Frances Gail Godinez
BUID: U31156004
CS591
HW 6

Twitter Popularity and Harassment (Data Collection)

After the recent harassment of several prominent women in the gaming community (infamously known as #GamerGate) and Twitter's inability to appropriately deal with said harassment, I wanted to try exploring if there was correlation between "Twitter popularity" and harassment.

The following metrics will be used for measuring a twitter user's "popularity":

- Number of followers
- Number of tweets made
- Number of favourites per tweet
- List membership
- Friends count

“Harassment” will be measured by using sentiment analysis by traversing through said user's mentions and measuring the average polarity ranging from -1.0 to 1.0, where more 'positive' tweets have higher polarities, while 'negative' tweets will have lower polarity.

For this assignment, I collected data from the following hashtags:

- Users using the hashtag #GamerGate
- Users using the hashtag #StopGamerGate

And the following hashtag used as control:

- Users using the #ThanksObama hashtag
    - I wanted to explore this tag since it's mostly used in a sarcastic way though it would be interesting to explore in terms of polarity since sarcasm detection is still a problem that NLP doesn't have an elegant solution for.
    - Essentially, I wanted a middle ground between negative tweets and positive tweets.

For each collection of users, I grabbed the aforementioned metrics (followers, tweets made, etc) as well as the average polarity of their mentions.

The final dictionaries for each hashtag can be found in the attached text files

** A problem I noticed while collecting data was that the “search” method in twitter’s API can only grab 100 tweets at a time, limiting the amount of users I could collect per call. I tried running it 10 times (to get 1000 tweets), and it only improved by just 10 more users. I tried

gradually increasing it by was eventually given a timeout error at 5000 tweets (even after adding sleep after each call). **

** For the actual project, I'll run it multiple times to gather more users but for the assignment I'll be working with how many users I can get per hashtag **

After gathering the data, I gathered these preliminary stats regarding mentions polarity:

Hashtag	Number of Unique Users (per 100 tweets)	Average Mentions Polarity
#GamerGate	65	0.036349059358269556
#StopGamerGate	90	0.06762642536646489
#ThanksObama	58	0.07386042403986838

- Interestingly enough, there appears to be more unique users per 100 tweets for #StopGamerGate
    - This is most likely due to the general #GamerGate "movement" is dying out since it started in August 2014.
    - This is also most likely due to more awareness of harassment on twitter and a larger movement towards preventing harassment and promoting awareness on internet harassment.
- As per my theory, the polarity of the #StopGamerGate was higher (more positive) than the average polarity of the #GamerGate hashtag with the former having a polarity of almost twice the latter.
    - This might also be skewed due to the fact that more unique users per 100 tweet were using the #StopGamerGate tag than #GamerGate
- Interestingly enough, the average polarity of the #ThanksObama tag appears to be the most positive out of all of them.
    - As mentioned earlier, this is most likely due to the hashtag being used more sarcastically than the former two.
    - This could also be due to the fact that it has the least amount of unique users

Next, I gathered the average polarity for their timeline - which are the tweets they make on their own twitter.

Hashtag	Average Timeline Polarity
#GamerGate	0.00549019303346
#StopGamerGate	0.0365904498884
#ThanksObama	0.053186128929

As expected, #GamerGate has the lowest timeline polarity when compared to #StopGamerGate and #ThanksObama.

Another issue I wanted to explore was the use of “throwaway” accounts on twitter that people might use to “anonymous” harass users without leaving a trace of their identity

To explore this further, I take a look at the average amount of followers, friends, and tweets per each hashtag with the belief that the more followers, friends, and tweets a user has the more likely it is that it is a “real” account as opposed to a “throwaway” one.

Hashtag	Average # of Followers	Average # of Friends	Average # of Tweets
#GamerGate	733.738461538	834.584615385	16625.6153846
#StopGamerGate	1089.67777778	725.044444444	24640.8333333
#ThanksObama	23533.3448276	7908.60344828	31758.637931

On average, users using the #GamerGate hashtag appear to have the lowest amount of followers and tweets, which is a possible indicator or a “throwaway” account. Interestingly enough, #StopGamerGate falls behind #GamerGate users in terms of average amount of friends.

