Sentiment Analysis on the Consequences of FB Data Breach, the Cambridge Analytica Scandal

Dinesh Kumar Dhamotharan, Khushboo Gupta and Sakthi Priya Rajendran
Department of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NewYork-13244
{ddhamoth, khgupta, sarajend }@syr.edu

***Abstract*—Since last decade the social media has emerged as a stronghold platform to raise users opinions and inclinations. For any organization, it is like a jackpot which can exploited in any way seem fit, gaining business intelligence or influencing the masses towards or against a particular idea. Hence Twitter with a user base of more than 240 million people provide a great opportunity to mine the opinion of the people to see how the propagation of information, impacts the lives of people. Sentiment analysis is a tool which determines the inclination or feelings of the world population towards a specific event. One such case effecting the lives of billion people around and threatening the privacy of people is the Facebook and Cambridge Analytica data breach scandal. It started in 2014 with the British political firm started collecting personal information on million users from Facebook, for electoral manipulation. When this became known, #DeleteFacebook and similar hashtags started trending followed by banning of Facebook and similar platforms. So, in an ironic way, in this project we are using Twitter to get the emotions, polarity of the public opinion on the same and generating the results as plots, graph and word cloud to assess the aftermath of the scandal.**

***Index Terms*—FB Data Leak, Cambridge Analytica, FB scandal, privacy leak, mark zuckerberg, Facebook, deletefacebook, cambridge scandal**

I. Introduction

Social media websites have emerged as a whole of the platforms to get users' opinions and influence the means any business is commercial. Opinion of individuals matters a great deal to research however the propagation of knowledge impacts the lives in a very large-scale network like Twitter. Sentiment analysis of the tweets confirm the polarity and inclination of large population towards specific topic, item or entity. These days, the applications of such analysis will be simply determined throughout public elections, flick promotions, whole endorsements and lots of alternative fields.

As net is growing larger, its horizons have become wider. Social Media and small blogging platforms like Facebook, Twitter, Tumblr dominate in spreading encapsulated news and trending topics across the world at a fast pace. A subject becomes trending if additional and additional users are contributing their opinion and judgments, thereby creating it a valuable supply of on-line perception. These topics are typically supposed to unfold awareness or to promote public figures, political campaigns throughout elections, product endorsements and diversion like movies, award shows. Giant organizations and corporations take advantage of people's feedback to enhance their product and services that any facilitate in enhancing promoting strategies.

People quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss about the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, rate about the performance of not only any movie but about other products and to know about whether it will be a success or not. This type of vast information on these sites can used for marketing and social studies. Therefore, sentiment analysis has wide applications and include emotion mining, polarity, classification and influence analysis.

The Facebook and Cambridge Analytica data scandal is the breach of data on a bunch of personal information of around 50 million users on Facebook that the firm Cambridge Analytica started collecting in 2014. Cambridge Analytica, a British political firm for consulting has combined data mining, data analysis and data brokerage with the overall aim of communication for the electoral process [1]. For the past few weeks the world of internet has been divided over this issue. Delete Facebook hashtag has been trending on twitter and other social media platforms after the news of the scandal broke. There are also section of people who claim that they knew about this privacy setting from the very beginning. So, our project aims at analyzing the aftermath of this scandal on the Twitter platform which would be a reflection of the public sentiment regarding the same. So for this project, we are using a number of related hashtags as inputs which are used to collect the data. The output is generated based on whatever hashtag we select from the dropdown menu. Also we went one step ahead and were able to generalize to input any hashtag and get the visualization based on the outcome of that particular hashtag. Also once one hashtag has been processed, it gets added to the dropdown menu, for later reference.

However, analysis of the expressed tweets is not a straightforward job. Quite a number of challenges are concerned in terms of key, polarity, lexicon and descriptive linguistics of the tweets. All these aspects of the tweets have an inclination to be extremely unstructured and non-

grammatical. It gets tough to interpret their actual intention like in the case with sarcasm and amusement. Moreover, in depth usage of slang words, acronyms and out of vocabulary words are quite common while tweeting on the micro-blogging website. The categorization of such words per polarity gets very complicated as per the natural processors are concerned.

II. Motivation - Why Twitter?

Twitter was introduced as a website for micro-blogging in the year 2006. This social media platform allows the people to put a piece of writing as status updates, which can be of 140 characters. These are known as tweets. Since its commencement, the social media platform twitter has gained a huge fan following and has now more than 300 million active users [2].

