Sentiment Analysis of Twitter Posts

Leverageing Hadoop for Large Scale Sentiment Analysis on a Static Dataset of Social Media Content to Model Public Opinion Over Time

Anuja Rayarikar
MS Information Technology
Purdue University
West Lafayette, Indiana
arayari@purdue.edu

Penghao Wang
MS Information Technology
Purdue University
West Lafayette, Indiana
Wang1128@purdue.edu

Robert Haverkos
MS Information Security
Purdue University
West Lafayette, Indiana
rhaverko@purdue.edu

***Abstract*—Acknowledging that social media is a prolific means of communication and self-expression in today's digital world, this paper presents a case study for carrying out sentiment analysis with the intention of observing changes in public opinion over time. The Syrian refugee crisis was chosen as the subject about which public opinion was to be observed do to its emotional charge and media presence at the time this study was undertaken.**

Keywords—big data, map/reduce, hadoop, sentiment analysis, social media, twitter, public opinion

I. INTRODUCTION

Europe has recently been faced with extreme migrant-crisis. According to statistics from International Organization for Migration (IOM), approximately 886,000 migrants have travelled to Europe this year from January 1 to November 2015 [1][2]. This is a significant increase from 2014 when about 280,000 migrants came to Europe. Amongst them, 80% were from countries like Syria, Afghanistan and Eritrea, countries suffering from civil war, violence or government repression [2]. The main reason for migration was thus raging conflict, terror and civil war.

The goal of this project was to find public opinion regarding Syrian refugee crisis over a specific period of time and determine if and when changes took place. For this purpose, a sentiment analysis of twitter posts was conducted over a period of 16 days.

II. MOTIVATION

Social media like Twitter, Facebook and MySpace has been used extensively for promoting various social and political campaigns. It gained more popularity during Barack Obama's successful US Presidential campaign [3]. Many analysts attributed Obama's victory to his extensive usage of social media and online strategy [3].

Social media provides a platform for people across the globe to share their opinion about issues related to their personal life or issues having global significance. As such, social media provides an interesting dataset for analyzing public opinion about a particular issue.

The motivation for this project was to gain insight about public opinion related to some global issue. The migrant influx in Europe was a highlight during this project timeline and hence was chosen as a topic of study.

Sentiment analysis is also known as opinion mining, which basically builds a system to collect and categorize opinions about a particular topic [5]. It is a very powerful tool for evaluating opinions and making predictions for better decision-making. The significance of sentiment analysis can be found in various fields such as election campaigns, product evaluation, and movie reviews [3][4][5][6].

III. DATA COLLECTION

Data collected for this project was limited to tweets extracted from Twitter API. The library used was a Python library called Tweepy. Tweepy is an open-sourced Python library, which communicates with Twitter platform and uses its API. It is very easy to install Tweepy from GitHub repository. It accesses Twitter through OAuth authentication.

Tweets were extracted in JSON format and save as text files. The search for tweets was based on the following keywords:

- "Syria migrants"
- "Refugees"
- "Refugee crisis in Europe"

There were several limitations of using Twitter API, such as:

- Limitation on date
  - Used stream listener in Tweepy
  - Tweets for an earlier date could not be extracted
- Limitation on volume
  - Varies every day

- Twitter locks user if data collection reaches a specific limit

Figure 1: Data sample

IV. Sentiment Analysis

Sentiment analysis is a method that is commonly used for gauging public opinion. It has been a popular method for finding public opinion prior to election polls [3][4][6]. Companies are also increasingly using it instead of traditional polls to analyze public opinion for product evaluation and marketing [7].

Various algorithms and tools are available in market for conducting sentiment analysis. For the scope of this project, a Python library called TextBlob was used. TextBlob processes textual data by using Natural Language Processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more [8].

The sentiment property returns named-tuple of the form Sentiment (polarity, subjectivity). The polarity score is a float, which provides sentiment score ranging from -1.0 to 1.0. The subjectivity is a float ranging from 0.0 to 1.0, where 0.0 is very objective and 1.0 is very subjective [8]. Only the polarity scores were considered for this project.