As mentioned earlier, I wanted to measure if there is a relationship between twitter “popularity” (measured by overall twitter presence, followers, etc) vs. “harassment” aka polarity of tweets. To get more insight on this could be related I calculated the correlation coefficients for the following metrics

- Followers vs. Mentions Polarity
- Followers vs. Friends
- Mentions Polarity vs. Timeline Polarity
- Mentions Polarity vs. Status Count

Correlation Coefficients

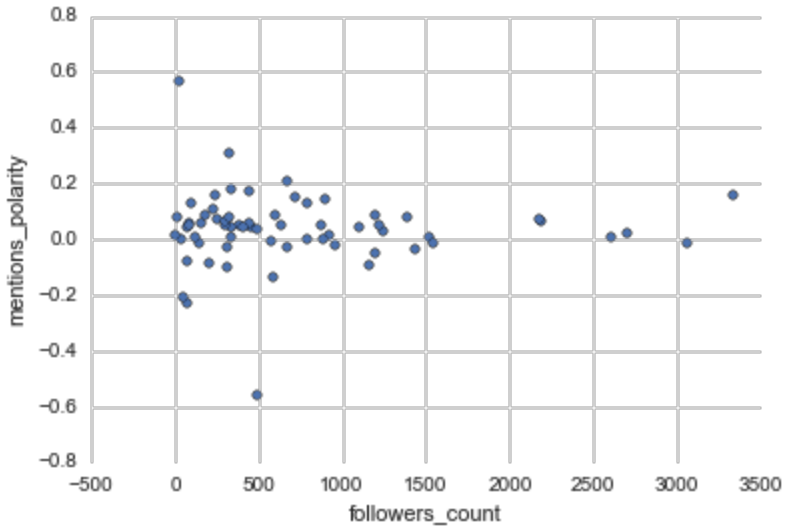
Hashtag	Followers vs. Mentions Polarity	Followers vs. Friends	Mentions Polarity vs. Timeline Polarity	Mentions Polarity vs. Status Count
#GamerGate	-0.0022757414855	0.743649087383	-0.0802402473903	0.0452157667676
#StopGamerGate	-0.0159337434801	0.292269495898	0.293868892763	0.325802114866

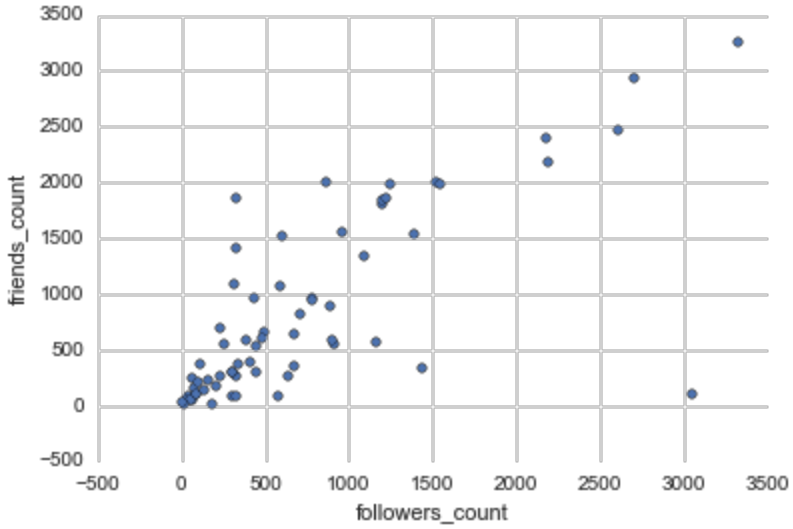
#ThanksObama	-0.0361130635362	0.960241023843	0.335783133598	0.0974486694264