Twitter tweets can be sent from various third party websites after appropriate authentication as well as the twitter website or twitter mobile application. The crowd also have the command over the privacy controls in a way that they can either make their statuses available for the public view which would make them visible to anybody who is not even a twitter user or make them less visible such that it can be accessed only by twitter users or just by the followers of that particular user. Following other users feature gives a person chance to view one person's tweets on their landing page after the twitter login .

Twitter has also various interesting features. Twitter platform enables the members responding to the tweets of the others by hitting on the "reply" button that is present alongside the tweet of the person who one wants to responds to. By this process, one can respond something back to a users tweet. Furthermore, members can also add the symbol @ to the user-name of another user in their tweets, to mention the other users in their tweets. A mention enables us to point to some other member. Twitter also has provided the ability of. A retweeting is a process of distributing a members tweet to our own friends/followers. This plays a vital role in the act of spreading some news on the twitter platform. Users can also include a symbol before the keywords to create a hashtag in their tweets. This can used to categorize the list of tweets to present them easily in twitter search. The trending topics are nothing but the most used hashtags on twitter.

A. Organization of the report

Our project report is structured as follows. In the technology stack section, we have mentioned the technologies and libraries used and some installation requirements. The data collection and the process flow section outline the detail processing of collecting, cleaning and processing the twitter data. Our Features section describe the most prominent features of our project. The next section shows how the visualization has been performed during the process. The visualization section has all the plots, graph and wordcloud generated from our sample data. Also, it has screen-shots of how our sentiment analysis application looks like at present. Challenges section enumerates all the restrictions and challenges we faced along the course of whole process. Further, in the future scope section, we are discussing what more can be accomplished given more time. Finally, our report concludes with result of our analysis along with the learning process we have gone through.

III. Technology Stack

- Anaconda (Python distribution 3.x) - We have used this for our basic programming. Also iPython is used instead of the usual python files as they help in demonstration every step separately in the cell
- Python 3.x or above
- Twitter - The data was gathered from twitter.com website. The first task is to build a data set. Twitter provides APIs that we can use to interact with their service. We used the Tweepy library for this purpose
- Tweepy - Tweepy provides support to access the Twitter data by Basic Authentication and the OAuth method. Twitter has blocked the process of accepting Basic Authentication. And hence, OAuth is the only method of using the Twitter API
- Bokeh - Bokeh library is used to generate the scatter plot to display the polarity of the tweets based on location in our project
- Matplotlib - Matplotlib library is used to generate the bar graph for categorization based on the emotions
- Numpy
- Wordcloud - It is used to generate the wordcloud for all the data collected from Twitter
- ipynb - This library made it possible for us to use one .ipynb file as an imported module in another .ipynb file
- Random - It generates a random float uniformly in the semi-open range [0.0, 1.0]. We have used this in the wordcloud generation for twitter mask manipulations
- Pandas - The pandas library played an important role in converting all the collected data from twitter into manipulatable data frames
- Geopy - This library enabled us to generate longitudes and latitudes from the locations of the tweets, for the tweets that lack the actual coordinates' information
- Ipywidgets - It is a Jupyters display system, which enabled us to make two differently functioning input tabs in the jupyter notebook

IV. Data Collection

We collect different types of data for the sentiment analysis in our project. In this part, we provide a short information on the data collection process which is followed by a section that describes the results that the collected data helps us to achieve. We collect twitter data for the following three types:

A. Geo-tagged tweets

The project involved collecting geo-tagged data from Twitter. Twitter provides the users with the feature of selectively adding the tweet location to the tweets by its - Tweet with Your Location feature. The users will be able to include their location details to new tweets that they tweet, if they choose to include location information. Some applications also give choice to the users to tweet with their exact coordinates of the geo-location with latitudes and longitudes, instead of just the name of the place which they tweet from . The following figure shows an example, of a tweet posted on twitter that has geo-tagged enabled.

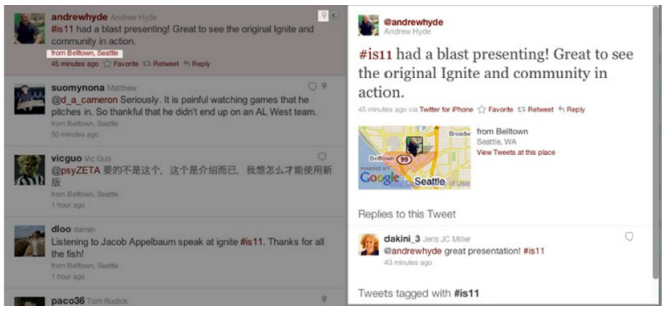


Fig. 1: An example of geo-tagged tweet

B. Non Geo-tagged tweets

The tweets that do not have the location in the form of coordinates are passed on to Geopy to get the corresponding longitudes and latitudes

C. Tweets with a particular hashtag

The next set of tweets that were collected were based on the hashtag passed as input.

