Université Jean Monnet

Machine Learning and Data Mining Master

Information Diffusion in Online Communities

Advisors:
Christine Largeron
Andrei-Marian Rizoiu

Author:
Pilli Yogesh Kumar

Contents

1 Introduction

1.1 Context

Social media have radically changed the paradigm of news consumption and their impact on information spreading and diffusion has been largely addressed,. In particular, several studies focused on the prediction of social dynamics[19],[4]. with a special emphasis on information flows patterns and the emergence of specific community structures. Indeed, social media are always more involved in both the distribution and consumption of news.

According to a recent report[29], approximately 51% of users get news directly via social media, and such information undergoes the same popularity dynamics as other forms of online contents, such as selfies or kitty pictures. In the current, disintermediated environment users actively participate to contents production and information is no longer mediated by journalists or experts. Communication strategies and how messages are framed and shared across the social networks have changed.

Indeed, online disintermediation elicits users' tendency to select information that adheres (and reinforces) their preexisting beliefs, the so-called confirmation bias. In such a way, users tend to form groups of like-minded people where they polarize their opinion i.e., echo chambers[25],[34]. Moreover, confirmation bias plays a pivotal role in informational cascades. Experimental evidence shows that confirmatory information gets accepted even if containing deliberately false claims [1], while dissenting information is mainly ignored [14]. Debating with like-minded people has been shown to influence users' emotions negatively and may even increase group polarization [32].

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 Motivation

The first goal of this work focuses mainly on detecting and predicting the behaviour of people involved in Online Social Networks [23] 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 of 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 [33] 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 [22], [24], [35], and in particular on the subject on Brexit [6], 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 X and Y share the same opinions, if person X 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 Y will also do the same, if exposed to this sequence of messages. X 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

- Homophily of user through discussions around Brexit on social media platform as Reddit
- Tool to visualizing the EdgeHomogeneity between the Nodes on discussion thread on Brexit
- Prediction of User Political Stance based on the 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, Homophily and Social Influence on social network.

Section 3 presents the datasets involved in the development of our proposed approach and the way the data was pre-processed. We detail the platform used for collecting the data.We use Reddit dataset to train our stance detection classifier(pro, against or neutral around Brexit subject). Next, this classifier is 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 Homiphily and How social influence is effecting the user to change the behaviours.In this method we can see how user polarization are defined.We calculate the Edge homogeneity between the nodes and visualize how Edge Homogeneity plays a role of information diffusion between the users.

In Section 5, we present the methodological contribution proposed during this internship regarding the political stance prediction. 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 [18]. 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.

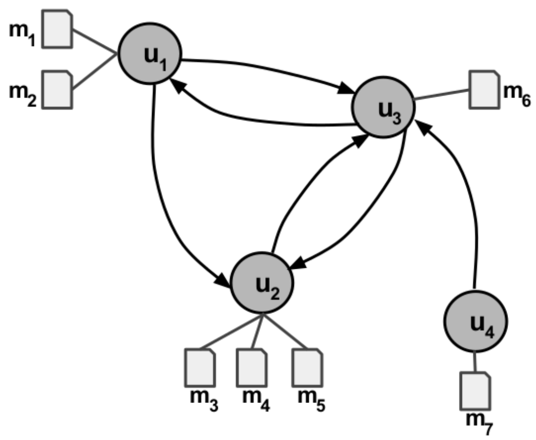

Figure 1: Online Social Network modelled as graph. Each vertex u_i is a user, while a directed edge (u_i, u_j) means user j exposes user i to a certain information, via message k, denoted as m_k Source: *Adrien Guille, Hakim Hacid, C. Favre, Djamel Abdelkader Zighed. Information Diffusion in Online So- 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 [16], [17]. On the other hand, the second category [20], [27] 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 Social Influence in the changing User behavior

In human societies homophily, the tendency of similar individuals getting associated and bonded with each other,is known to be a prime tieformation factor between a pair of individuals[26]. This association and bonding can be related to one or more features including gender, race, age, education level, economic and social status, and many more. Consequently homophily has a large impact on a number of fundamentally important social phenomena like segregation, inequality, perception biases, and the transmission of information between groups of individuals. The mechanism of focal closure, i.e., the process of forming links to others with shared characteristics but without common acquaintances can be considered as a typical manifestation of homophily[31].

2.3 Brexit Position Classification of Twitter Users

Another important piece of work that represents the starting point for our research is represented by [3]. 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, [8] 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 Dataset

In this section, we will detail about the dataset that were used during this study. We will present the Reddit dataset which was used for defining the partitioning of users with regard to the Brexit problem and how reddit dataset was helpful to detected the users with alike minded people while discussing various topics regarding the problems and benefit over the brexit.

3.1 Reddit Data

Reddit is a social media platform for online discussions to users share news,articles and opinions with eachother on the particular areas of user interest.The areas of topical interests in Reddit are called "subreddits", each of which serves as an independent community. A subreddit can be created by any user who is interested in any particular topic, e.g., game, politics, or sports. Each subreddit is managed by several"moderators"who are responsible for moderating and policing conversations among members. In each subreddit, users can (i) submit content (i.e., write a post), (ii) write a comment to a post, or (iii) write a comment to another comment.

Eventually, we find hierarchies of posts, in a tree like structure, as we can observe in Figure 2, 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:

- **Comment** = the chronologically subsequent messages posted as replies to a thread.

