Twitter Data Analysis

David Wilson & Grant Haataja

May 3, 2019

Contents

Introduction

A growing concern of social media is the amount of profanity and inappropriate content appearing to the public. This paper will address this issue by outlining data analysis for profanity on Twitter.

Goals of Analysis

The goals for this analysis were to test the occurrence of profanity from a sample of tweets, to determine whether Twitter was more than 50% profane or not. But more interestingly, the goal was to find the proportion of tweets that contain profanity.

Methodology

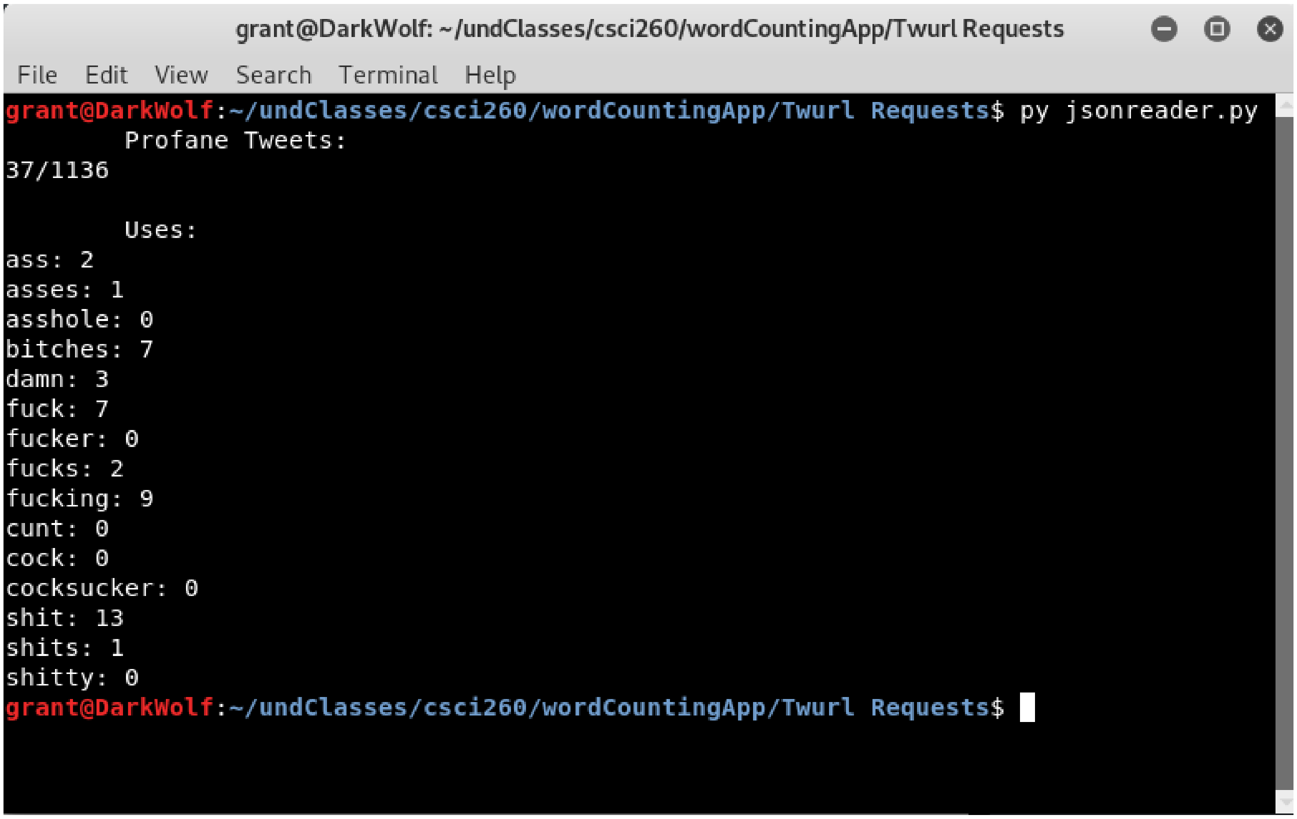
In order to collect the tweet data for the analysis, this study chose to use the Twitter API [2]. This required a twitter developer account and security keys to access the API using HTTP requests. A twitter developer was created, and the project was registered under Twitter to get the access keys.

After acquiring the access keys, the project utilized the command line program ”Twurl” to send Twitter search requests to the API. Results were returned in a json file format. Unfortunately, because the project didn’t acquire the premium form of the Twitter API, each search request was limited to 15 results.

Because of this, the project was instead composed of a hundred separate search requests, using Wikipedia’s list of the most common words in English [3]. Then, a python script was created to parse the json files retrieved, and search the tweets for the profane words. Finally, the collected data was used to create tables and support our conclusion.