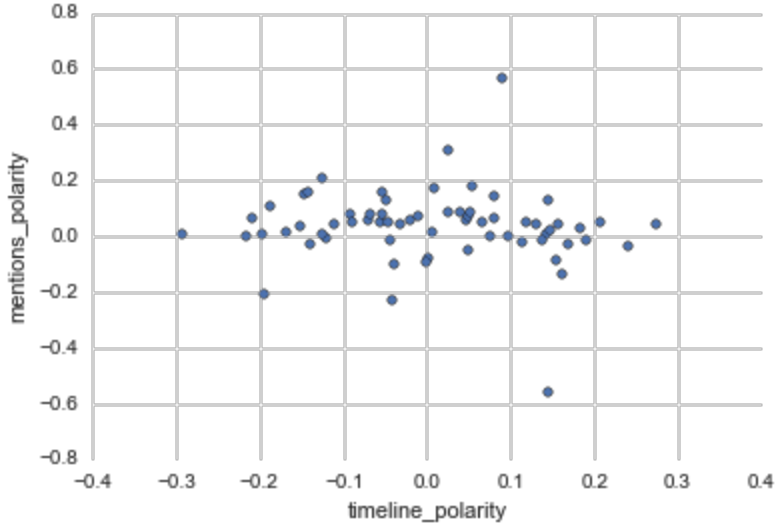
- Interestingly enough, the number of followers a user has does not seem to have a great effect in the number of followers vs. mentions polarity. I assumed that there would be a higher correlation under the belief that more followers might result in more negative tweets.
- As expected, there is a high positive correlation between number of followers and number of friends. However, I was expecting there to be a lower correlation for #GamerGate users under the assumption that most users using that hashtag were "throwaway" accounts. Though, it appears that #StopGamerGate follows this trend.
- For Mentions Polarity vs. Timeline Polarity I wanted to see if there was a correlation between having a "negative" timeline would result in having "negative" mentions as well. It appears that for #GamerGate, this doesn't seem to be the case. Though more highly correlation than #GamerGate, there is little correlation for this for the other two hashtags as well.
- Lastly, I wanted to see if there was a correlation between mentions polarity and status count under the belief that more statuses might result in lower mentions polarity. As expected, this was more highly correlated for #StopGamerGate than any of the other hashtags, although not strongly.

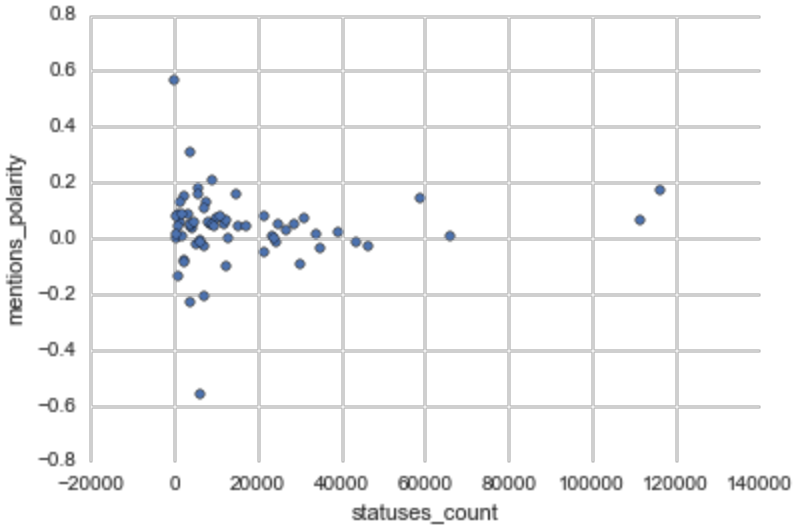
Next, I graphed these metrics against each other to get a visual on how they might be related before I run linear regression on each data frame.