The dataset was collected using the Pushshift API [30], 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 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,submission id (this is unique for each thread),parent id (useful for building the tree structure),**

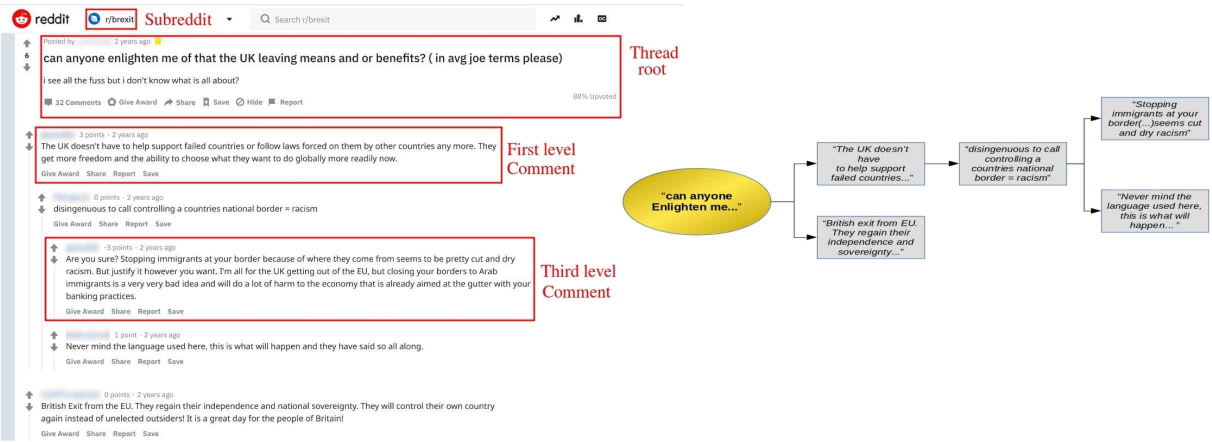

Figure 2: a) Elements of the Reddit platform. Structure of a discussion thread, with multi-level comments, inside a subreddit. b) Logical structure used for analyzing the data.

score and number of comments (specific for the root of the threads)[9].From submission id we can relate to find 22010 different threads with each thread has different number of participates in the discussion.

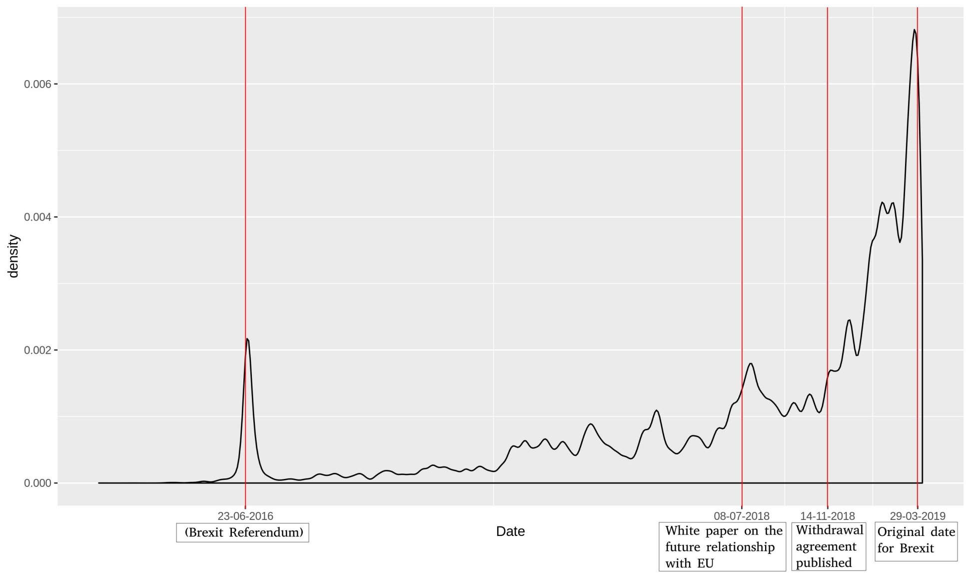

Figure 3: Time distribution of the submissions collected from Reddit (subreddit brexit), between November, 2015 and April, 2019.

We have time distribution of the collected message which is represent in Figure 3. This Figure illustrate an ascending trend,which points the growing importance of this brexit subject as it is perceived by usual people using social media platforms.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[5].However, 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.

Majority of the messages are comments as depicted in Figure 4 (a), as opposed to initial, thread starting messages. In terms of unique authors, Figure 4 (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[9]. The majority of the unique users are **only** commenters, meaning that they never start discussions, but usually engage in them.

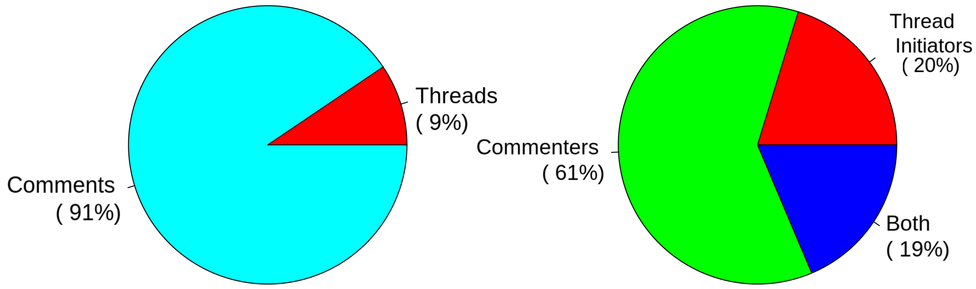

Figure 4: (a) distribution of submitted messages in terms of types of messages. (b) distribution of unique users in terms of roles.

