The Data Breach at Facebook

Scandal & Fallout

An Exploration of Twitter Users' Common Concerns and Sentiments

A REPORT PRESENTED BY

Sahil Aggarwal
Brian Hudnall
Hye Kim
Neha Palsokar

IMT 547
SOCIAL MEDIA DATA MINING AND ANALYSIS

INFORMATION SCHOOL
UNIVERSITY OF WASHINGTON
SEATTLE, WA

I. INTRODUCTION

BACKGROUND ABOUT FACEBOOK

Facebook was founded by Mark Zuckerberg, Eduardo Saverin, Dustin Moskovitz, Andrew McCollum, and Chris Hughes in 2004. (Facebook, N.D.). Facebook is available via the web and mobile devices and provides an online platform for user to share their day-today lives through personal posts and photos, ability to organize events, and share others' content. Furthermore, Facebook expands the boundaries of what is said and received in the whole world and serves effectively as a marketplace of ideas which are much broader and diverse than ever before. It also enhances the reach of non-mainstream political movements especially in places with stringent controls over political rights. Under any condition or circumstance, Facebook has fully embodied its mission statement "to give people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them." (Facebook, N.D.). However, it appears Facebook's business model also includes connecting third party companies with users' very personal data.

In 2017, CEO Mark Zuckerberg stated that the platform had 2 billion users each month in 2017 (Balakrishnan, 2017). From these users, Facebook collects realms and realms of data about a user which is then used for active advertising generating billions of dollars. Categories like a user's login information, where a user logged in from, the browser used to send the request to its server, users IP address, the device a user logged in from, their contacts, the messages sent by the user, biometric facial data of a user, their likes etc. get dumped into its database every time a user logs in (Singer, 2018). Facebook then uses sophisticated artificial intelligence to analyze a user's behavior and other non-transparent analysis and profiling. Further, many websites and applications use Facebook to enhance their services (Baser, 2018). Facebook offers services like social plugins, Facebook logins, Facebook analytics and ads measurement tool to websites which share their data with Facebook essentially helping it build an information goldmine. (Baser, 2018)

THE CAMBRIDGE ANALYTICA SCANDAL

Data collection practices by apps and websites are very common however, in this instance, Cambridge, Kogan, and Facebook may have crossed a line. Starting in 2010 Facebook made Open Graph available to third party apps which allowed them to request access Facebook users' personal data such as user's name, gender, location, birthday, education, political preferences, relationship status, religious views, online chat status, private messages, and even their friends' personal data.

In 2013, Akeksandr Kogan (and his company Global Science Research) created the thisisyourdigitallife app which paid users to take the psychological test as it collected the users' personal data (from the test and their Facebook profiles via Open Graph). However, in 2014, Facebook required that the third-party companies must gain direct permission from users' friends to access their data (whereas before they could access friends' data through one person's permission). Despite the rule change, Kogan did not delete the data that was previously collected without permission.

In 2015, it was reported that Cambridge Analytica was helping politicians gain an advantage over their opponents by using the personal psychological data of Facebook users. When Facebook discovered how the data was being use, they pressured Kogan and Cambridge Analytica to remove all the personal data, which the two companies certified had been done. Suspiciously, in 2016 Trump's campaign team begins to invest in Facebook ads and teams up with Cambridge Analytica.

On March 17, 2018, it was exposed that roughly 87 million Facebook users' profile data were collected for Cambridge Analytica to develop psychographic profiles of people and deliver pro-Trump material to them. However, Cambridge Analytica denies that Kogan's data collect from his thisisyourlife app was used in connection with Trump's campaign. Regardless, the Federal Trade Commission opens an investigation into Facebook regarding violation of privacy protections on March 20, 2018 and on April 10, 2018 and April 11, 2018, Mark Zuckerberg testifies in Senate Judiciary and Commerce committee, and House Energy and Commerce Committee hearings. During the hearings, he was asked questions about Facebook's privacy protection practices.

RESEARCH QUESTION

With the recency of the Facebook scandal, this research investigates Twitter users' common concerns as members of Facebook. For example, users may be most concerned about data privacy. We will quantify topics by using word counts and topic modeling. From there, sentiment scores will be calculated to quantify positive, negative or neutral perception of facebook. Therefore, the primary question we aim to answer is <u>how concerned are Twitter social media users with data privacy in relation to the Facebook scandal and what topics are most concerning to them?</u> Entities that could benefit from this research include any organization that retains and manages user information. By understanding the primary concerns of users, they can adjust their communication strategy by highlighting the primary tactics they use to properly manage user information as well as manage the aftermath of a potential scandal.

RESEARCH/PROJECT SCOPE

Due to the lack of resources ($0.00) and limited time (approx. > 1 month), the project operated within a narrow and tight scope. We were only able to access data from the Search API within 7 days of the request, therefore the amount of collected data was limited. These restraints played into the team's (in)ability to review, clean, and sort a large mass of data. Also, the team lacked sufficient hardware or processing power to handle more data.

II. METHODS

DATA COLLECTION

The team used Twitter Search REST APIs to collect relevant Tweet data. The API returned a collection of relevant Tweets matching a specified query. We used a comma separated values of ***keywords***, with a maximum of 5 keywords total. This way, we could specify the desired data. We experimented with this parameter to find values related to the Facebook controversy on Twitter. Some keywords the team used are: #deletefacebook, Facebook, Cambridge Analytica, Mark Zuckerberg, and Aleksandr Kogan.

The tools used to collect and assist with analyzing the data were:

- Python as the language to create the API calls, extract, and munge the data.
- Python Libraries:
    - Pandas: used to easily shape the data into a functional DataFrame object.
    - SKlearn: to convert the text data into vectorized arrays and a corpus to run NLP sentiment models with.
    - NLTK: used for sentiment analysis on the corpus.

