Université Jean Monnet

Machine Learning and Data Mining Master

Information Diffusion in Online Communities

Author:
Andrei Mardale

Advisors:
Christine Largeron
Marian-Andrei Rizoiu

16 March 2019

Contents

1 Introduction

1.1 Context

Internet has met an unprecedented growth of popularity in the last years, thus it is not surprising that many aspects of our lives have been influenced by the improvements of technology. The way information is spreading nowadays is vastly driven by the development of online social networks. Hundreds of millions of Internet users have access to novel information and various points of view [2], which they can share themselves, leading to emerging online communities. The diffusion of information in these communities has been often the subject of numerous studies as it turned out that the outcome of major events such as the 2008 presidential elections in the United States of America [22] or the decision of leaving the European Union made by English people [21], was influenced by the transfer of information through such kind of networks.

It is of particular importance to understand the intrinsic mechanisms of such social networks, in order to efficiently detect possible changes in the users' attitudes around the different problems they are debating. By aggregating textual information that defines users with the dynamics of the diffusions they are part of, the goal is to predict future stances of the individuals when they interact in the Social Network with their peers. This is of particular usefulness, because it provides a hint of how people change their stances after having interacted with other people, thus leading a to a better understanding of the networks.

1.2 Goals and Interests

The first goal of this work is to perform a longitudinal analysis of the discussions around Brexit, mainly by focusing on the different topics that are debated and how do these topics evolve in time. Moreover, this study aims to correlate the evolution of the topics discussed with the behavioural evolution of participants, by providing a tool for visualizing the dynamics of discussion topics and the trajectory of users.

Furthermore, this work focuses on detecting and predicting the behaviour of people involved in Online Social Networks [28] when they are exposed to certain kind of information and interact with people already supporting different ideas. For accomplishing this goal we use as case study the Brexit. By interpreting the position of participants on the subject of the withdrawal of the United Kingdom from the European Union, we are interested in predicting behavioral changes in the discourse of individuals, who are beforehand proved to be part of different communities.

We are also interested in detecting if there are other factors that might cause shifts in the opinions of participants in the online debates, such as the number of messages they exchange, with what kind of persons they do exchange messages, the popularity of their messages, etc. Nonetheless, the social influence [52] and the homophily [11] are also aspects we are trying to take advantage on in the process of predicting the trajectory of only participants.

Even though there are quite a lot of papers on topics such as information diffusion in online social networks [24], [29], [57], and in particular on the subject on Brexit [3], these works mostly try to detect the communities (leavers and remainers) and find the topics that are hot among these two parties. In our study, we are starting mainly from these works and try to advance them, by looking for patterns among the different diffusions and we aim to predict future changes in positions of persons who have been exposed to certain kinds of information. For instance, if two persons A and B share the same opinions, if person A is exposed to a certain sequence of messages of a particular type, let's say pro-brexit, and consequently changes the tone of the discourse in the following messages, what are the chances that person B will also do the same, if exposed to this sequence of messages. A second difference between our work and other studies published so far consists of the kind of information used: while other works use textual information for classifying users, we rely only on information derived from the social interaction they have on online platforms.

1.3 Main contributions

- longitudinal analysis of discussions around Brexit on two social media platforms: Twitter and Reddit;
- tool for visualizing the dynamics of discussion topics and trajectory of users;
- prediction of future political stance based on features defined using the structure online diffusions;

1.4 Outline

The subsequent sections are structured as follows: Section 2 describes the state of the art in the field of community detection, information diffusion in Online Networks, topic detection, followed by the description of a similar work to the one presented in this report, in which the authors have a tangent goal, but use a different methodology, relying on textual information. Our work starts from this paper, and enhances it through a number of changes.

Section 3 presents the datasets involved in the development of our proposed approach and the way the data was pre-processed. We detail the two platforms used for collecting the data. These two platforms yield two different datasets, one of which is collected by us from the Reddit platform, while the other represents Twitter data. We use the Twitter data to train our stance detection classifier (pro, against or neutral around Brexit subject). Next, this classifier is used on Reddits to determine the two main communities from this Online Social Network and by aggregating different features, we predict future trends in the network.

Section 4 details the strategy behind the longitudinal analysis of topics around Brexit: the splitting method used to divide submissions temporally, the methodology used for topic extraction and the analysis. We perform clustering of the users represented in the topics

space and offer a two dimensional representation by applying t-SNE, for visualizing the dynamics of the discussion topics and the evolution of the users.

In Section 5, we present the methodological contribution proposed during this internship regarding the political stance prediction. The two threads of this study are presented. Firstly, the main Reddit study which aims to predict the stance of a user taking into account its interactions with other members of the community. This research path unveiled a second path, due to the necessity of labeling and partitioning the users. Section 5.2 depicts this classifier and the way it was trained.

Finally, Section 6 presents the results obtained after conducting this study, whereas in Section 7, we draw conclusions and enumerate the next steps which have to be made in order to improve our results.

2 State of the Art

2.1 Information Diffusion in Online Networks

Many studies focused their attention on the way social networks information flow can be modelled in a way which allows exploring, understanding and predicting the information diffusion. A thorough survey of the different methods related to this subjects is performed by [19]. Since the beginning, the authors are modelling the Online Social Networks using elements from graph theory. Thus, each vertex represents a user from the network, while an existing arc between two users means that the source vertex is exposed to information coming from the destination vertex.

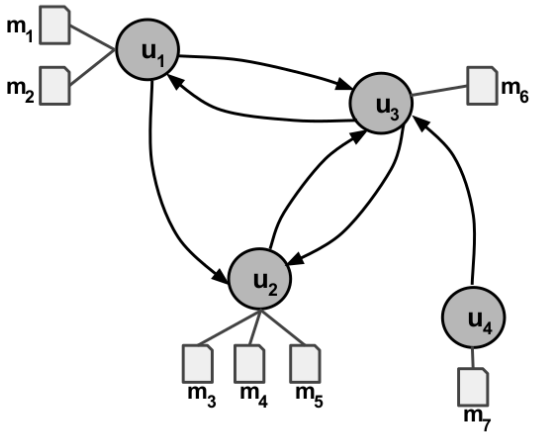

Figure 1: Online Social Network modelled as graph. Each vertex u_i is a user, while a directed edge (u_i, u_j) means user j exposes user i to a certain information, via message k, denoted as m_k Source: *Adrien Guille, Hakim Hacid, C. Favre, Djamel Abdelkader Zighed. Information Diffusion in Online So- cial Networks: A Survey.*

A diffusion is therefore defined by a sequence of vertices u_i, $i = 1 : D$, which are ordered

in function of the time t_i the user u_i was exposed to a message defining the topic of the diffusion. There are two kind of predictive models when it comes to information diffusion. The first category is defined by approaches which take into account the graph structure of the Online Social Network, preserving the properties of the relationships between vertices, such as in [15], [16]. On the other hand, the second category [20], [44] is not based on the graph structure and it aims at classifying the nodes in several states (such as susceptible, infected, recovered) and quantifying the proportions.

2.2 Topic Detection in Online Communities

There are many studies regarding the current state of the art concerning the topic detection task. Among them, there are methods that propose a real time detection technique which allows obtaining the most trending topics within a community, such as [7]. The messages in the community are split according to time criteria, into consequent time frames. The authors define the notion of aging for a term, from which they derive the life cycle of a term. Emerging topics are formed by terms which have a high frequency in the current time-frame, but a low frequency in the past. In others, a keyword is considered emergent, if it is extensively used in the current time-frame, but not in the previous ones. Moreover, by analysing the relations between users, they rank their authority and propose a model for determining the terms that are correlated the most, under specific topics. Finally, a graph of the most emergent terms is computed which leads to the birth of topics.

2.3 Brexit Position Classification of Twitter Users

Another important piece of work that represents the starting point for our research is represented by [4]. They analyze messages sent on Twitter, especially sent in the period of the Brexit referendum. Their main goal is to be able to accurately predict by the text sent in the message, if the author is pro or against Brexit. In the second part, they take into account the two parties and analyze the most important subjects that they had engaged into. The first part of their work is a precious tool for our research: they provide a recipe for classifying Remain versus Leave accounts, by using a supervised Machine Learning algorithm, namely Naive Bayes Classifier. The task is particularly difficult because the set up of the problem does not provide labels or ground truth for the collected replies. However, their main contribution is in building such a ground truth. They find the top 200 most mentioned accounts in that period and manually assess their membership to one of the above mentioned classes. Then, for the two found groups, they look into the hashtags and create two lists of hashtags: pro and against brexit. Next, they aggregate the tweets of each account and by looking into all accounts texts, they pick the ones that are using those hashtags, compute a leave index, which is the number of leave hashtags minus the number of remain hashtags and sort the accounts according to the index. Finally, they pick the first 10% as pro brexit, and the first 10% as against brexit. Then, they train a NB classifier and label all other accounts.

2.4 Reddit Studies

Although at the beginning, research on Reddit datasets has caught up in the last years. For instance, [10] performed a study on the conversations hosted on Reddit online network, focusing on the quantitative and qualitative aspects of the messages. The main target of the study is classifying the conversations and finding properties in terms of volume, responsiveness of the users and the tendency of becoming very popular and spreading very quickly. Thus, the authors concluded that a viral diffusion tends to have more difficult text, whereas cascades that will remain not so unknown to the large public have simpler, shorter messages. Moreover, the authors detail how a large conversation is actually made up of an inter-twinning of large number of messages, most of which are sent by a small set of unique users. Perhaps not surprisingly, each sub-community has different characteristics, while subreddits containing media content like photos and videos, news and discussions sub-communities have a tendency to lead to viral conversations.