From Figure 5 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[12]. 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.

4 Information Spread around Brexit on Reddit

Nowadays, social media platforms constitute a major component of an individual's social interaction. These platforms are considered a robust information dissemination tool to express opinion and share views. People rely on these tools as the main source of news to connect with the world and to get instant updates[28]. They have a major beneficial side that allows individuals to explore various aspects of an emerging topics, express their own points of view, get instant feedback and exploring the public views. The huge dependency of users on these platforms as the main source of communication allowed

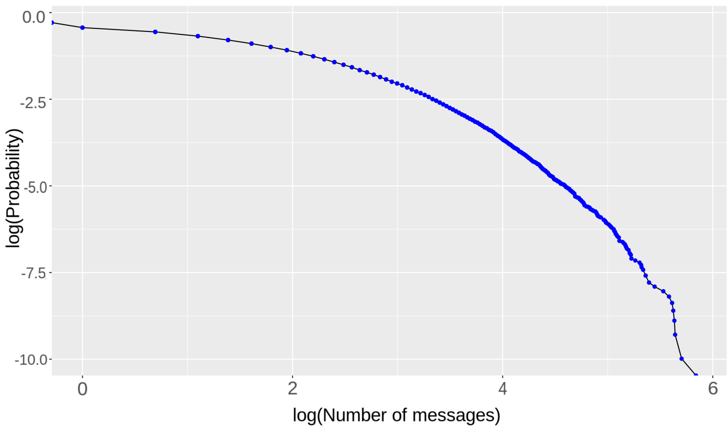

Figure 5: Complementary Cumulative Distribution Function of the number of messages sent by each user.

researchers to study different aspects of online human behavior, including public stance towards various social and political aspects[10]. Stance is defined as the expression of the speaker's standpoint and judgment toward a given proposition Stance detection plays a major role in analytical studies measuring public opinion on social media, particularly on political and social issues[2]. The nature of these issues is usually controversial, where people express the opposing opinions towards differentiable points[7].

4.1 Polarization and Friendship Network

A tree is an undirected simple graph that is connected and has no simple cycles. An oriented tree is a directed acyclic graph whose underlying undirected graph is a tree.we define a comment tree as an undirected tree, T = (V,E), where V is the set of all messages, which includes the original post (root) and all the follow-up comments in the thread, and E is the set of edges, each of which connects two messages that are linked by commenting. Figure 6 illustrates a comment tree that has one post and nine comments. A Comment tree, in the context of our research, is an oriented tree made up of the successive comments on the reddit discussion thread. The root of the comment tree is the node that performs the post on the reddit platform.

Previous research studies on Stance detection was obtained on using the classifier trained on the Twitter studies by [9].After applying this classifier on Reddit dataset. We obtained a categorical variable which can be either 0, 1 or 2, meaning Against,Brexit or Neutral and also we get the probability p which lies in $0 \leq p \leq 1$, that a user is supporter of Brexit. We define the user polarization $\sigma = 2p - 1$, where $p$is probability, and hence $-1 \leq \sigma \leq 1$. From user polarization we define edge homogeneity[34] , for any edge e_{ij} between the nodes i and j, as

$$\sigma_{ij} = \sigma_i \sigma_j$$

From the user polarization[25] we again divided the polarization to get the Stances labelled to the user.

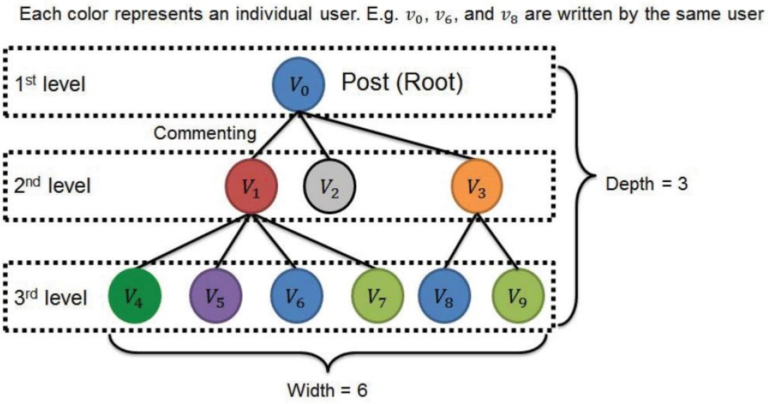

Figure 6: A comment tree is illustrated for a post that has 9 comments

- $-1 \leq \sigma \leq -0.2 \implies$ **A**gainst brexit
- $-0.2 < \sigma < +0.2 \implies$ **N**eutral
- $+0.2 \leq \sigma \leq +1 \implies$ **T**owards brexit

Firstly we take the polarization σ which lies in $-1 \leq \sigma \leq 1$. we divided the polarization to get the Stances. From -1 to -0.2 we assign this to Against brexit stance. In between -0.2 to +0.2 we assign them Neutral stannce. Finally from +0.2 to 1 we group it to ProBrexit stance.

From the Dataset Reddit, we get around 22010 different discussion threads where each discussion thread has unique submission id.