#GamerGate Scatter Plots

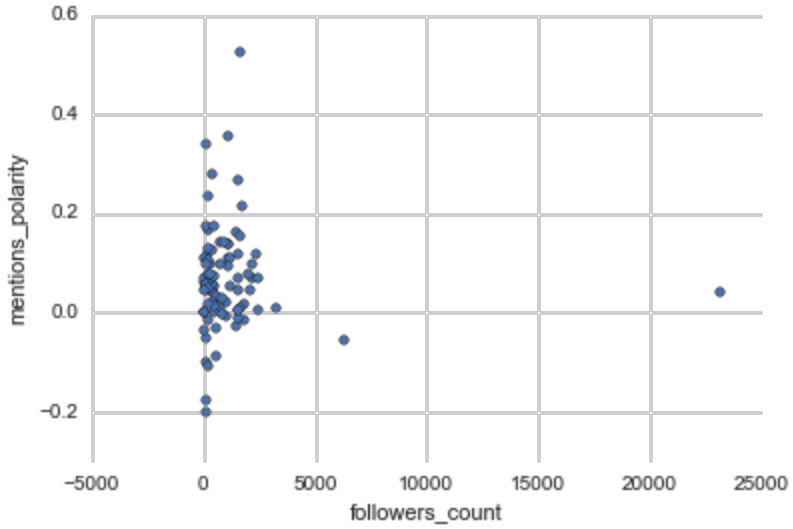


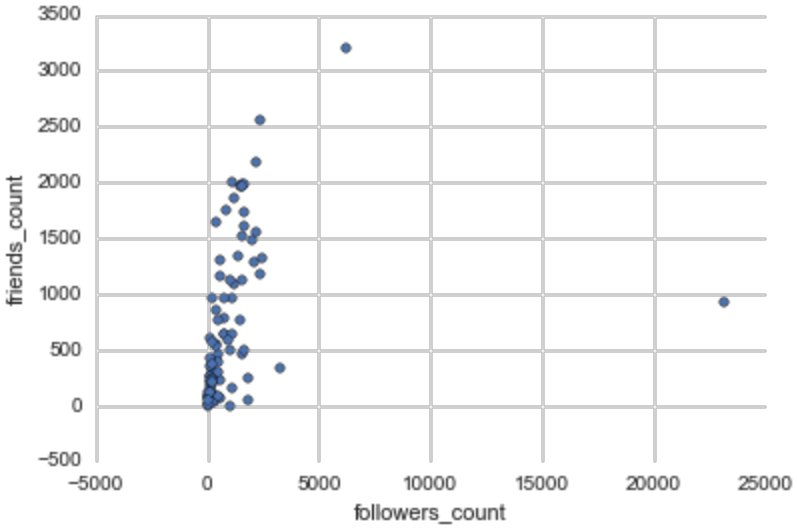

friends_count
followers_count
3500
3000
2500
2000
1500
1000
500
0
−500
−500
0
500
1000
1500
2000
2500
3000
3500