3 Datasets

In this section, we will detail the two datasets that were used during this study. In the first part, we will present the Twitter dataset which was used for defining the partitioning of users with regard to the Brexit problem, while in the second part, we will introduce our Reddit dataset, used for performing the topic detection and prediction of the users' behavior.

3.1 Twitter Data

Twitter is an online social networking platform where its users can post messages named *tweets*, share other peoples' messages, action called *retweet* or tag other users, so that the information, although publicly available, is targeted to the mentioned user. This platform is mainly used for news sharing and thus, communities are often formed around people who share the same interests. This platform played an important role in crucial political moments in the last years, Brexit being one of them. Actually, there is a study claiming that mining the opinions among twitter users allowed some researchers to correctly predict the outcome of the public vote, with a better accuracy than conventional polls [8].

The dataset we are using was collected and used [4] by K. Benoit and A. Matsuo, who were kind to share their dataset with us. They collected the dataset using the Twitter "firehose" [13], which allowed them to have access to the live streaming of data from January 6, 2016 to July 2016. They filtered the captured data based on a set of elements (keywords, hashtags, user names) related to Brexit, as presented in Figure 2.

A total of 26 millions of Tweets were collected in the before mentioned period of time, out of which 10 millions were original tweets. The data has a set of 5 variables: **Tweet ID**, **Date**, **Retweet**, **Text**, **User Screen Name**.

We are using this dataset captured from Twitter mainly because of the key property of

Search Criteria	Terms
Simple words	brexit
Hashtags	#betterdealforbritain #betteroffout #brexit #euref #eureferendum #eusummit #getoutnow #leaveeu #no2eu #notoeu #strongerin #ukineu #voteleave #wewantout #yes2eu #yestoeu
User screen names	@vote_leave @brexitwatch @eureferendum @ukandeu @notoeu @leavehq @ukineu @leaveeuofficial @uk-leave_eu @strongerin @yesforeurope @grassroots_out @stronger_in

Figure 2: Hashtags and usernames used to collect Tweets related to Brexit. Source: *Celli, F., Stepanov, E., Poesio, M., and Riccardi, G. (2016). Predicting brexit:Classifying agreement is better than sentiment and pollsters*

Twitter Data: users tend to use a lot **#hashtags** which makes it easier to detect pro and against brexit groups and turns the unsupervised learning environment into a supervised one, thus allowing us to learn a classifier for later purposes.

3.2 Reddit Data

Reddit is a platform for online discussions where members can submit news content, opinions or articles in the form of text, link or media. Unlike Twitter, it does not provide a hash-tagging mechanism which makes it more difficult to classify users of different parties. However, the content is more structured, since it provides sub-reddits, which are organizatory elements to group discussions having common subjects together. Inside these chambers, users can start threads similarly to forum like environments, and in these threads they can also post comments to the initial messages. Thus, often, we find hierarchies of posts, in a tree like structure, as we can observe in Figure 3, where we present both the real structure of a reddit and the logical tree structure used for analyzing and extracting non-textual features.

Throughout this study, the following terms will be used:

- **Thread** = the discussion starting post, being the root of any further submissions.
- **Comment** = the chronologically subsequent messages posted as replies to a thread.
- **Diffusion** = Thread + Comments.

The dataset was collected using the Pushshift API [45], which is developed within the Pushshift Project. This is a big data storage and analysis project which allows downloading a large quantity of reddits, from a certain period, respecting different imposed criteria, like the sub-reddit. The data is totally free of use and can be accessed very easily, via

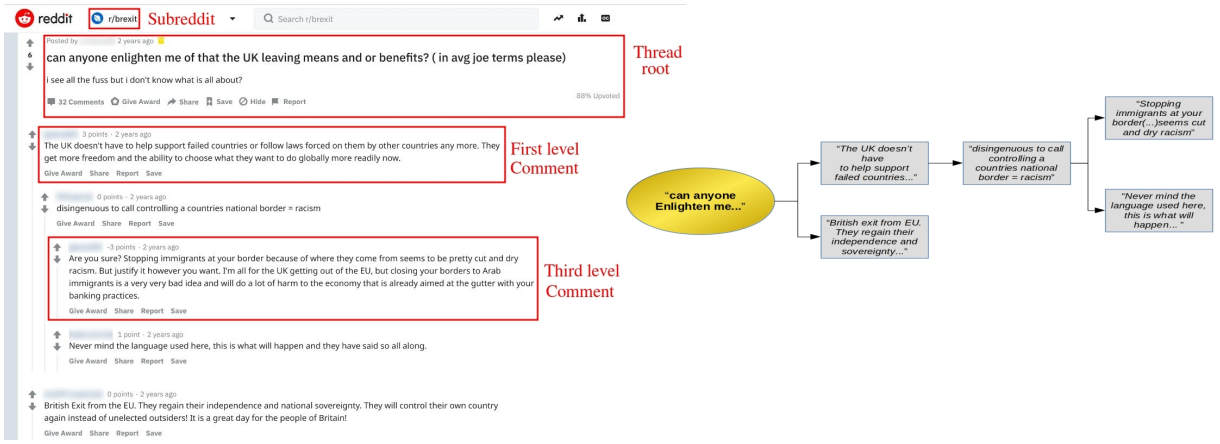

Figure 3: a) Elements of the Reddit platform. Structure of a discussion thread, with multi-level comments, inside a subreddit. b) Logical structure used for analyzing the data.

a Python script using a pre-defined API. Thus, we collected all reddits from **November, 2015** to **April, 2019**, which were part of the **brexit** subreddit. In this period of time, a total of **229619** submissions were collected, each entry having the following variables **identification information, text, timestamp, author, parent id (useful for building the tree structure), score and number of comments (specific for the root of the threads)**.

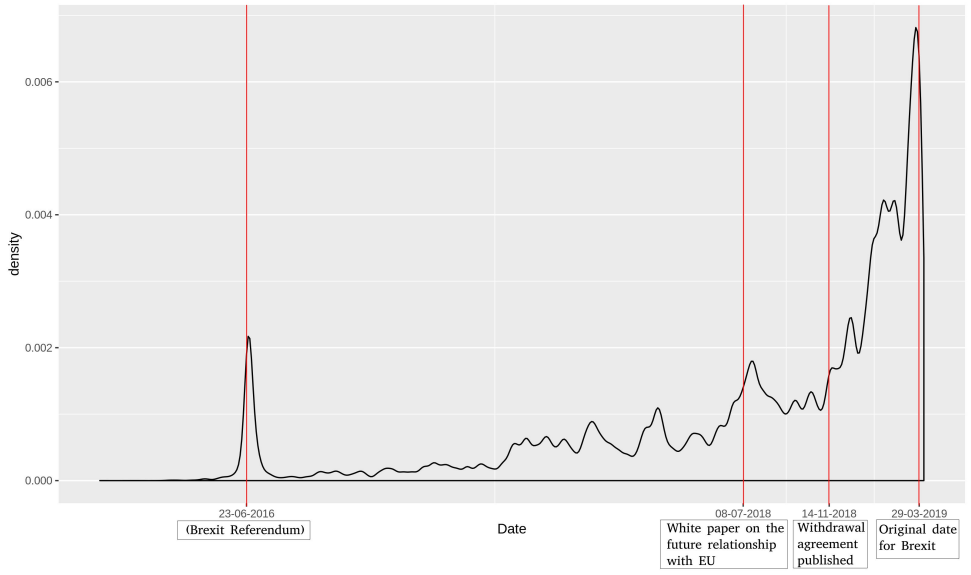

Figure 4: Time distribution of the submissions collected from Reddit (subreddit brexit), between November, 2015 and April, 2019.

The time distribution of the collected messages is presented in Figure 4. In this figure, we can observe that generally there is an ascending trend, which may give a clue of the growing importance of this subject as it is perceived by usual people using online social

platform for discussing the news. Even though the monotonicity is increasing, we can observe a spike in June, 2016. This is generated by the fact that on the 13th of June, 2016, British people were expected to vote for the national referendum. This event enjoyed a considerable attention from British media and news channels, which was also reflected in the activity of online social networks users. On the other hand, the peak in terms of number of submitted messages is in February - March 2019. This is a consequence of the official schedule of the Brexit process, which should have completed in March, 2019.

As far as the statistical structure of the submitted messages is concerned, the very vast majority of them are comments, as depicted in Figure 5 (a), as opposed to initial, thread starting messages. In terms of unique authors, Figure 5 (b) shows that 20% of all the unique authors are **only** thread initiators. This means they only send a single message, starting a discussion thread, in which they never post again. On the other hand, 19% of the authors, are both thread starters and commenters, meaning that they start threads and take part actively in the discussions, posting answers in their own started thread or getting involved in other discussions. The majority of the unique users are **only** commenters, meaning that they never start discussions, but usually engage in them.

