# Social Media Analytics: Twitter Users Responses to Financial Events

Saeed Karimabadi
Carleton University
1125 Colonel By Dr., Ottawa, Ontario K1S 5B6, Canada
Saeed.Karimabadi@Carleton.ca

Shaowei Pu
Carleton University
1125 Colonel By Dr., Ottawa, Ontario K1S 5B6, Canada
Shaowei.Pu@Carleton.ca

## ABSTRACT

Twitter data can be used to study social response to political and economical events and major political or economical leader's decisions. we have analyzed tweets and their relationship to relevant news events, press releases, and important political or banking individuals. We have shown how machine learning methods can be used to extract social emotional responses. using the natural language processing algorithms, Naive Bayes Classifier and opinion lexicon we have studied tweets structures and by Using visualization and sentiment analysis, we've found how people has responded towards the heads of the central banks.

## GENERAL TERMS

Social Media, Sentiment Analysis, Central Banking, text analysis

**ACM Reference format:**
Saeed Karimabadi and Shaowei Pu. 2017. Social Media Analytics: Twitter Users Responses to Financial Events. In *Proceedings of ACM Conference, Paderborn, Germany, September, 2017 (3rd International Workshop on Software Analytics)*, 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Social media have gradually become integral part of the daily life for many people. Out of many social media platforms, Twitter, due to its limited length, anonymity and accessibility, has risen to its prominent popularity. Indeed, as of the end 2016, Twitter claims that it averaged 319 million monthly active users.[7] With this multitude of the high usage, many have tried to utilize it to extract information and spread influence.

Efforts to collect information and use it to influence Twitter users' behaviour can be seen in various ways. There are personalities who use this platform to promote personal image. There are also corporations that take advantage of the quick response time and the openness of Twitter to interact with the public including providing customer services and making corporate announcements. Moreover, Twitter's anonymity nature offer users a unique platform to engage in often controversial and political debates. Politicians often now take advantage of this fact to promote their campaigns. Donald Trump's surprising victory in the 2016 U.S. presidential election was often attributed to his effective use of Twitter as a communication tool.

Although there are many success cases of effectively use of Twitter information, the study on information extraction of Twitter is still in its exploratory stage. This is partly due to that fact that Twitter, founded only merely a decade ago, is still a fairly new product. Just like everything new in the Internet age, its growth has been exponential. Currently, over 100 TB of data are processed daily[7], a speed of information accumulation never been witnessed by mankind before.

There certainly has been many attempts to extract information and analyze the tweets. On the industry level, Mosley [5] analyzes insurance related Twitter posts by identifying keywords and applying clustering and association of the keywords. He is able to identify the trend and relationships between industry concepts. One of many methods is sentiment analysis. Bollen et al[2] conduct a sentiment analysis of tweets by using a psychometric instrument to extract six mood states and correlate aggregated mood of each day to popular events, therefore they conclude that Twitter posts can offer a good prediction on collective mood trends. As an international social media platform, Twitter posts in foreign language are also studied. For instance, Aldayel and Azmi [1] combine semantic orientation and machine learning techniques to capture public sentiment, especially during tumultuous times in the Middel-East.

In this project, we attempt to analyze a dataset of Tweets collected by the Bank of Canada with a filter of twenty-nine central banking related keywords. This is done by first performing data cleaning with various tools and techniques then conducting sentiment analysis including polarity scoring. We also investigate the most popular Twitter accounts in terms of retweets. Finally, we attempt to visualize financial events with histograms of some keywords.

The rest of the project report will be organized as follow. In the next section, we will discuss the research questions. It will be followed by methodology which includes data collection, data cleaning and data processing. We then move on to data analysis. Then finally, we conclude and offer some suggestions for future improvement.

## 2 RESEARCH QUESTIONS

Based on the dataset we received and previous research on similar topic, we would like to investigate if there is any useful information we can extract from the collected tweets. If so, how we can do that and what kind of information is attainable with our limited resources.


*3rd International Workshop on Software Analytics, Paderborn, Germany*

DOI: 10.1145/nnnnnnn.nnnnnnn

# 3 DATA

## 3.1 Brief Description of Twitter and tweets

Created in 2006, Twitter is a social media platform that enables users to post and browse short texts, called "tweets". Each tweets has a 140 character limit on the length of the text. Reading tweets is open to anyone with access to the Internet. However, posting tweets required an account which is free to sign up.

When you create an account, you will choose a unique username. When it is proceeded by an "@" symbol, it is called a Twitter "handle" which provides a unique identifier that can identify a user on Twitter. When a user posts a tweet with another user's handle in it, the other user is notified that they are "mentioned". A user's tweets can be subscribed by another user by "following" and becoming a "follower". The tweets posted by followed users will be displayed on the follower's Twitter front page. This series of tweets is known as the "Twitter feed".

When a user find another user's tweet worth sharing, they can "retweet" it and "RT" is added to the tweet in a process similar to forwarding an email. Another feature is the use of "hashtags (#)" which can be used to signify a topic or a group.

There are also other information in each tweets, including time and date of the tweet, geographic location that is collected if the tweet is sent from a mobile device and the feature is enabled, what types of device from which the tweet is sent and many others.