Fig. 2: An example tweet with #deletefacebook

These set of tweets again serve the goal of obtaining the tweets based on a particular hashtag around the world which would help us assess the popularity of that topic at different points in time.

V. Process Flow

In the implementation of the project, a semi-supervised machine learning approach was adapted, as there was a larger proportion of unlabeled data compared to the labeled data. Broadly, following tasks were accomplished:

1) Tweets extraction by query search performed based on the hashtag(s) as input from a dropdown menu. Response is the list of the tweets that contain the hashtags - "FB data leak, "Cambridge Analytica" related to the FB data scandal.
2) The user can either initiate a search on a new topic in the "Process new Hashtag" tab or get instant results in the "Process existing set" tab

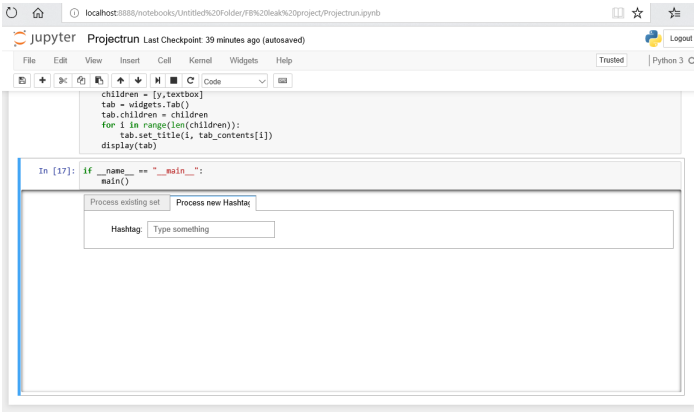


Fig. 3: Input text box

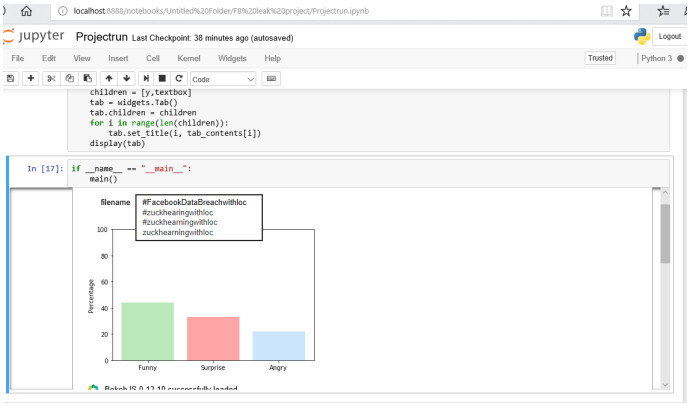


Fig. 4: Input dropdown populated

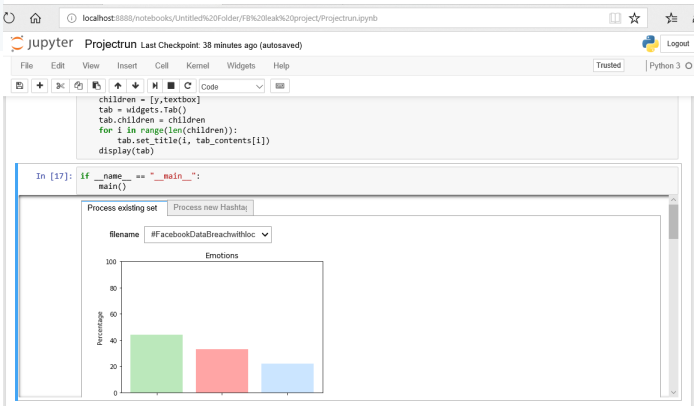
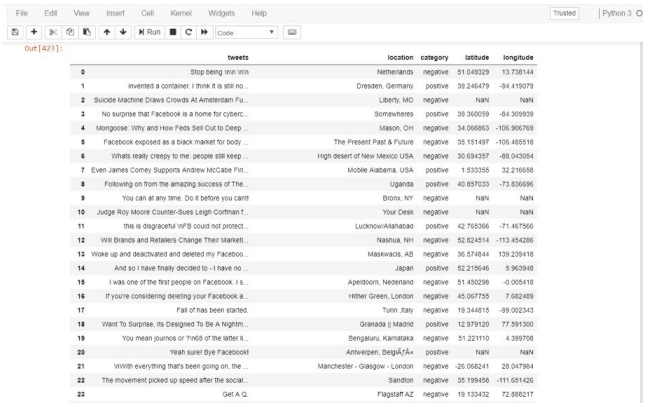