V. Map/Reduce Implementation

A. Methods

In the project, our team use Hadoop Streaming. Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer. [9] Because we chose python as our programming language, Hadoop streaming help us to run the mapper and reducer.

B. Mapper

There are two mappers. First one is based on the time. That means, the key is the date and time that the twitter was created. The value is the value of sentiment analysis. The rank is from -1 to 1. -1 means the twitter is negative. 1 means the twitter is positive. 0 means, the twitter is neutral. The second mapper is based on the location. The value is the rank of the sentiment.

1) Explanation of the Logic of the Mapper

To extract the data, we need to clean the data at the first, which means read the data and extract the useful data we want by tags. In the following part, the logic of the mappers is explained:

1. Import different libraries.
2. Then use a for loop to analysis each tweet in the file.
3. Then clean the data by using strip() function and split() function.
4. Extract the target data by using if function. In this part, several parts are selected. The “Create_at” and “text” are selected. Then store cleaned data into a list. Also, the count of tweets of the same date and time are stored into the list.
5. Use a for loop to analysis the clean data.
6. In each step, analysis each tweet and return a value.
7. The output is (date and time, rank)

The two mappers have the same logic. The differences of them are one is based on date and time, another is based on the location.

C. Reducer

There are two reducers. First one is sum all the rank of sentiment analysis based on the key. The second one is calculating the average score of that. The logic of reducers is simple. First one is basically a sum function. In the beginning, the split() functions are used to clean the data. Then the float() function is used to convert the rank from string to float. The average one count the amount of the tweets which have the same key, and divide the sum of the rank by the amount of the tweets.

D. The Problems and Thoughts

The debugging part in Hadoop is painful. In the beginning, the mapper and reducers run well in the Linux. However, something always went wrong about the mapper and reducers in Hadoop. It only showed there is a bug, but never tell you what went wrong. The problem we faced is that the problem about installing python packages. If a programmer want to run the python program and import different libraries, he or she needs to install all of them in both master and slaves. The problem we faced is a little bit complex than that. After the packages are installed in all machines, we also need to guarantee that it works under the “Hadoop” users in slave machines, but not root users.

The mapper and reducers are not complex. The hard part is debugging and solve different problems in Hadoop. This project did give us a lesson of problem solving. I learnt a lot in process of solving the problems.

VI. Observations

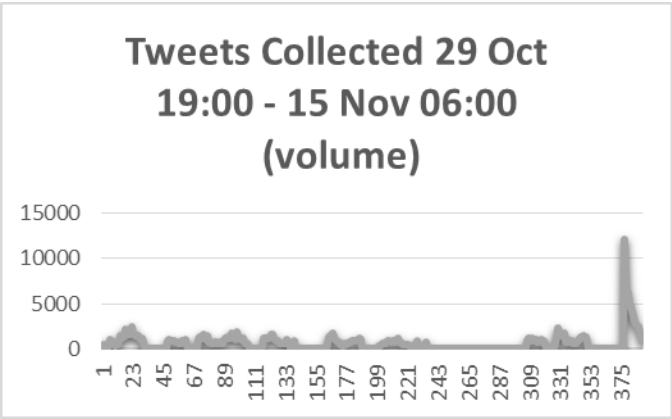
The core of the analysis for this study was conducted on tweets collected over a 13 day period between 29 October

2015 and 11 November 2015. During this time roughly 250,000 tweets were collected taking up about 2gb of storage on the server. During this time however, there are eight distinct intervals where no tweets were collected, this was likely due to limitations in the collection mechanism selected. For this study, three relationships within the data were examined. These relationships were Volume per Hour, Average Sentiment per Hour, and Total Sentiment per Hour. With these observations, the hope was to gain insight to answer two fundamental questions, how much are people talking about the subject, and what are they saying about it. Before moving on, the in-article graphs, while a ready aid, obscure much of the detail, larger versions will be made available in the appendix.