DATA CLEANSING

After obtaining the data, we used several methods to clean the tweets and prepare them for analysis. First, we removed stop words that do not add value to the meaning of the tweet. These primarily consist of conjunctions and pronouns, like: ‘i’, ‘and’, ‘but’. Second, the words were converted to lowercase and the symbols will be removed. Third, we deployed stemming and lemmatization of words to convert words to their derivational form. We then removed duplicate retweeted tweets from the lateg corpora to strengthen our analysis to measure sentiment on unique tweets.

ANALYSIS METHOD

Once the data was collected, cleansed, and organized, the team deployed the following to analyze the data:

- ***N-grams****:* This calculates word counts to understand word usage pre/post Facebook events.
- ***Topic Modeling***: The Latent Dirichlet Allocation can be used to do topic modeling and understand what are common topics that are relevant posts regarding Facebook. The team will gather the topic common topics from each period of time (pre-scandal, post-scandal, post-hearing).
- ***Sentiment Analysis****:* With the assistance of the python Vader sentiment analyzer, tweets collected that match the search parameters will be processed to determine if the collection of words are positive, neutral, or negative.

III. ANALYSIS/RESULTS

PRIMARY TAKEAWAYS

- Twitter users' often use the platform as a place for news updates. Within our data set, many of the topics we uncovered in relation to the Facebook scandal were based on New York Times or Guardian articles, like the NYT article 'Justice Department and F.B.I. Are Investigating Cambridge Analytica'. Although newspapers, and older content creators are often viewed as antiquated, they are proliferated on Twitter.
- Cambridge Analytica is a lead driver in negative sentiment. Of the keyword topics, Cambridge Analytica had the second highest tweet total within the data set. The topics most discussed within tweets related to Cambridge Analytica were related to the FBI and Justice Department Investigation. Of the 4.4K total tweets, the phrase 'Justice depart fbi' was present in 15% of all tweets. Relative to the other keyword categories, Cambridge Analytica had the highest negative sentiment at 10.5% and the lowest positive sentiment score of 5.6%.
- There were not consistent topics in general when users tweeted #deletefacebook. In general, this topic had the highest negative sentiment score at 11%. At 1% of tweets related to this keyword category also included the hashtag #internetbilllofrights, potentially referring to more user control of their online information.

EXPLORATORY DATA ANALYSIS

Before we began topic modeling or sentiment analysis, it was important that we did Exploratory Data Analysis (EDA) to understand the nuances of the data. This also helped to ensure if we need to conduct further data cleansing if needed to assist in the downstream analysis process. We began EDA by understanding the grain of our data table, and further dimensions that could be extracted within the tweet text data for enrichment. The columns we have within the data set are datetime, tweet, and keyword. Datetime is the timestamp the tweet occurred, tweet is the cleaned text of the tweet including the username, and keyword is the keyword that was used within the search API.

Some further cleansing was conducted to make for easier analysis. As datetime makes the data harder to aggregate across days due to the lower granularity, we added a new column that is the date version of datetime. After inspecting the tweet text further, we noticed that there was a large number of tweets that are 'Retweets'. On twitter, a retweet is when users repost tweets of other users. Retweets made up 76% of our total tweet data. Naturally, retweets create duplicates within the data. This is a strong indicator that the majority of content in relation to the Facebook scandal within our dataset is repurposed content. Of all keywords used, 'Cambridge Analytica' had the highest relative counts of retweets at 83%, followed by 'Mark Zuckerberg' at 78%, 'deletefacebook' at 71%, and 'facebook' at 47% (Figure 1).

As a political consulting firm, Cambridge Analytica was a relatively unknown entity to the general public before the scandal. Their part within the controversy is also very specific in comparison to other entities we used as keywords, such as Facebook, and Mark Zuckerberg. It's also possible that the general public has a hard time clearly grasping their role within the controversy. These reasons may have led to a higher proliferation of retweets related to

Cambridge Analytica. Without data to investigate reasoning, this hypothesis would need to be validated with a subsequent study.

The top retweet was 8% of all retweets. Retweets inherently add value as they convey importance of a tweet across more than one user, though we needed to separate tweets and retweets to ensure that a small percentage of tweets did not dominate the overall takeaways. A distribution of the top 10 retweets is described in Figure 2.

After removing retweets, the final dataset contained 15,928 total unique tweets with the distribution by search keyword as described in Figure 3.

The tweets were collected across May 4, 2018 to May 26, 2018. After generating word counts from the data, we also noticed it was important to remove keyword that were used within the API search from the tweet text itself. If we were to leave these words in, the large majority of topics we discovered would be related to the keywords we used to extract the data from Twitter (Figure 4).

It is surprising that the issue is about whether or not the data improperly collected by aleksandr kogan was used by Cambridge Analytica or not but yet, aleksandr kogan's name was mentioned a mere 46 times. Which suggests that the public does not really understand the whole story of the scandal at hand and its causes. Also it is difficult in general for people to understand the value of data, how it is collected, or in general how it is used let alone all the nuances of the all the players and their specific roles in the timeline.

This could suggest that companies should be more worried about perception rather than actual level of involvement in scandals. Therefore, when companies are faced with data scandals, it is important they share the whole story so that companies (external and internal) and people are more aware of how data was managed and how it should be managed in the future, in whole a more transparent process is encouraged.

TOPIC MODELING

Naturally, the dataset contained massive amounts of information relative to the Facebook controversy. We needed a way to be able to synthesize the data and understand the key topics relative to the scandal. With this information, we are able to understand the pieces that are most glaring/important to Twitter users. We used two primary topic modeling methods to help establish key topics, including generating N-grams and Latent Dirichlet Allocation.

Our first attempt relied on N-grams. This method combines adjacent keywords into word combinations. We were then able to aggregate these word combinations to see which phrases are most prevalent within the data. The number of word combinations we used ranged from 3-4. It was also important that we then checked the percentage of tweets that contained these key phrases that were established through N-gram aggregation. It’s possible that keyword search can contain a wide range of unique topics, with no one topic making up a larger percentage of the story. Inversely, there are keyword searches that all pertain to one specific topic and there is a clear takeaway.