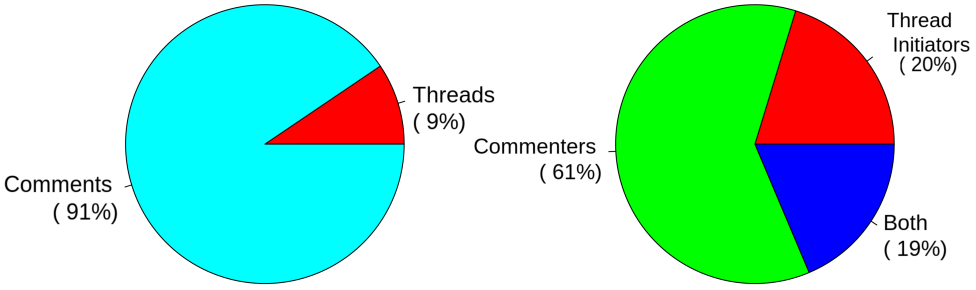

Figure 5: (a) distribution of submitted messages in terms of types of messages. (b) distribution of unique users in terms of roles.

Figure 6 presents another worthy of note statistical feature of this dataset. We can observe the long tail of the graphic presenting the number of messages per user. In this figure, the Complementary Cumulative Distribution Function of the number of messages shows that a very large number of users send only a few messages, whereas there are a few users sending many messages in the observed interval. These users can either be opinion formers or simply paid social bots. One of the tricky tasks of this work was to identify the bots and remove them, as they do not bring new information, but most of the time reshare news and posts.

Data formalization

In Figure 7, we present the formal structure of a diffusion. A diffusion d_i can be defined as:

- d_i = sequence of temporally ordered triplets n_j

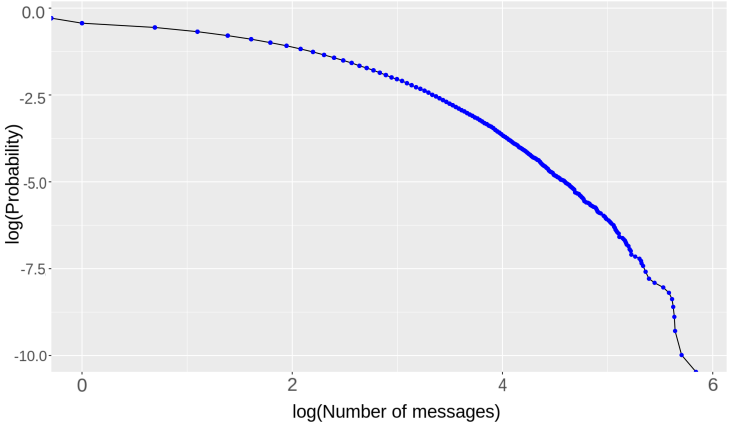

Figure 6: Complementary Cumulative Distribution Function of the number of messages sent by each user.

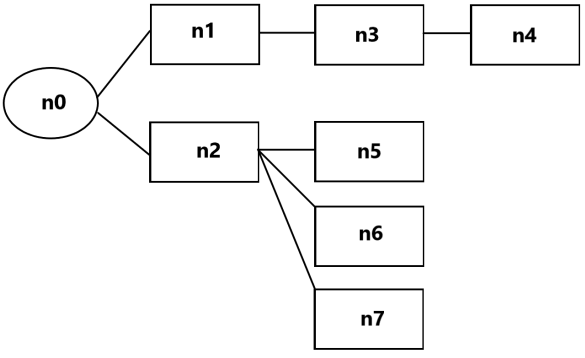

Figure 7: Formal structure of a diffusion.

- $n_j = (u_j, \mathrm{c}(u_j), t_j)$, where: u_j = user name, $\mathrm{c}(u_j)$ = the content published by user j in the current submission, t_j = the time stamp of the submission j

Thus, a subreddit j with N total diffusions (initial threads and comments) on the Reddit platform can be defined as: $S_j = \bigcup_{i=1}^{N} d_i$

4 Longitudinal analysis of discussion topics around Brexit on Reddit

4.1 Temporal split and distances between periods

The most important preprocessing step for building the tool for visualizing the dynamics of discussion topics and the trajectory of users, consists of splitting the set of messages in subsets in order to better capture the underlying trends of the users. Every such time period can be formally seen as an interval:

$$T_k = [ts_k, te_k]$$

Figure 8 presents the results obtained after applying the splitting strategy based on real events. In order to better visualize the sizes and distribution in time of the resulting time frames, a second splitting strategy was tried in which the cutting points are spread at equal time intervals. From Figure 8 we can observe that at the beginning of the analyzed overall period, the cutting points are rarer and span over a longer period of time (first blue period lasts from November 16, 2015 to June 25, 2016) because the density of events was little. However, towards the end of the analyzed period of time, even though red T15 has the same time length as T1, there are 4 events-based periods as in the last months a lot of different actions were taken.

Another observation worth making is the number of posts which increased a lot towards the end of the studied interval, thus leading to the decision of cutting the intervals in a more densely way.

Once both the Equal-Interval and the Event-Based splitting strategy are performed, a mandatory condition for using the Event-Based splitting was to prove that if we use the Event-Based splitting, no artifacts are introduced. Thus, for each of the two cutting methods, we built a Document Term Matrix considering the split dataset as input corpus. Therefore, we aggregate every message belonging to a specific time-frame in order to form the Document associated to that time-frame and obtain 15 Documents.

Date	Equal Cut	Event Based	Posts	Users	Events
2015-11-16	T1	T1	3367	1268	**23.06.2016** The UK holds a referendum on whether to leave the European Union. 51.9% of voters vote to leave. **24.06.2016** David Cameron announces his resignation as Prime Minister
2016-02-12	T2				
2016-04-29	T3				
2016-06-25		T2	6265	1623	**13.07.2016** T. May accepts the invitation to form a gov.
2016-07-14		T3	3084	810	**27.07.2016-** The European Commission nominates French politician Michel Barnieras European Chief Negotiator for the United Kingdom Exiting the European Union. **7.12.2016-** The UK House of Commons votes 461 to 89 in favor of May's plan to trigger Article 50 by the end of March 2017
2016-07-21	T4				
2016-10-11	T5				
2016-12-08		T4	1466	320	**24.01.2017-** UK Supreme Court rules that Parl. must pass legislation to authorize the trigger of Art. 50. **26.01.2017-** The UK Gov. introduces a 137-word bill in Parl. to empower May to initiate Brexit by triggering Art. 50.
2017-01-01	T6				
2017-01-27		T5	2300	431	**16.03.2017-** The bill receives Royal Assent. **29.03.2017-** A letter from May is handed to President of the EU Council D. Tusk to invoke Article 50, starting a 2 year process with the UK due to leave the EU on 29 March 2019.
2017-03-25	T7				
2017-03-30		T6	4102	557	**19.06.2017** - Brexit negotiations commence.
2017-06-15	T8				
2017-06-20		T7	54505	2339	**6.07.2018** - A UK White paper on The future relationship between the United Kingdom and the European Union is finalized. **8.07.2018** - Davis resigns as Secretary of State for Exiting the European Union. Dominic Raab is appointed as his successor the following day.
2017-09-06	T9				
2017-11-27	T10				
2018-02-17	T11				
2018-05-11	T12				
2018-07-09		T8	23067	1479	**21.09.2018** - EU reject the UK white paper.
2018-08-01	T13				
2018-09-22		T9	15385	1480	**14.11.2018** - Brexit withdrawal agreement published. **15.11.2018** - Raab resigns as Secretary of State for Exiting the EU. S. Barclay is appointed as his successor.
2018-10-23	T14				
2018-11-16		T10	3718	732	**25.11.2018** - Other 27 EU Member endorse the Agreement
2018-11-26		T11	25468	2485	**15.01.2019** - First vote on the Withdrawal Agreement in the UK House of Commons. The Government is defeated
2019-01-13	T15				
2019-01-16		T12	54850	4489	**12.03.2019-**2nd vote on the Agreement. Gov. loses.
2019-03-15		T13	9119	1836	**20.03.2019** - May obtains an extending of Art. 50
2019-03-22		T14	13414	2444	**29.03.2019-** The original planned date for Brexit.
2019-03-30		T15	9509	1840	**5.05.2019-** May requests a 2nd extending of Art. 50 period until 30.06.201
2019-04-05					

Figure 8: Time periods used splitting the dataset.

$$D_k = \bigcup_{i=1}^{N_k} c(u_i),$$
$$k = 1 : 15,$$
$$t_i \in T_k,$$
$$N_k = \#users \in T_k$$

We use Term Frequency as weighting method and apply the usual text preprocessing steps: lower-casing, removing punctuation, stopwords and numbers and stemming the words. Moreover, as some periods have a quite large number of terms, we also apply a sparsity threshold. Finally, the Document Term Matrices have a number of documents equal to the number of periods of each strategy and a reasonable number of terms. Next, we compute the cosine similarity between each time period and build the heatmaps presented in Figure 9.

