Twitter Sentiment Analysis: Georgia State Election

Anant Tripathi
University of Georgia
at41132@uga.edu

Priyank Malviya
University of Georgia
pm05969@uga.edu

I. Introduction:

Sentiment analysis is the detection of attitudes "enduring, affectively colored beliefs, dispositions towards objects or persons". Since long time ago, human beings have used the media to express their needs, preferences and emotions. Internet has been one of the most used media to make the communication possible and find information of interest. The growth of social networks like: Twitter, Facebook, LinkedIn, and others, has generated a large amount of information about the preferences and behavior of the users. Most of the data that is constantly generated in the social networks could contain valuable information like point of view and tendencies of the users.

Due to exponential growth of social network, sentimental analysis has been applied to analyze public opinion. In this project we will be analyzing sentiments of the people on the Georgia state election 2018. We will make use of the twitter data to do our analysis. Using the data collected we will be analyzing the sentiments by dividing the data set into two groups of Stacey Abrams, which would be having the tweets addressing Stacey Abrams, and Brian Kemp, which would be having tweets addressing Brian Kemp.

II. Background, Motivation and Related work:

The motivation behind taking this project is to work with huge amount of data. We started our project with collection of tweets from twitter. For collection of data from twitter we made a developer account on twitter. We were provided with Access Token, Access Token secret, Consumer key, Consumer Secret. Then we used tweepy API to download the tweets from the twitter using the key provided to us. To get the tweets related to the elections we used hashtags which were relevant to the elections recently held, such as #iamwithstacy #teamrepublican etc. Working with huge amount of data is one of the major part of this project. For pre-processing of the collected data, we removed emoji, regular expression and hyperlink. To get the hidden information from the meaningful data obtained after removal.

III. Technical Details:

- **Data collection:**
  To collect the data form twitter we made developer account on twitter using the access token and is provide to us we downloaded the tweets form twitter using Tweepy. We used scripting languages like Python for developing the essential server-side codes. We integrated our code with tweepy to download the tweets from twitter

- **Data Processing /filtering**:
  We used R for data processing. First and foremost, we got rid of the undesired fields from the dataset. We removed emojis, hyperlinks from the tweets for cleaning the data. The most important point of this project was to remove the biased users as well from the dataset. For this we did two level of processing. In the first level we simply looked at the number of

tweets per user and if the number of tweets by any user was outside the 1.5*IQR range, we deemed that user as being biased. Basic reasoning behind this is that any user who is unbiased would not make such huge number of tweets in such a short time span. For the second level of removal of biased users we calculated the sentiment value for each user based on the tweets he/she made. Then followed the same procedure of removing all the users whose tweet score was not in the 1.5*IQR range.

- **Sentiment Analysis :** Once we had biased users removed from our dataset. We calculated the sentiment score of each user using library syuzhet in R. After getting the sentiment score for each user we found out the maximum score, the minimum score and the mean score for both brain kemp's and stacey abrams' tweets.

IV. Performance Evaluation:

- Figure 1 shows the tweet per person in favor of Stacy Abrams. We did this to remove the biased users. The x-axis of the graph represents the tweet count and the y-axis represents the number of users corresponding to each tweet count. From the graph we can see most of the users tweeted only once but there were few users that tweeted more. We found that there were two users who made more than 1000 tweets. We removed those users and their tweets as biased tweets which would have affect our result. In Figure2 we did the same thing for the tweets pertaining to Brian Kemp. Using histogram we observed the same pattern again and removed the users having more than 1000 tweets.