For better understanding, we have taken a smallest thread from the dataset with 12 comments made by several users involved in this discussion.We can see the discussion thread for the smallest thread with 12 comments in the Figure 7. The initial post is not considered as a comment. Comments correspond to the other nodes. Thus in Figure 7, we have a thread which is started by a post – also known as submission in Reddit terminology, the root of the tree – and all 12 other nodes denoted as comments – chronologically subsequent messages posted as replies to a post or to other comments in the same thread.

From Figure 7 we can say the each discussion thread is combined with all the User Stance and defined according to their polarization score.However, A thread begins by a root. A diffusion corresponding to a path from the root to a leaf in the tree. We can see the Jaygernaughty is the user who is source and all the other four are target node in this case.However the JaygerNaughty user acts as Target node to rbnc in one of the edges. Similarily JaygerNaughty user acts as source node when there is edge between JaygerNaughty user and uB166ERu user. Eventually source node and Target node keeps on changing according the edge

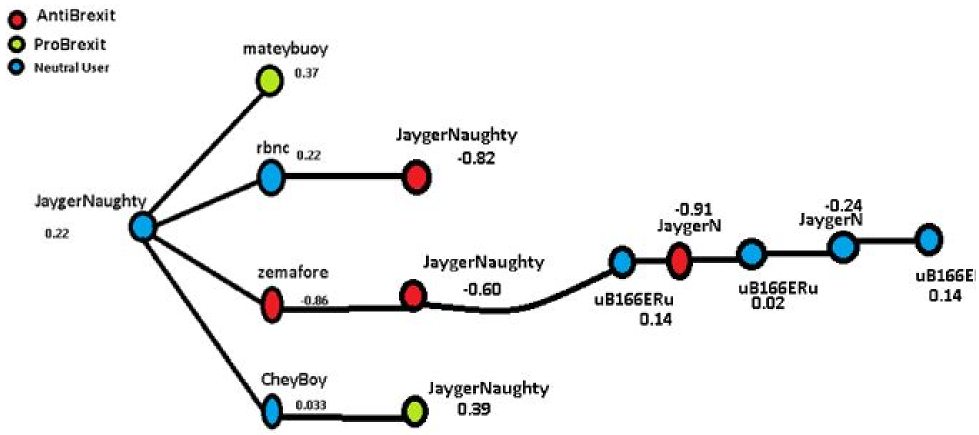

Figure 7: Smallest Discussion Thread in Reddit Dataset.

4.2 Homiphily: Engagement and friends

In this section we want to understand if the engagement in a specific kind of comment is a good proxy to detect group of users with similar attitudes. Homophily – i.e., the tendency of users to aggregate around common interests – has been already pointed out as a factor in changing the user behavior/mood swing[14]. Indeed, online content mingle cognitive,behavioral, and social aspects of a user community. The resulting ecosystem allows to investigate the various processes at play in the interactions of individuals, and to study the ways in which users relate with the kind of information they interact with. Thus, users' liking activity across contents of the different categories [34] may be intended as the preferential attitude towards the other user behaviour

4.3 Social Influence on changing user behaviour

Social influence a force that person A (i.e., the influencer) exerts on person B to introduce a change of the behavior and/or opinion of B Influence is a causal process.where the action of a user is triggered by others opinion. An example of this scenario is when a user buys a product because one of his/her friends has recently bought the same product This refers to the phenomenon that the action of individuals can induce their friends to act in a similar way. This can be through setting an example for their friends (as in the case of fashion), informing them about the action (as in the case of viral marketing), or increasing the value of an action for them (as in the case of adoption of a technology). In our context we would be interesting to know weather user behaviours are influences to change the user stance from the one stance to others.

Inorder to study the Homiphily and Influence , we take reddit dataset and define two cases as Case1 and Case2 . In Case1 we fix one stance as source node and target node is combined with all other stances.By this we can see which Target node is attracted towards the source node.Similarly we do viceversa of the Case1 with fix one stance as Target node and sources node to be combined with all other stances.By this we can say that which source node is attracted most by the target node.

To analyse the information transmission , we take two cases and each case has 3 different scenarios. Considering Case1 - Scenario 1 ,In this scenario the Source node is Neutral and Target node is combined with the all the other user stances(AgaintBrexit, Neutral and ProBrexit). This defines as the all the other user stances who are commenting towards the neutral stance user and by this we can see the which are the majority of the users attracted to the Neutral stance is this particular scenario. By calculating the count of the user who are interested to comment on the Neutral stance node through Pie-diagram in Figure 8

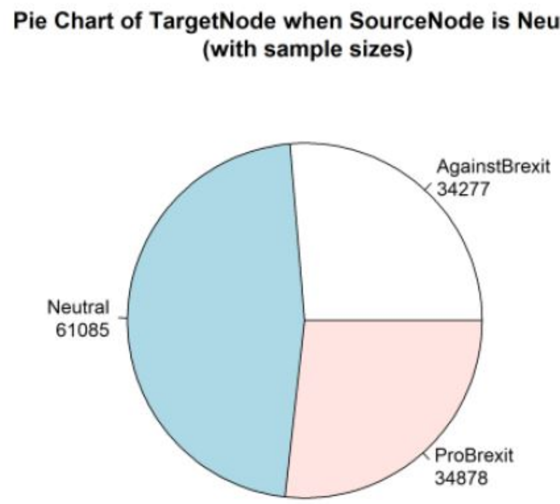

Figure 8: Pie-Chart of the Target Node Stance when Source Node is Neutral Stance

Majority of the Target node is Neutral stance in this case.Not only the Neutral stance node interacts with the to the neutral stance node, but also the AgainstBrexit stance and ProBrexit stance node interacts as well.

