Report for Week 9

I. Summary

1. General Data Profiling for Twitter Dataset

2. Periodical Data Profiling

II. Actions performed

1. General Data Profiling for Twitter Dataset

Name	Value	Description
#Users	3,064,983	Total number of users
#Tweets	26,525,351	Total number of entries in the dataset
#Diffusion-Starting-Tweets	2,108,598	Tweets that started a diffusion (were retweeted)
#Re-Tweets	18,844,382	Non-original tweets; endorsements of original tweets.
#D.S.T + #Re-Tweets	20,952,980	The total set of tweets that are part of diffusions
#Non-Retweeted-Tweets	5,572,371	Tweets that do not appear in any diffusion
Quantile of diffusions size	0% 25% 50% 75% 100% 2 2 3 6 45764	

2. Periodical Data Profiling

No.	Period	#Users	#D.S.T	#Re-Tweets
1	2016-01-06 - 2016-03-04	218,803	149,685	1,076,556
2	2016-03-05 - 2016-05-02	307,352	233,940	1,768,516
3	2016-05-03 - 2016-06-30	2,948,873	1,745,800	16,202,825

In the image below, we can see the trend is ascending and the peak is obtained around the date of the referendum which is 13th of June 2016. Also it’s important to notice the relatively short time extent, starting in January 2016 and ending in July 2016.

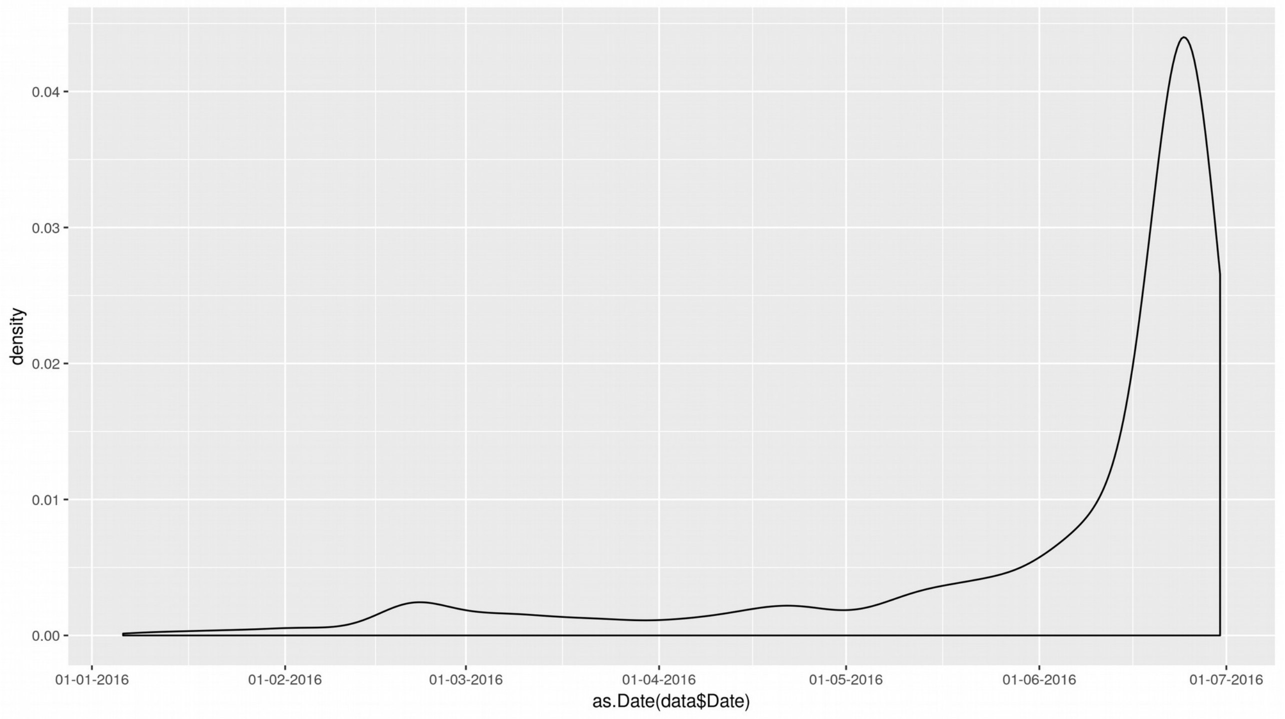


Figure 1: Temporal distribution of the posts in time.