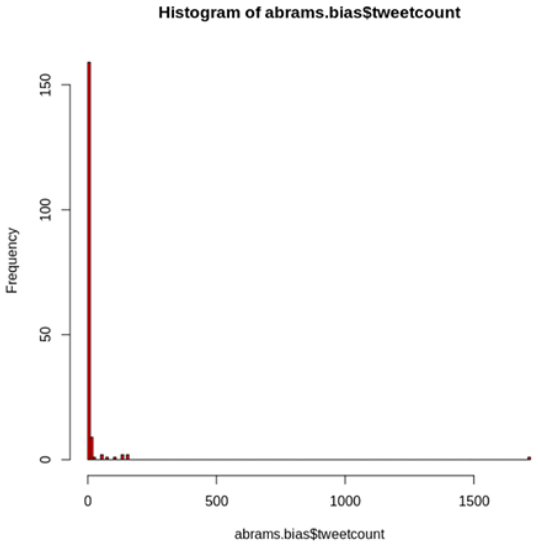


Figure 1: Histogram of Stacy Abrams tweets

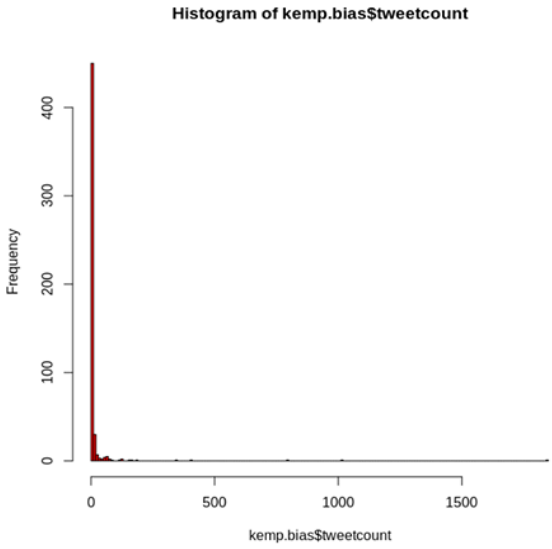


Figure 2: Histogram of Brian Kemp tweets

- Figure 3 and 4 is a density graph showing the number of tweets per user. The y axis shows the number of user and x axis shows the number of tweets. We did both of this to plot the see the distribution of number of tweets of tweet per user so that we can separate biased and unbiased user.