Considering the Case1- Scenario2 , In this scenario Source node is AgainstBrexit stance node and Target node is combined with all the other stance nodes.Technically this states the target node user who are interest and attracted towards the AgainstBrexit stance node.However we can observe the number of Target nodes who are attracted towards the AgainstBrexit stance in Figure 9. We can see that the when source node is Against majority of the users are attracted towards the Source Node are Neutral stance and followed by AgainstBrexit Stance and ProBrexit stance. User do interact with other stance when they are on discussion.

Moving forward to the Case1 - Scenario3, where in this scenario Source node is ProBrexit stance node and Target node is combined all the other stances nodes.Similarly as other scenario we can observe the majority of the target are Neutral followed by ProBrexit in Figure 10.However the user interaction results to the similarity as in scenario2 and scenario3 as ProBrexit stance node attracted by all the other stance users

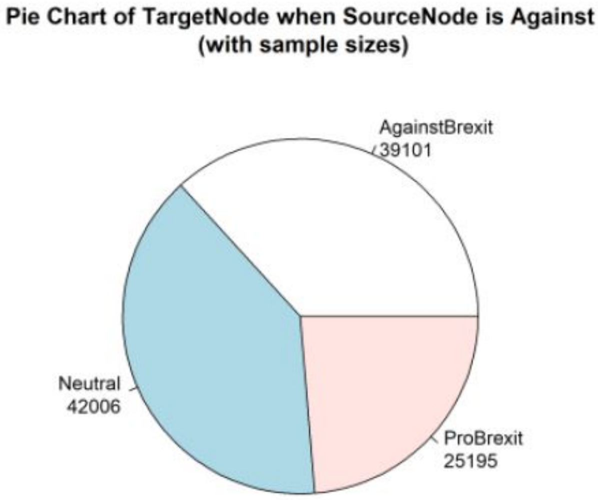

Figure 9: Pie-Chart of the Target Node Stance when Source Node is AgainstBrexit Stance

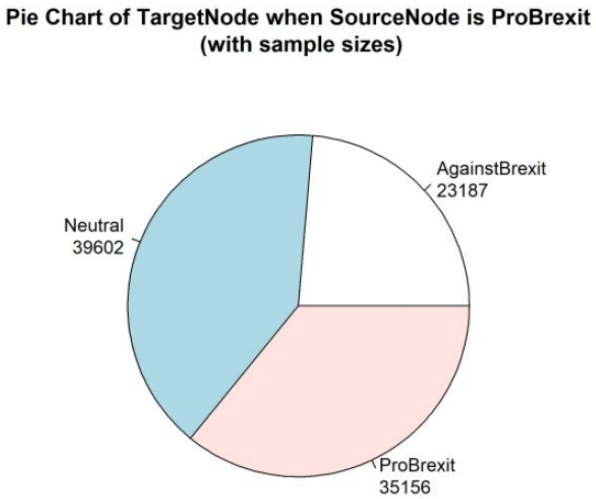

Figure 10: Pie-Chart of the Target Node Stance when Source Node is ProBrexit Stance

Getting into the details of Case2 with 3 different scenarios, Firstly we take Case2 - Scenario1 where Target node is Neutral and Source node is combined with all the other stances. Interesting part to be to see the behaviour of the which source node stance got attracted by Target node when it is neutral.We can see the results in Figure11. From the results we can say Neutral stance nodes are majority and it states that the neutral node attracted are mostly in this scenario.Also the other two stances got attracted too.

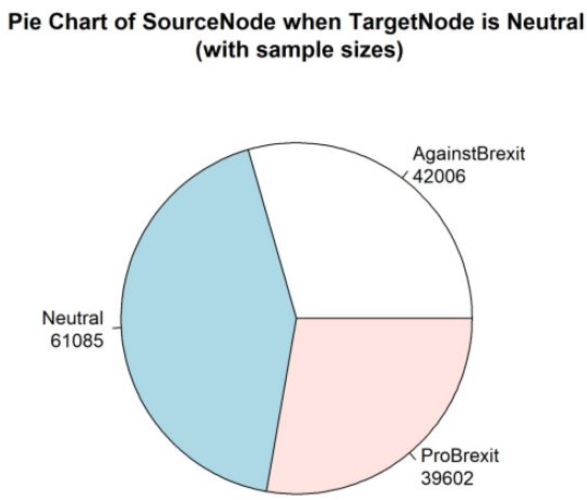

Figure 11: Pie-Chart of the Source Node Stance when Target Node is Neutral Stance

Coming to the Case2 - Scenario2, where we have Target node as AgainstBrexit stance and Source node as combined with all other stances.Likely to the other scenarios we also need to check the attracted source node towards the Target node stance.In Figure12. However here as we can see the majority of the Source node which got attracted towards the target node as AgainstBrexit are highly scale on the AgainstBrexit stance users.