Fig. 5: Input given and after the run

3) Importing the required libraries
4) Creating a Twitter App

5) Creating a (pandas) DataFrame
6) Adding relevant data (here - tweets, category and location) are saved into our panda data frame for further processing



Out[423]:

	tweets	location	category	latitude	longitude
0	Stop being \n\n\n	Netherlands	negative	51.049329	13.738144
1	invented a container. I think it is still no...	Dresden, Germany	positive	39.246479	-94.419079
2	Suicide Machine Draws Crowds At Amsterdam Fu...	Liberty, MO	negative	NaN	NaN
3	No surprise that Facebook is a home for cyberc...	Somewheres	positive	39.360059	-84.309939
4	Mongoose: Why and How Feds Sell Out to Deep ...	Mason, OH	negative	34.066863	-106.906769
5	Facebook exposed as a black market for body ...	The Present Past & Future	negative	35.151497	-106.465518
6	Whats really creepy to me: people still keep ...	High desert of New Mexico USA	negative	30.694357	-88.043054
7	Even James Comey Supports Andrew McCabe Fir...	Mobile Alabama, USA	positive	1.533355	32.216658
8	Following on from the amazing success of The...	Uganda	positive	40.857033	-73.836696
9	You can at any time. Do it before you can!!	Bronx, NY	negative	NaN	NaN
10	Judge Roy Moore Counter-Sues Leigh Corfman f...	Your Desk	negative	NaN	NaN
11	this is disgraceful \nFB could not protect...	Lucknow/Allahabad	positive	42.765366	-71.467566
12	Will Brands and Retailers Change Their Marketi...	Nashua, NH	negative	52.824514	-113.454286
13	Woke up and deactivated and deleted my Faceboo...	Maskwacis, AB	negative	36.574644	139.239418
14	And so I have finally decided to - I have no ...	Japan	positive	52.215646	5.963948
15	I was one of the first people on Facebook. I s...	Apeldoorn, Nederland	negative	51.450296	-0.005418
16	If you're considering deleting your Facebook a...	Hither Green, London	negative	45.067755	7.662489
17	Fall of has been started.	Turin ,Italy	negative	19.344615	-99.002343
18	Want To Surprise, Its Designed To Be A Nightm...	Granada || Madrid	positive	12.979120	77.591300
19	You mean journos or ?\n68 of the latter li...	Bengaluru, Karnataka	negative	51.221110	4.399706
20	Yeah sure! Bye Facebook!	Antwerpen, BelgiÃ¢	positive	NaN	NaN
21	\nWith everything that's been going on, the ...	Manchester - Glasgow - London	negative	-26.056241	28.047964
22	The movement picked up speed after the social...	Sandton	negative	35.199456	-111.651426
23	Get A Q.	Flagstaff AZ	negative	19.133432	72.886217

Fig. 6: A snapshot of our dataframe

7) Developed a classifier to perform Sentiment Analysis of the twitter data based on the three emotions - Anger, Fun / Happiness, Surprise
8) Determined the Polarity of the tweets, by classifying the same into positive and negative categories
9) Visualized the tweets classified based on emotions as a bar graph and polarity as a scatter plot
10) Generated wordclouds based on the tweets collected

A. Pre-processing

1) When the tweets are collected, they may be noisy in their original form. Many Twitter users use clumsy, ungrammatical sentence structures, and we would have to weed out the irrelevant parts of the tweet, like the URLs, mentions etc., to be able to process them for the sentiment analysis
2) A data cleanup of the twitter data by tokenizing the tweets was done after removing the following:
   - hastags
   - retweets
   - stopwords
   - URLs
   - mentions
   - punctuations that had little or no semantic meaning towards a tweet

VI. Sentiment Analysis

A. Classification

For the classification of the twitter data, the Nave Bayes Classifier was used.

Nave Bayes: Nave Bayes is a simple yet effective probabilistic technique to construct a classifier that is based on the Bayes theorem. The Nave Bayes classifier makes an assumption that the presence of a specific feature of a class is in no way related to the presence of any other feature. Nave Bayes models use parameter estimation pertaining to the method of maximum likelihood. These classifiers typically outperform the other classifiers when the sample sizes are small. [3].