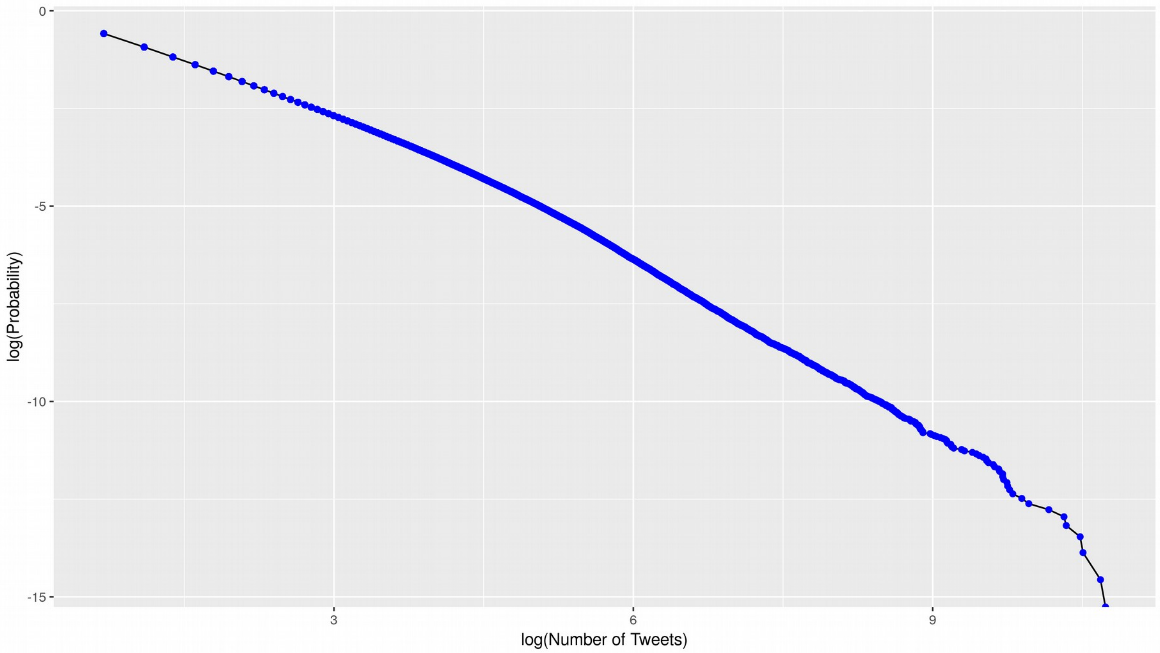


Figure 2: CCDF - Number of tweets per diffusion

In Figure 2, we can observe the Complementary Cumulative Distribution Function for the number of Tweets per diffusion in log scale. We can see that most of the diffusions have a small number of tweets.