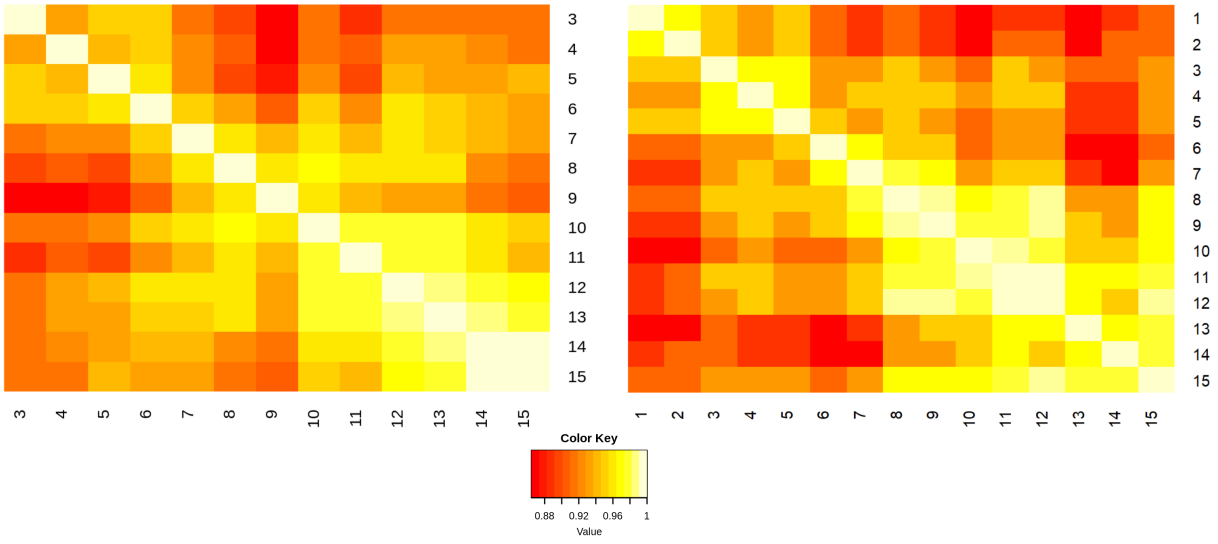

Figure 9: Cosine similarity between different periods of the two time splitting strategies. a) Equal Interval. b) Event Based

Figure 9 a) shows the heatmap of cosine similarities between all except the first 2 periods. This is caused by the fact that, when splitting the set of messages according to the Equal Interval rule, the first two time frames have only 3, respectively 171 messages, compared to the first two time frames of the Event Based splitting. Due to this difference, the first heatmap would have been skewed and it would have hidden the true underlying of similarities.

Finally, Figure 9 reveals that even not perfectly similar, the evolution of the used topics and words in both cutting scenarios is very similar. We can observe that the closer we are to the end of the studied period, the different the words are. Moreover, a lean gradient

of the topics can be deduced from the figure and perhaps more importantly, the trend is very similar in the two heatmaps. In conclusion, using the Event Based splitting does not affect nor introduce source of skewness to the data. It is preferable to the previous one, because explaining the trends of topics is easier, as they can be correlated with the events that happened in those periods of time.

4.2 Topic extraction and analysis

In order to perform topic analysis on a specific time period, from the 15 periods defined above, the first step consists of building a Document Term Matrix for that period, in the same manner which was described in the previous section, in terms of textual preprocessing. However, the main difference is that the documents represent the aggregated texts of different users, while the terms represent the most frequent words in their utterances.

$$D_k = w_{T_i}(u_k)$$

where $w_T(u)$= the aggregation of the messages of user u in the period T.

In addition to the usual preprocessing steps, an extra action was performed which is removing words that essentially do not bring important meaning, are too general or are used too often so they hide other more meaningful terms: "just", "one", "can", "like", "get", "now", "voter", "voted", "vote", "brexit", "people", "want", "think", "know", "say", "even", "time", "year", "still", "thing", "let".

Once the Document Term Matrix is built, we use it for determining important topics and their distribution over the users' speeches. We retrieve these topics using Latent Dirichlet Allocation [6], which is the state of the art in the field of topic mining. LDA is a probabilistic model which starts from the important assumption that each document can be represented as a combination of topics and furthermore, each topic can be represented as a combination of terms. Thus, a LDA model provides two sets of probability distribution: the first one is the probability distribution of the topics over the documents and the second one is the probability distribution of words over topics. LDA models use an Expectation Maximization algorithm in order to infer these two distributions. In our situation, the documents represent the aggregated texts of different users, while the terms represent the most frequent words in their utterances. The most important hyper-parameter which needs to be tuned when using Latent Dirichlet Allocation is the number of topics sought. There is no exact way of choosing this hyper-parameter as it is difficult to asses the quality of the obtained models, because the setup is fundamentally unsupervised. After trying several values for the number of topics, over several time-frames, we concluded that 10 would be a suitable value for the number of sought topics.

Having the probability distribution of the topics over the users in a certain time-frame, we know exactly for each user the ratio of the topics he had been talking about, as discovered by the LDA model. In other words, each user can be described by a set of 10 features, representing the probability distribution over the 10 topics.

Our goal is to further process this representation so that we can obtain a better visualization and understanding of the social dynamics. First we compute the distances between

all users using Kullback-Leibler divergence [27]. This measure is not exactly a distance, because it lacks symmetry. However, an average of the KL values between two distributions of the users can be used as a distance in situations where we want to compute distances between probability distributions, as following

$$dist(u_i, u_j) = \frac{KL(u_i, u_j) + KL(u_j, u_i)}{2}, u_i, u_j \in \mathbb{R}^{10}$$

Next, we use this KL distance matrix to compute a lower dimensional representation of the data points. For this, we use t-Distributed Stochastic Neighbor Embedding [30] (i.e t-SNE). This is a probabilistic machine learning algorithm used for embedding a representation in high dimensional feature space into a lower dimensional feature space, most often 2D, with the purpose of better visualizing the data. The technique builds a probabilistic distribution over all pairs of higher dimensional points in a way that similar entries will have a higher probability, thus they will be more likely to appear closer to each other. The representation obtained using t-SNE is not deterministic, but it provides a very good tool for visualizing and understanding the data.

Moreover, we use the original KL Divergence matrix to apply a k-Medoids algorithm in order to cluster the users represented in the topics space. It's important to notice that the clusters obtained are not linked to the topics discovered by the LDA model, yet. These groups are only obtained due to the relative distance between users, in terms of the topics they approach. However, the number of medoids we aimed was also 10, as the number of topics.

Figure 10 presents the 2D representation of the users in the topic space for two different time-frames, namely 2 and 14, after applying the clustering algorithm. In order to associate a topic to each cluster, we determined the top 10 most frequent words of each cluster and then compared these sets of words with the words obtained by the distribution of probability of words per topics obtained from the LDA model. The longitudinal character of this study is highlighted by the ability to compare different groups of users clustered according to their interests, from different time-frames. Thus we can observe the evolution of topics around Brexit.

4.3 Topic change between periods

Table 1 presents the most frequent terms for each topic. By aggregating these terms with the corresponding cluster in Figure 10, we can observe a clear evolution of the topics discussed around Brexit. While it is true that some subjects are common, such as economy, traits of democracy (majority, decision, vote, referendum) or employment, there are also some subjects specific to each period of time. For instance, in Period 2, people talked more about the agreement with the European Union, immigration and how it affects British population or campaigns lead by some pro brexit political leaders, such as Boris Johnson. In Period 14, which is chronologically located towards the end of the studied period, other topics emerged such as the need of an extension of the exit period, a petition for remaining in the EU which would cancel the initial referendum. These lead to other topics such as the importance of the first referendum in the context of a democratic

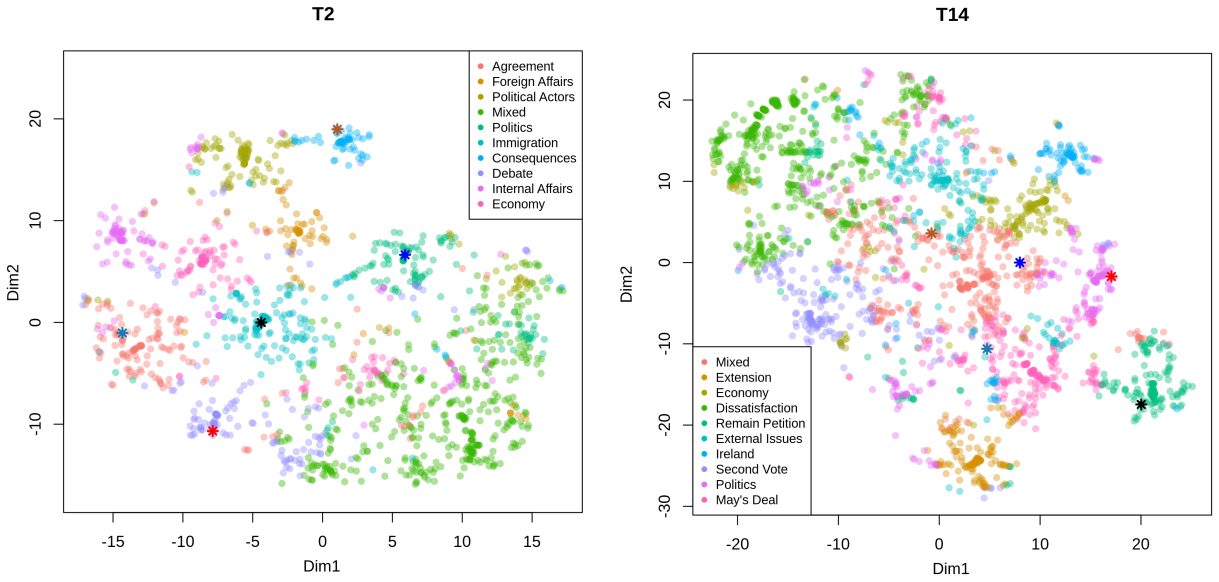

Figure 10: T-SNE embedded representation of clustered users in Topic Space for Timeframe 2 and Timeframe 14.