A. Volume Per Hour

The first measurement to be examined is the volume per hour, that is to say the number of tweets collected each hour over the observation period of the study. While a simple observation, this measure has the potential to provide a great deal of insight. In fact it answers one of the two aforementioned questions, specifically, how much are people talking about the subject. While this may seem like vague information, conducting this sort of traffic analysis helps to gauge how much of the public's interest that the subject in question holds. If one were to determine that the mapping between public attention and twitter traffic was 1 to 1, then by analyzing the twitter sphere then one could determine exactly what the general public was interested in at any given time. In addition to being a bit of a stretch however this proposition is well beyond the scope of the current experiment. However, if one accepts that this traffic is a viable approximation as proposed by Bermingham and Smeaton, then we at least can gain some insight. [6] While it would have been wonderful to have omniscient access to the twitter traffic to gauge what percentage was directed at the Syrian refugee crisis, such comparisons were beyond the capabilities of the team. This, however, does not devalue entirely the measurement of traffic volume as we can at least approximate relative interest over time on the subject with regards to itself.

Figure 2



As shown in the figure above, for the majority of the observation period, there was little difference in the traffic volume. If we only account for the times when data was collected then the average number of tweets per hour that met the observation criteria was roughly 1145 with the volume seldom exceeding 2000 tph (tweets per hour). The extended runs of 0 tph are likely a result of deficiencies in our collection capabilities and there is one spice on the last day of observation that netted nearly 12000 tweets.

B. Average Sentiment Per Hour

The second measure to be taken was the average sentiment per hour. This measure is rather self-explanatory and was mapped as the average of all of the sentiment analysis scores for the tweets collected in the given hour. This is important because it answers to a degree the second fundamental question of interest being, what people are saying about the subject. To be fair, this is still a very course measure, being limited not only be the effectiveness of our sentiment analysis methods, but also in the fact that it only determines whether or not the statements appear to be positive and rating them in that regard between -1 and 1. However, it was the hope of this study that this would at least provide some high level, if coarse, insight as to the public opinion of the Syrian refugee crisis.

Figure 3

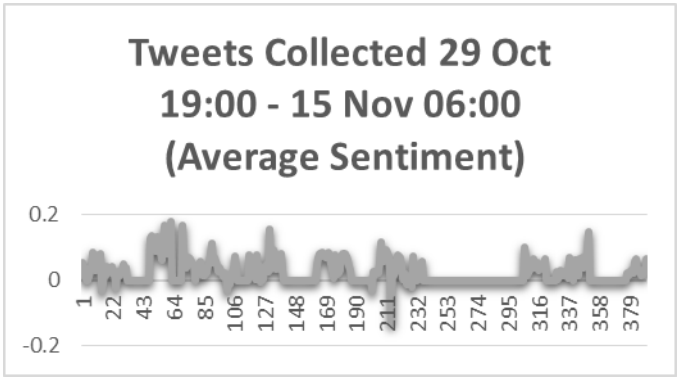
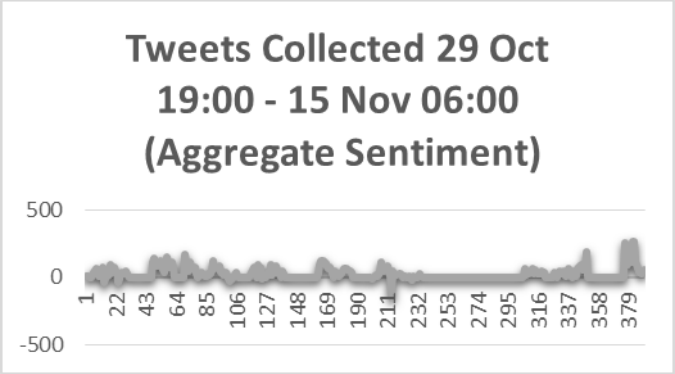


Figure 3 shows the graph of the data collected as the average sentiment per hour over the course of the observation perriod. While it is apharent that the sentiment is generally pisitive, the average sentiment never extends beyond 0.2 or -0.15, in short, only using up about 18% of the available spectrum. It has been suggested by part of the team that this pheonomena may be a result of a large number of the tweets referencing news articles which tend towards neutral sentiment and it is also possible that the sentiment ananlysis suite that was wmployed for this study was to precisely tuned to the current "lingo" of twitter, leading to less words being associated with a given sentiment. However, it is also possible that many of the posts were inheritly neutral or evenly distributed between positive and negative sentiment.

