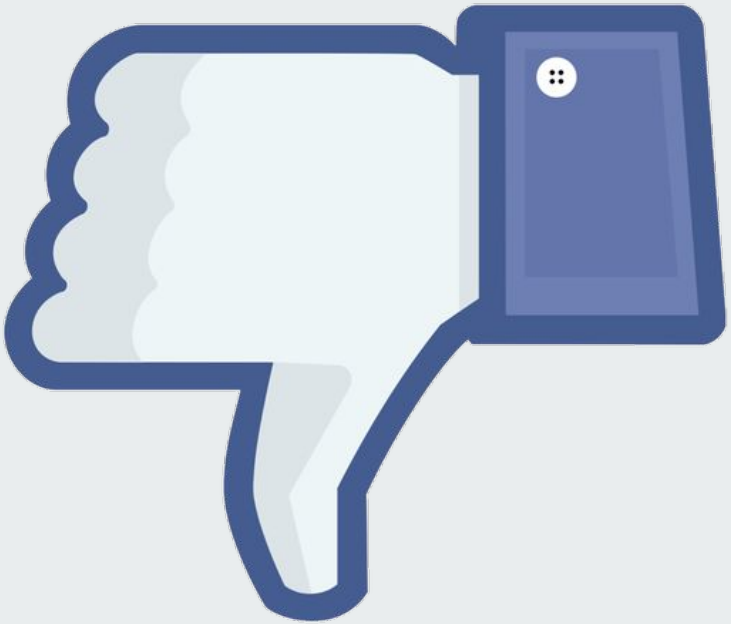
Facebook's Data Breach Scandal and Fallout

An Exploration of Twitter Users' Common Concerns and Sentiments

(Data Analysis Plan)



Group 1

Sahil Aggarwal
Brian Hudnall
Hye Kim
Neha Palsokar

Agenda

Research question

Data Collection

Data Cleaning

Summary Statistics

Analysis Plan

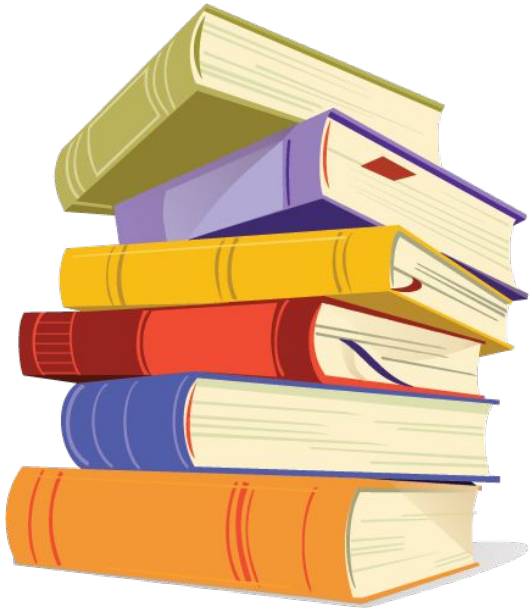
Moving Forward

Research question

“How concerned are Twitter social media users with data privacy in relation to the Facebook scandal and what topics are most concerning to them ”