Using the visuals below, we used the top chart to rank the top phrases within a keyword relative to their overall frequency within that keyword category. For example, ‘justic depart fbi’ was the leading topic within that keyword category based on frequency. We then used the chart in Figure 5 to get a relative sense of the percent of total tweets that contain each key phrase.

As the words used within the N-gram have been stemmed and lemmatized, we needed to manually review each topic and also determine if there was overlap across key phrases. Through this process we established the following topics relative to each keyword category:

Keyword Category	Cambridge Analytica	Facebook	Deletefacebook	Mark Zuckerberg
Topic 1	Justice department and the FBI investigation	Post new video	Internetbillofrights	Set up fraudulent weapons scheme (guardian article)
Topic 2	Mark Zuckerberg meeting with Europe	Suspending 200 apps	Googlegestapo	
Topic 3	Sheera Frenkel New York Times coverage		Social Media censorship	
Topic 4			Class action lawsuit	

Table 1. Trending topics in per keyword

In general, we see legal actions as a meta topic across each keyword categories. For Cambridge Analytica, 25% of tweets contain the word ‘FBI’. More than any other keyword category, it’s clear that the most important topic related to Cambridge Analytica, is the FBI investigation.

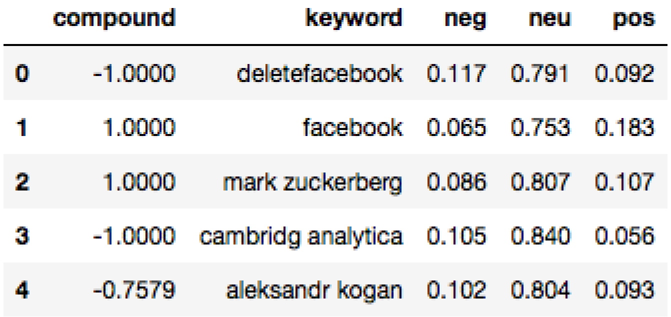
The most relevant meta topic within #deletefacebook is related to digital censorship. The most frequent topic within tweets is the Internet Bill Of Rights (Link 2), social media censorship is also frequently tweeted relative to other topics. The words ‘censorship’ and ‘internetbillofrights’ are contained within less than 1% of deletefacebook related tweets. Although it is the most frequent topic within the deletefacebook keyword category, it’s not an overwhelmingly high percent of the conversation. As expected, the Facebook keyword category has many different topics similar to deletefacebook. It was hard to find topics relevant to the scandal, though we did see roughly 1% of of tweets within this category contain the words ‘suspend 200’. This is most likely related to suspending 200 apps that had misused Facebook in accessing customer data.

SENTIMENT ANALYSIS

Sentiment analysis is an active area in the field of Natural Language Processing and is used to analyze text data to decipher emotions, opinions and attitudes. Valence Aware Dictionary for sEntiment Reasoning (VADER) is an approach which uses quantitative and qualitative methods to perform sentiment analysis. The process begins by constructing a list inspired by examining existing well-established sentiment word-banks (example-Linguistic Inquiry and Word Count(LIWC) and Affective Norms for English Words(ANEW)). Lexical features which are common to sentiment expressions and emoticons in microblogs are added to this. These features are combined with consideration for five basic standards that exemplify linguistic and grammatical conventions that people utilize when communicating or emphasizing sentiment intensity. This leaves us with a little more than 7,500 lexical highlights with approved valence scores that show both positive and negative sentiments and the sentiment intensity on a predefined scale. The sentiment analysis accuracy is greatly improved in several contexts by these inclusions and it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domains. VADER has a high accuracy and outperforms individual human raters at correctly classifying the sentiment of tweets into positive, neutral, or negative classes.

FINAL SENTIMENT ANALYSIS RESULTS

We measured the sentiments of Twitter users' across the different keywords that were used to pull data from the Twitter Search API.



	compound	keyword	neg	neu	pos
0	-1.0000	deletefacebook	0.117	0.791	0.092
1	1.0000	facebook	0.065	0.753	0.183
2	1.0000	mark zuckerberg	0.086	0.807	0.107
3	-1.0000	cambridg analytica	0.105	0.840	0.056
4	-0.7579	aleksandr kogan	0.102	0.804	0.093

Table 2. vaderSentiment results per keyword

Overall, we noticed deletefacebook tweets are most negative, followed by Cambridge, then Mark Zuckerberg, and then Facebook.

WEB BASED VISUALIZATIONS

The team designed interactive visualizations on tableau, a business intelligence and analytics software. This really helped us to see and understand the data and dig into some interesting insights. The team created a story of the analysis that was conducted, describing the common concerns people have regarding the Cambridge Analytica Scandal.
In sum, we planned to answer the following questions with visuals:

1. What is the Cambridge Analytica Facebook Scandal?
2. Did it go viral?
3. What are people retweeting with the respect to the data breach at Facebook?
4. How do twitter users feel about the scandal?
5. What topics are people most concerned about regarding the scandal?

A link to the Tableau Story has been provided in the appendix (Link 1).

IV. LIMITATIONS AND ETHICAL CONSIDERATIONS

GENERALIZATION, VALIDITY, AND RELIABILITY

The Twitter API structure sets restrictions on how analysts can build their examinations and what they can research through APIs as a methodological device. The APIs of social media companies appear to offer restricted access to their underlying databases without providing thorough documentation of how the API filters the database (Morstatter et al. 2013). Therefore, searching for relevant data using keywords like "Facebook" may not yield the desired data (in the context we are searching for in relation to the scandal only). Lack of understanding regarding how data is structured/filtered can influence how we chose our keywords.

Users may take on different actor roles in their engagement with Twitter- Many simply connect but never participate actively by posting tweets and some barely log onto the site. Subsequently, online users are not a homogeneous gathering, and the utilization of APIs to reap behavioral information from web-based social networking has critical impediments as far as representing the diverse client parts.