## 3.2 Data Collection

The dataset that we use was provided by the Bank of Canada. It was collected by using the Twitter Application Program Interface (API). With Developer's Access, this API gives users the ability to collect tweets based on certain keywords in a specified period of time. The dataset is collected using 29 central banking related keywords in three three categories:

- Central Banks:
  Federal Reserve, Bank of Japan, Bank of Canada, Reserve Bank, ECB, european central bank, bank of England, FOMC, fed
- Head of Central Banks:
  Yellen, Draghi, Poloz, glenn stevens, graeme wheeler, phillip lowe, mark carney, kuroda
- Policy terms:
  boc rate, boc inflation, boc monetary, boc financial, boe rate, boe inflation, boe monetary, boe financial, boj rate, boj inflation, boj monetary, boj financial

The final dataset contains 4.7 million records collected from January 1st, 2016 to August 1st, 2016. In each record, there is a unique tweet ID number, user name who make the tweet, content of the tweet, time and date, user name to which are tweeted (if available), geographic location in latitude and longitude (if available), and information on the device from which the tweet is sent.

## 3.3 Data Pre-Processing

A tweet has no restrictions except the 140 character limit. It can be about anything. According to Mosley [5], more than half of all content on Twitter is non-information fillers. In order to analyze the content of the tweets, we need to perform pre-processing so

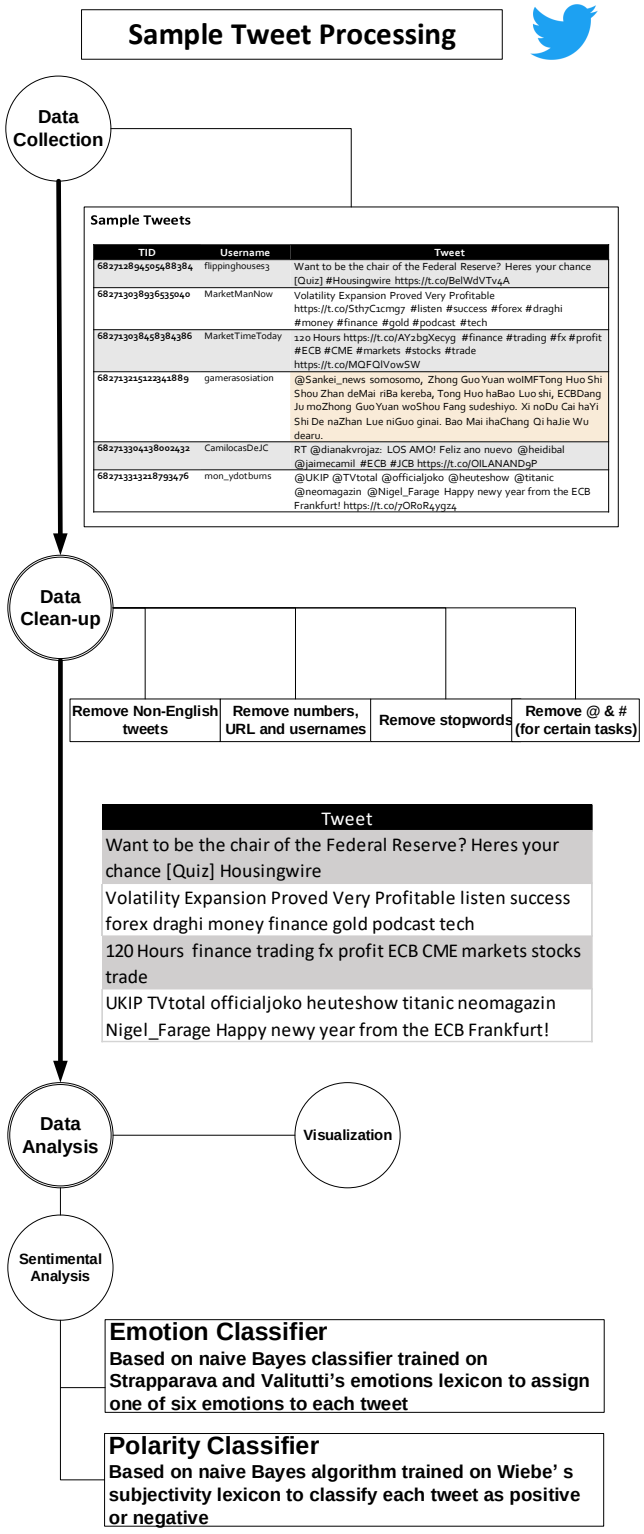


**Figure 1: Procedures to process sample tweets**

the tweets are clean enough so it can be analyzed. Our approach has the following steps, which is also illustrated in Figure 1:

*3.3.1 Removing non-English tweets.* In our analysis, we would like to focus on English language only since it is the language known to the authors and the language that some association with the keywords can be made. To remove non-English tweets, first we need to detect the language of the tweets. We do this by using the "textcat" package in R. It is developed by Hornik et al [4] based on the N-Gram based text categorization. More specifically, it is based on Cavnar and Trenkle's 1994 approach [3]. In their approach, an algorithm is performed on each text, producing an n-gram profile using the options in the category profile database. Then, the text is categorized into the category of the closest profile based on the distance between its computed profile and the category profiles (NA is assigned if no unique profile is found).

*3.3.2 Removing Stop Words.* In many languages, there are some words that do not convey much meaning or information but are a natural part of daily use of the language. By eliminating them, we could focus on finding the most important part of the information. These stop words in English include: and, is, are, we, should, via, will, per, get, says, just, dont, one, ...