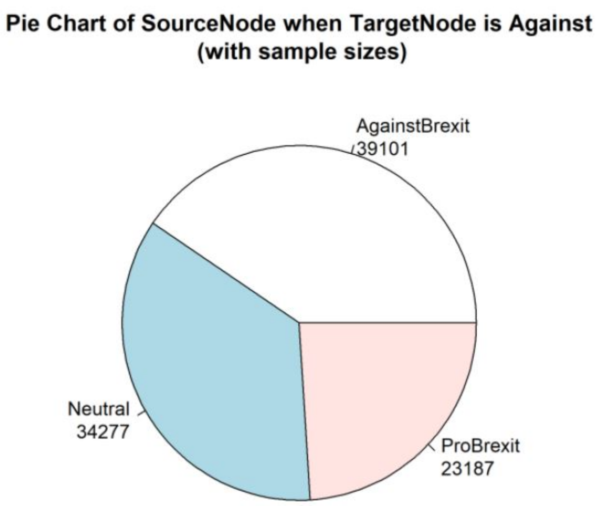

Figure 12: Pie-Chart of the Source Node Stance when Target Node is Against Stance

Lastly we have Case2 - Scenario3 were Target Node stance is ProBrexit stance nodes and Source node are combined by all other stance.Moreover we can see from Figure13 which node is attracted towards the target node.From the result we can clearly say the ProBrexit stances are more in number as the number of sample size is higher.Though other stances can also be seen in the Figure 13

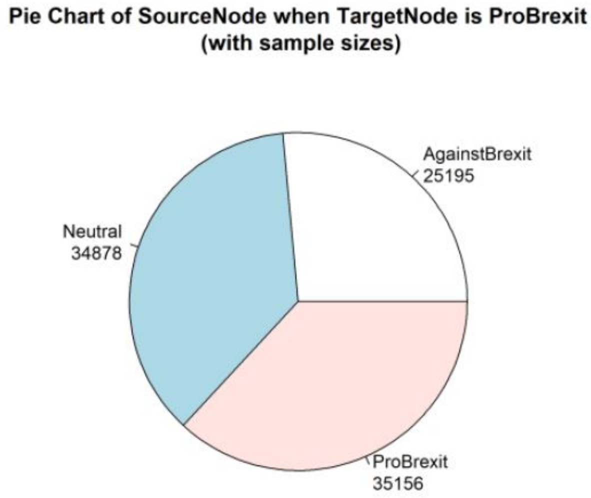

Figure 13: Pie-Chart of the Source Node Stance when Target Node is ProBrexit Stance

As of now, we can relate results in the different Cases with several scenario and analysis the user stance behavior according to the result shown.To get the information diffusion through comments and user interaction chambers, we have to plot the Probability Density Function of Edge Homogeneity[25],[34] which we discuss in results section later

5 Model Description

5.1 Methodological approach

In this subsection, we will describe the methodology proposed for solving the problem ofpredicting 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.We have considered the Long Short - Term Memory Networks (LSTM) as basemodel for the predicted the stance of the user

5.2 Long Short-Term Memory Networks

Recurrent neural networks (RNNs) are able to propagate historical information via a chain-like neural network architecture. While processing sequential data, it looks at the current input x_t as well as the previous output of hidden state h_{t-1} at each time step. However, standard RNNs becomes unable to learn long-term dependencies as the gap between two time steps becomes large.To address this issue,LSTM was first introduced in [21] and reemerged as a successful architecture since Ilya et al. [15] obtained remarkable performance in statistical machine translation. Although many variants of LSTMwere proposed, we adopt the standard architecture [21] in this work.

The LSTM architecture has a range of repeated modules for each time step as in a standard RNN.In In Figure14 at each time step, the output of the module is controlled by a set of gates in Rd as a function of the old hidden state h_{t-1} and the input at the current time step x_t: the forget gate f_t, the input gate i_t, and the output gate o_t. These gates collectively decide how to update the current memory cell c_t and the current hidden state h_t.

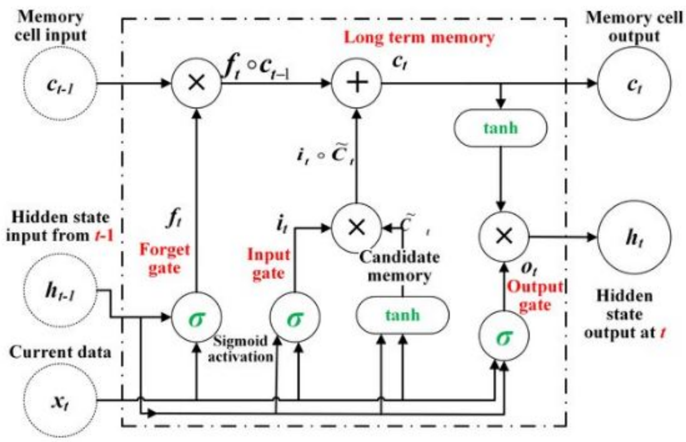

Figure 14: Architecture of a Long-Short Term Memory (LSTM) unit.

We used to denote the memory dimension in the LSTM and all vectors in this architecture share the same dimension. The LSTM transition functions are defined as follows:

- **Forgetgate**: This is a sigmoid layer that takes the output at t-1 and the current input at time t and then combines them into a single tensor. It then applies linear transformation followed by a sigmoid. The output of the gate is between 0 and 1 due to the sigmoid. This number is then multiplied with the internal state, and that is why the gate is called forget gate. If ft =0 ,then the previous internal state is completely forgotten, while if ft =1, it will be passed unaltered.

$$f_t = \sigma(W_f \; [h_{t-1}, x_t] + b_f)$$

- **Inputgate**: This state takes the previous output together with the new input and passes them through another sigmoid layer. This gate returns a value between 0 and 1. The value of the input gate is then multiplied with the output of the candidate layer.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