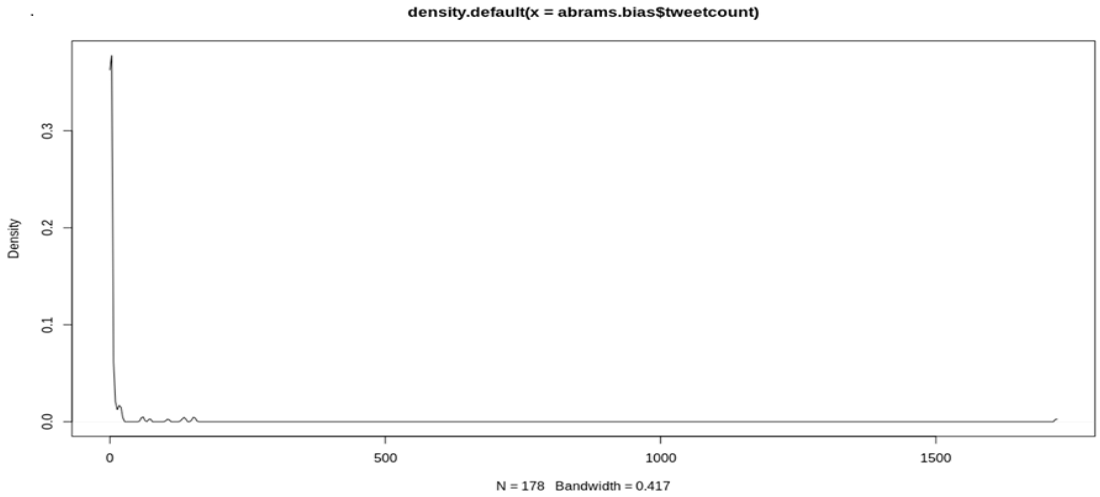


Figure3: Density graph of Stacy Abrams tweets