Table 1: Most frequent terms used in each topic in Period 2 and Period 14.

Cluster	Period 2	Period 14
1	*trade, deal, market, agreement, free, work, negotiate*	*leave, make, point, way, see, reason*
2	*germany, europe, greece, france, union*	*parliament, withdraw, extension, may, deal, article, agreement, vote, european*
3	*leaver, referendum, cameron, boris, campaign, article*	*government, work, money, pay, business*
4	*see, way, reality, leave, country*	*f*ck, country, need, right, s*it, much*
5	*referendum, democracy, parliament, government, vote, majority, decision*	*petition, sign, remain, signature, million, article, email, vote*
6	*immigrants, f*ck, work, live, racist, job, education*	*country, state, govern, power, nation, european, member, law, manifestation*
7	*country, leave, work, problem, manifestation*	*border, ireland, deal, leave, custom, trade, northern irish*
8	*leave, remain, make, research, argument, based, interest*	*referendum, leave, result, democracy, second, vote, election, change*
9	*european, union, british, leave, nation, way, right*	*may, british, britain, million, theresa, tori, minister, london, political*
10	*pound, economy bank, market, value, currency, price, drop, money*	*deal, may, option, vote, mps, negotiation, party, parliament, option*

choice. Another emerging topic is the relationship England has with the other countries in the United Kingdom, namely Ireland or Scotland. Finally, events such as Theresa May trying to get passed the deal in the Parliament are also reflected in the topics people discussed on Reddit.

4.4 Users changing neighbourhoods

In Figure 10, the colored stars represent common users between the two time-frames. We can observe from this figure that not only topics evolve in time, but also users themselves. For instance, the user represented by the black star appears to be more interested in the immigration problems in the first time-frame, whereas in the second one he is more involved in the discussions related to the petitions that were proposed as a solution from the remainer side. Even if we do not know its stance on these problems, we have a clue about its interests.

Another example is the red star, which in the first period is more into the discussions about informative decisions, involving research and argument, before making a decision. In the second period, he is more into the discussions about Theresa May's political decisions.

The blue star user is a good example of someone to migrated from concrete subjects like the referendum and democracy related topics to a more general area, represented by the red cluster in the right figure. Nonetheless, we can also observe that both figures have clusters of generalities: the green cluster in the left figure and the red cluster in the right figure.

5 Political stance classification

5.1 Methodological approach

In this subsection, we will describe the methodology proposed for solving the problem of predicting the stance of a participant in the Brexit debate on Reddit. In other words, by being given a user who posts messages in the *brexit* subreddit, either initial thread starting messages or comments in other cascades, we aim to predict the character of his future message, which can be either pro Brexit, against Brexit or neutral, by taking into account various crafted features which do not consider the text itself, but the composition of the diffusions in which the three categories of users engage.

The first step in the proposed solution consists of cleaning the data acquired from Reddit, after which the Event-Based 15-slice time splitting procedure is applied. This process will be detailed later in the following sections. After, this step, we build the features required for the predictive models. Even though most of the features are defined based on the structure of the diffusions users take part in, a feature that would describe the stance was needed. We needed labels for each user, labels that could tell us if he is pro or against Brexit. In fact, based on the value of this variable at the current time-frame, we aim to predict its value at the following time-frame. In order to obtain this partition of the users, more approaches were tried: clustering in terms space using cosine distance, clustering in

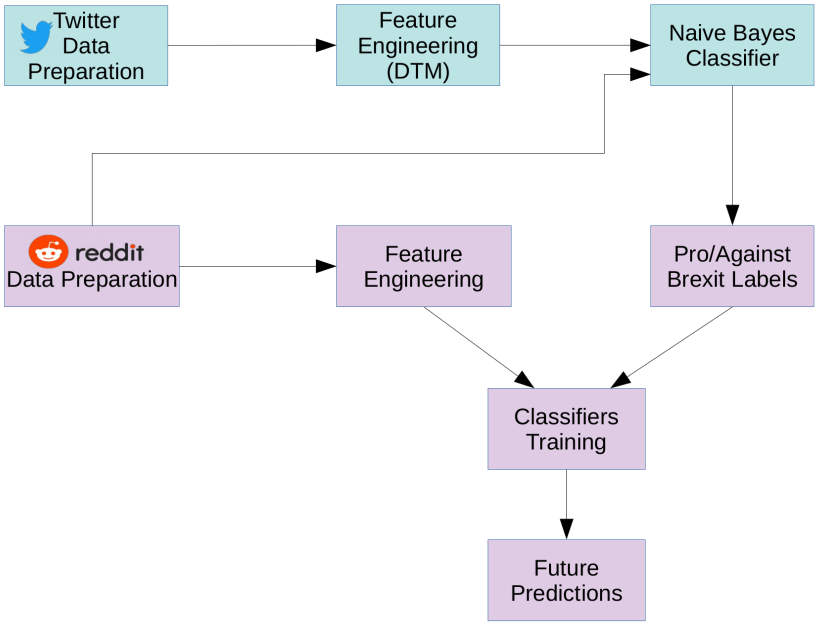

Figure 11: Schema of the proposed strategy for solving the stance prediction problem on Brexit topic.

topic space using Kullback-Leibler Divergence and sentiment detection. However, even though the results obtained with the first 2 methods are interesting and worth mentioning, we found another solution for obtaining our labels.

A final solution was engaged: a classifier was built on a Twitter dataset, taking into account the hashtags as clues for the membership of users to different partitions. After this Naive Bayes classifier is obtained, we use it on our initial Reddit dataset in order to get an automatic partition of the Reddit users in three categories: Brexit, Against-Brexit, Neutral.

At this point, we have the messages split in 15 groups, according to the date they had been sent. We aggregate them by the authors and label the membership to different categories in every time frame. A family of three different features is computed for each author, which take into account the consistence of the diffusions the authors had been active in. Based on this training data, we build 5 different models (Logistic Regression, KNN, Random Forest, Gradient Boosting [14], XGBoost [9]) to try to predict the future stance of a user, in a consequent time-frame.

5.2 Political stance textual classifier

In order to build the model which would be able to predict the stance of a person on the topic of Brexit, based on previous activity and interactions, a solution to quantify and assess the membership of a participant to a certain group was developed. We built a multi-class classifier using the Twitter Dataset, presented in Section 3.1, following the methodology presented in [4]. The main reason for using the dataset collected from

Twitter platform is the ease of classification which characterizes Tweets, because of the high usage of hashtags and mentions. The aim is to use this classifier in order to label the entries in our final Reddit dataset, based on the posts they had previously submitted, as **A**gainst brexit, **B**rexit, and **N**eutral. Using these labels enables us to predict the future side a participant will be on.

Data Preparation

To preprocess the Twitter data and build the classifier, we used exhaustively the Quanteda framework [5]. It provides very good functions for dealing with textual data: preprocessing tweets, building the training corpus, as well as fitting a Naive Bayes model, given a document term matrix and training labels.

The first step which was made was to label the training data, by creating two categories: pro Brexit and against Brexit. To do this, the following methodology was employed. Out of the 26.5 millions of tweets, we removed retweets and duplicate text. On Twitter platform, a frequently observed habit is to *retweet* a tweet. This means that you re-post a message, previously submitted on the platform by another user. This can be regarded as a way of giving credit and agreeing with their position, reinforcing their point of view. Such kind of a message can also contain text added by the person who shares the initial tweet, in which situation it turns into a *quote*. We aim to remove tweets which are not unique, because our classifier will not gain any insights from already seen messages. After performing this step, 7.6 millions of unique tweets remain for our study.

Next, we aggregate the tweets by authors, so that we can continue the filtering process, with a more compact dataset. After this step, we obtain a set of 1.5 millions of unique Twitter users along with their unified texts. The data is very skewed in terms of number of posts per user. Less than 10% of all users sent more than 10 tweets in the period which the data has been collected.

As hashtags give important clues and hints about the membership of a person to a certain group, the following step naturally consists of filtering out the tweets which do not contain relevant hashtags for our study. These hashtags are presented in Table 2. After performing this operation, the size of the dataset drops significantly to 136000 entries.

Table 2: Hashtags used for splitting Twitter users in two categories, to train the Naive Bayes Classifier.

Stance	**Hashtags**
Pro Brexit	#voteleave, #inorout, #voteout, #takecontrol, #borisjohnson, #lexit , #independenceday, #ivotedleave, #projectfear, #britain, #boris, #go, #projecthope, #takebackcontrol, #labourleave, #no2eu, #betteroffout, #june23, #democracy
Against Brexit	#strongerin, #intogether, #infor, #votein, #libdems, #voting, #incrowd, #bremain, #greenerin

In order to keep only the most vocal users and at the same time to discard occasional

users, who may not have a very well contoured opinion on the Brexit problem, we further filter the dataset, preserving only the users who sent at least 50 messages in the studied period. This step reduces our initial dataset to 11277 unique users.