*3.3.3 Miscellaneous Data Cleaning.* For most of our analysis below, we deleted numbers and URL's in the tweets since they do not provide information relevant to our project.

*3.3.4 Data Cleaning for Specific Analysis.* There are some other cleaning process that needs to be done in order to help some of the analysis produce better results. For instance, "@" symbol, hashtags # and usernames are integral elements of Twitter communication. The "@" symbol is used to "mention" others while hashtags # is used to refer to a topic, but we only keep some of them when a specific analysis requires them. This is done to have a more responsive analysis and reduce unnecessary noise since we have a dataset of 4.7 million records.

Figure 2 shows the result of language detection of some sample tweets after data cleaning is performed. Cavnar and Trenkle [3] state their algorithm can achieve 80% to 99.8% correct classification rate depending on the tested newsgroup articles. While it's an impressive accuracy rate, it was done on news items which for the most part, uses proper use of the specific language. However,on Twitter, users are not bounded by editors and proofreaders therefore more prone to grammatic and spelling errors. Combined this with the fact that users often create new phrases and abbreviations, similar to other online platforms, this makes Cavnar and Trenkle's language categorization technique less reliable. We can see the evidence of this from Figure 2. All of the sample English tweets are correctly identified except the No.4 and No.8 where they are identified as French and Frisian. French and English share similar terms specially in Finance and Economics, thus it is understandable such error can occur. English at its foundation is a Germanic language. It can be explained that its similarity with Frisian, a western German language, could cause the the inaccuracy in language detection. In order to alleviate such inaccuracy and at the same time, avoid deleting useful information, we run the same language detection tool again on a data set of 100, 200, 500 tweets to detect if there is any patterns with which certain language is consistently categorized as English. We then keep the consistently mis-identified tweets and delete all other non-English tweets. We understand this is not an ideal solution but with limited resource, we believe this is the best solution we can achieve.

| No | Detected Language | Tweet |
|---|---|---|
| 1 | english | Want the chair of the Federal Reserve? Heres your chance [Quiz] Housingwire |
| 2 | english | Federal Reserve Bank:Creature From Jekyll Island Revisited FED BANKSTER |
| 3 | english | Volatility Expansion Proved Very Profitable listen success forex draghi money finance gold podcast tech |
| 4 | french | Hours finance trading fx profit ECB CME markets stocks trade |
| 5 | polish | Zhong Guo Yuan woIMFTong Huo Shi Shou Zhan deMai riBa kereba |
| 6 | middle_frisian | somosomo |
| 7 | slovenian-iso8859_ | RT LOS AMO! Feliz ano nuevo ECB JCB |
| 8 | frisian | Happy newy year from the ECB Frankfurt! |
| 9 | italian | RT *nomi di band tradotte in italiano*\n\n"immaginare draghi" |
| 10 | italian | RT *nomi di band tradotte in italiano*\n\n"immaginare draghi" |
| 11 | english | Bank of Canada Newfoundland c Bank Note Coin Half Penny Cent c c Lot |
| 12 | english | RT The Reserve Bank has no clue...can only count to . like its BOSS. |

**Figure 2: Language detection using "textcat" after data cleaning**

## 4 ANALYSIS

### 4.1 Most Frequent Terms

Our dataset is collected with a filter of 29 keywords related to central banking. We would like to see which word is most frequently used. It's worth noting here that we are investigating all processed text including those are not among the 29 keywords. In Figure 3, using randomly selected 10,000 tweets, we produce a bar chart that shows the 22 most frequent terms.

For most of the terms, they are expected. There are keywords we used, such as "yellen", "reserve", "ecb", "graghi", "interest" and others. Besides these keywords, the data collection also captured some other finance related terms, for instance; "stock", "markets", "economy" and "brexit". Given that "stock" and "markets" have very similar frequency, we suspect that they are used together as a phrase "stock markets". It is not surprising this phrase appears in central banking related tweets due to the volatile and reactionary nature of the stock markets to outside factors, This is especially for news on central banks as it is often regarded as an indicator of future trajectory of the economy. And this also explains the reason the word "economy" is a frequently tweeted word. Our data collection dates coincides the campaign and referendum about "brexit" or the U.K.'s exit from the European Union. "Brexit" being among the most frequently tweeted terms shows that the economic ramification of the "brexit" is hotly discussed on Twitter. The most frequent term is the term "fed". This result shouldn't come as a surprise since "fed" is often colloquially referred to the numerous U.S. Federal Government departments and agencies, including the Federal Bureau of Investigation (FBI), the Internal Revenue Service (IRS), the Federal Reserve and many others. When the tweets are collected using the keywords "fed", it is possible to incidentally sweep up tweets unrelated to central banking, our study topic. However, we do not have a solution to this problem at this stage without risking losing information. Another questionable result is the fact that "https" and "amp" appears on the list. "Https" stands for Hypertext Transfer Protocol with Security; it is the standard prefix