- **MemoryCell**: This layer applies hyperbolic tangent to the mix of the input and previous output, returning the candidate vector. The candidate vector is then added to the internal state, which is updated with this rule:

$$q_t = \tanh(W_q \cdot [h_{t-1}, x_t] + b_q)$$

  The previous state is multiplied by the forget gate, and then added to the fraction of the new candidate allowed by the output gate

$$c_t = ft \odot c_{t-1} + i_t \odot q_t)$$

- **OutputGate**: This gate controls how much of the internal state is passed to the output and works in a similar manner to the other gates.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot tanh(c_t)$$

Here, σ is the logistic sigmoid function that has an output in [0, 1], tanh denotes the hyperbolic tangent function that has an output in $[-1, 1]$, and $\odot$ denotes the elementwise multiplication. To understand the mechanism behind the architecture, we can view ft as the function to control to what extent the information from the old memory cell is going to be thrown away, it to control how much new information is going to be stored in the current memory cell, and o_t to control what to output based on the memory cell c_t LSTM is explicitly designed for time-series data for learning long-term dependencies, and therefore we choose LSTM upon the convolution layer to learn such dependencies in the sequence of higher-level features.

5.3 Data Preparation for LSTM Models

Datapreparation for LSTM model without Textual features

In general, LSTM model are good at learning when input features are more and it also requires large data to process to result good output. Our problem is classification based problem,In which we are interested in predicting the future stance of the user. Firstly in our model we use only user stance feature to feed into the network.In our Reddit brexit dataset we have 220110 samples. We create the X, Y pairs by taking X as the Stance of the user and Y as the 1-shift sequence of the X user stance .After creation of X and Y pair with 220109 samples each as we have a 1-sequence shift for creating Y pair.We change this Y pair to the categorical type and to change to one-hot-encoded values.

As per the user stance we have three different user stance AgainstBrexit, Neutral, Pro-Brexit user stance.However data set is imbalanced with majority of AgainstBrexit. To overcome the problem of data imbalancy were the prediction will effect due to the imbalanced classed data is normalized to get better prediction. As we are using LSTM model, we need to reshape our data into 3D format expected by LSTMs,namely (samples,timesteps,features).After reshaping of data we use this data as input data to the network.We define number activation function as ”categorical crossentropy’as this suits best for the classification problem.

Before training the model,we split into two subsets for training and testing was preformed, with a ratio of 80% training and 20% testing data. We train over model with cross validation methodolody with a 5- fold cross validation. Due to high computation constraints, we use 5 - fold.We report F1 score, Accuracy, Precision and Recall in the result section.

To Compare the results of LSTM model without Textual features added to it, we also build the another model with adding the textual features and check the differt outcomes and select which model gives us better results.

Datapreparation for LSTM model with Textual features

To start of, the dataset was balanced by the number of categories. The baseline was the category with the least entries. Thus, from each category this baseline amount was sampled to create a balanced set of data. However, the unbalanced dataset was also kept to compare against models trained on the balanced dataset. As balancing the datasets lead to a significant loss in training data, this could and has resulted in decreased accuracy.

For data cleaning, abnormalities and left-behind components in the commented text, specifically in the form of URLs were tracked and removed.Finally, text to word sequencing was used to clean the remaining data by removing all punctuation and unecessary characters before stripping it into a vector indexing individual words[13].

Traditionally labelled categories (simple conversion into integers) would lead the algorithm to assume higher values to be ”better” leading to in- accurate output. This error was mitigated with a combination of text to word sequence for the news headlines and one-hot encoding for the classes resulting in a dataset sorted by categorical values, as opposed to ordered indices.

Lastly, all the vectorized text were padded for uniform input to ease the training process.The models were then trained using the 50,000 most commonly occurring words in the word dictionary, as including the rest of the words did not yield to improved results.We Report the F1 score, Accuracy, Precision and Recall in the result section.

6 Results

In Section 4, We have already seen the behaviour of the user interaction in the discussion thread.In order to interpret on information diffusion through the online discussion and interaction to the other user stance which can be clarified by plotting Probability Density Fucntion (PDF) of Edge Homogeneity between the user stances and considering the behaviours change of the user.We take Case1 scenarios where all the source node is combined to get the behaviour of the users. Even tough Neutral users have peak density where as there is no much change in the other two stances, we have scaled the density to plot the PDF of Edge Homogeneity in order fit all Stance together.

In Figure15 We can see the AgainstBrexit and ProBrexit users are having similar notion