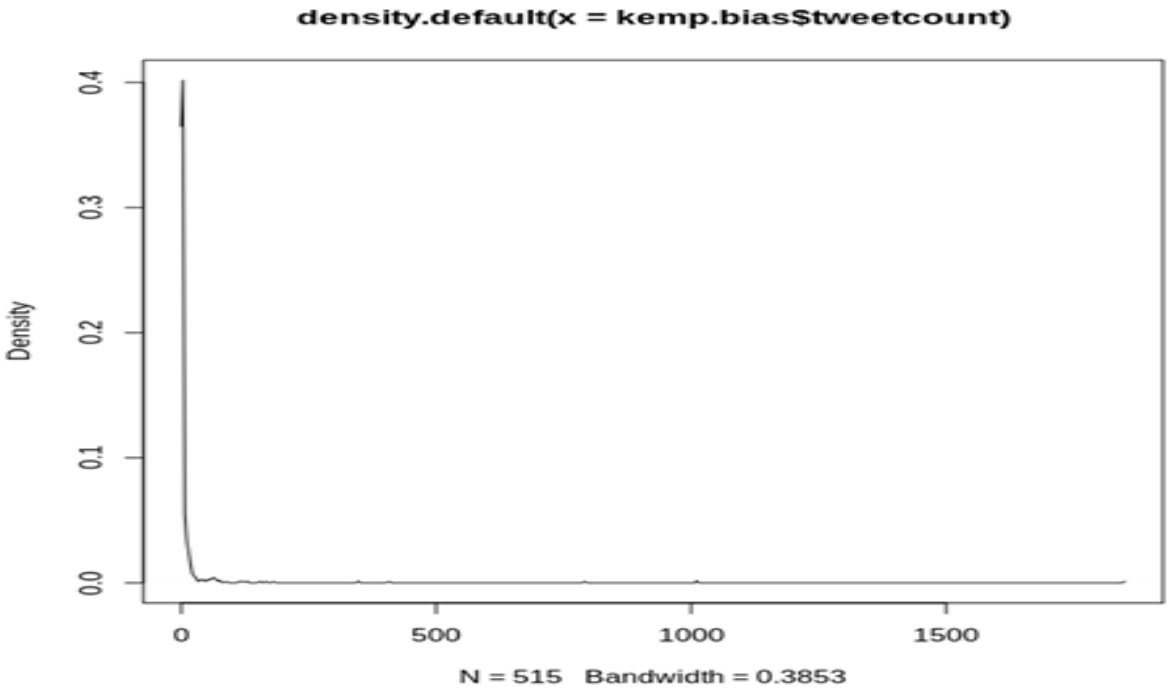
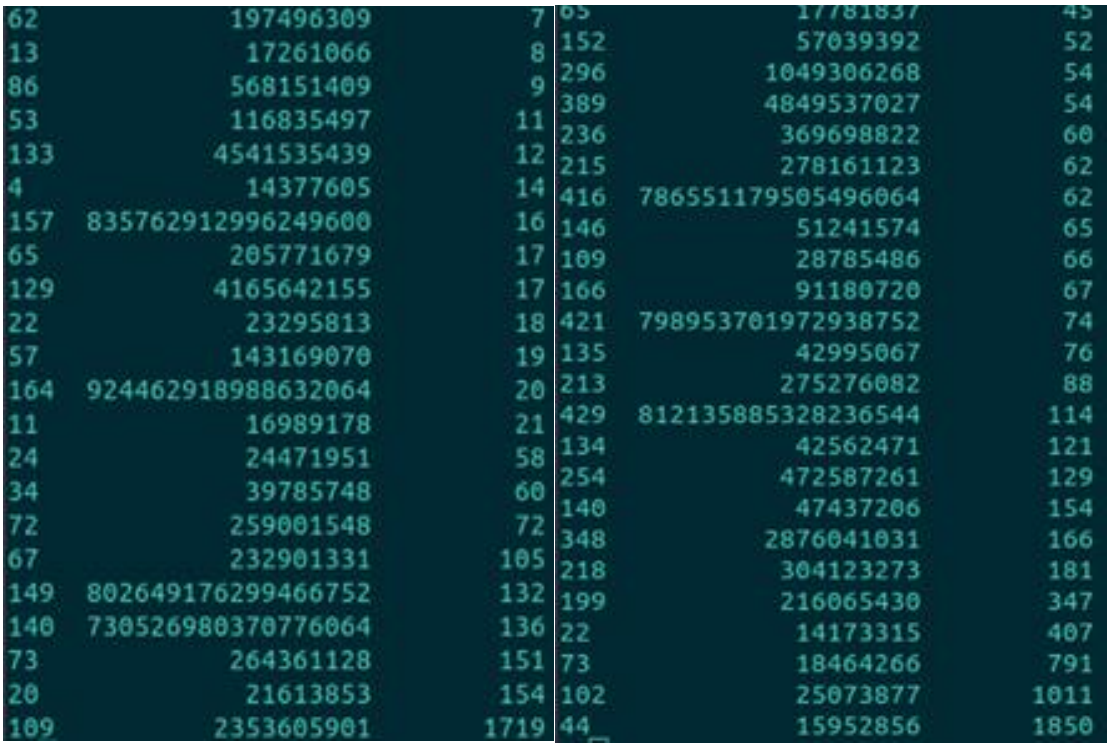