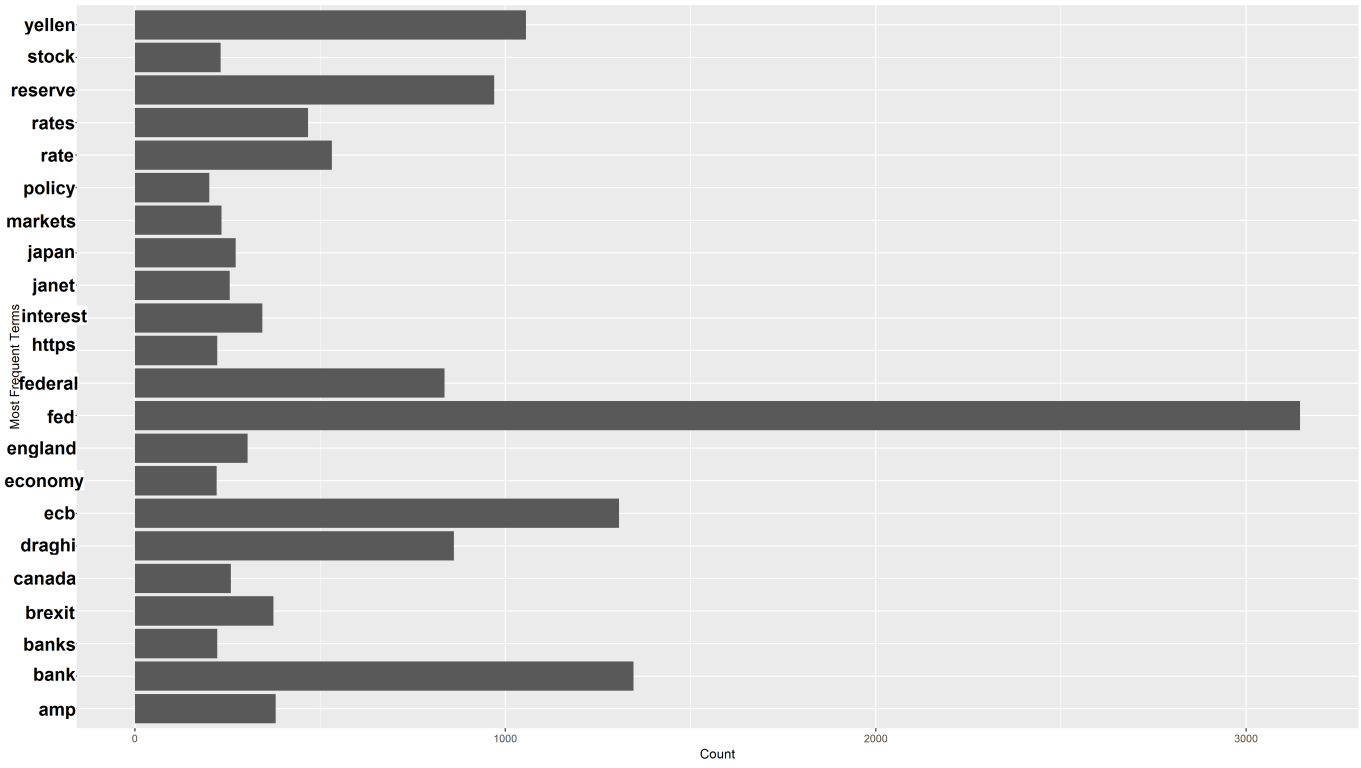


**Figure 3: Most Frequent Keywords**

proceeding a URL while "amp" most likely stands for ampersand "&" and "&amp" is often the code for "&". The appearance of these two terms certainly signify the failure in our data cleaning process in which URL's are supposed to have been removed. One possible reason for this problem is that some URL's are broken either when they are posted or when they are collected, therefore our cleaning tool failed to detect some of the URL's. The solution to this problem is similar to the potentially problematic term "fed" in that we do not yet have an effective tool to exclude the noise without risking remove relevant information.

## 4.2 Most Retweeted Accounts

On Twitter, there are many indicators of popularity of a user. The number of followers of a user shows how many other users are interested in what this user has to say while the number of retweets in each tweet demonstrates how much other users are willing to interact with that tweet and sharing with other users. In our context of central banking, we believe the number of retweets a user get shows how useful their information is and how interactive other users are towards the original tweet. In Figure 4, it shows, in all 4.7 million tweets, the top 10 most retweeted users.

Among the most retweeted users, most of them are expectedly financial and economic news related, although it represents a great deal of diversity of information sources. Regular news sources include "business" (Bloomberg), "wsj" (Wall Street Journal),"reutersbiz" (Reuters Business) as well reporters like "Schuldensuehner" (real name: Holger Zschaepitz, a financial editor for German newspaper "Die Welt") and "anamariecox" (Ana Marie Cox, a reporter for the New York Times). There are also unconventional news/opinion sources such as "zerohedge" (Zero Hedge, a financial opinion blog) and "philstockworld" (a stock analysis opinion blog). "ecb" (European Central Bank) is the only financial institution on the list. We believe that this is due to the fact that people are