Further on, a leave score, as described in [4], is computed for each user i, using formula:

$$LeaveScore_i = \#LeaveHashtags - \#RemainHashtags$$

This score allows us to rank users from the most vehement *leavers* to the most vehement *remainers*. This ranking is used to pick the top 10% most enthusiastic Brexit supporters and the top 10% most enthusiastic remain supporters. Thus, we obtain a training corpus of **2257** documents, representing the aggregated text of these users.

Feature Engineering

Dealing with textual data, a bag of words model is employed to build the training features. When building the Document Term Matrix, the weighting strategy is the well-known term frequency approach. As we are aiming to use this classifier on Reddit Data, which fundamentally does not contain hashtags or mentions, we remove all the terms features related to these elements, specific to Twitter. Website URLs, punctuation signs, numbers and English stopwords are also removed, after which all terms are lowercased and stemmed. Moreover, only terms with a frequency higher than 5 are kept in order to reduce the size of the learning vocabulary. After performing all these steps, a final Document Term Matrix having 2257 documents (aggregated Tweets belonging to the same author) and 19842 terms is obtained.

Model

The previous Document Term Matrix is used to estimate a Naive Bayes model which will output two sets of probabilities: the probability of a term to belong to one of the two classes, Leave or Remain and the probability of a document to belong to the two classes. For this study, the second sets of probabilities is particularly interesting. The output probability is converted then into a discrete category: if the leave probability is below 0.25, then the label is Remain, if the leave probability is greater than 0.75, the label is Leave, otherwise, the label is Neutral. The model is built using a multinomial distribution, with uniform prior. To build the model a splitting in two subsets for training and testing was performed, with a ratio of 80% training and 20% testing data.

Table 3: Accuracy, precision, recall and F1 score for the estimated Naive Bayes Model used for predicting user's stance based on their submitted posts.

Set	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Train	0.9419	0.9269	0.9572	0.9418
Test	0.8936	0.8826	0.8910	0.8868

In order to asses the quality of the classifier, both Accuracy and F1 Score have been reported. As Table 3 shows, the estimated model tends to perform well both on the

training set and on the test set. As this is just an intermediate phase and not the final goal of this work, an accuracy of 89.36% and a F1-Score of 88.68% are acceptable results for this step.

5.3 Reddit Stance Predictor

The third main contribution of this study is represented by the Reddit Stance Predictor. The aim is to train a model that is able to predict future stances of online platforms users, by analyzing and understanding the structure of the diffusions they have been part of and also the information they have been exposed to, in a temporal manner. The working pipeline of this part of the study is depicted in the second row of Figure 11.

Once the reddit is preprocessed accordingly, it is fed to the Naive Bayes Classifier which offers for every user a label reflecting its attitude towards the Brexit Issue. At the same time, the aggregated Reddit data is used to build features based on the structure of diffusions. These features will be described in the second subsection. Having the labels and the features, five different models are trained for predicting the future stance

The evaluation metric used is F1 score, because the dataset turns out to be highly imbalanced in favor of neutral participants, while pro and against Brexit are less present. These models allow making predictions about the future positions of the participants in the issue of Brexit.

Data Preparation

As Figure 11 shows, the first step consists of preprocessing the data, when all the submissions, initial thread starting messages or comments, are split temporally according to the heuristic detailed in Section 4.1. This splitting is needed as the topics of discussions evolve over time and the different behavioral characteristics and traits of the participants can be learned more effectively if each time-frame is analyzed individually.

Another step in the data preparation is the clean-up of the user aggregated texts, in terms of replies. Reddit allows users to reply to other users by quoting. When this happens, the reply automatically incorporates the previous text. This chunk of text needs to be removed as it can lead to misleading results when the Political stance textual classifier is applied in order to obtain the polarity of the user replying.

After splitting, aggregating and cleaning the utterances, the actual training set is built. Consequent time-frames are taken two by two, and common authors are extracted. Based on the text in the first time-frame we compute a leave score using the Political Stance Detector trained on the Twitter Dataset, applied on each author aggregated text. This score will be used for building the current stance, a variable used for learning. At the same time, we use the second time-frame to compute the future stance, the position around Brexit in the following period of time. This variable will be the learnable label, which we aim to predict on new, unknown data. Moreover, the first time-frame of the pair is used to build the features which will be described in the following subsection.

Feature Engineering

Four classes of features are proposed. They are built using different strategies and all of them include information about the stance of a user at the current time-frame.

1. **FS1 - User activity**
   - number of initiated threads;
   - number of submitted comments;
   - number of received replies per comment;
   - stance at current time-frame;
2. **FS2 - User activity per group**
   - number of initiated threads;
   - number of submitted comments;
   - number of received replies per comment from each group (**A**gainst, **B**rexit, **N**eutral);
   - number of submitted comments to users from each group (**A**, **B**, **N**)
   - stance at current time-frame;
3. **FS3 - Structure of diffusion**
   - ratio of comments from each group (**A**, **B**, **N**), in the diffusions the user takes part in;
   - stance at current time-frame;
4. **FS4 - All features**
   - FS1 + FS2 + FS3

FS1 - the first set of features focuses on the activity of the user around Brexit. We count the number of initiated threads (original posts in a diffusion), the total number of submitted comments and the post success, quantified by the total number of received replies at each comment. If a user sends 100 comments, the number of replies for each comment are counted, thus resulting in 100 values. To summarize this information, we compute the 5 quantiles (0%, 25%, 50%, 75%, 100%), which leads to a synthesized set of 5 values.

FS2 - this set of features focuses on the activity of user at a more granular level. We are particularly interested in the number of replies from **A**gainst side a user receives for his comments, number of replies from **B**rexit side a user receives for his comments and the number of replies from **N**eutral side. We apply again the same quantile based summarizing. Moreover, for this feature set, we also take into account the number of submitted comments to posts belonging of users from every side.

FS3 - this set of features aims to describe the structure of the overall diffusions a users takes part in. The key difference from the first two sets, FS1 and FS2, is the fact that here we consider diffusions as a whole. We count the number of posts of each kind from the diffusions a user takes part in and perform the quantile summarizing.

FS4 - this is the aggregation of the above mentioned feature sets.

All 4 feature sets have one common variable: the stance at the current time-frame. This is a label obtained using the classifier trained on the Twitter data and applied on the Reddit data. It is a categorical variable which can be either 0, 1 or 2, meaning **A**gainst, **B**rexit or **N**eutral. However, the Political Stance Detector trained on the Twitter dataset is a Naive Bayes Classifier, which outputs the probability P, that a user is a supporter of Brexit. In order to compute the categorical value, the following bounds have been used:

- $0 \leq P \leq 0.25 \implies$ **A**gainst brexit
- $0.25 < P < 0.75 \implies$ **N**eutral
- $0.75 \leq P \leq 1 \implies$ **B**rexit

After building the new training set, using the common authors between every two consecutive time-frames, the features described above and the current stance we obtain duplicate entries in the training set which have different target labels. For instance, for a training element, we can have 5 appearances, having the ground truth (the future stance) 0, 0, 0, 1, 2. This means that in 3 out of the 5 cases, users attitude as revealed by their aggregated posted messages in the immediately following time-frame changed to being Against Brexit, while one 1 user's attitude out of 5 changed to pro Brexit. To deal with this situation, a majority vote is applied and the ground truth will be 0 (stance at next time-frame).

Moreover, as we are dealing with a highly imbalanced dataset in favor of neutral users, we perform a guided sub-sampling, by removing training entries who have a stance transition of Neutral to Neutral between two consequent time-frames and appear only once. In other words, we want to keep only those training entries which have a transition from Neutral to Neutral and appear at least twice in the training set, when comparing the training features.

After applying these filtering techniques for reducing the imbalance, our training set has 1753 entries for FS4, 1039 entries for FS3, 1044 entries for FS2, and 385 entries for FS1.

Models

Five different Machine Learning algorithms were trained using the features described before: Logistic Regression, KNN, Random Forest, Gradient Boosting, XGBoost. The aim is to predict the stance around Brexit in a future time-frame by using information from the current time-frame.

All models have been trained using a double Cross Validation methodology. First, we use a 10-fold Cross Validation, to split the data into training and testing set. Then, the 9

parts picked for training are again subject to an inner 5-fold Cross Validation for tuning the hyper-parameters. This second Cross Validation is repeated for 500 iterations, while the first Cross Validation is repeated 10 times. Therefore, for each model, we have 10 values corresponding to the 10 main repetitions for each evaluation metric: F1, Accuracy, Precision and Recall. We report the mean and the standard deviation for the obtained Accuracy and F1-Score.

6 Results

To evaluate the models, we computed the accuracy and the F1 score for each classifier with the proposed feature sets. The training set is considerably imbalanced in favor of Neutral stances. It is for this reason that the F1 score was computed along with the accuracy. The results can be observed in Figure 12.