Researchers can gather information from APIs about user location, demographics, news feed, uploaded material, the social chart,etc. Clearly, the users who generate most of this content are not representative of the entire population. The data collected from APIs has an in-built bias toward those types of users that are the most “active” content contributors. This issue can be dealt with by sampling the entire Twitter population, however, this arrangement requires a huge server limit, or server-side access, and is not really a choice accessible to general researchers such as this research team. (Bechmann, 2014).

Another test of the utilization of APIs in producing rich understandings of online networking use is epistemological and needs to do with the nature and nature of behavioral information. While such information is exceptionally legitimate and dependable as far as detecting patterns, the behavioral information tells us little in regard to the actual intent or context associated with Twitter use. For instance, if a positive word is used ironically, the python library will simply categorize it as positive based on the word and not the content.

METHODS, ANALYSIS, AND RESULTS

The team collected data based on five keywords and then latter visualized results for only four keywords (excluding Aleksandr Kogan) as the tweets returned by the API were significantly less in contrast to the other keywords. The team did not see much value in including the outlier keyword in the analysis as it would significantly skew the entire data set, its visualizations, and analysis. However, if given more time or a more robust set of keywords, the team may have had more data which could have shown a different pattern where Aleksandr Kogan is not the only outlier.

Our team visualized the intensity of emotions and opinions by delving into a large corpora of text capturing shifts and trends in sentiment intensity. Among these results are however biased by tweets from the Facebook Twitter marketing representatives like Mark Zuckerberg and Facebook company itself. They would most likely be loyal and not share negative sentiments about their brand which may slightly skew the data towards a more positive direction. However, the team did not collect individual Twitter user usernames therefore it will be difficult to show at what level or percentage of the tweets are generated from the stakeholders themselves.

Also, as part of the initial phase of the research, the team performed basic data cleaning operations to remove capitalization and symbols/punctuations. This may have resulted in slightly different results than if the cleaning as not done before processing it through the VADER for sentiment analysis. This is due to the fact that VADER is attuned to text from social media which utilize different emoticons and capitalization to assign different positive, neutral, or negative values.

Reliability of information assembled through the API are difficult to keep up with and test as organizations constantly roll out changes to the API. For instance, if the companies can commercialize on specific sets of data, they will likely remove these data from the API as seen with Twitter's increasingly restrictive API access following the explosive user growth of the service (Gonzales-Bail ´ on et al. 2013). Therefore, even with a solid method or plan for data collection, it will be difficult to evaluate whether every pertinent data has been gathered or if there are blind spots, attributable to server downtime or control for this research project.

LEGAL AND ETHICS

A focal lawful and moral part of API research is the utilization of content that might be viewed as private by the users. The data may have an alternate status, as individual information is seen upon distinctively crosswise over national and social settings. The lack of clarity as to the sensitivity of personal data demands that researchers should be careful concerning the lawful methods of information accumulation and handling, yet additionally calls for ongoing reflection and transparency concerning the moral techniques and decisions that are a part of the project. Challenges concerning research ethics arise from the questions of informed consent and anonymization of data in the processes of collecting and analyzing data and publishing findings. Unfortunately, it is nearly impossible with our (lack of) resources and time to gain consent from all users to use their Tweets and/or be transparent, it was even more imperative to at least ensure the highest level of anonymization our team could provide. Thus, we did not aim to collect any usernames/identifications and decided the results of the research paper will not be published to further decrease the risk of revealing any identifiable data.

V. **CONCLUSION**

After the Cambridge Analytica data breach scandal at Facebook, it was interesting for the team to explore and evaluate 1) how concerned Twitter social media users are with data privacy in relation to the Facebook scandal and 2) what topics are most concerning to them? The team was interested in comparing public perception and/or understanding of the issues as well possibly use the research to help any entity(ies) that deal with personal information.

The research included accessing Tweets from the Twitter API, cleansing it (stemming, lemmatization, lowercase, and removing symbols) then using topic modeling and sentiment analysis to find the most common topics associated with the five keywords (deletefacebook, Aleksandr Kogan, Mark Zuckerberg, Facebook, and Cambridge Analytica).

The question is two part, the first part of the question was answered using sentiment analysis. The results of the sentiment analysis show a negative sentiment for #deleteFacebook, Cambridge Analytica and Aleksandr Kogan. However, interestingly enough, the overall sentiment for Facebook and Mark Zuckerberg remained positive. The results are quite odd as Facebook and Mark Zuckerberg are the headlining names associated with the scandal however, they receive more positive score. Could it be the short memories of the public? Or could brand power play a role? Further research is required to explore exactly why that may be.

The section part of the question, “what topics are most concerning to them” was answered by topic modeling. The team found that in general, legal actions are seems as a meta topic across each keyword categories. For Cambridge Analytica, the most important topic related to Cambridge Analytica, is the FBI investigation. As for #deletefacebook, it was digital censorship. The most social media censorship is also frequently tweeted relative to other topics. The Facebook keyword category has many different topics similar to Deletefacebook. However, roughly 1% of of tweets within this category contain the words ‘suspend 200’ which is most likely related to suspending 200 apps that had misused Facebook in accessing customer data.

Though the results of the research were interesting, it only skims the top. In the future, it would be worthwhile to make some adjustments to this research by:

1. Deploying sentiment analysis on raw data to see if there are any point differences between the raw data and the cleansed data.
2. Purchase a service to access data from Twitter APIs for more than 7 days from the request date to make a pre-scandal, post-scandal, and post-senate hearing comparison of the topics and sentiments of Twitter users.
3. Compare other data breach scandals to compare with the Facebook and Cambridge Analytica scandal to see if there are any overlapping concerns and sentiments that appear across similar scandals/issues.
4. Compare different levels of Tweets (macro, meso, and micro) as they may produce differences in sentiments.