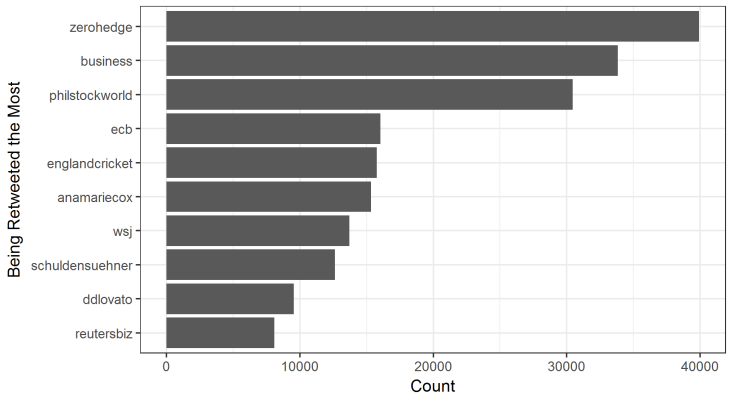


**Figure 4: Users that are Retweeted the Most**

more interested in other's opinions on certain subject, than reading official statements. The noisy nature of Twitter give us another surprise; we noticed that the username "ddlovato" which belongs to singer Demi Lovato is among the list of most re tweeted twitter accounts. In Mid June, Demi Lovato asked her Fans to Know More About Wars Than 'Celebrity Feuds'. so she tweeted: "Has society fed into it or is this just easy business with everyone from blogs to magazines trying to cash in?" where she tweeted about tabloid culture. this tweet made lots of buzz and has re tweeted by her fan approximately 10,000 times. as she is using word "fed" in her tweets, our result has categorized her tweets among the "fed" related tweets. from another point of view this reflects the power of celebrities to raise people awareness about social events.

We believe that despite of the flaws, the number to retweets is a adequate way to measure the popularity and influence, at least on the Twitter.

## 4.3 Histograms

In this section, we produce histograms of all 4.7 million tweets and for each keywords to see patterns. Figure 5 is the histogram of all tweets collected.

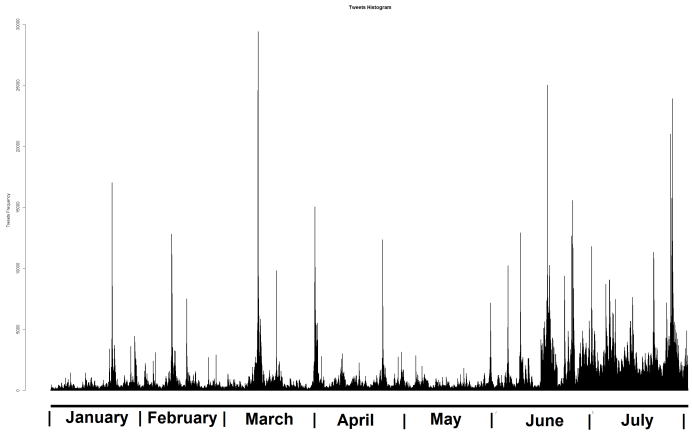


**Figure 5: Histogram of all tweets**

In Figure 5, each bar is the the number of tweets in one day. Overall, we can notice that after around mid-June, compared to before this day, the intensity of the frequent tweets increases significantly. That is, Twitter users collectively tweeted more about central banking after mid-June. Based on the results from the most frequently tweeted terms and this particular day, we can surmise that "Brexit" or the U.K. referendum on its exit from the European Union which happened on June 23rd, 2016 is the cause of the heightened twitter use. One of factors in the "brexit" debate is the role of the Bank of England and European Central Banks in complicated relationship between the U.K and the EU. The possible loss of single market in Europe leads to many speculations on the impact of "brexit" on U.K.'s economy especially the financial sector. This is reflected in the significant spike around mid- and late-June. Following that, the aftershock from the "yes" outcome of the referendum cause the discussion on the central banking to be persistent. This suspicion can be somewhat verified by the histogram of the term "european central bank" in Figure 6.

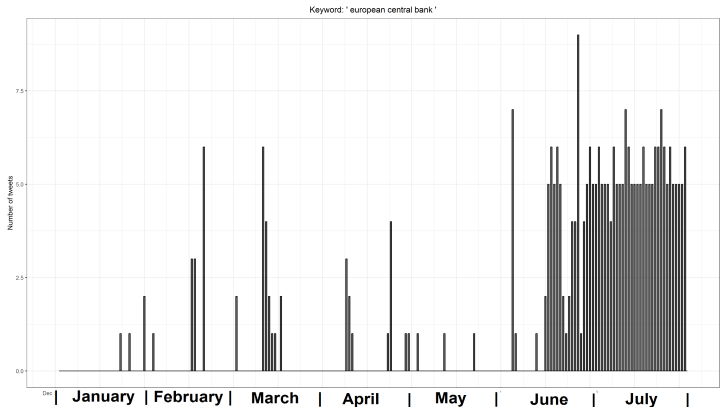


**Figure 6: Histogram of the term "european central bank"**

In Figure 6, we can observe similar intensification of discussions related to the European Central Bank. This term only has sporadic usage before this.

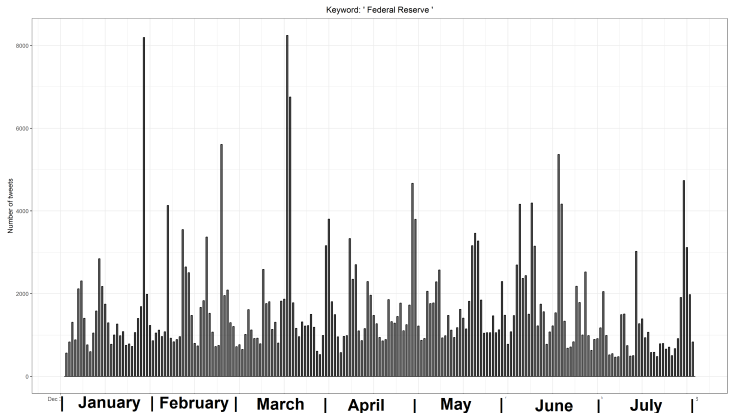


**Figure 7: Histogram of the term "federal reserve"**

Here we also include another histogram to demonstrate the increase of the tweets is due to "brexit" and the discussion it brings. Figure 7 is the histogram of the term "federal reserve". We can see that outside the U.K. and Europe, the discussion on central banking is unaffected and stay consistent, at least regarding the U.S. Federal reserve.