Using Bayes' theorem, the conditional probability can be decomposed as:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Fig. 7: The conditional probability using Bayes theorem

The above equation is a Bayes classifier which is basically a function that performs the task of assigning y=C (a class label) for some k.

Using Bayesian probability terminology, the above equation can be written as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

1) Polarity Classification:: A prime task in sentiment analysis is the classification of polarity of a given tweet/text at the feature/aspect, sentence or document level whether the conveyed opinion in that feature/aspect, sentence or document is positive, negative, or neutral. The current approaches mainly depend on the personal feelings, tastes, or opinions part of text in which sentiment is expressed explicitly through specific words, called sentiment words. We have classified the tweets into two categories here positive or negative.

2) Emotion Classification:: Advanced, "beyond polarity" sentiment classification looks, at emotional states such as "angry", "surprise", and "happy/funny" in the pre-processed tweet data [4].

VII. Features

A. Bag of Words

In a Bag of Words model, the representation of text is in the form of a word-multiset, where we dont pay any heed to the order or the grammatical structure, but we take the multiplicity into consideration. This model is widely used in text/document classification methods, where the word frequency is typically utilized as a feature for classifier training. Hence, the bag of words model is an important form of information representation in areas like Information retrieval and Natural Language Processing (NLP) [5].

B. Wordcloud Generation

A tag cloud (Wordcloud, or weighted list in visual design) is a unusual visual representation of tweet/text data, usually used to represent keyword metadata (tags) on websites, or

to visualize free form text. Tags are usually just one word, and the weightage of each tag is shown with font size or color[6]. This format is mainly useful for a quick perception of the most important/prominent terms and for discovering a term alphabetically to find its relative prominence. When used as an aid to ascertain a website's position, the terms are hyper-linked to items associated with the tag.

Type of Wordcloud:

There are three main types, in social software, of tag cloud applications , distinguished by their appearance but their meaning. In the first type, the frequency of each unit has a tag, whereas in the second type, global tag clouds are present in which over all items and users, the frequencies are grouped. In the third type, categories are present in the cloud, in which the size indicates the subcategories count.

Frequency type:

In this type, used in this project, as a presentation of each unit's state or condition of being liked, size depicts the number of units to which a tag has been applied.

VIII. Visualizations

Following are a few visualizations and analysis on the data collected using the locations and keywords that allow us to draw certain simple inferences from the above tweet data.

A. Visualization of tweets based on emotions - Bar Graph:

Bar graphs can be used to show how something changes over time or to compare items. Here, the emotions are on the x-axis and the percentage of the twitter crowd is on the y-axis. Matplotlib library is used to generate the bar graph for categorization based on the emotions. This graph shows the comparisons among the categories - angry, surprise and happy/funny.

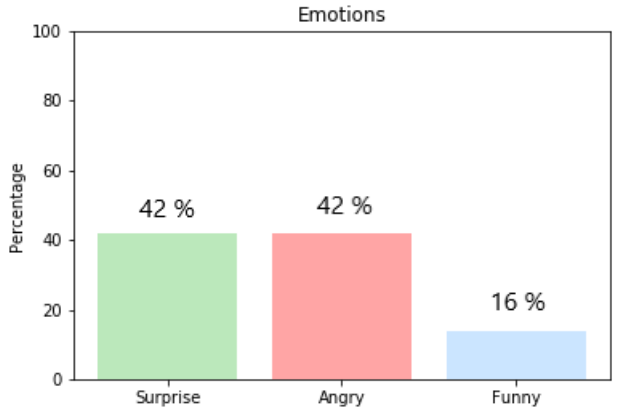


Fig. 8: Bar Graph - Polarity classification

B. Visualization of tweets collected by location - Scatter Plot on a map:

Scatter plots on the world map here mark geographic areas and are colored based on the polarity. The positive category is represented by blue and the negative by red. Here, Bokeh library is used to generate the scatter plot to display the polarity of the tweets based on location. And, Geopy is used to acquire the latitudes and the longitudes for the tweets that lack the coordinates' information.