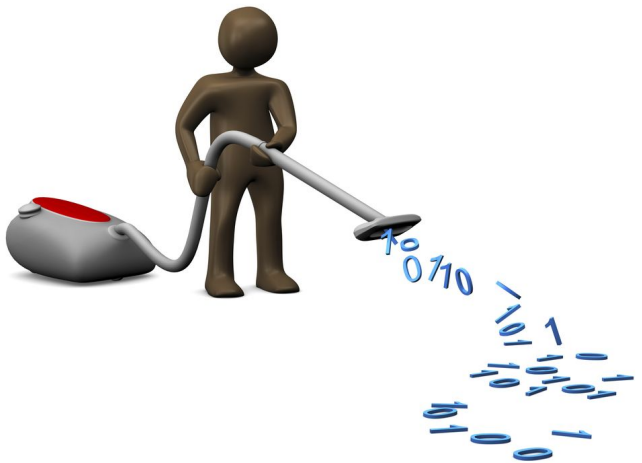
Data Collection

- We created a Python script that leverages the Tweepy library to connect to the Twitter search API.
- The script accepted a search term as a parameter passed to it from the command line on launch.
- The script also did early stage cleaning.
- A powershell script was used to call the script and pass the keyword value to the script.
- Windows scheduler on a remote server automated the launch of the python script.
- The script ran every hour for 7 total days and appended the raw data to a .txt file, and the cleansed data to a .csv



Data Cleaning

- Stemming and lemmatization
- Removed symbols
- Removed duplicate tweets
- Converted the words to lowercase



Summary Statistics

In the dataset, the keywords are associated with the following number of tweets over a period of seven days:

Keyword	Number of Tweets Associated With Keyword
Aleksandr Kogan	68
Cambridge Analytica	3928
deletefacebook	2099
facebook	2454
Mark Zuckerberg	2854

Summary Statistics

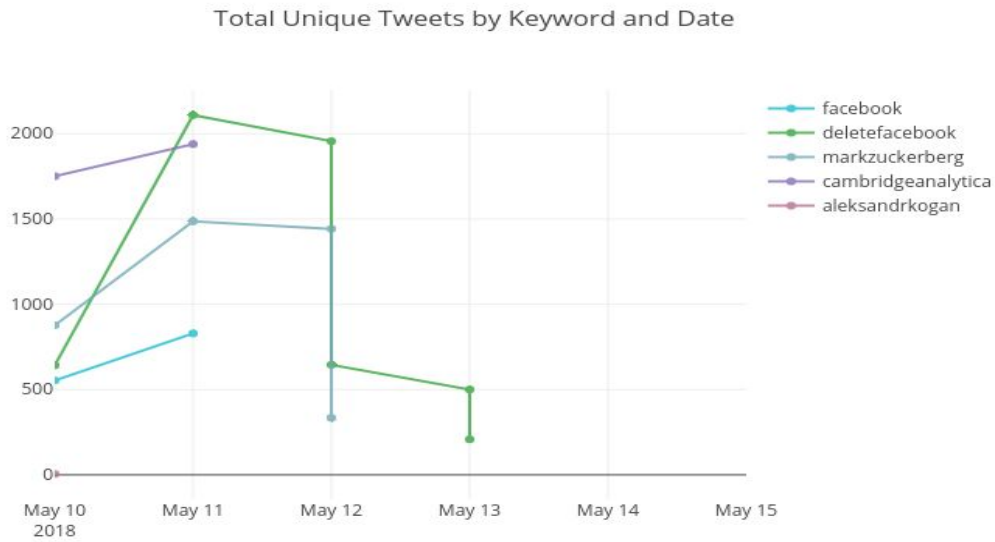
For the entire data set, the top 15* words appear the following number of times:

data	1020	whistleblow	603	use	461
investig	940	justic	582	today	291
fbi	77.3	new	572	know	291
wa	637	say	509	may	279
us	636	depart	481	birthday	277

* The report has 100 top words but for the purposes of the presentation, only 15 were included. Please see the Data Analysis report for the complete list.

Summary Statistics

Frequency of Keyword by date



Analysis Plan

The goal is to look at high level topics that describe user concerns as well as compare sentiment for each keywords and (also) see overall sentiment.

So with the above summary statistics and our data set, we aim to tell a story that is about how and to what degree different stakeholders/players are affected by the scandal based on the keywords and their associated words and sentiments.

Moving Forward

If there are more resources and time in the future, we hope that the research can expand to actual responsibility of each player and compare it with the public's perception of the responsibility of the players.

This information can later be used to help stakeholders to help manage their image after scandals. Also, it would show them the consequences of their roles and decisions regarding their handling of users' data.


The End