From the example above and histograms of all other keywords we have included in the appendices, we can conclude that although Twitter is international, the discussions regarding central banking are mostly concentrated to the their own local governments. Therefore in order to effectively analyze the Twitter users' reactions to financial events, we need to focus on specific keywords and their affected regions instead of the overall picture.

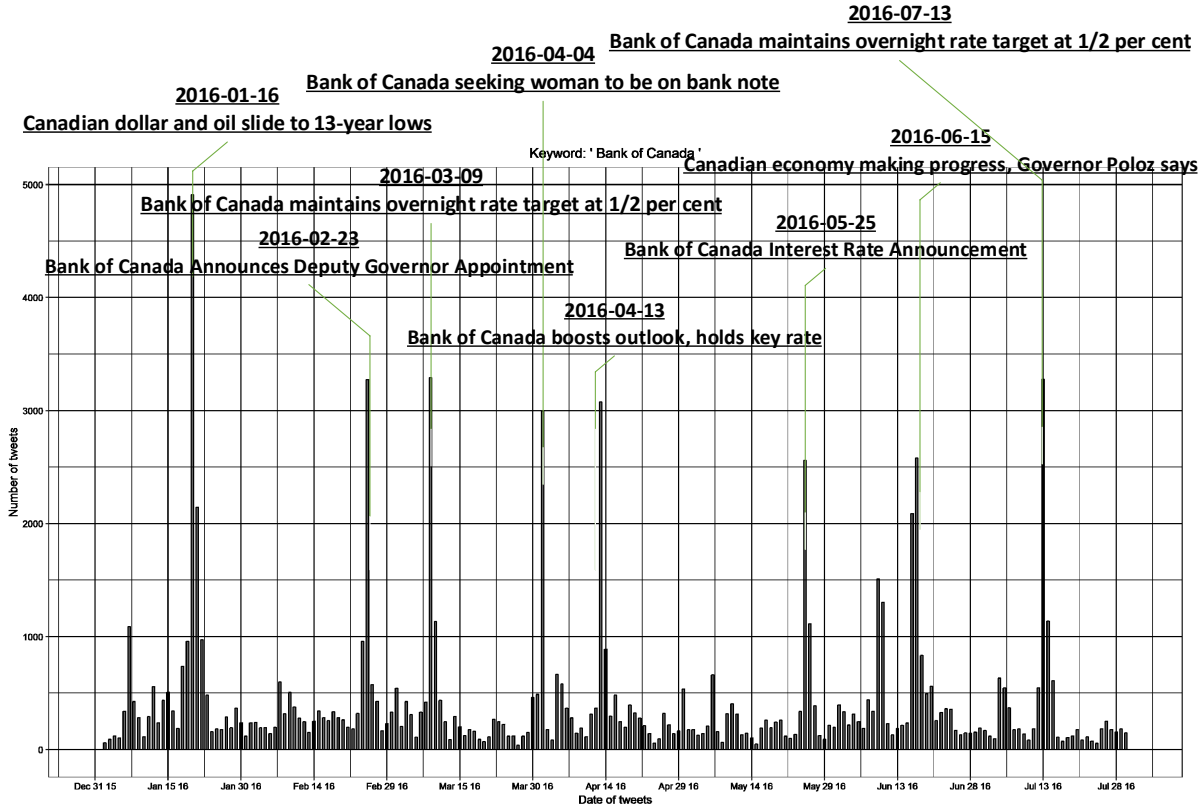


**Figure 8: Event Mapping for "Bank of Canada"**

## 4.4 Event Mapping

The previous section attempts to find patterns in the overall frequencies of the tweets. In this section, we would attempt to coordinate the spikes of the keywords with financial events, including central bank announcements and news reports.In order to study the correlation of the news and people's reaction, we need to pick the peak frequency of the tweets and find the corresponding news item. With numerous news taking place everyday and our huge dataset of 4.7 million tweets. We believed that we have to find a way to automate this process.

After some study on related studies, we realize that this task is much more difficult than it seems. There are several hurdles to overcome to complete a convincing result. First, Twitter is full of noises. According to Mosley [5], 44% of all Tweets all just pointless babble.This situation is even worse for keywords that is prone to be the topic of a discussion, such as "federal reserve" where there could be more than 20 spikes. Combining this with the fact there is always a delay from the time news breaks to the time people react to it, it makes detecting legitimate spikes in histogram very difficult. We can simply pick tweet spikes above certain frequency, the concern is we are arbitrarily cut off information.

Another issue is that how the machine pick the news item. We can find the news item of the day by the tweet frequency using some simple tool, but the results from this are often not ideal and sometimes very wrong. This is due to the fact that we lack the tool to train our machine to distinguish different news items that are about the same event but are reported in different ways. This would lead to the situation that the machine failed to pick up the real news items due to low frequency of the many variations of it.