Figure 4: Density graph of Brain Kemp

- Figure 5 shows is the centre the user ids and on the extreme right the number of tweets made by each one of them. The figures on the left hand side are nothing but just the serial number of each user id, since here they are arranged in the ascending order of the number of tweets made by them the number seems a bit haphazard. As we can see from the figure that there were two users who made 1719 and 1850 tweets respectively. The same has been reflected in the histograms and density graphs shown above.



62	197496309	7
13	17261066	8
86	568151409	9
53	116835497	11
133	4541535439	12
4	14377605	14
157	835762912996249600	16
65	205771679	17
129	4165642155	17
22	23295813	18
57	143169070	19
164	924462918988632064	20
11	16989178	21
24	24471951	58
34	39785748	60
72	259001548	72
67	232901331	105
149	802649176299466752	132
140	730526980370776064	136
73	264361128	151
20	21613853	154
109	2353605901	1719

65	17781837	45
152	57039392	52
296	1049306268	54
389	4849537027	54
236	369698822	60
215	278161123	62
416	786551179505496064	62
146	51241574	65
109	28785486	66
166	91180720	67
421	798953701972938752	74
135	42995067	76
213	275276082	88
429	812135885328236544	114
134	42562471	121
254	472587261	129
140	47437206	154
348	2876041031	166
218	304123273	181
199	216065430	347
22	14173315	407
73	18464266	791
102	25073877	1011
44	15952856	1850

Figure 5: Number of tweets per user for Kemp and Stacy's respectively

- To better understand the spread of this data we calculated the maximum number of tweets by user, mean and standard deviation. The results are shown in figure 6. We found that on an average there were 17 tweets made for Stacey Abrams and 15 tweets for Brian Kemp by each user we had in our dataset, during the time period of collection of tweets. Standard deviation was 130 and 102 respectively.

	Stacey Abrams	Brian Kemp
Mean	17.02	14.92
Standard Deviation	130.46	102.90
Maximum no tweet/user	1719	1850

Figure 6: Mean, Standard deviation and maximum no of tweets for Stacy and Brian.

- After removing the biased users we calculated the scores of each user. We found the minimum, maximum, SD, and mean score for both Stacey Abrams and Brian Kemp. The results are shown in Fig 7. The negative score corresponds to the sentiments against the candidate while the positive score corresponds to the sentiment in favor of the candidate. Interesting thing to note from this is that mean score for both of them is around 0. Which says that on an average after removing biased users we had almost same emotions from the public, for and against, both of them. Although slightly higher negative on the Brian Kemp's side. The same is represented in the form of histograms in the figures 8 and 9.

	Stacy Abrams	Brian Kemp
minimum	-514.45	-514.15
maximum	385	136.9
S.D.	36.18	40.14
mean	-1.5	-4

Figure 7: Minimum, maximum, standard deviation and mean of sentiment for Stacy and Brian.

