Report for Week 2_V2 – 08 April – 12 April

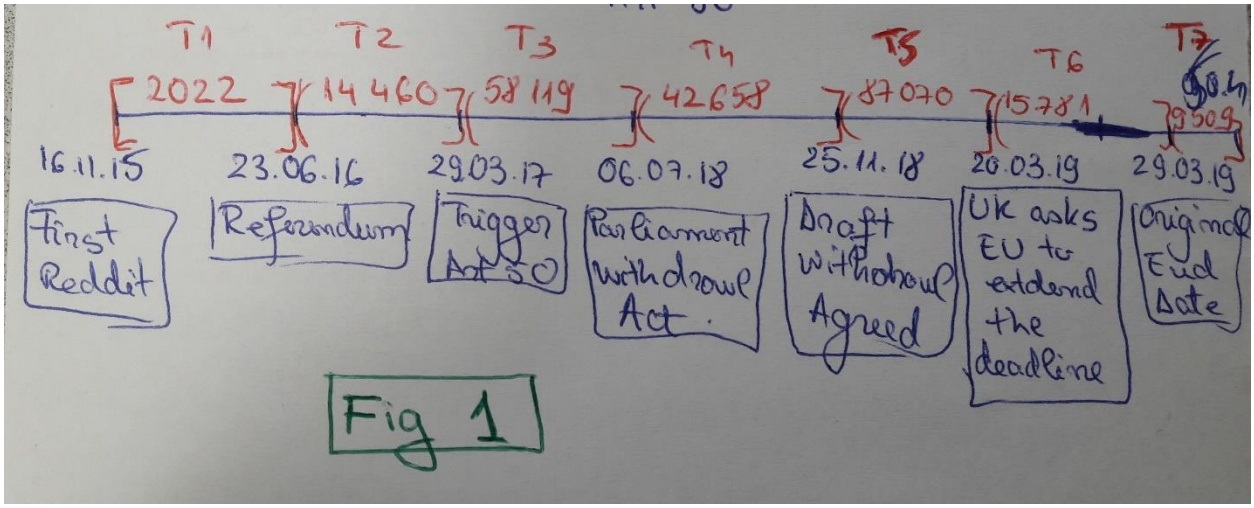
I. Actions performed

1. Downloaded the Data for Twitter from subreddit: https://www.reddit.com/r/brexit/ (to do this, I implemented a script in Python, to crawl all necessary data)
2. I mapped the structure of the database in data frames and defined the following:
    a. **Submission** = the first node of a thread (i.e. The first message posted in a diffusion) – the root of the diffusion tree
    b. **Comment** = the chronologically subsequent messaged posted in a thread, after the submission
    c. **Diffusion** = Submission + Comments
3. I defined two cutting point series
    a. the Wikipedia https://en.wikipedia.org/wiki/Brexit#Timeline_June_2016_to_April_2019 series with > 20 points for high granularity -> however for our purpose, I merged the consequent days events => **16 time periods**

Number of diffusions:

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16
3367	6265	478	2606	1466	2300	4102	54505	23067	15385	3718	25468	54850	9119	13414	9509

    b. my cutting points based on the major events of the timeline, enhanced with the extra two (see Fig 1 on paper) => **7 time periods**



Number of diffusions:

T1	T2	T3	T4	T5	T6	T7
2022	14460	58119	42658	87070	15781	9509

4. Statistics:
    a. number of diffusions: 229 619
    b. number of comments: 207 894
    c. number of submissions: 21 725
    d. number of different users: 14362
5. Searched papers about Brexit -> found a website with all the timeline
6. Plotted Wordclouds for each period (the 7-cut) and for the overall (see figures **timeframe_*i)***
    a. We can clearly see that the word "brexit" tends to appear in all of them as a central element.
7. Following discussion with Andrei, about evolution of thematic, we decided that we want to know if thematics do evolve over time. It is important to see the distance between where a person is at the beginning of the time study (T1) and where it is after 2, 3 periods or at the end (T7). To do this, we need to measure 2 distances: the interperiod distance (where it was at T1 and where it is at T7) – this periods is affected by the evolution of thematics. And the interperiod distance (the distance between pro party and counter party).
8. To measure the evolution of thematics: for each time frame T_i, we compute an embedding vector R_i, having N elements (N = the first N most frequent terms in the whole data) – R_i is a Tf based embedded vector.
9. Then we can build a heatmap to see how thematics evolve from T1 to T_N. (see figure **heatmaps.png**)
10. In this heatmap, things seem to be really fine: thematics do evolve, with the exception of line 1 column 6 and line 3 column 6 (color of red is more powerful). To find out what's wrong, I used LDA topic modelling independently on periods T1, T3 and T6. (see fig. **topics_Ti.png**)
11. The discoveries are: T1: 16.11.15-23.06.16 (period pre-Referendum -> topics: **Europe**, **brexit**, **vote**. (this period also contains the day of the referendum)

12. T3: (Trigger Art. 50 -> Parliament Withdrawal act) -> they're talking about **trade**, **deal**, **referendum**, **vote**
13. T6: (Uk asks Eu to extend deadline -> Original End Date) -> I don't see a kind of main topic...
    Maybe the problem is with T6, which causes the situations of outlier, as (perhaps) people are not talking about something, but just random stuff?

14. Same analysis for topic evolution was performed on the (16-cut) set. (see folder More periods) - > outlier line 3, col 8 and 14. I plotted the wordcloud => in period 3: Theresa May accepts the Queen's invitation to form a government. David Davis is appointed the newly created Secretary of State for Exiting the European Union to oversee withdrawal negotiations.
    Period 8: Brexit negotiations commence.

    Period 14: May requests the EU extend the Article 50 period until 30 June 2019.
15. Started working on t-SNE in user space -> polarization of users (using 16-cut) I've done from T1 to T4

II. Encountered Difficulties

III. Plans for next week

1. Go on with polarization of users (maybe use word embedding)
2. Tipologies of users
3. Perform clustering based on simple Tf-Id and partition the users
4. Analyze duration of diffusions
5. - probrexit / antibrexit - how much time
6. - inside a group, how do people interact (C, A, C, A, M, C)