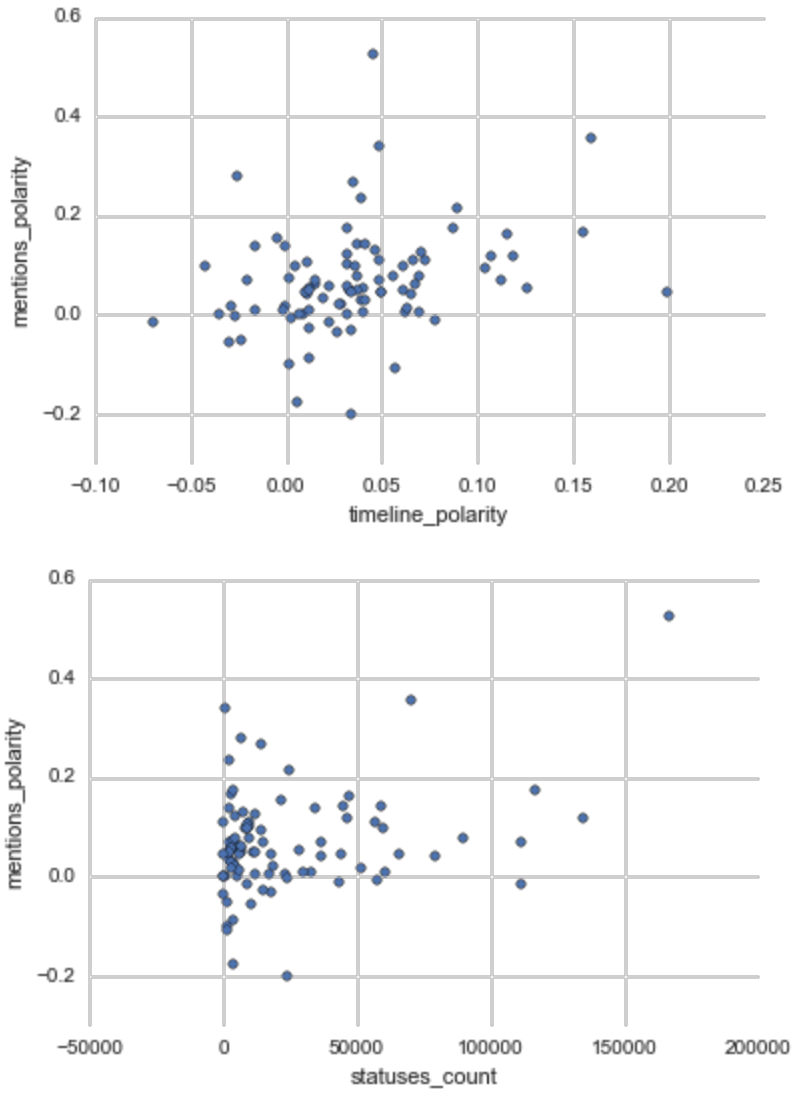

mentions_polarity
timeline_polarity
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4


mentions_polarity
statuses_count
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−20000
0
20000
40000
60000
80000
100000
120000
140000