VI. APPENDIX

FIGURES

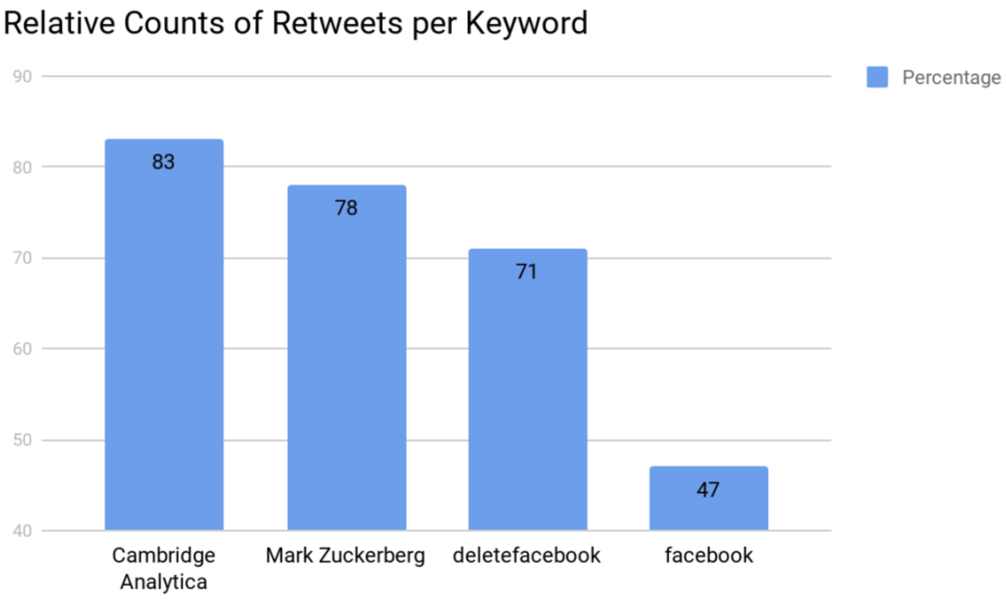


Figure 1. Retweet count per Keyword

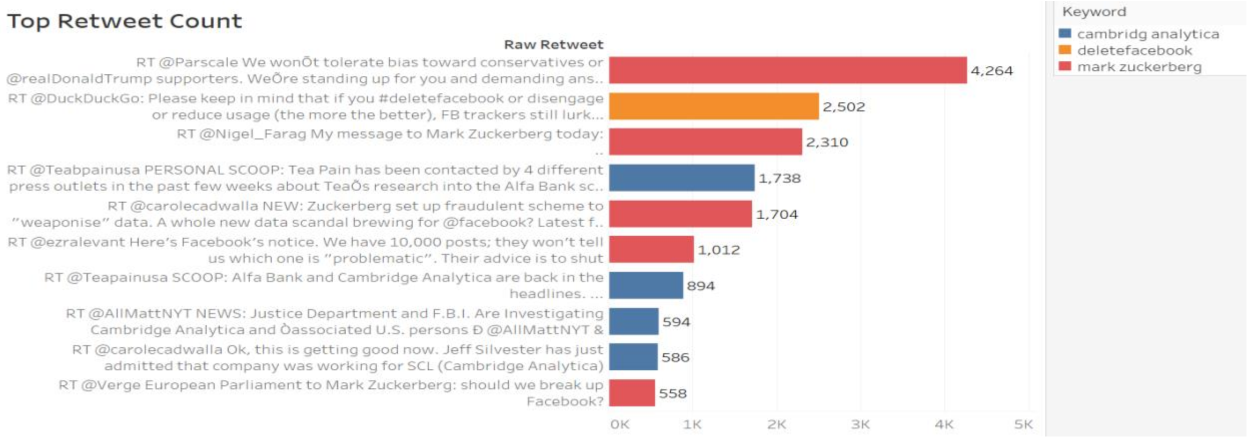


Figure 2. Top retweets from the tweet dataset

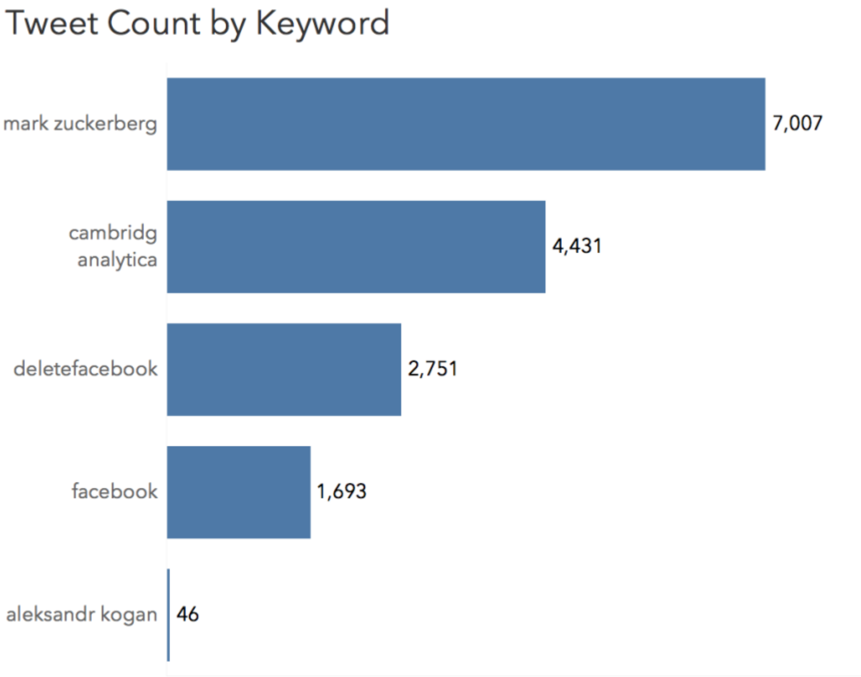


Figure 3. Number of tweets per keyword

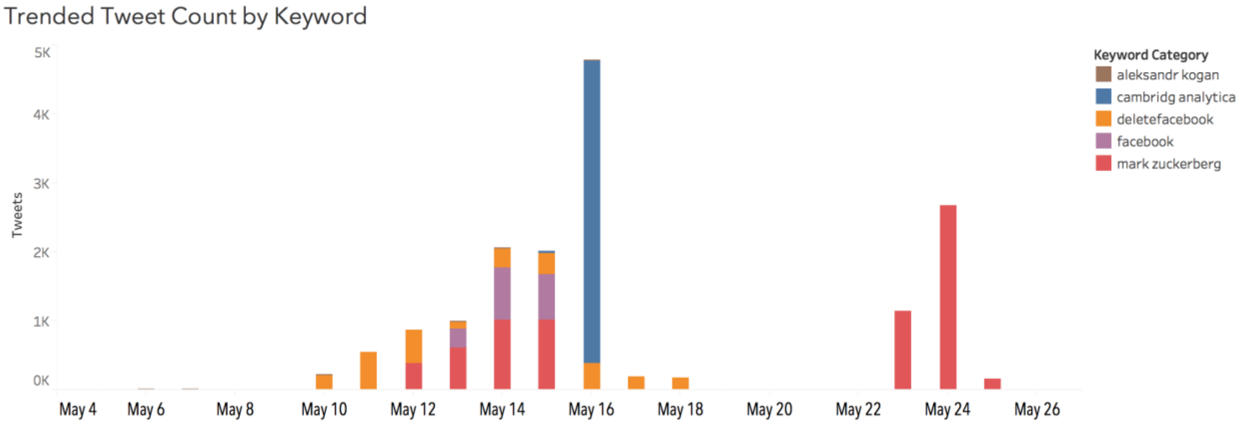


Figure 4. Trended tweet count per Keyword

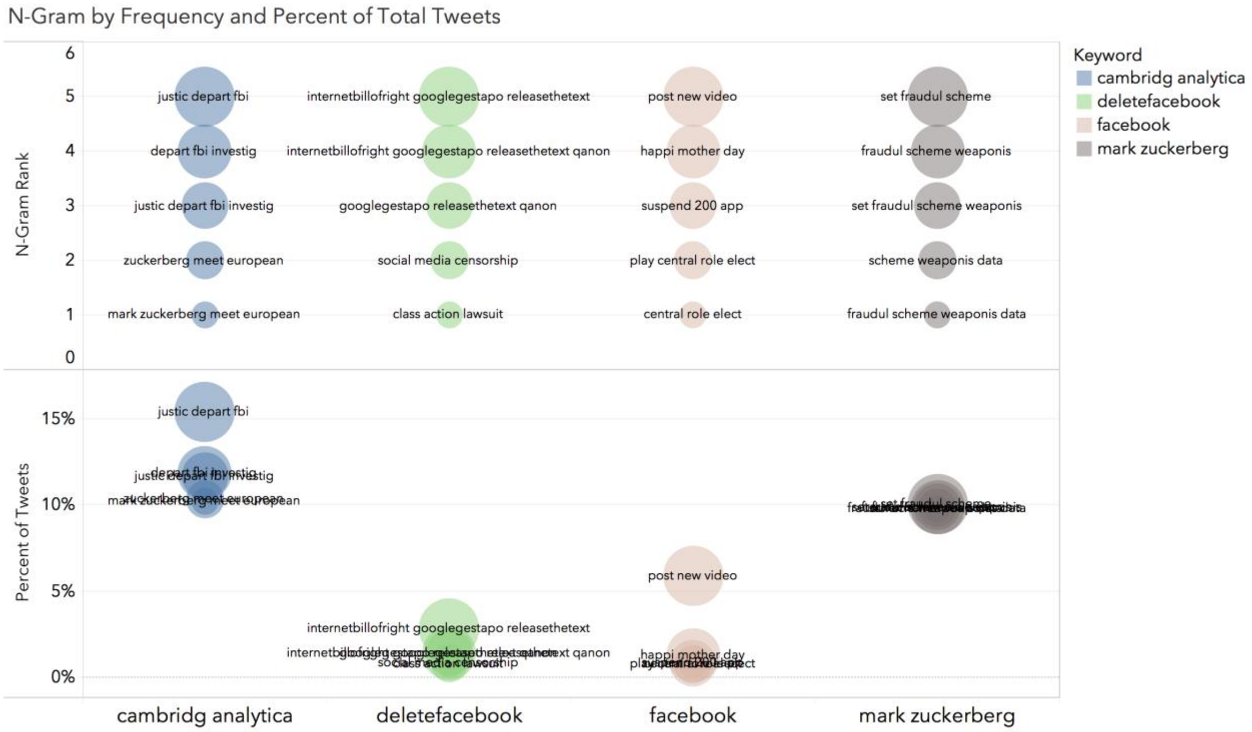


Figure 5. N-Gram by frequency and percent of total tweets

LINKS

Link 1
https://public.tableau.com/shared/FC6H8NWM4?:display_count=yes

Link 2
https://petitions.whitehouse.gov/petition/internet-bill-rights-2

PYTHON CODE

API REQUEST

```
import tweepy
import re
import pandas as pd
import datetime as dt

import nltk
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from nltk import ngrams
from nltk.sentiment import SentimentAnalyzer
from nltk.sentiment.util import *
from collections import Counter
import sys

stop = stopwords.words('english')
lemmatizer = WordNetLemmatizer()
stemmer=nltk.PorterStemmer()

# python twitter_pull.py deletefacebook


# API information
CONSUMER_KEY = 'lYnWAWNd8k2SYeRFHzOa2wFLM'
CONSUMER_SECRET = 'fsqoNOlT6FfnRKzhBbyeZozCfOmqgwkjhF7pOIei7LRDkCrr78'
ACCESS_KEY = '848393672248786945-7MtRKf8BYGKmw7h3mrf7VRqXIGPcyCW'
ACCESS_SECRET = 'h1RwCKVsaUEwiyn97y6bSlGyIVLkbdyCYCBGOpi59Aduu'

auth = tweepy.auth.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_KEY, ACCESS_SECRET)
api = tweepy.API(auth)
keyword_arg = sys.argv[1]


def get_tweets(keyword):

    '''use the api to get tweets and write them'''
    keyword = str(keyword)

    try:

        tweets = {keyword: {}}
        for tweet in tweepy.Cursor(api.search,
                                 q=keyword,
                                 rpp=100,