C. Total Sentiment Per Hour

With the two fundamental questions addressed, an attempt was made to consolidate the results down into a single visual for easy digestion by a wider audience and a more efficient way of adding value to the analysis. The result of this endeavor was to aggregate the sentiment scores into a sum, or total score, for each hour of the observation. While no new data was added to this representation, the average sentiment is essentially expanded with the magnitude of the traffic for that hour. Again, this was an attempt to summarize the results of the first two measurements and shoe the general bias and interest in one graphic.

Figure 4



Here we see much the same pattern as with the average sentiment with the addition of some artifacts being made more pronounces such as a valley at around -150 sentiment which was shown as an average sentiment of -0.15 in the previous figure and a peack of around 250 sentiment which occurred during the observed spike in traffic on the last day of the observation period.

VII. Analysis

Doing detailed analysis proved challenging for the observations collected, and this issue was exasperated by limitations of the team in the relevant areas of statistics. Beyond that, however, it was determined that the data observed was likely limited by the collection mechanisms that were able to be employed and that any authority added to the observations be detailed statistical analysis would be a façade for only having been able to collect a small amount of data. While not necessarily impacting the average sentiment, such a flaw has serious implications for our traffic analysis and aggregate scores. However, if one assumes that the volume of data streaming in from the collection api is a sound sample, then meaningful analysis can still be conducted on the average sentiment data.

To that end we can observe that the average sentiment stays fairly consistent throughout the course of the study. There were no apparent major long term shifts in the sentiment of the content which would have indicated a turn in public opinion and allowed the team to investigate further to identify a cause for such a change. There are however two particular elements of interest in the data, the first of which being a period of relatively high sentiment at roughly 45-70 hours in the study. While trending high, the sentiment only averaged about 0.1 higher than the surrounding areas and we were not able to link this change with any specific events. There was also a one hour dip to an average sentiment of -0.15 about halfway through the study which was likewise unlinked to a cause.

VIII. Limitations

There were a number of limitations that this study faced. Perhaps most apparent and mentioned several times throughout the course of the paper were those imposed by the collection mechanism. While tweepy was determined to be the best option available given the time, budget, and resource constraints of the project, the limitations on data volume and lapses in data collection heavily devalue the volume and aggregated sentiment observations. These project constraints were also a limiting factor in the length of time that the study was conducted. Ideally the study would have run longer, but the time constraints of the semester project coupled with long spin up time restricted the bulk of the data collection.

There was also a self-imposed limitation in the regard that English words and phrases were chosen to key the data collection. While this tradeoff seemed reasonable given the scope of the project, it severely limits the demographic coverage of the observations made to being of those people who post on twitter in English, or in reference to articles in English. As this study was focusing on a subject that is largely of European interest, this may have impacted the results more so than would be desired. Additionally, social media and twitter perhaps especially, seems to have its own unique and rather protean dialect. The language involved can be esoteric to a human reader and the difficulty grows exponentially when trying to evaluate the meaning with a computer algorithm. The effectiveness of the sentiment analysis system that we were using may have been degraded by the circumstance.

IX. Future Directions

This study served primarily to act as a foundation in the field for the research team to build experience and explore the possibilities of big data analysis. To that end, a number of directions were proposed as potential avenues for future investigations. Many of these thoughts evolved out of discussions regarding how to overcome some of the challenges and limitations that became apparent throughout the course of the project. The first of these would be to repeat the study on a larger scale, to collect content over a longer period of time and with a more powerful collection mechanism. This would serve to answer our initial question with more accuracy. Additionally, it could be beneficial to conduct a similar study using a different event or theme as a

focus and see if the trends of near-neutral sentiment and high degree of noise within that spectrum are maintained or if they were specific to the topic of this study. It would also be productive to see if there is a sentiment analysis scheme which has been specifically tuned to work with a high level of effectiveness on twitter posts or see if there is any way that an existing methodology could be tuned in such a way to attain this effectiveness. This would allow for a greater ability to tap the twittersphere as a resource for information regarding public opinion. It would also be interesting to weight the results of the sentiment analysis based on the number of retweets that each post received to see if one side or another is "going viral." Additionally it would be interesting to organize the opinions be geographic area if an efficient way to rationalize that data could be established.