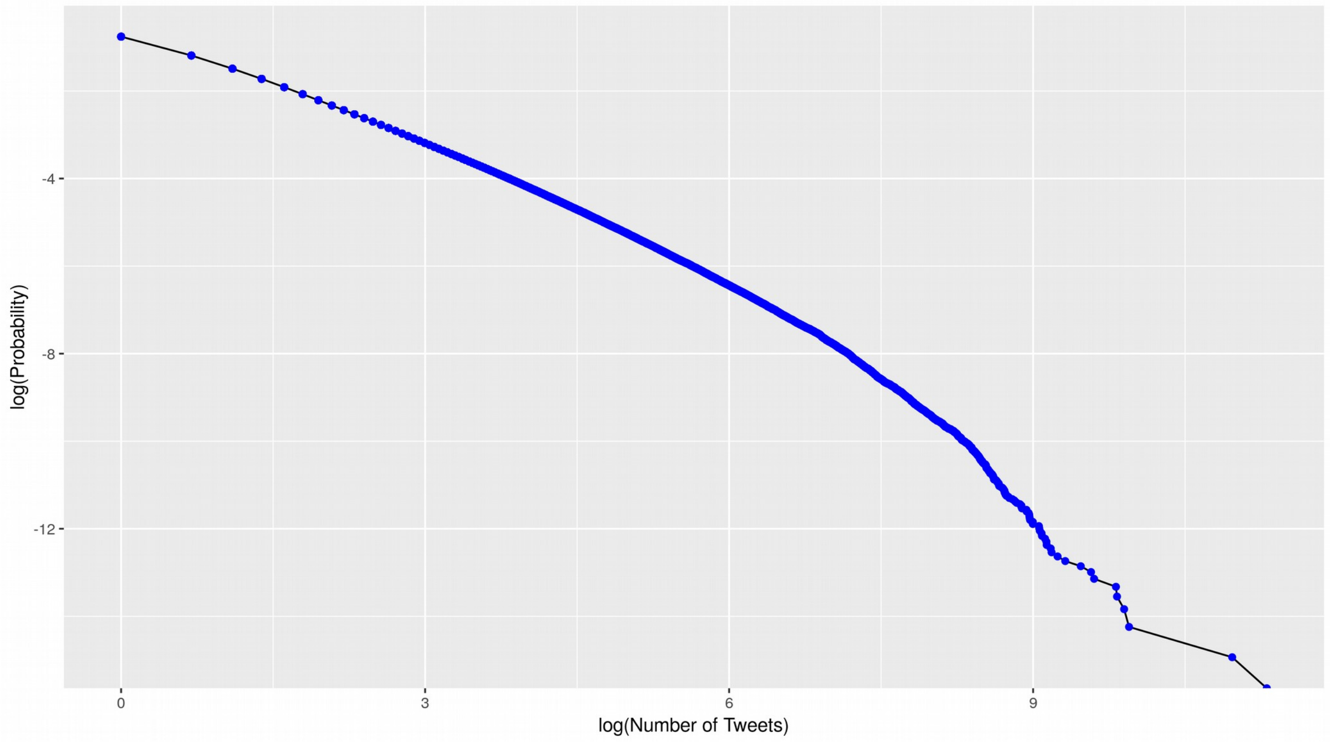


Figure 3: CCDF - Number of Tweets per author

In Figure 3, we have the CCDF for the number of Tweets per author in the studied period of time. (Both Figure 2 and Figure 3) are computed for the entire period of time). Again, same as in the previous figure, there is a high number of users posting a small number of tweets.

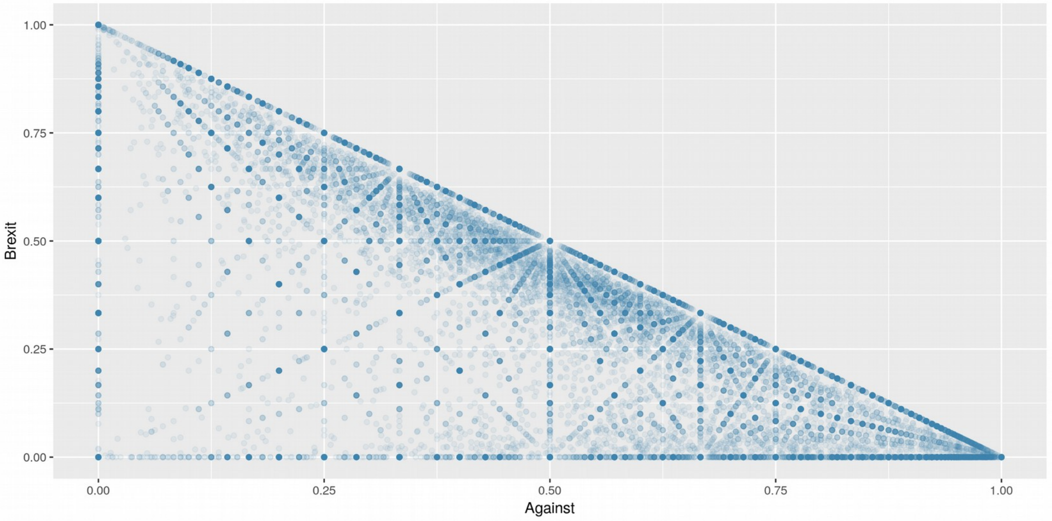
3. Periodical Data Profiling

I perform a equal size splitting and obtain 3 periods. In the following set of 3 figures (Figure 4,5,6) I describe the structure of the diffusions in each of the 3 periods. For each period, I aggregate the replies of every user and apply the stance classifier on the users' discourses. Then, I analyse the structure of the diffusions, concerning these stances. (I am intreseted to see the % of users pro, against and neutal around Brexit). In the Figures, every point is a diffusion (represented in 2 dimensions – the % of users pro brexit and against brexit). We are interested to check the corners of the triangles as they show diffusions with only one polarity in them.