                                 result_type="recent",
                                 include_entities=True,
                                 lang="en").items():
```

```
          dc = {tweet.created_at: tweet.text}
          # append to tweet dictionary
          tweets[keyword].update(dc)

          # write to the tweet file
          with open('tweet.txt', 'a') as myfile:
              myfile.write(str(dc))

    except tweepy.TweepError:
        pass

    return tweets


def df_convert(ls):

    '''convert the dictionary into a dataframe'''
    tweet_list = []
    for key, val in tweets.items():
          for key2, val2 in val.items():
              tweet_list.append([key, key2, val2])

    df = pd.DataFrame(tweet_list)
    df.columns = ['keyword', 'date', 'tweets']
    df.date = pd.to_datetime(df.date, infer_datetime_format = True, format='%Y-%m-%d')

    return df


def df_clean(df):

    '''clean up text'''
    text_list = []
    for item in df.tweets:

          if item is not None:
              item = [word.lower() for word in item.split(" ")]
              item = [re.sub(r'[^\w]', '', word) for word in item]
              item = [re.sub(r'^https.+', '', word) for word in item]
              item = [stemmer.stem(word) for word in item]
```

```
            item = [lemmatizer.lemmatize(word, pos = 'v') for word in item if word
not in stop and word != '']
        else:
            item = []

        text_list.append(' '.join(item))

    df.tweets = text_list

    return df


if __name__ == "__main__":
    tweets = get_tweets(keyword_arg)
    df = df_convert(tweets)
    df_cleaned = df_clean(df)
    df_current = pd.read_csv("tweets_table.csv")
    df_master = pd.concat([df_current, df_cleaned])
    df_master.to_csv("tweets_table.csv", index = False)
```