Finally, on the more technical end of the spectrum, this study focused on conducting sentiment analysis of data in a static environment, that is to say, the data was all collected in one phase and then loaded into the cluster for analysis. While, simple, this method is resource intensive and scales poorly for use on longer term studies or those with a larger data stream. Perhaps the most interesting improvement would be to use the Hadoop cluster as a buffer. To have an implementation such that content is streamed strait into the cluster from the collection for analysis and obliterated as soon as an analysis segment is completed to make room for more incoming data. This would allow for a more efficient use of resources and the ability to conduct this type of analysis more quickly.

References

[1] IOM: 350,000 Migrants Have Fled to Europe This Year, *News Europe, Voice of America,* 2015. Available at http://www.voanews.com/content/hungary-closes-train-station-stops-stream-of-migrants/2940431.html

[2] Europe's Migrant Crisis. *Human Rights Watch.* 2015. Available at https://www.hrw.org/tag/europes-migration-crisis

[3]] A. Tumasjan, T. O. Sprenger, P. G. Sander, I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Fourth International AAAI Conference on Weblogs and Social Media*. 2010. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852Predicting

[4] B. O'Connor, R. Balasurbramanyan, B. R. Routledge, N. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Fourth International AAAI Conference on Weblogs and Social Media*. 2010. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842/

[5] M. Rouse, I. Barber. Opinion Mining (Sentiment Mining) Definition. *Business Analytics, TechTarget.* 2010. Available at http://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining

[6] A. Bermingham, A. F. Smeaton. On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Dublin City University*. 2011. Available at http://doras.dcu.ie/16670/1/saaip2011.pdf

[7] W. Chamlertwat, P. Bhattarakosol, T. Rungkasiri. Discovering Consumer Insight from Twitter via Sentiment Analysis. *Journal of Universal Computer Science vol 18 no 8*. 2012. Available at http://jucs.org/jucs_18_8/discovering_consumer_insight_from/jucs_18_08_0973_0992_chamlertwat.pdf

[8] Tutorial: Quickstart, *TextBlob* Available at https://textblob.readthedocs.org/en/dev/quickstart.html

[9] Hadoop Streaming. (2013, August 4). Retrieved December 4, 2015, from https://hadoop.apache.org/docs/r1.2.1/streaming.html#Hadoop Streaming