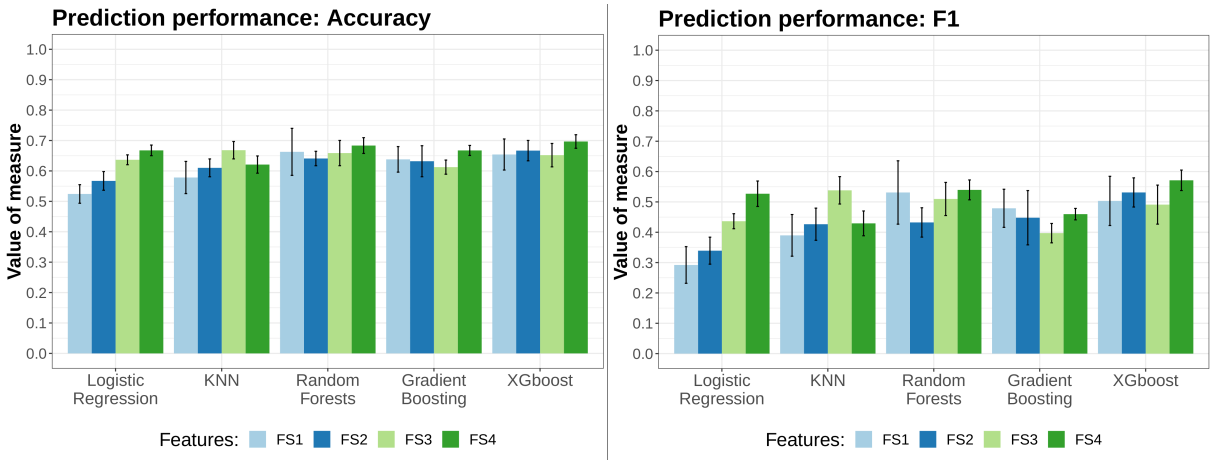

Figure 12: Evaluation metrics for the developed models: a) Accuracy (left). b) F1-Score (right).

We consider the **baseline F1 score and accuracy to be at most 33% as the problem is an imbalanced multi-class classification problem with 3 classes to predict**: Against, Neutral, Brexit. We can observe in Figure 12 that the best classifier reached about 57% F1 score, which is double the random guessing score. On the other hand, as the data is imbalanced, the accuracy is a bit higher, on average reaching values of 70%.

In general, we can observe that most classifiers have an ascending F1 score when going from FS1 to FS4 because complexity of the features and the expressivity grows. FS1 mainly describes the general activity of the user, while FS2 and FS3 deal also with the membership to the three different classes, Against, Brexit or Neutral. Finally, FS4, which is a combination between the 3 previous sets performs best on 4 out the 5 models, the only exception being KNN. Perhaps this is due to the fact that for FS4 accumulates 34 features in total, which may be a bit too much for the amount of training examples we have for KNN.

In order to better understand the outcome obtained and presented in Figure 12, we performed a deeper analysis of the results. We select the best trained classifier, the XGBoost, and the best feature set, FS4, as showed by the previous figure. Then, we split the overall, general F1 score obtained in 9 different F1 scores, corresponding to the 9 possible transitions between the 3 stances. The values are reported in Table 4.

Table 4: F1 score for every transition between the stances of the users.

Currently \ *Following*	**Against**	**Brexit**	**Neutral**
Against	0.26	0	0.98
Brexit	0	0	1
Neutral	0.73	0.63	0.45

As depicted in Table 4, our classifier performs notably well on the transitions from initially Neutral stances to pro or against Brexit stances in consecutive time-frames, obtaining 0.73 respectively 0.63 F1 scores. This result is particularly important as the key of the study is defined by predicting the future positions of undecided participants in online debates and for this, the XGBoost obtains very good results.

Furthermore, in order to explain these results, we compute volume of users corresponding to each of the 9 transitions. The results are shown in Table 5. This table explains the values from Table 4. Firstly, the number of users having a Neutral - Neutral trajectory is considerably small, because of the applied filtering presented in Section 5.3 - Feature Engineering, namely the removal of entries having translation Neutral - Neutral which appear only once, for a given set of input features. Next, we obtain very low scores for the transitions from a pro Brexit stance to an against Brexit stance and vice-versa because the number of examples of users in these situations is very small compared to the other categories, 35, respectively 33.

Table 5: The volume of users in the testing set for each category of transitions.

Current Stance	**Following Stance**	**Number of Users**
Against	Against	158
Against	Brexit	33
Against	Neutral	371
Brexit	Against	35
Brexit	Brexit	60
Brexit	Neutral	332
Neutral	Against	387
Neutral	Brexit	350
Neutral	Neutral	27

Indeed, the chances that a user will change his position in two consecutive time-frames from totally Brexit supporting to totally against Brexit are low and this is revealed by the

distribution presented in Table 5. In most situations, there will be a transition through the intermediate Neutral state, also shown by Table 5, the number of users going from Against and Brexit to Neutral being 371 and respectively 350. This allows our classifier to understand the underlying structure of the dataset, leading to high F1-scores for these translations.

The XGBoost classifier obtains an overall F1-score of 0.57, using the FS4 feature set. We can identify two main sub-components of this score: the first one is represented by the low scores obtained when predicting transitions Brexit - Against or Brexit - Brexit, which lowers the overall score. The second sub-component is defined by the predictions involving the Neutral state, namely from Neutral to the other three states.

Predicting the following stance a user will arrive into turns out to be a difficult task, as we learned from the previous results. In Figure 13 we show the volume of users translating from neutral state to a Brexit supporting state or Against Brexit state, in consecutive time-frames. The left figure shows the percentage of users, while the right figure shows the exact number of users going to the two polarized states. We can observe that except the starting time-frame, where there was a strong campaign for Brexit which is reflected in the ratio of users who migrated to a Brexit stance in the second time-frame, usually the trend is in favor of the Against side. In translations from T2 to T12, neutral users tended to move to Against Brexit. However, in T12, there is a change in the trend.

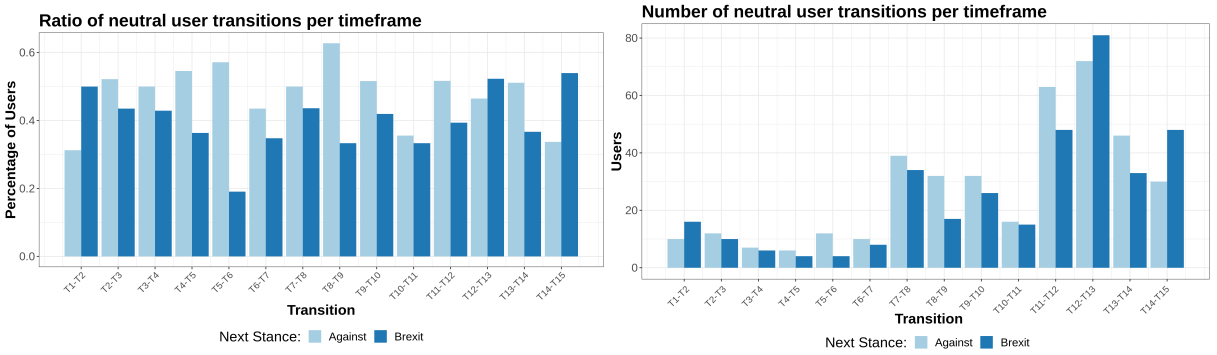

Figure 13: a) Ratio of users transitioning to the other stances in consecutive time-frames. ratio. b) Number of users transitioning to the other stances in consecutive time-frames.

We performed some analysis on this period and found out that the main event in this period of time was the second negative vote in the Parliament for the Withdrawal Agreement negotiated by Theresa May. UK would have had to pay the European Union 39 billion pounds, which was upsetting and disappointing for people, as revealed by some of the utterances we checked (we checked utterances with high leave probability): "*It's taken far too much time, we should leave hard and deal with the consequences. It will be tough for a time but there's no price not worth paying for freedom from the globalist overlords... "It's **better** to **die free**, **than live** as a slave." F. Douglass*" or "*Democracy? We roam the world dismantling dictatorships to install democracy, leaving failed states, but we cannot deliver the democratic will of our own population!* ". We assume people

tended to be disappointed by the incapability of the politicians to deliver the Brexit, thus the raise in Neutral people from time-frame 12 transitioning to Brexit in time-frame 13.

We tested our best predictor, the XGBoost using FS4, the feature set comprising all the other sets, on users on this period of time. Thus, we considered time-frame 12 and tried to predict how many of them will have a trajectory towards the Brexit side and how many will migrate towards the Against Brexit side. The results are shown in Table 6.

Table 6: Number of neutral users in time-frame 12 predicted to migrate to each of the other stances in time-frame 13.

Against	**Brexit**	**Neutral**
71	83	4

Even though the task of predicting the future stance of groups of users is difficult, we manged to correctly predict towards which group will neutral users migrate in a consequent period of time. This shows that our predictor correctly captured the trend and the underlying dynamics of the online population, despite the sudden change of tendancy, after a long series of similar inter-period transitions. Such kind of results can be very interesting for polling agencies trying to discover before-hand the outcome of important events debated online. What is more surprising, these results are obtained without the involvement of textual information at all in the feature sets, but just with taking advantage of the interaction between people and the diffusion of information.

7 Conclusion and future work

In this project we aimed to perform a thorough analysis of the way information diffusion affects participants in social media platforms. In order to clearly capture the dynamics and the intertwinnings of the online communities, we chose Reddit platform. Not only does it offer structured information, but also a complete range of opinions, often expressed in antithesis. The main subject of our study is Brexit, due to its polarity character.

Firstly, we performed a longitudinal temporal analysis of the threads of discussions around Brexit and built a tool that enhances the visualization of the dynamics of discussion topics and tracks the users as they tackle different topics.

Secondly, we build a future political stance predictor, based on the online diffusions structure. Again the target platform where we perform the analysis is Reddit, however, we needed to train a side model on Twitter data, in order to provide ground truth labels for the initial Reddit training data.