T1

Single stance diffusions:

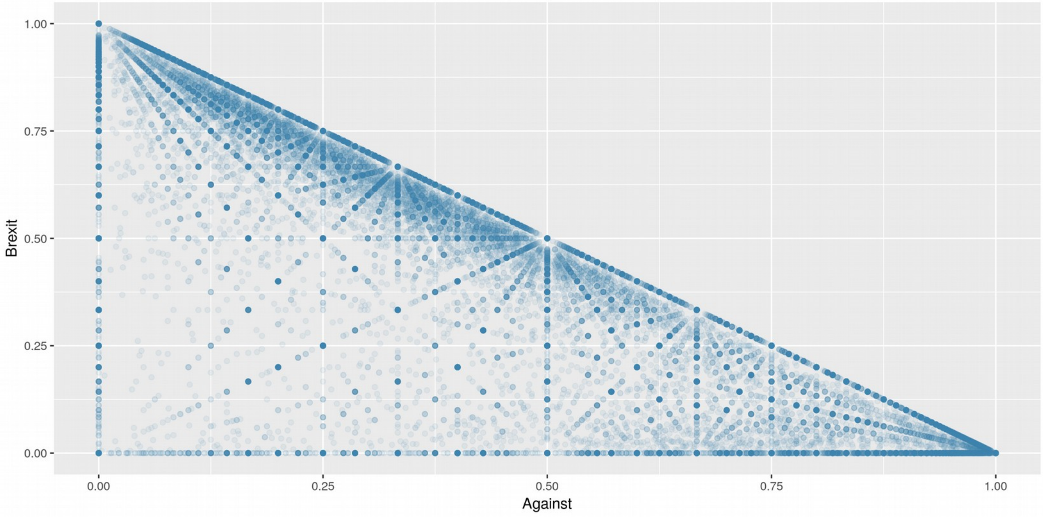
83960 of 149685



T2

Single stance diffusions:

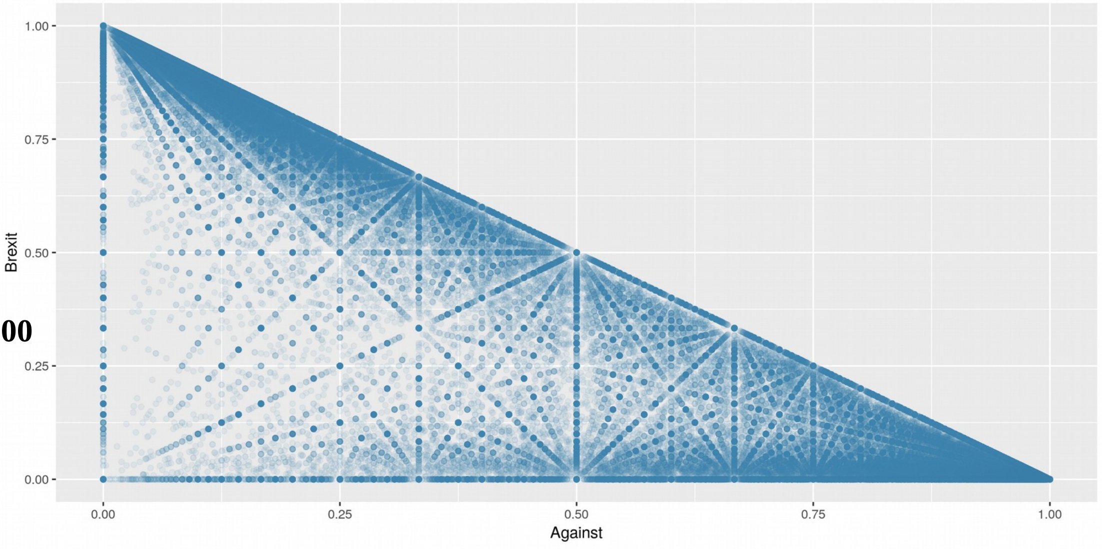
136025 of 233940



T3

Single stance diffusions:

1135323 of 1745800


Brexit
Against