Facebook's Data Breach Scandal and Fallout:

An Exploration of Twitter Users' Common Concerns and Sentiments

Data Analysis Plan

Sahil Aggarwal
Brian Hudnall
Hye Kim
Neha Palsokar

(IMT 547-Spring 2018)

I. Research Question

Our current research question is and will remain as: How concerned ar Twitter social media users with privacy in relation to the Facebook scandal and what topics are most concerning to them?

The research question still makes sense, however we are unable to pull data from the API (Before scandal, after scandal, and after the hearing) as the Twitter Search API provides data only up to seven days before the request date. Since we just started collecting the data from the API about a week ago, it will be difficult to pursue our original research question and research plan to the fullest extent. Therefore, there will no comparison between pre and post scandal as originally planned. However, we are still able to deliver what the sentiments are post-scandal in general.

II. Data Collection

We created a Python script that leverages the Tweepy library to connect to the Twitter search API. The keywords that were searched for the research are 'deletefacebook', 'Mark zuckerberg', 'cambridge analytica', and 'Alexandr Kogan', as these resonate closely to the scandal The script accepted a search term as a parameter passed to it from the command line on launch and also did early stage cleaning, like removing symbols, and lemmatizing the words. A powershell script was used to call the script and pass the keyword value to the script. We used the Windows scheduler on a remote server to automate 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 file.

As far as data collection, so far it has been successful thus we will not be making any changes regarding our data collection method.

III. Data Cleaning

As mentioned in the above section, the python script removed symbols, and applied stemming and lemmatization to the tweets/words that were collected. The script also converted the words into lowercase. The second phase of the data cleaning process involved removing duplicate tweets from the dataset.

One issue we ran into was that the hashtag was not removed for #deletefacebook. So as a part of the data cleaning process, we will have to remove all punctuations and symbols an additional time. Also, some words were stemmed/lemmatized to the point

where it was unrecognizable. We will attempt to resolve this issue as we pursue this research further.

IV. Summary Statistics

FIGURE 1: 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

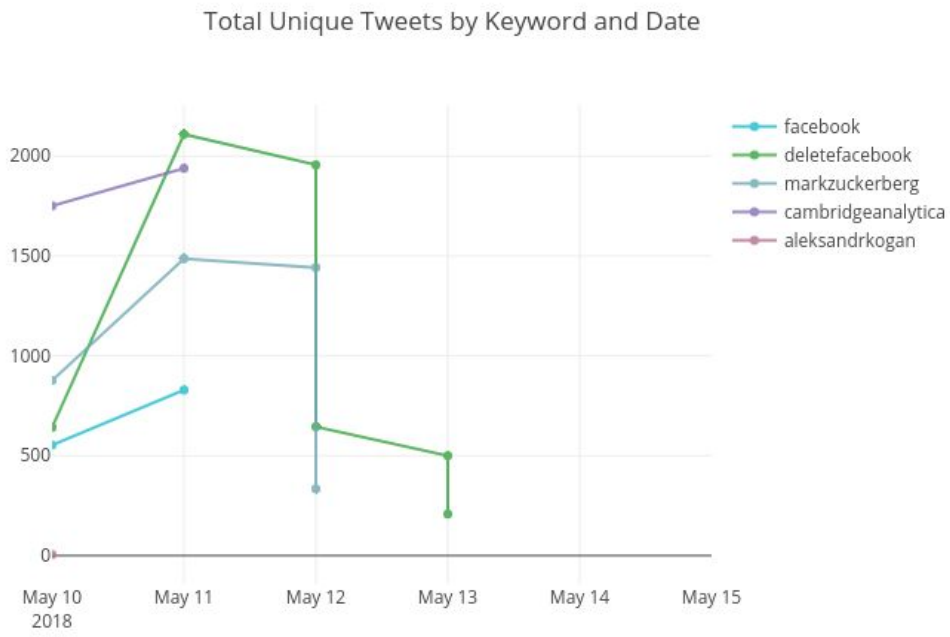
FIGURE 2: For the entire data set, the top 100 words appear the following number of times:[1]

Tweet Word	No. of Appearances	Tweet Word	No. of Appearances	Tweet Word	No. of Appearances	Tweet Word	No. of Appearances
data	1020	today	291	testifi	206	media	144
investig	940	know	291	give	205	ad	144
fbi	773	may	279	year	204	account	143
wa	637	birthday	277	watch	203	post	140
us	636	one	275	come	201	war	140
whistleblow	603	live	270	question	200	voter	140
justic	582	want	263	social	194	good	135

[1] Excludes the keywords and 'rt' which stands for retweet. This is to gain a sense of what the important topics are regarding this data set. Also ambiguous words such thi, e, and ha were also removed as they do not add value to the analysis and cannot be deciphered to add value.

new	572	user	255	person	185	find	135
say	509	hear	254	meet	184	polit	134
depart	481	privaci	253	video	184	nameshameblockzionist	134
use	461	christoph	247	wherepond	180	thank	133
amp	452	ceo	245	see	179	firm	133
like	444	dont	244	world	178	love	133
via	405	befor	243	uk	177	youtub	132
get	398	think	243	tell	177	would	130
internetbillofright	392	compani	241	campaign	177	200	129
time	379	take	237	need	173	vote	127
senat	359	happi	225	russia	171	elect	127
trump	340	scandal	225	cambridgeanalytica	165	thing	125
peopl	332	report	225	bannon	164	fake	124
googlegestapo	324	share	224	fb	162	anoth	124
day	301	work	214	york	154	becaus	123
make	301	look	213	call	147	committe	122
go	300	app	212	twitter	145	russian	121
news	295	million	211	delet	145	show	121

FIGURE 3: Frequency of Keyword by date



V. Analysis Plan

The goal is to look at high level topics that describe user concerns as well as compare sentiment analysis for each keyword and (also) see overall sentiment for each topic. So, with the above summary statistics and our data set, we aim to tell a story that talks about how and to what degree different stakeholders/players were affected (aka put under the spotlight) after the data breach at Facebook

VI. 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.[2] 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.

[2] This plan is the ideal, however it may require some intimate information from the companies themselves, which may or may not be available to researchers due to the sensitive and confidential nature of the data and situation.