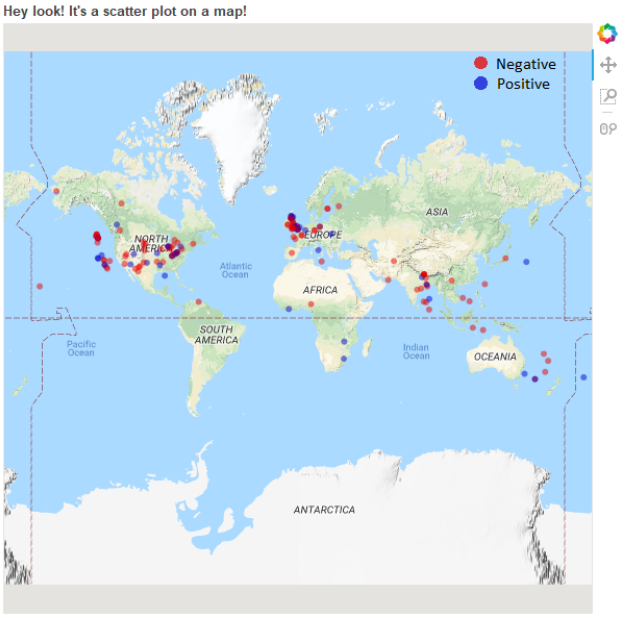


Fig. 9: Scatter Plot - Emotion classification

C. Visualization of tweets as a Wordcloud:

A word cloud is a pretty traditional, and maybe already old fashioned way to depict the content of a text or a corpus (a set of texts). Nevertheless, it is still a good way to convey the general idea of the text. This form of communication can be further improved by generating a word cloud image which resembles the general idea of the text. For this purpose, Wordcloud was generated using the twitter logo image as mask. This well reflected the basic idea behind our project.

1) Wordcloud with a twitter logo mask: The wordcloud library using an image as a mask fills all available space of the given image and can utilize arbitrary masks. This also has simple algorithm (with an efficient implementation) that can be easily modified in Python.

Following is the wordcloud that gets generated with the use of the twitter logo image as mask.

Fig. 10: Masked Wordcloud for the collected twitter data

2) Wordcloud without a mask: Following is the default wordcloud that gets generated by the use of wordcloud package.

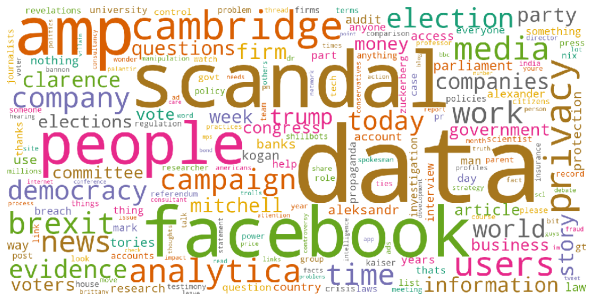


Fig. 11: Unmasked Wordcloud for the collected twitter data

IX. Challenges Faced

- Geopy restricts the user re-sending the same data, which hindered the processing. This needed to be handled in our process, in such a way that the search is done for unique data (location) during every call to Geopy and saving it locally to use for the further processing and eventual visualization
- Tweets are highly unstructured and also non-grammatical - with links, mentions, hashtags etc., so we had to pre-process every tweets to make it structured for the processing
- Non Vocabulary Words - grammatically challenged words are of no use to the sentiment analysis as the libraries and the algorithms will not recognize them
- Lexical Variation - use of words such as coz, tmrw etc., users often shorten the words they tweets to fit their ideas within the character limit and these are not helpful in our processing
- Removing re-tweeted tweets: Retweets are the same tweets shared by many users on twitter. While we can keep count of these retweets to assess the popularity of that particular tweet or the sentiment emitted by that tweet, we cannot have them in processing and had to be get ridden of to avoid redundant requests

X. Future Scope

It is never guaranteed to know whether the particular tweet emits a positive sentiment or negative as any machine or even a human cannot read another human's mind. We just have to operate on the assumption and tag the opinions under the categories that closely resemble that sentiment. Thus our future scope involves making this process of categorizing better by making improvements to the algorithm and our model thus. We also are working on our model to broaden the categorization by adding more emotions. We also plan to make the UI more user friendly and interactive. We also plan to make the geographic maps interactive.

XI. Conclusion

The most popular communication tool to share everyday opinions and life events is social media. Twitter is one such major online social networking service. We can analyze the tweets on any topic in real time, and infer meaningful relationships. When aggregated, these tweets reflect public sentiment towards that topic. We can extract the sentiment from them and look at the general correlation between these sentiments and a topic. A large collection of such tweets could be leveraged to provide a useful reflection of public sentiment towards sports.

As we have shown, we can perform a very accurate Sentiment Analysis on any topic. In our case, we have chosen the recent FB data leak scandal. This kind of analysis can be made on just not on the hashtag search but on mentions, word search as well using our model.