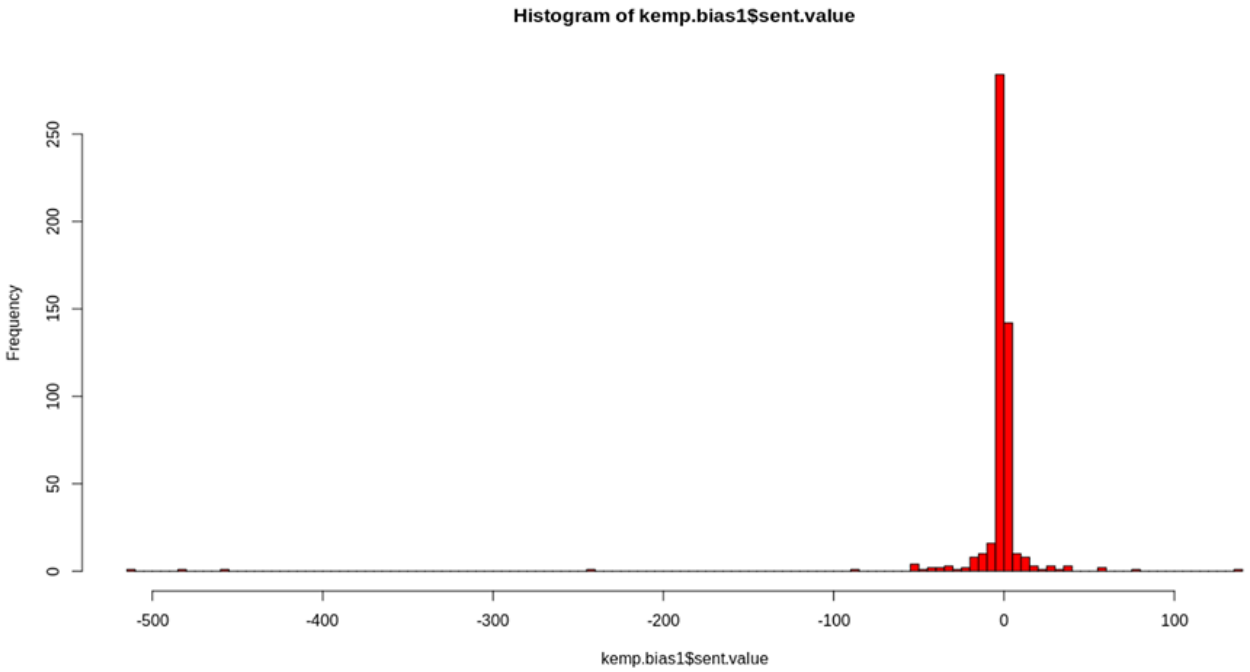


Figure 8: Histogram of kemp after removal of biased

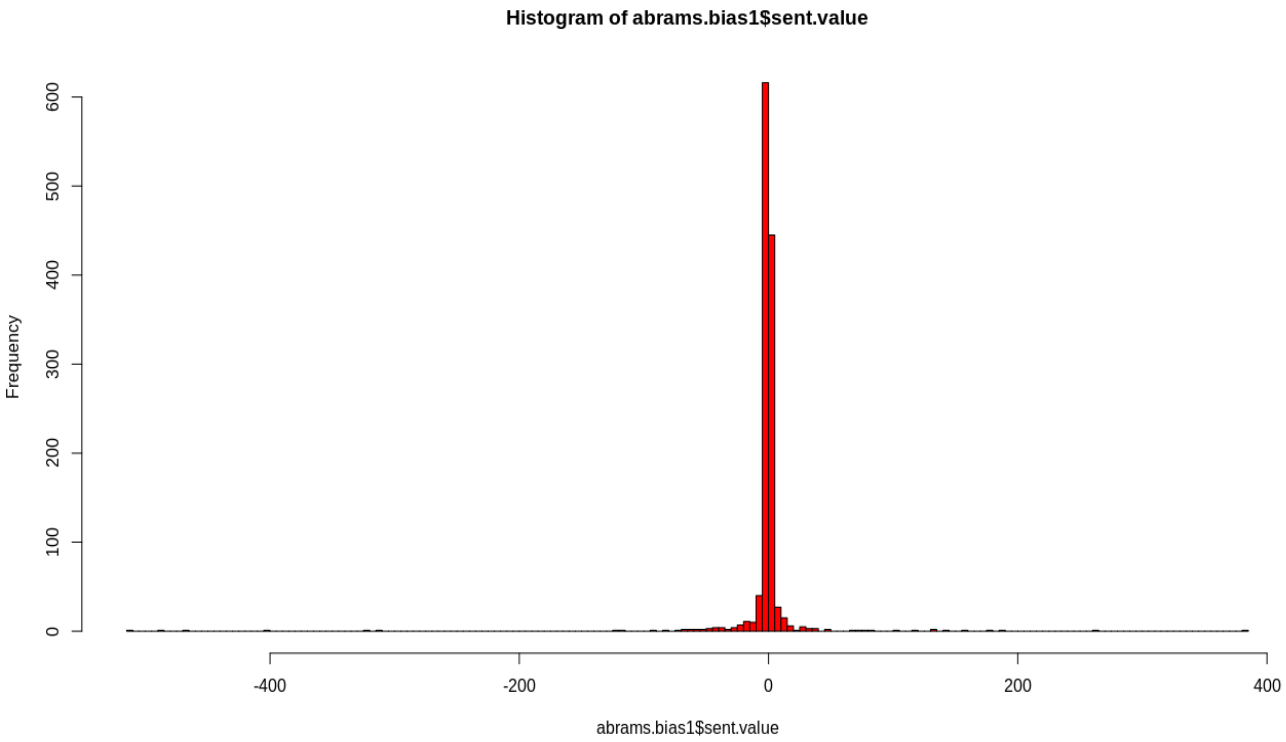


Figure 9: Histogram of Stacy Abrams after removal of biased

- After analyzing tweets without biased users, we saw that the emotions were same for both and negative as well, so we thought of analyzing the whole data set that we got and see how the results change if we include biased users as well. We again did the whole thing on the dataset with the biased users and the results we got were really interesting. It turned out that biased users affected the sentiment score of Brian Kemp very heavily and it reached a positive score of 1.787 from a negative score of 4. Which means that if we were to trust this data, we would conclude that Brian Kemp would win the elections. Thus, we can say that biased users affect the predictions heavily in such circumstances. The maximum, minimum, SD and mean scores of both Stacey Abrams and Brian Kemp are given in the following figures along with the histograms.

	Stacy Abrams	Brian Kemp
Maximum	3687.25	3687.25
Minimum	-3582.1	-694
Mean	-1.576	1.787
S.D.	150.56	170.37