At the moment of writing the report, the results are intermediate and can be regarded as a proof of concept, as important improvements will be made in the remaining time of this internship. Namely, we aim to replace the political stance detector trained on Twitter, with one that will be trained on Reddit data. With no doubts, the underlying distribution and structure of the two platforms differ, so labelling training elements on Reddit with a classifier trained on Twitter brings inexactity.

In the following month, we will build a proper Reddit training set for political stance detection, by buying labelling services. Even if the number of Reddits is very large, we plan to use the leave probability output by our current Twitter classifier as a **propensity score**. More exactly, we will use the classifier we trained on Twitter to sort the users according to the leave probability, then select a number of users spanning all over the spectrum, from Against to Pro Brexit and we will have these users' texts labelled again by humans. Next we will train a Naive Bayes classifier on this dataset, for detecting political positions.

At the same time, improvements will be made on the features sets proposed. We aim to integrate not only users' activity and interactions, but also important textual contents and information derived from their usernames. For this, we will build features based on the Document Term Matrices and where the terms will be the highly used terms by the two sides.

Bibliography

[1] Aboufarw, K., Grigorev, A., and Mihaita, A. (2022). Traffic accident risk forecasting using vision transformers,. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*.

[2] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM.

[3] Bastos, M. T. and Mercea, D. (2019). The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54.

[4] Benoit, K. and Matsuo, A. (2018). Network analysis of Brexit discussion on social media. -.

[5] Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., and Matsuo, A. (2018). quanteda: An r package for the quantitative analysis of textual data. *J. Open Source Software*, 3(30):774.

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[7] Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the tenth international workshop on multimedia data mining*, page 4. ACM.

[8] Celli, F., Stepanov, E., Poesio, M., and Riccardi, G. (2016). Predicting brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 110–118.

[9] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

[10] Choi, D., Han, J., Chung, T., Ahn, Y.-Y., Chun, B.-G., and Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 acm on conference on online social networks*, pages 233–243. ACM.

[11] Colleoni, E., Rozza, A., and Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332.

[12] Dawson, N., Rizoiu, M.-A., Johnston, B., and Williams, M. A. (2019). Adaptively selecting occupations to detect skill shortages from online job ads. In *Proceedings* -

2019 IEEE International Conference on Big Data, Big Data 2019, pages 1637–1643, Los Angeles, CA, USA. IEEE.

[13] Firehose, T. (2019). Twitter compliance firehose api. `https://developer.twitter.com/en/docs/tweets/compliance/api-reference/compliance-firehose`.

[14] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

[15] Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3):211–223.

[16] Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.

[17] Grigorev, A., Mihaita, A., Saleh., K., and Picardi, M. (2022a). Traffic incident duration prediction via a deep learning framework for text description encoding. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*.

[18] Grigorev, A., Mihaita, A.-S., Lee, S., and Chen, F. (2022b). Incident duration prediction using a bi-level machine learning framework with outlier removal and intra–extra joint optimisation. *Transportation Research Part C: Emerging Technologies*, 141:103721.

[19] Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28.

[20] Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.

[21] Howard, P. N. and Kollanyi, B. (2016). Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*.

[22] Hughes, A. L. and Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4):248–260.

[23] Issa, F., Monticolo, D., Gabriel, A., and Mihăiţă, A. (2014). An intelligent system based on natural language processing to support the brain purge in the creativity process. *IAENG International Conference on Artificial Intelligence and Applications (ICAIA'14) Hong Kong*.

[24] Kimura, M. and Saito, K. (2006). Tractable models for information diffusion in social networks. In *European conference on principles of data mining and knowledge discovery*, pages 259–271. Springer.

[25] Kong, Q., Rizoiu, M.-A., Wu, S., and Xie, L. (2018). Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pages 175–178, Lyon, France. ACM Press.

[26] Kong, Q., Rizoiu, M. A., and Xie, L. (2020). Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *International Conference on Information and Knowledge Management, Proceedings*, pages 645–654, New York, NY, USA. ACM.

[27] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[28] Kumar, R., Novak, J., and Tomkins, A. (2010). Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer.

[29] Lou, T. and Tang, J. (2013). Mining structural hole spanners through information diffusion in social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 825–836. ACM.

[30] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[31] Mao, T., Mihaita, A., and Cai, C. (2019). Traffic signal control optimisation under severe incident conditions using genetic algorithm. *Proc. of ITS World Congress (ITSWC 2019), Singapore*.

[32] Mao, T., Mihăită, A.-S., Chen, F., and Vu, H. L. (2022). Boosted genetic algorithm using machine learning for traffic control optimization. *Trans. Intell. Transport. Sys.*, 23(7):7112–7141.

[33] Mihaita, A., LI, H., and Rizoiu, M. (2020a). Traffic congestion anomaly detection and prediction using deep learning.

[34] Mihaita, A. S., Benavides, M., Camargo, C., and Cai, C. (2019a). Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models. *International Journal of Intelligent Transportation System Research (IJITSR)*, 17(2):125–141.

[35] Mihaita, A. S., Dupont, L., Cherry, O., Camargo, M., and Cai, C. (2018). Air quality monitoring using stationary versus mobile sensing units: a case study from lorraine, france. *Proc. of ITS World Congress (ITSWC 2018), Copenhagen, Denmark*.

[36] Mihaita, A.-S., Li, H., He, Z., and Rizoiu, M.-A. (2019b). Motorway Traffic Flow Prediction using Advanced Deep Learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1683–1690, Auckland, New Zealand. IEEE.

[37] Mihaita, A.-S., Liu, Z., Cai, C., and Rizoiu, M.-A. (2019c). Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. In *Proceedings of the 26th ITS World Congress*, pages 1–12, Singapore.

[38] Mihaita, A.-S., Papachatgis, Z., and Rizoiu, M.-A. (2020b). Graph modelling approaches for motorway traffic flow prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, page 1–8. IEEE Press.

[39] Mihăiţă, A., Camargo, M., and Lhoste, P. (2014). Evaluating the impact of the traffic reconfiguration of a complex urban intersection. *10th International Conference on Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014*.

[40] Mihăiţă, A. S., Tyler, P., Menon, A., Wen, T., Ou, Y., Cai, C., and Chen, F. (2017). An investigation of positioning accuracy transmitted by connected heavy vehicles using dsrc. *Transportation Research Board - 96th Annual Meeting, Washington, D.C.*

[41] Mihăiţă, S. and Mocanu, S. (2011). An energy model for event-based control of a switched integrator. *IFAC Proceedings Volumes*, 44(1):2413–2418. 18th IFAC World Congress.

[42] Mishra, S., Rizoiu, M.-A., and Xie, L. (2018). Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Con-*

ference on Web and Social Media (ICWSM '18), pages 1–10, Stanford, CA, USA.

[43] Monticolo, D. and Mihăiţă, A. (2014). A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication (IJFCC)*, 3(1):66–70.

[44] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

[45] Pushshift (2019). Pushshift. `https://pushshift.io/`.

[46] Rizoiu, M. A. and Velcin, J. (2011). Topic extraction for ontology learning. In Wong, W., Liu, W., and Bennamoun, M., editors, *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, pages 38–60. IGI Global.

[47] Rizoiu, M.-A. and Xie, L. (2017). Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International AAAI Conference on Web and Social Media (ICWSM '17)*, pages 182–191, Montréal, Québec, Canada.

[48] Rizoiu, M. A., Xie, L., Caetano, T., and Cebrian, M. (2016). Evolution of privacy loss in Wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 215–224, New York, New York, USA. ACM, ACM Press.

[49] Shaffiei, S., Mihaita, A., and Cai, C. (2019). Demand estimation and prediction for short-term traffic forecasting in existence of non-recurrent incidents. *Proc. of ITS World Congress (ITSWC 2019), Singapore*.

[50] Shafiei, S., Mihaita, A., Nguyen, H., Bentley, C. D. B., and Cai, C. (2020). Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.*

[51] Shafiei, S., Mihăiţă, A.-S., Nguyen, H., and Cai, C. (2022). Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, 14(6):629–639.

[52] Tang, J., Sun, J., Wang, C., and Yang, Z. (2009). Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM.

[53] Unwin, J. T., Routledge, I., Flaxman, S., Rizoiu, M. A., Lai, S., Cohen, J., Weiss, D. J., Mishra, S., and Bhatt, S. (2021). Using hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17(4):e1008830.

[54] Wen, T., Mihăiţă, A.-S., Nguyen, H., Cai, C., and Chen, F. (2018). Integrated incident decision-support using traffic simulation and data-driven models. *Transportation Research Record*, 2672(42):247–256.

[55] Wu, S., Rizoiu, M.-A., and Xie, L. (2019). Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25.

[56] Wu, S., Rizoiu, M. A., and Xie, L. (2020). Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 715–725.

[57] Yang, J. and Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining*, pages 599–608. IEEE.

[58] Zhang, R., Walder, C., and Rizoiu, M.-A. (2020). Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6803–6810.

[59] Zhao, D., Mihaita, A., Ou, Y., Shafiei, S., Grzybowska, H., Qin, K., Tan, G., and Li, M. (2022). Real-time attention-augumented spatio-temporal networks for video-based driver activity recognition. In *Proc. of the IEEE Intelligent Transport Systems Conference 2022, Macao, China*.