Appendix I: Figures

Figure 1


```
{"created_at":"Thu Oct 29 19:44:32 +0000 2015","id":
659818118697656320,"id_str":"659818118697656320","text":"Check out @HeartOMfilm, the new project
focused on the EU refugee crisis from the makers of \"Not My Life\" - https:\/\/t.co\/
J7Ghwk3pgY","source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client
\u003c\/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_us
er_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":
87754597,"id_str":"87754597","name":"Tedeschi Trucks
Band","screen_name":"DerekAndSusan","location":"Jacksonville, FL","url":"http:\/\/
TedeschiTrucksBand.com","description":null,"protected":false,"verified":true,"followers_count":
66520,"friends_count":8672,"listed_count":1181,"favourites_count":2140,"statuses_count":
2984,"created_at":"Thu Nov 05 18:50:20 +0000 2009","utc_offset":-14400,"time_zone":"Eastern Time (US &
Canada)","geo_enabled":false,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_ba
ckground_color":"ACDED6","profile_background_image_url":"http:\/\/pbs.twimg.com\/
profile_background_images\/123177368\/background.jpg","profile_background_image_url_https":"https:\/\/
pbs.twimg.com\/profile_background_images\/123177368\/
background.jpg","profile_background_tile":false,"profile_link_color":"038543","profile_sidebar_border_c
olor":"FFFFFF","profile_sidebar_fill_color":"F6F6F6","profile_text_color":"333333","profile_use_backgro
und_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/621351457699311616\/
iZAbBXaz_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/
621351457699311616\/iZAbBXaz_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners
\/87754597\/
1427682026","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent
":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_s
tatus":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https:\/\/
t.co\/J7Ghwk3pgY","expanded_url":"http:\/\/igg.me\/at\/hom","display_url":"igg.me\/at\/hom","indices":
[108,131]}],"user_mentions":[{"screen_name":"HeartOMfilm","name":"Heart of the Matter","id":
4003432061,"id_str":"4003432061","indices":[10,22]}],"symbols":
[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"en"},"i
s_quote_status":true,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":
[{"url":"https:\/\/t.co\/giT9CP1UNv","expanded_url":"http:\/\/twitter.com\/musicenergy1\/status\/
659819439865638912","display_url":"twitter.com\/musicenergy1\/s\u2026","indices":
[3,26]}],"user_mentions":[],"symbols":
[]},"favorited":false,"retweeted":false,"possibly_sensitive":false,"filter_level":"low","lang":"und","t
imestamp_ms":"1446148480785"}
```