Figure 10: Result without removal of biased user

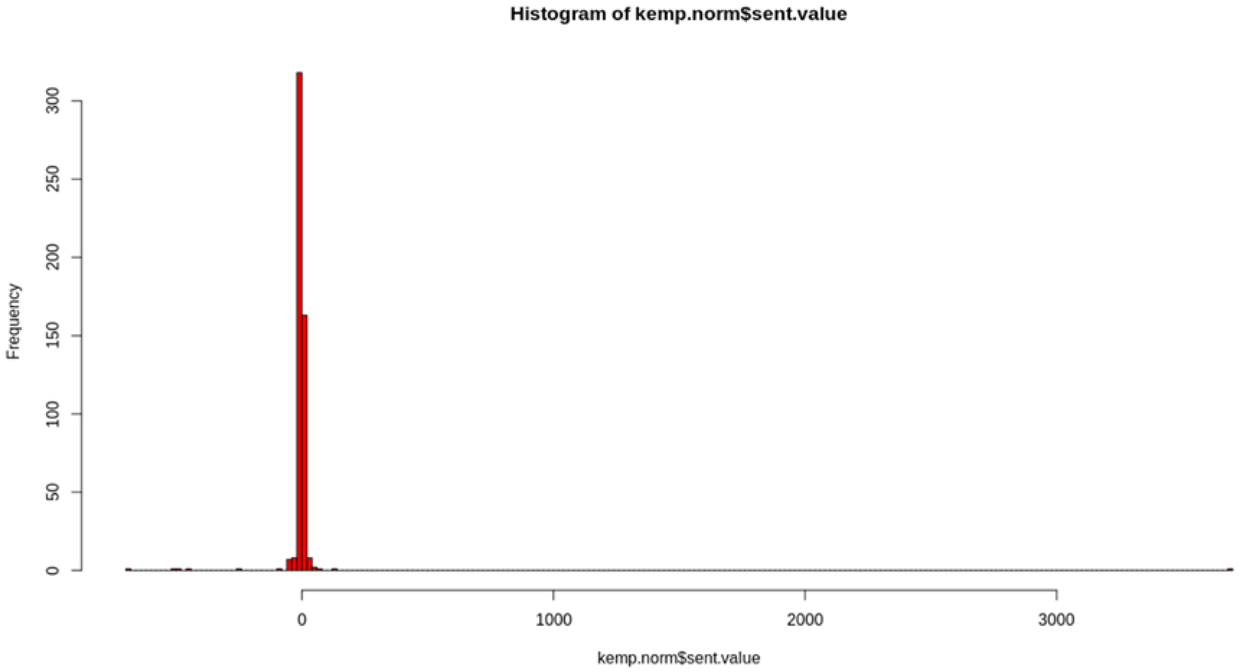


Figure 11: Histogram of Kemp without removal of biased user

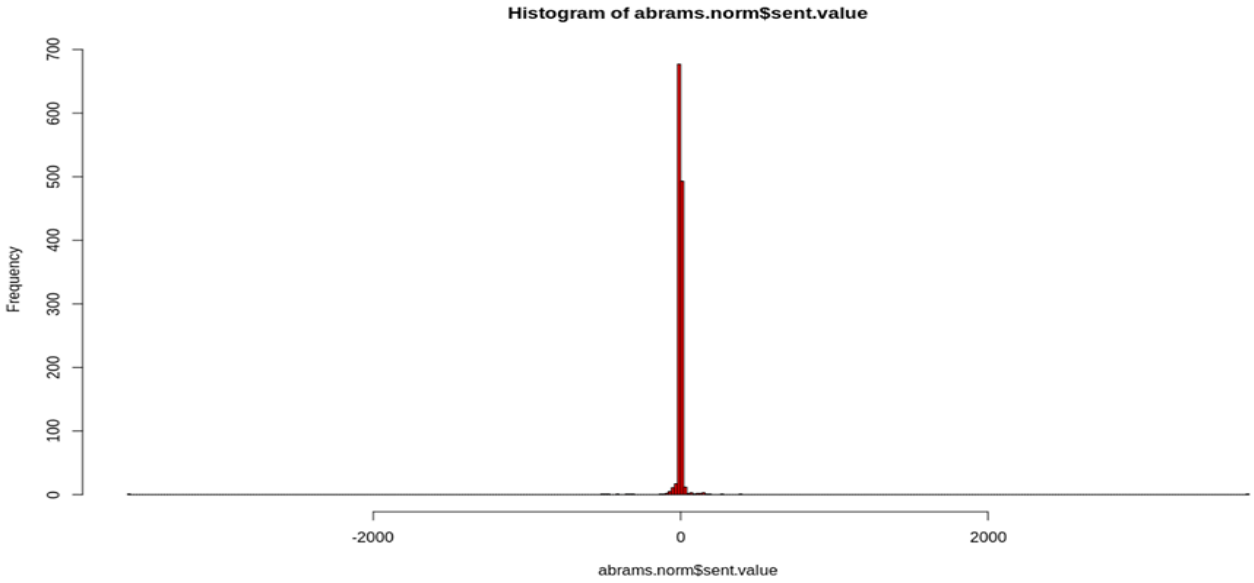


Figure 12: Histogram of Abrams without removal of biased user

V. Conclusion/Lesson Learnt:

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. We used tweepy in extracting live streams of data. For data filtering we used R. Data analysis makes it possible not only for business organizations to keep track of their

services and generates opportunities to promote, advertise and improve from time to time but also allow to capture the sentiment of people on the issue related to politics which if taken in proper consideration can affect the formation of government. Since data science and data analytics is growing at a very fast pace so to work on the project like this helped us to get hands on in dealing with huge amount of data. It helped us in knowing the method preprocessing and then come up with some meaning full data that can help in understanding the behavior of the user. By working on this project, we learned a new language R and the package that can be used for pre-processing and analysis.

VI. References:

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.

[2] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[3] Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836. C. Fellbaum. 1998. Wordnet, an electronic lexical database. MIT Press.

[4] Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics.

[5] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[6] David Haussler. 1999. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

[7] Sentiment Analysis of Twitter Data by Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Department of computer science Columbia University.