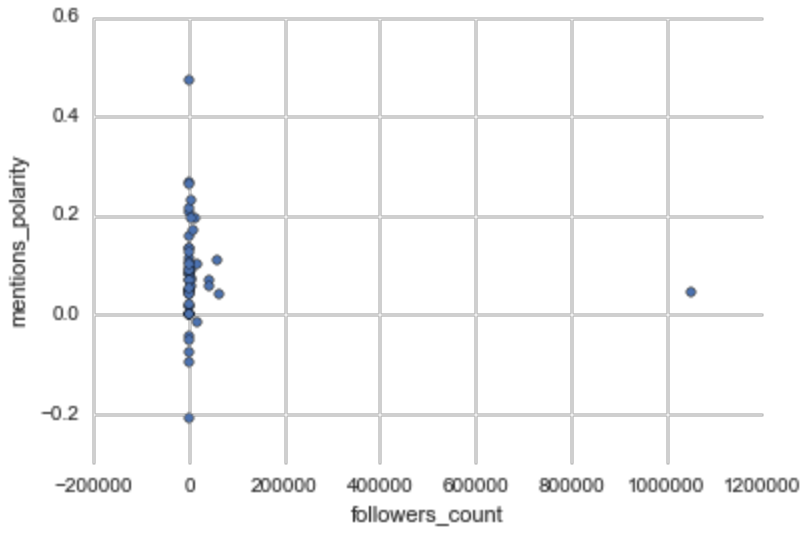
#StopGamerGate Scatter Plots

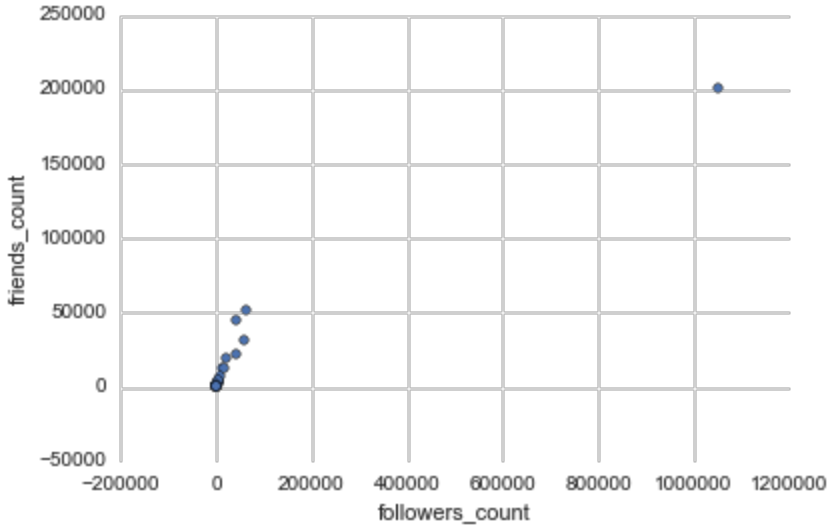


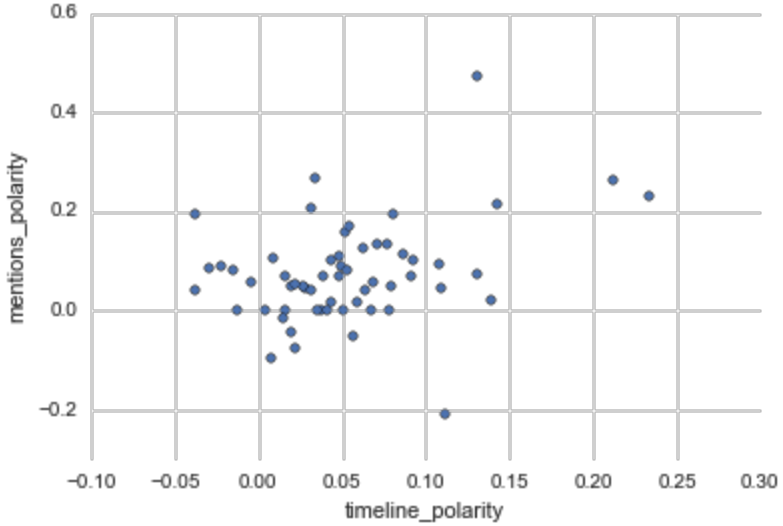


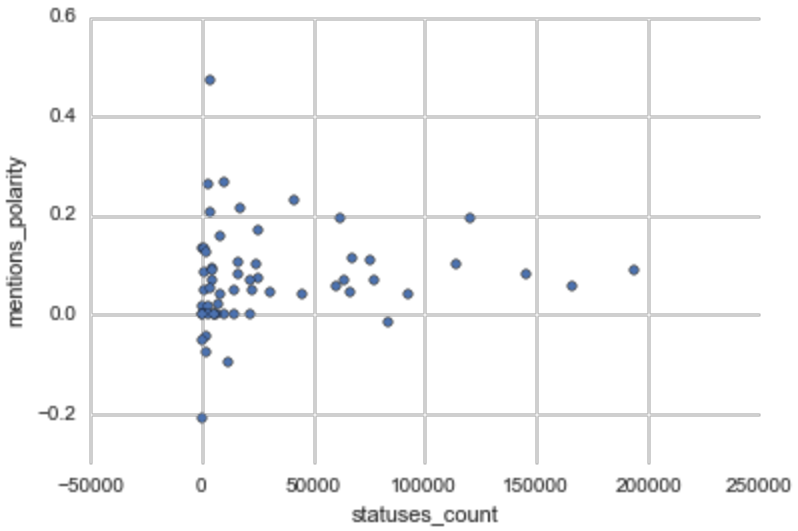


#ThanksObama Scatter Plots


0.6
0.4
0.2
0.0
−0.2
mentions_polarity
−200000
0
200000
400000
600000
800000
1000000
1200000
followers_count


250000
200000
150000
100000
50000
0
−50000
friends_count
−200000
0
200000
400000
600000
800000
1000000
1200000
followers_count


0.6
0.4
0.2
0.0
−0.2
mentions_polarity
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
timeline_polarity



For the rest of the project, I want to run linear regression against all these features to get a better understanding of how strongly related these metrics are to mentions polarity.

Finally, I'm going to run logistic regression on each individual data frame, then again on a combined data frame (combining #GamerGate, #StopGamerGate, and #ThanksObama) to predict mentions polarity based on all the metrics that contribute to "popularity".