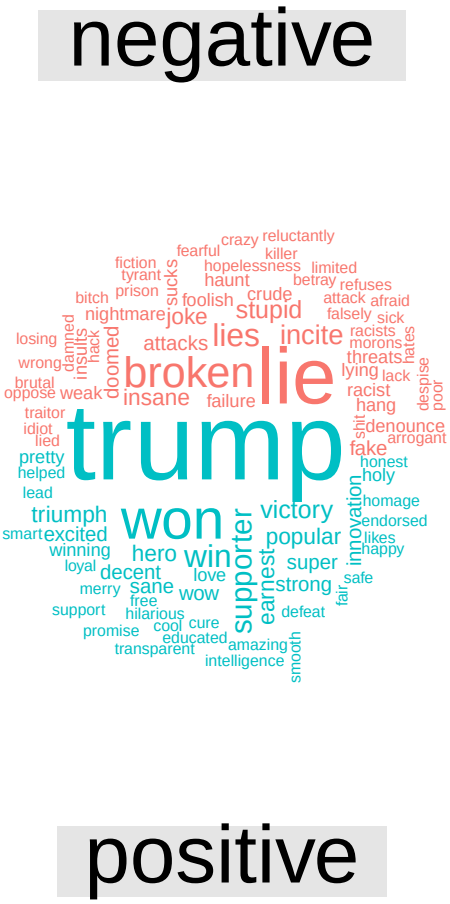
Using bing lexicon, we create bar graph representing different positive and negative sentiments.

### 2.2.2 Positive Negative Words

This section segregates positive and negative words from the tweet data set.

```
# A tibble: 2,024 × 2
             word     n
            <chr> <int>
1          #trump   925
2    @dumptrump22   217
3             lie   137
4           #maga   131
5           trump   125
6           https   122
7            hope    83
8            life    78
9     #truepundit    77
10 @trumpsuperpac    74
# ... with 2,014 more rows
```

### 2.2.3 Wordcloud positive negative words




### 2.2.4 Positive Negative Words

Finally we count total number of negative and positive words in New York state. We see that bing lexicon has total of 4782 negative words and 2006 positive words. Therefore, number of negative count will always be greater than positive count. To mitigate this effect, we will multiply the negative count by a factor of 2006/4782. This multiplication will scale down negative counts to the level of positive counts.

```
[1] 505
```

```
[1] 338
```

In similar fashion, we analyse positive and negative word counts for different states and we find overall shift towards positive sentiments, which validates our conclusion of part I.

```
[1] 503

[1] 287

[1] 161

[1] 170

[1] 164

[1] 98

[1] 496

[1] 336

[1] 290

[1] 236

[1] 304

[1] 245

[1] 1005

[1] 134

[1] 369

[1] 255

[1] 102

[1] 59

[1] 47

[1] 20
```

# References

1. https://github.com/mayank93/Twitter-Sentiment-Analysis/tree/master/docs
2. https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf
3. https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html