ANALYSIS

```
# data pull and shaping
import tweepy
import re
import pandas as pd
import datetime as dt
import numpy as np
import ast

# models
import nltk
import twython
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from nltk import ngrams
from nltk.sentiment import SentimentAnalyzer
from nltk.sentiment.util import *
from collections import Counter
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import PCA, LatentDirichletAllocation, NMF
from sklearn.feature_extraction.text import TfidfVectorizer

# plot
import plotly.plotly as py
import plotly.graph_objs as go
import cufflinks as cf
import plotly.tools as tls
cf.set_config_file(offline=False, world_readable=True, theme='ggplot')
credentials = tls.get_credentials_file()
from IPython.display import Image

# data cleansing
stop = stopwords.words('english')
lemmatizer = WordNetLemmatizer()
stemmer = nltk.PorterStemmer()
analyser = SentimentIntensityAnalyzer()


df = pd.read_csv("datas/tweets_table.csv")
df = df.drop_duplicates()

#remove keywords from the tweets
In [159]:
clean_keyword = []
for keyword in df.keyword:
   keyword = re.sub(r'[^\w]', ' ', keyword)
   keyword = keyword.strip()
   keyword = stemmer.stem(keyword)
   keyword = lemmatizer.lemmatize(keyword, pos = 'v')
   if keyword == 'cambridge analytica': keyword = 'cambridg analytica'
   clean_keyword.append(keyword)

df['clean_keyword'] = pd.Series(clean_keyword).values
clean_tweets = [str(tweet).replace(keyword, '') for tweet, keyword in df[['tweets',
'clean_keyword']].values]
df['clean_tweets'] = pd.Series(clean_tweets).values
```

```
def get_scat(keyword, color):
   scat = go.Bar(x=df_tweet_totals.clean_date,
               y=df_tweet_totals[df_tweet_totals.clean_keyword == keyword].tweets,
               name = keyword,
               marker = dict(color = color),
               #mode = 'lines',
               opacity = 0.8)
   return(scat)

keyword_list = {'facebook': '#000000',
               'deletefacebook': '#17BECF',
               'mark zuckerberg': '#6EA9B5',
               'cambridg analytica': '#846EB5',
               'aleksandr kogan': '#B56E8D'}

scats = [get_scat(keyword, color) for keyword, color in keyword_list.items()]

layout = dict(
   title = "Total Unique Tweets by Keyword and Date",
   xaxis = dict(range = ['2018-05-11','2018-05-25'])
)

fig = dict(data = scats, layout = layout)
py.iplot(fig)



#calculate trigrams - use for topic modeling

def top_n_grams(keyword, min, max, col):
   """calculate the n-grams"""
   df = df_no_rt[df_no_rt.clean_keyword == keyword]
   word_vectorizer = CountVectorizer(ngram_range=(min, max), analyzer='word')
   sparse_matrix = word_vectorizer.fit_transform(df[col])
   frequencies = sum(sparse_matrix).toarray()[0]

   df_out = pd.DataFrame(frequencies,
            index=word_vectorizer.get_feature_names(),
            columns=['frequency']).reset_index().sort_values(by     =     ['frequency'],
ascending = False)
```

```
    df_out['percent_frequency'] =
df_out['frequency']/len(df_no_rt[df_no_rt.clean_keyword == keyword])
    df_out['keyword'] = keyword

    return(df_out.head(5))


In [356]:
camb_n_gram = top_n_grams('cambridg analytica', 3, 4, 'clean_tweets')
face_n_gram = top_n_grams('facebook', 3, 4, 'clean_tweets')
del_face_n_gram = top_n_grams('deletefacebook', 3, 4, 'clean_tweets')
mark_n_gram = top_n_grams('mark zuckerberg', 3, 4, 'clean_tweets')

n_grams = pd.concat([camb_n_gram, face_n_gram, del_face_n_gram, mark_n_gram])

#Use LDA to do further topic modeling
In [142]:
def generate_df(model, feature_names, n_top_words, keyword):
    topics = []
    for topic_idx, topic in enumerate(model.components_):
        topic = [" ".join([feature_names[i] for i in topic.argsort()[:-n_top_words -
1:-1]])]
        topic.append(topic_idx + 1)
        topics.append(topic)
    df = pd.DataFrame(topics)
    df.columns = ["topic", "topic_numer"]
    df["keyword"] = keyword
    return df

def lda_model(df, keyword, n_topic = 5, n_word = 5, max_features = 1000):
    '''model for latent dirichlect allocation'''
    lda = LatentDirichletAllocation(n_components=n_topic, max_iter=10,
learning_method='online', learning_offset=10., random_state=42)