XII. Appendix

Sentiment Analysis on the Consequences of FB Data Breach, the Cambridge Analytica Scandal

College of Engineering and Computer Science
Syracuse University

Dinesh Kumar Dhamotharan Khushboo Gupta Sakthi Priya Rajendran

April 26, 2018

Outline

What is our project about?

- The Facebook and Cambridge Analytica data breach is the data breach of a collection of personally identifiable information of about 50 million Facebook users that Cambridge Analytica began collecting in 2014
- 'Delete Facebook' hashtag has been trending on twitter and other social media platforms after the news of the scandal broke. There are also section of people who claim that they knew about this privacy setting from the very beginning
- So, our project aims at analyzing the aftermath of this scandal on the Twitter platform which would be a reflection of the public sentiment regarding the same

Why did we choose Twitter for Sentiment Analysis?

- Popular micro-blogging site
- Short text messages of 140 characters
- 240+ million active users
- 500 million tweets are generated everyday
- Twitter audience varies from common man to celebrities
- Users often discuss current affairs and share personal views on various subjects
- Tweets are small in length and hence unambiguous

Libraries and Languages used

- Anaconda (Python distribution 3.x)
- Python 3.x or abaove
- Tweepy
- Bokeh
- Matplotlib
- Numpy
- Wordcloud
- Random
- Pandas
- Geopy

Extracting twitter data

- Importing our libraries
- Creating a Twitter App
- Tweets extraction by query search performed based on the hashtag(s) as input from a dropdown menu. Response is the list of the tweets that contain the hashtags
- Creating a (pandas) DataFrame
- Adding relevant data (here - tweets and location) are saved into our panda data frame for further processing
- Also saved as a csv file for backup
- Visualization of the tweets classified based on emotions and polarity
- Generating word cloud

Data Cleaning

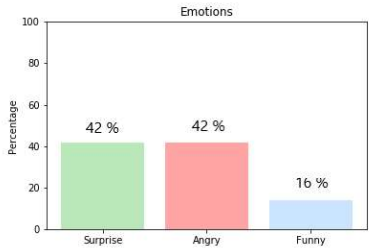
- Data cleanup and the pre-processing of the twitter data by removing the following and tokenizing the tweets for further processing
  1. hastags
  2. retweets
  3. stopwords
  4. URLs
  5. mentions

Sentiment Analysis and Word Cloud Generation

- **Sentiment Analysis:** Sentiment analysis is a text mining technique to analyze the sentiment of the writer or to the topic written about overcoming social slang and lingos, non-textual expressions and language
  Emotion Classification:
  Mapping the tweets to the categories - Anger, Surprise, Fun/Happiness
  Polarity Classification:
  Tweets are also classified to one of the two categories of polarity - positive or negative, using **NLTK library**
- **Word Cloud Generation**:
  A word cloud visualized using **matplotlib**, with words sized based on their frequency within the Pandas dataframe

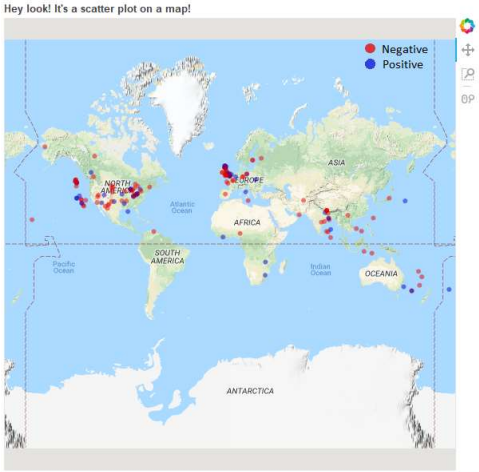
1. Bar graph for the tweets classified based on emotions

- Bar graphs are used here to present the tweets of different groups of emotions that are being compared with each other



2. Scatter plot on map for the tweets classified based on locations

- Twitter allows its users to provide their location when they publish a tweet which would be used to get latitude and longitude coordinates from Geopy
- With this information, we are ready to create some nice visualization for our data, in the form of maps