Figure 2

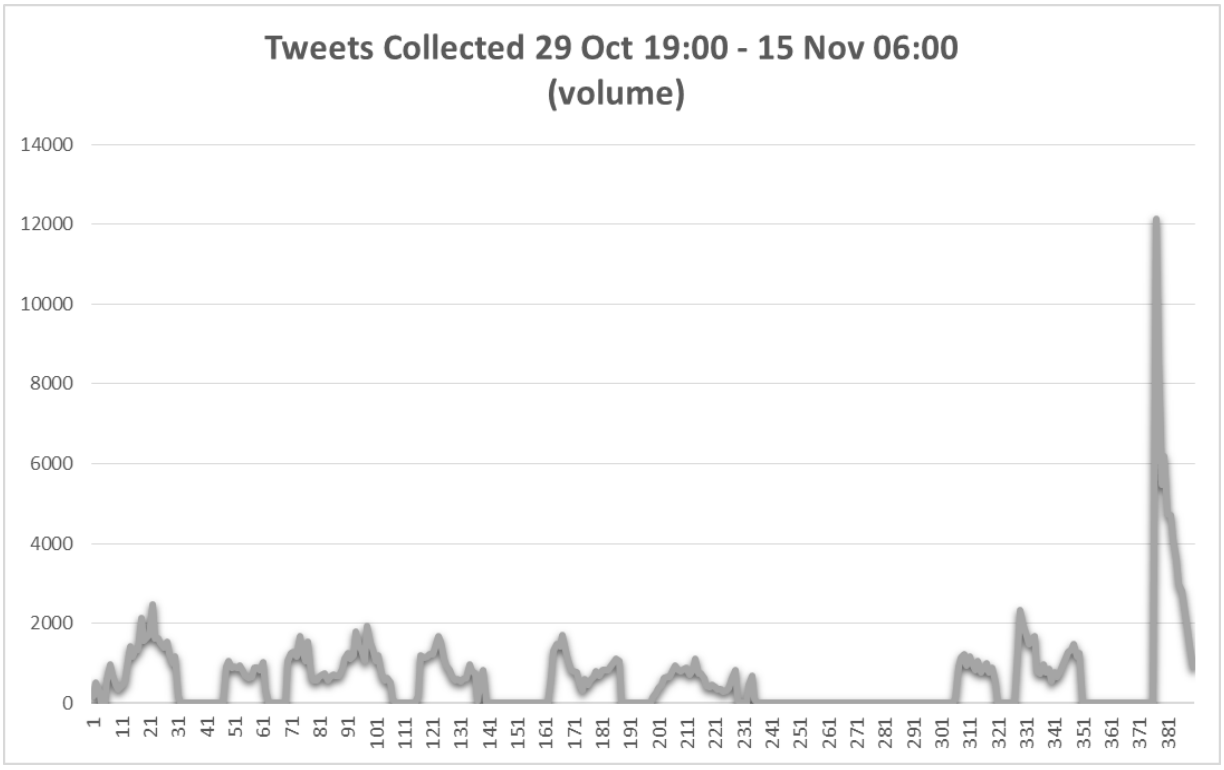


Figure 3

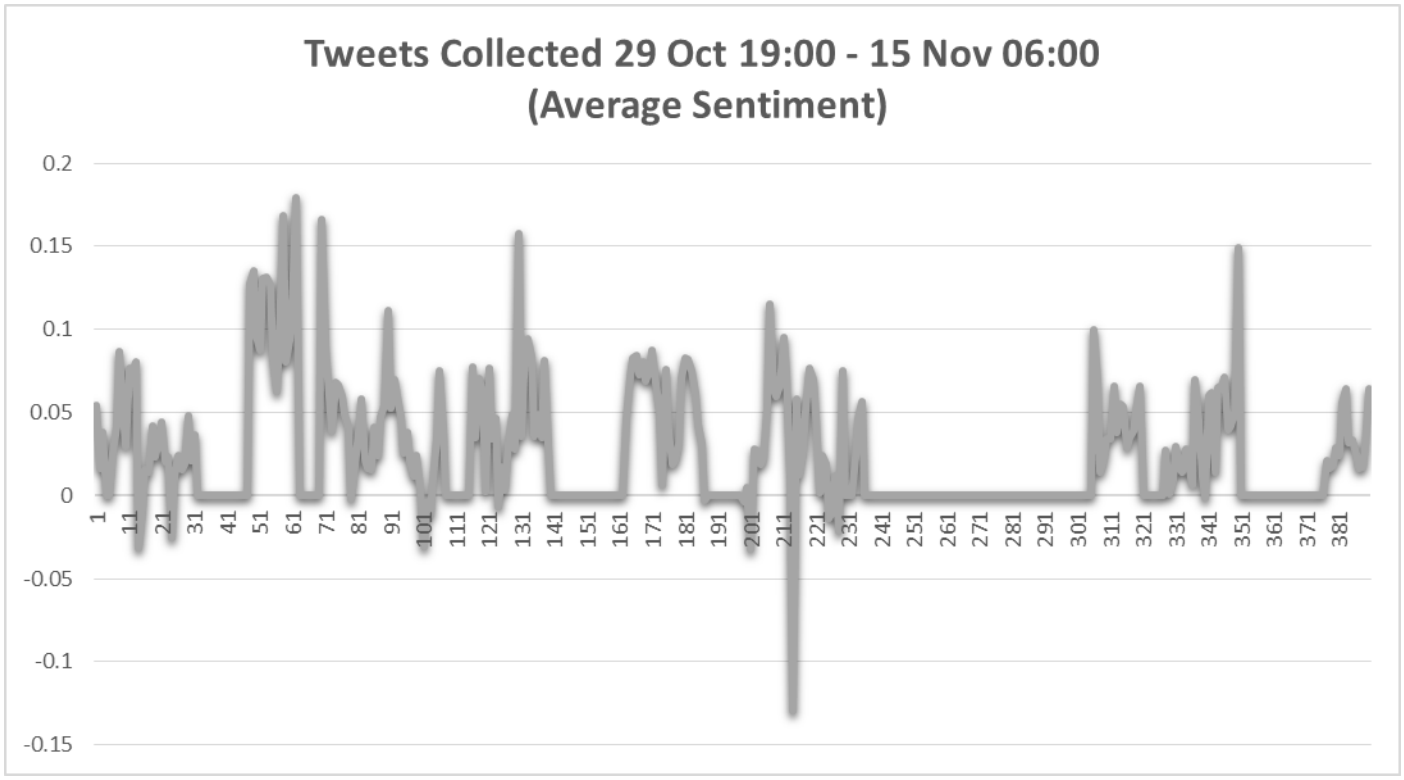


Figure 4