

# 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

### Introduction

Project Idea

Motivation - Why Twitter?

### Requirements

#### Design

Tweets Collection

Text Pre-processing

Data Cleaning

Sentiment Analysis and Word Cloud Generation

### Results - Data Visualization

Emotion classification

Polarity classification

Word Cloud Generation

### Challenges

### Conclusion



## 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 above
- ▶ 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



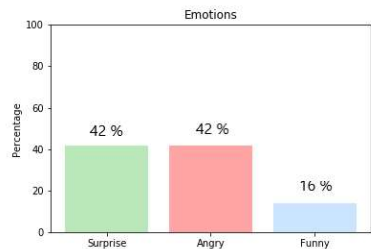
## Sentiment Analysis and Word Cloud Generation

- **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

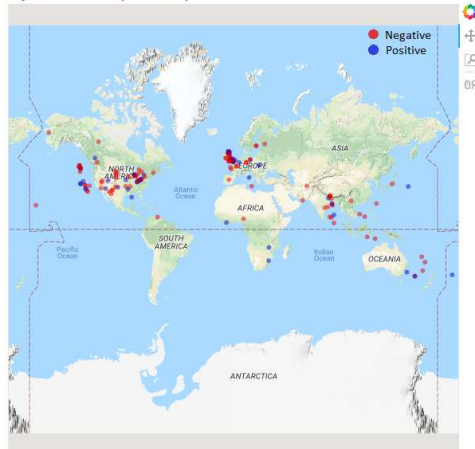


## 2. Scatter plot on map for the tweets classified based on locations

- 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



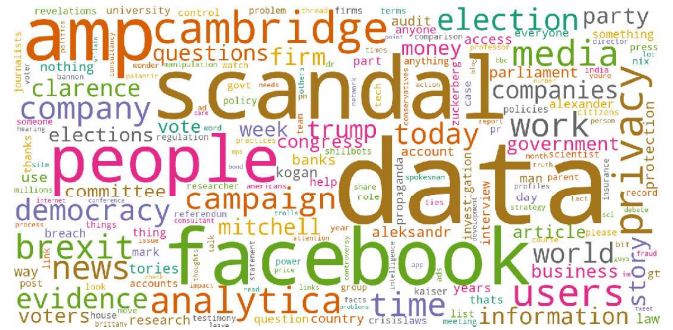
Hey look! It's a scatter plot on a map!



## 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

A word cloud in the shape of a Twitter bird, filled with terms related to the Cambridge Analytica scandal. The most prominent words are 'data', 'scandal', and 'facebook'. Other visible words include 'privacy', 'cambridge', 'analytica', 'election', 'company', 'brexit', 'story', 'work', 'evidence', 'people', 'government', 'party', 'money', 'today', 'clearance', 'campaign', 'time', 'users', 'anonymity', 'part', 'year', 'day', 'project', 'article', 'news', 'investigation', 'documents', 'access', 'information', 'data', 'company', 'brexit', 'story', 'work', 'evidence', 'people', 'government', 'party', 'money', 'today', 'clearance', 'campaign', 'time', 'users', 'anonymity', 'part', 'year', 'day', 'project', 'article', 'news', 'investigation', 'documents', 'access', 'information'.



- ▶ 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

- ▶ 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.,

-  Facebook and Cambridge Analytica data breach,  
[https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Bag-of-words_model)
-  Emotion Classification,  
[http://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon/](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???