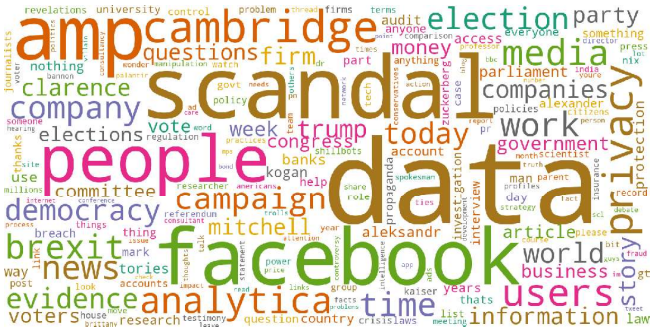


Word Cloud of the tweets

- Generating word cloud from the CSV file with the processed tweets created from the previous steps, with the word frequencies
- Word Cloud is a good way to convey the general idea of the text. So, we have tried to improve this form of communication further by generating a word cloud image which resembles the general idea of the text which could be accomplished using masks

Word Cloud for #cambridgeanalytica - with twitter logo mask

Word Cloud for #cambridgeanalytica - without image mask



Challenges faced

- Geopy restricts the user re-sending the same data, which hindered the processing. This needed to be handled in such a way that the search is done for unique data (location) during every call to Geopy and saving it locally to use for the visualization
- Tweets are highly unstructured and also non-grammatical - with links, mentions, hashtags etc.,
- Out of Vocabulary Words - grammatically challenged words
- Lexical Variation - use of words such as coz, tmrw etc.,
- Removing re-tweeted tweets

Conclusion

- Twitter sentiment analysis comes under the category of text and opinion mining
- It focuses on analyzing the sentiments of the tweets using python libraries and visualize them using bar graphs and geographical maps
- It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model
- Also its also tested for topics (other inputs) other than the one in consideration
- Hence, our sentiment analysis can be used for other input types like querying for mentions, particular text in tweets, other hashtags etc.,

For Further Reading I

- Facebook and Cambridge Analytica data breach, https://en.wikipedia.org/wiki/Facebook_and_Cambridge_Analytica_data_breach
- Bag of words in NLTK package, https://en.wikipedia.org/wiki/Bag-of-words_model
- Emotion Classification, http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- Sentiment Analysis, http://dspace.thapar.edu:8080/jspui/bitstream/10266/4273/4/4273.pdf

THANK YOU

QUESTIONS???

ACKNOWLEDGMENT

The authors are grateful to Professor Martin Harrisson for his unremitting encouragement throughout the coursework.

REFERENCES

[1] Facebook and Cambridge Analytica data breach, Wikipedia, 30-Apr-2018. [Online]. Available: https://en.wikipedia.org/wiki/Facebook_and_Cambridge_Analytica_data_breach

[2] Using Twitter as a data source: an overview of social media research tools (updated for 2017), Impact of Social Sciences, 09-May-2017. [Online]. Available: http://blogs.lse.ac.uk/impactofsocialsciences/2017/05/08/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-updated-for-2017/

[3] "Naive Bayes classifier, Wikipedia, 26-Apr-2018. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[4] "Subjectivity Lexicon," — MPQA. [Online]. Available: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[5] "Bag-of-words model, Wikipedia, 10-Apr-2018. [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model

[6] "Tag cloud, Wikipedia, 25-Apr-2018. [Online]. Available: https://en.wikipedia.org/wiki/Tag_cloud

[7] Study.com. [Online]. Available: https://study.com/academy/lesson/ekmans-six-basic-emotions-list-definitions-quiz.html

[8] Understanding, Analyzing, and Retrieving Knowledge from Social Media, Understanding, Analyzing, and Retrieving Knowledge from Social Media - CUCIS. [Online]. Available: http://cucis.ece.northwestern.edu/projects/Social/. [Accessed: 01-May-2018]

[9] T. Ohbe, T. Ozono, and T. Shintani, Developing a Sentiment Polarity Visualization System for Local Event Information Analysis, 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2016

[10] "10 useful ways to visualize your data(with examples), Big Data Made Simple - One source. Many perspectives., 27-Jun-2016. [Online]. Available: http://bigdata-madesimple.com/10-useful-ways-to-visualize-your-datawith-examples/

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proc. ACM EMNLP ACL-02 Volume 10, 2002

[12] Saima Aman and Stan Szpakowicz, "Identifying Expressions of Emotion in Text", 1 School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada 2 Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland

[13] R. Mehta, D. Mehta, D. Chheda, C. Shah, and P. M. Chawan, Sentiment analysis and influence tracking using twitter, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE), vol. 1, no. 2, p. pp72, 2012

[14] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, Election result prediction using Twitter sentiment analysis, in Inventive Computation Technologies (ICICT), International Conference on, 2016, vol. 1, pp. 15

[15] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10), May 2010.