    tfid = TfidfVectorizer(max_df=0.95, min_df=3, max_features = max_features)
    tfidf_text = tfid.fit_transform(df[df.clean_keyword == keyword].clean_tweets)
    lda_text = lda.fit(tfidf_text)
    tfidf_feature_names = tfid.get_feature_names()
    lda_df = generate_df(lda_text, tfidf_feature_names, n_word, keyword)
    return lda_df

lda_cam = lda_model(df_no_rt, 'cambridg analytica', 4, 4)
```

```
face_lda = lda_model(df_no_rt, 'facebook', 4, 4)
mark_lda = lda_model(df_no_rt, 'mark zuckerberg', 4, 4)
delface_lda = lda_model(df_no_rt, 'deletefacebook', 4, 4)

lda_df = pd.concat([lda_cam, face_lda, mark_lda, delface_lda])


#Use polarity scores for sentiment analysis

def return_sentiment_scores(keyword):
   sentence = ' '.join(df_no_rt[df_no_rt.clean_keyword ==
keyword].clean_tweets.values)
   snt = analyser.polarity_scores(sentence)
   snt['keyword'] = keyword
   return(snt)
TOPICS: generated based on scan of n-grams and LDA scores
Cambridge Analytica:
Justice department and the FBI investigation
Mark Zuckerberg meeting with European
New York Times coverage from Sheera Frenkel
Facebook
Post new video
Suspending 200 apps
WhatsApp playing a central role
Deletefacebook
Internetbillofrights
Googlegestapo
Social Media censorship
Class action lawsuit
mark zuckerberg
Set up fraudulent weapons scheme (guardian article)

#% counts for cambridge
len(df[(df.tweets.str.contains('fbi')) & (df.clean_keyword == 'cambridg
analytica')])/len(df[df.clean_keyword == 'cambridg analytica'])

len(df[(df.tweets.str.contains('zuckerberg')) & (df.clean_keyword == 'cambridg
analytica')])/len(df[df.clean_keyword == 'cambridg analytica'])

len(df[(df.tweets.str.contains('sheera')) & (df.clean_keyword == 'cambridg
analytica')])/len(df[df.clean_keyword == 'cambridg analytica'])
```

```
len(df[(df.tweets.str.contains('zuckerberg | fbi | sheera')) & (df.clean_keyword ==
'cambridg analytica')])/len(df[df.clean_keyword == 'cambridg analytica'])

#% counts for facebook

len(df[(df.tweets.str.contains('200')) & (df.clean_keyword ==
'facebook')])/len(df[df.clean_keyword == 'facebook'])

len(df[(df.tweets.str.contains('video')) & (df.clean_keyword ==
'facebook')])/len(df[df.clean_keyword == 'facebook'])

len(df[(df.tweets.str.contains('whatsapp')) & (df.clean_keyword ==
'facebook')])/len(df[df.clean_keyword == 'facebook'])

#% counts for deletefacebook

len(df[(df.tweets.str.contains('internetbillofrights')) & (df.clean_keyword ==
'deletefacebook')])/len(df[df.clean_keyword == 'deletefacebook'])

len(df[(df.tweets.str.contains('censorship')) & (df.clean_keyword ==
'deletefacebook')])/len(df[df.clean_keyword == 'deletefacebook'])

len(df[(df.tweets.str.contains('lawsuit')) & (df.clean_keyword ==
'deletefacebook')])/len(df[df.clean_keyword == 'deletefacebook'])

#% counts for mark zuckerberg

len(df[(df.tweets.str.contains('fraudul')) & (df.clean_keyword == 'mark
zuckerberg')])/len(df[df.clean_keyword == 'mark zuckerberg'])
```

VII. REFRENCES

Balakrishnan, Anita. (2017). 2 billion people now use Facebook each month, CEO Mark Zuckerberg says. Retrieved from https://www.cnbc.com/2017/06/27/how-many-users-does-facebook-have-2-billion-a-month-ceo-mark-zuckerberg-says.html

Baser, D. (2018). Hard Questions: What Data Does Facebook Collect When I'm Not Using Facebook, and Why? Retrieved from: https://newsroom.fb.com/news/2018/04/data-off-facebook/

Bechmann, S. L. (2014). Using APIs for Data Collection on Social Media. *The Information Society*.

NLTK Project, (2017).https://www.nltk.org/api/nltk.sentiment.html

Twitter Search API documentation (2018). https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

Facebook. (N.D). Investor Relations. Retrieved from https://investor.fb.com/resources/default.aspx

Morstatter, F., Pfeffer, J., and H. L. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with.

Meredith, S. (2018). Facebook-Cambridge Analytica: A timeline of the data hijacking scandal. Retrieved from https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

González-Bailón, S., and N. W.-H. (2013). Assessing the Bias in Communication Networks Sampled from Twitter.

Singer, N. (2018). What You Don't Know About How Facebook Uses Your Data. *The New York Times*.