To avoid the first problem and to demonstrate the second problem, we manually produced Figure 8. This event mapping graph can be done because we first pick the histogram for the term"Bank of Canada" due to its small number but easily-identifiable frequency spike, then we manually research central banking related news items around the date of the spike and plot them. Figure 8 is also a great example of how the second problem can be difficult to solve. Through our research, we learn that every month, the Bank of Canada releases a routine statement to announce that whether they change the over night rate. Although majority of the times, they are just exactly the same statement of keeping the rate the same, those statements do cause Twitter user to react to them thus the spikes. These spikes can be found on March 9th, April 13th, May 25th and July 13th. However, we can see that among the four news items, there are three different titles. If we let the machine pick the news, they might not be able to tell that they are the same news. This is especially problematic for a headline mixed with other information like "Bank of Canada boosts outlook, holds key rate". So

each uniquely worded news would be counted as different, thus indistinguishable with other low frequency news items.

Unfortunately, we have not found a solution to this problem within our resource limitation. If we need to do another event mapping, human involvement is inevitable.

Despite the potential costly human involvement, when an event mapping is done, it does offer some interesting insight on the relationship between central banking related news and Twitter users' reaction. Besides the over night rate announcement, we can also see that Twitter users have a strong reaction to the news of "Bank of Canada seeking woman to be on bank note".

## 4.5 Sentiment Analysis

Finally, we conduct a sentiment analysis to understand what Twitter users' collective sentiments are towards certain central banking concepts including personnels and institutions.

We use the "sentiment" R package to create a bag of words for each keyword and also create a word cloud for each set of tweets related to different keywords.

There are different methods that can be used to classify tweets based on their emotions and their polarity. We use a naive Bayes classier trained on the emotional lexicon of Strapparava et al[6] as well as a simple voter procedure to perform the sentiment analysis task. For the sentiment analysis, we use an emotion dictionary of 1500 words where all words has been labeled with their corresponding emotion category. We process each words the "bag of words" and will find its corresponding emotion from our lexicon dictionary and assign a score based on number of occurrences of that word in the bag of words. As a result of this step we have each group of tweets labeled with six different scores for six different category of Joy, Disgust, Anger, Fear, Sadness and Surprise. Then we let the naive base classifier categorize each corpus and choose the fit. To demonstrate this approach, Figure 9 is produced.

In Figure 9, we can see that for every tweet there is a score for each emotion. When certain word or a combination of words increase one or more scores to become the dominant emotion, this tweet is assigned that emotion. (NA is assigned if no dominant emotion is detected.)

For the polarity we use the same method. we use an opinion lexicon dictionary which categorizes approximately 6,800 words as positive or negative. then a naive bays classifier , categorizes each tweets bag of word as positive, negative or neutral.Figure 10 shows a sample output of this algorithm.

In Figure 10, there are only two normalized scores for positive and negative. When the ratio of positive/negative is less than one, negative is assigned, while positive is assigned if the ratio is larger than two. If it is in between, it is neutral.

We choose this approach due to its ease to implement and the fact that it provides two different approaches to the same problem. We believe that, with this approach, we can compare and contrast two outcomes so we can arrive a more comprehensive conclusion.

With our cleaned dataset, we are able to obtain two outcomes for each keywords. Some of the keywords however, has only very few results therefore the power of the sentiment analysis would not be robust. As a result, we focus on head of two central banks that are reasonably popular and of our interest. To start, we get the

| | ANGER | DISGUST | FEAR | JOY | SADNESS | SURPRISE | BEST_FIT |
|---|---|---|---|---|---|---|---|
| 1 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | NA |
| 2 | 1.46871776464786 | 7.34083555412328 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | disgust |
| 3 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | NA |
| 4 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | NA |
| 5 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 7.34083555412328 | 1.7277074477352 | 2.78695866252273 | joy |
| 6 | 1.46871776464786 | 7.34083555412328 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | disgust |
| 7 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | NA |
| 8 | 1.46871776464786 | 3.09234031207392 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | NA |
| 9 | 1.46871776464786 | 7.34083555412328 | 2.06783599555953 | 1.02547755260094 | 1.7277074477352 | 2.78695866252273 | disgust |

**Figure 9: Emotion classification for sample tweets**

| | POS | NEG | POS/NEG | BEST_FIT |
|---|---|---|---|---|
| 1 | 1.03127774142571 | 8.78232285939751 | 0.117426534862834 | negative |
| 2 | 1.03127774142571 | 8.78232285939751 | 0.117426534862834 | negative |
| 3 | 1.03127774142571 | 8.78232285939751 | 0.117426534862834 | negative |
| 4 | 1.03127774142571 | 17.1191924966825 | 0.0602410272345239 | negative |
| 5 | 17.2265151579293 | 8.78232285939751 | 1.96149873259283 | neutral |
| 6 | 1.03127774142571 | 8.78232285939751 | 0.117426534862834 | negative |
| 7 | 9.47547003995745 | 17.8123396772424 | 0.531961000724885 | negative |
| 8 | 9.47547003995745 | 17.1191924966825 | 0.553499824351745 | negative |
| 9 | 24.9775602759011 | 26.1492093145274 | 0.955193710657424 | negative |
| 10 | 1.03127774142571 | 17.1191924966825 | 0.0602410272345239 | negative |
| 11 | 8.78232285939751 | 8.78232285939751 | 1 | neutral |

**Figure 10: Polarity classification for sample tweets**

results for two heads of the central banks: Janet Yellen, the Chair of the Federal Reserve and Stephen Poloz, the Governor of the Bank of Canada. In each graphs below, there is a bar chart of frequency and a word cloud for both emotion the polarity classifications.