Code to parse .json files

```
import json

jsonNum = 100
tweets = []
profanity = []

# Load profanity list
fprofanity = open('Profanity.txt', 'r')
for text in fprofanity:
  profanity += [text.strip().lower()]

# Load tweets
for i in range(1, jsonNum+1):
  temp = open('word' + str(i) + '.json', 'r')
  temp = json.load(temp)
  for tweet in temp['statuses']:
    tweets += [tweet['text'].lower()]

profaneTweets = []
uses = {}
for prof in profanity:
  uses.update({prof:0})

for tweet in tweets:
  for word in profanity:
    if ' ' + word + ' ' in tweet:
      uses.update({word:uses[word]+1})
      if tweet not in profaneTweets:
        profaneTweets += [tweet]

print('\tProfane Tweets:')
print(str(len(profaneTweets)) + '/' + str(len(tweets)))
print('\n\tUses:')
for use in uses:
  print(use + ': ' + str(uses[use]))
```


grant@DarkWolf: ~/undClasses/csci260/wordCountingApp/Twurl Requests
File
Edit
View
Search
Terminal
Help

```
grant@DarkWolf:~/undClasses/csci260/wordCountingApp/Twurl Requests$ py jsonreader.py
        Profane Tweets:
37/1136

        Uses:
ass: 2
asses: 1
asshole: 0
bitches: 7
damn: 3
fuck: 7
fucker: 0
fucks: 2
fucking: 9
cunt: 0
cock: 0
cocksucker: 0
shit: 13
shits: 1
shitty: 0
grant@DarkWolf:~/undClasses/csci260/wordCountingApp/Twurl Requests$
```

The above picture shows the output of the above code when ran on the 100 json files that were created from the queries.

Tweet Examples

In this section, several of the tweets that were searched are shown. The program did not output the entirety of some of the longer tweets. Note that many of them have multiple occurrences of different profane words, so some of them may be counted twice.

——————

rt @quackityhq: fuck koalas why can those little shits sleep 22 hours a day and be seen as cute but when i do it im seen as a lazy piece of...

——————-

rt @braindumptweets: trying to figure out how sex works but it doesn't seem plausible, how am i supposed to fit my entire ass in a woman's...

——————-

adding arya to my muted words because you bitches can't be trusted.

——————-

rt @iamvee_: i haven't been this happy in about 3 fucking years boy

——————-

he appears healthy but he seems just to be very damn tired. i'll keep an eye on him and call an animal sanctuary later

——————-

if i could get my tubes tied today i fucking would. i'd sell every single one of my eggs. fuck dem kids.

——————-

waw i already miss them ;((my spoiled ass need to stop i keep missing them even if they feed me so much

——————

Code to Create Graphs

Shown below is the Python code used to generate the pie chart used to display the data results.

```python
from matplotlib import pyplot
import numpy

#define the pie chart
labels = ['Ass', 'Asses', 'Bitches', 'Damn', 'Fuck', 'Fucks', 'Fucking', 'Shit', 'Shits']
sizes = numpy.array([2, 1, 7, 3, 7, 2, 9, 13, 1])
colors = ['yellow', 'orange', 'aquamarine', 'purple', 'dodgerblue', 'gray', 'red', 'green', 'mediumvioletred']

#do the plot
```

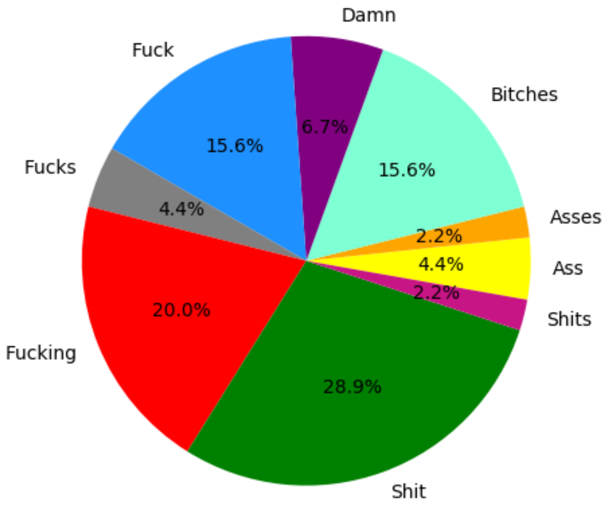
```
pyplot.pie(sizes, labels=labels, colors=colors,
autopct='%1.1f%%', startangle=350)

pyplot.axis('equal')
pyplot.show()
```

The following is the code to generate the bar chart containing the actual number of tweets with the queried profanities.