**Figure 11: Sentiment Analysis of the term "yellen"**

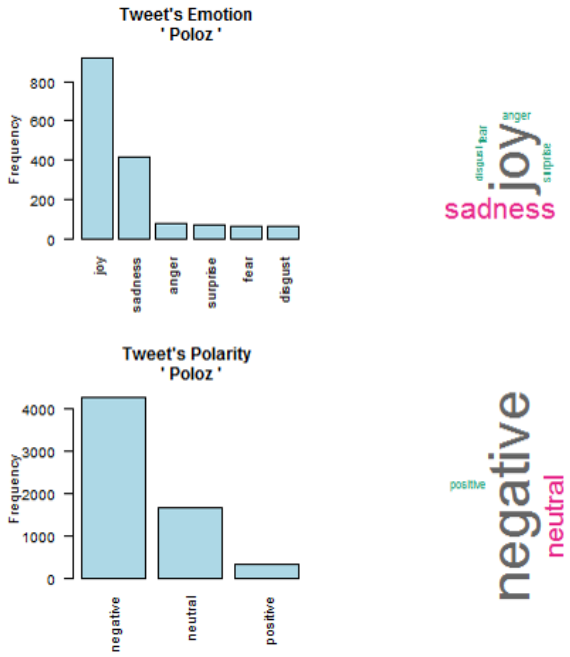


**Figure 12: Sentiment Analysis of the term "poloz"**

From the Figure 11, we see that more Twitter users express "disgust" and negative attitude towards Ms. Yellen. This is not in any way surprising, since many people express more honest and more extreme opinion due to Twitter's anonymity nature. Noticeably, "joy" is the second most frequently expressed emotion. "Joy" is the most expressed emotion towards her counterpart in Canada . The general attitude towards Mr. Poloz is still largely negative though. This seemingly contradictory result in fact demonstrates the benefit of using two different metrics to measure sentiment. This allows us to have understanding the nuance behind these classification and take them with caution. To investigate this issue in depth, we can find some sample tweets such as the following:

| | |
|---|---|
| 1 | : Yelen/Draghi/Lagarde: when zero interest leveraged debt has wiped out all resources we will request the #vatican to #pray for #miracles |
| 2 | Stephen Poloz says "We're going to need a bigger C$ devaluation": Canadian Manufacturing Hits Record Low |

**Figure 13: Sample tweets related to the term "yellen" and "poloz"**

we believe that the potential cause the conflicting results is how each lexicon categorize certain words. In the some tweets, there are words that might be categorized as negative but other words may give a "joy" result. For instance, in the first sample tweet, "zero" might be categorized as negative, words like "pray" and "miracle" out of context would lead to a "joy" emotion. Or even some exactly same words could be categorized as both negative and joy or other positive emotions. An example is the in the second sample tweet, the word "low" might be consider both negative and joy if that's the lexicon on which the naive Bayes classifier is trained.

## 5 CONCLUSIONS

For the last few years, the use of social media including Twitter, has grown significantly. The information generated through this growth is unprecedented. Analytics of the information could help researchers, corporations an society as a whole understand human behaviour. In this project, we attempted to understand the patterns in Twitter user's reaction to central banking related keywords. We also analyze their sentiment and find their mostly negative sentiment towards heads of central banks.

## 6 SUGGESTIONS FOR FUTURE STUDIES

In the process of completing our project, we encounter several difficulties and we would like to give some suggestions for future studies.

**Choice of Software** We used R during our project. While this is a competent statistical software, it does not handle big dataset well due to its unique way of loading the entire dataset to the memory. We would suggest other software that stores data on the local disk.

**Detection and Treatment of Bots** Like any other software programs, the process of tweeting can also be automated, creating "bots". Bots are usually very new accounts that automatically tweet certain content and retweet certain other users. The existence of bots introduce bias and distort analytical results. Future studies should attempt to remove tweets by bots.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Olga Baysal for her continuous support of our project and her encouragement throughout the course and Data Day 4.0. We would also like to thank Johan Brannlund from the Bank of Canada for his precious advice and suggestions.

## REFERENCES

[1] Haifa K Aldayel and Aqil M Azmi. 2016. Arabic tweets sentiment analysis – a hybrid scheme. *Journal of Information Science* 42, 6 (2016), 782–797. DOI: http://dx.doi.org/10.1177/0165551515610513

[2] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. (????). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2826/3237/

[3] William B Cavnar and John M Trenkle. 1994. N-Gram-Based Text Categorization. (1994). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.9367&rep=rep1&type=pdf

[4] Kurt Hornik, Patrick Mair, Johannes Rauch, Wilhelm Geiger, Christian Buchta, and Ingo Feinerer. 2013. The textcat Package for n-Gram Based Text Categorization in R. *Journal of Statistical Software* 52, 6 (2013), 1–17. DOI: http://dx.doi.org/10.18637/jss.v052.i06

[5] Roosevelt C Mosley. 2012. Social Media Analytics: Data Mining Applied to Insurance Twitter Posts. *Casualty Actuarial Society E-Forum* 2 (2012). https://www.casact.com/pubs/forum/12wforumpt2/Mosley.pdf

[6] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. The Affective Weight of Lexicon. (????). http://hnk.ffzg.hr/bibl/lrec2006/pdf/186_pdf.pdf

[7] Twitter. 2017. TWITTER USAGE / COMPANY FACTS. (2017). https://about.twitter.com/company