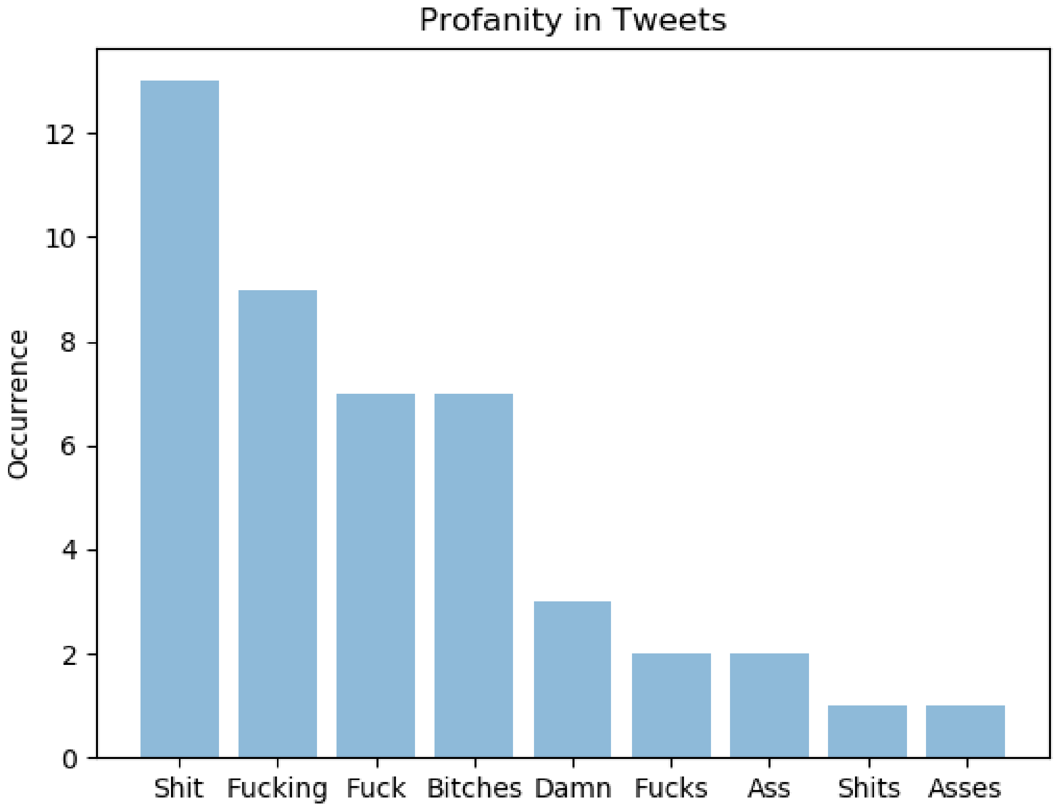
```
from matplotlib import pyplot
import numpy

#define the bar chart
labels = ['Shit', 'Fucking', 'Fuck', 'Bitches', 'Damn', 'Fucks', 'Ass', 'Shits', 'Asses']
y_pos = numpy.arange(len(labels))
sizes = [13, 9, 7, 7, 3, 2, 2, 1, 1]

#do the plot
pyplot.bar(y_pos, sizes, align='center', alpha=0.5)
pyplot.xticks(y_pos, labels)
pyplot.ylabel('Occurrence')
pyplot.title('Profanity in Tweets')
pyplot.show()
```

The Data

Below are the recorded percentages of various profanity from the 1136 tweets in the sample size. A total of 37 tweets were found to be profane.



Shown in the next chart are the actual numbers of occurrences of the various profanity in the 37 profane tweets found from the sample.



Conclusions

The data collected show that Twitter is much less profane than 50%. In fact, only about 3.3% of tweets collected were found to be profane. However, those that did contain profanity often had multiple occurrences of it. The question of the safety of a platform such as Twitter must ultimately be answered by the individual.

Further Research

This small data study could be expanded into a much larger research project quite easily by collecting profanity proportions from other social media platforms, and analyzing possible correlations between variables. It would be interesting indeed to see which of the major social media platforms had the most occurrences of profanity, which had the largest total number of profane words, or how the amount of profanity of a platform was correlated to users' opinion of the platform.

References

[1] Twitter. (2019). It's what's happening. Retrieved from https://twitter.com

[2] Twitter. (2019). Developer Documentation. Retrieved from https://developer.twitter.com/en/docs.html

[3] Wikimedia Foundation. (2018, December 12). Most common words in English. Retrieved from https://en.wikipedia.org/wiki/Most_common_words_in_English