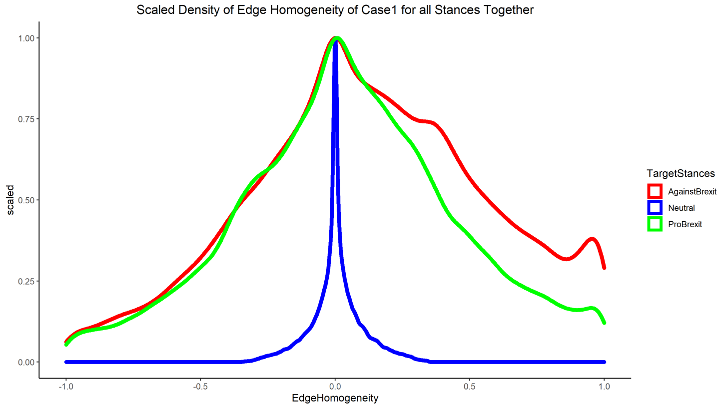

Figure 15: PDF of Edge Homogeneity of Case1 combined with all Source Node

of interaction.However we are interested in seeing how AgainstBrexit and ProBrexit user are attracted towards the source node. In Figure16 we have PDF of Edge Homogeneity between the AgainstBrexit and ProBrexit user stance.Most interesting analysis from Figures 9, 10 we know the majority of the user behaviours individually.But when we see the PDF of Edge Homogeneity, We can say that user discusses and interacts more with the alike minded user,such as AgainstBrexit user interact most with AgainstBrexit user.On the other hand they also interacts with other users as well.Similarly ProBrexit users interacts mostly with the ProBrexit users and they do involve in discussion with otheruser stance.From the Peak of the density plot we can say in this Figure16 majority of the user are ProBrexit users.Majority of the user are ProBrexit but the density towards the similar user is in AgaintBrexit user.

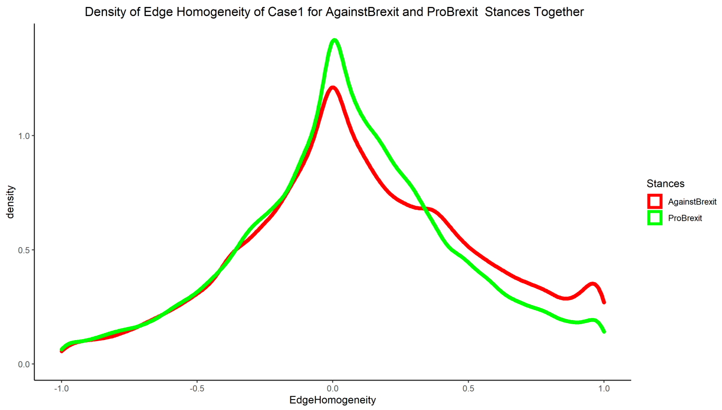

Figure 16: PDF of Edge Homogeneity of Case1 combined with AgainstBrexit and ProBrexit Source Node

It would be interesting to see the Case2 results when all the Target Node users combined together,moreover how the edge homogeneity plays role of information diffusion and analyse any change in the user behavior. Firstly we see PDF of Edge Homogeneity of all the Target Node combined in Figure17.Even though results seems to be similar to Case1 but they are not equal to eachother.When in Figure14 we can find the user interaction with

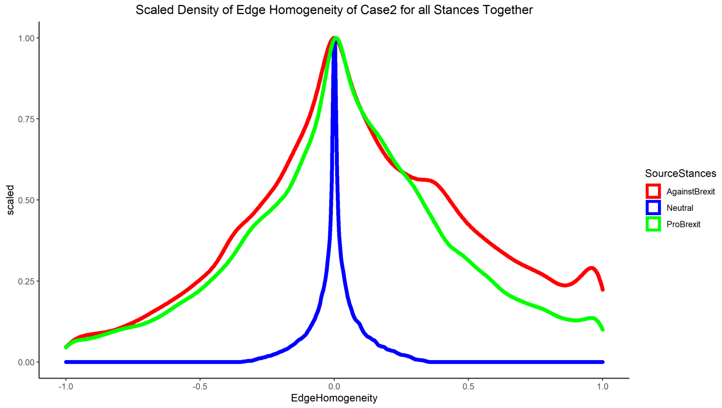

Figure 17: PDF of Edge Homogeneity of Case2 combined with all Target Node

same echo chambers is more but in Figure17 when the Target Node is combined with all other stance.Users interacts with other users here.Moreover we can identify ProBrexit and AgainstBrexit are having similarities.To dig deeper into the comparision we plot PDF of Edge Homogeneity of other than Neutral Stance to check the behaviour of the users.

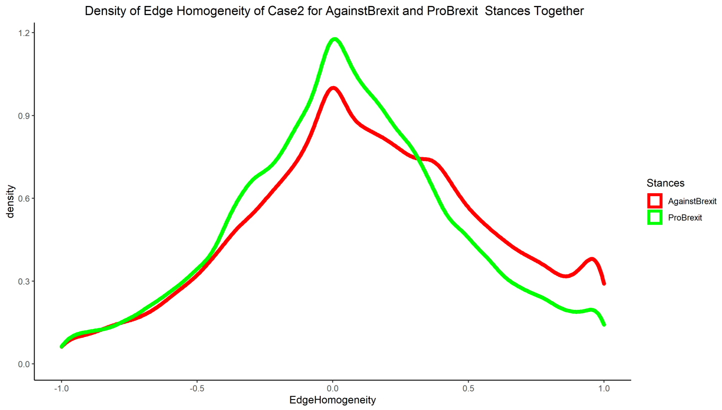

Figure 18: PDF of Edge Homogeneity of Case2 combined with AgainstBrexit and ProBrexit Source Node

In Figure18 we can see the AgainstBrexit is having more density indicating it interacts with alike minded people.However, ProBrexit user are involved in discussion with other stance users which defines edge homogeneity plays major role in the user changing their behaviours and interacting the other users. Interesting would be to combine the AgainstBrexit Stance of Case1 and Case2 and check how is user behavior changing in this aspect.

Moving forward to the comparison to Case1 and Case2 with all the three different user stance,we can see in Figure 19 Neutral Stance users is identical in both the cases.Moreover Neutral users do not involve much in discussion with other users. In Figure 20 we can clearly observe the AgainstBrexit users always tends to interact with the similarly opinions users. It states that the information is diffused in the echo chambers. Interesting part

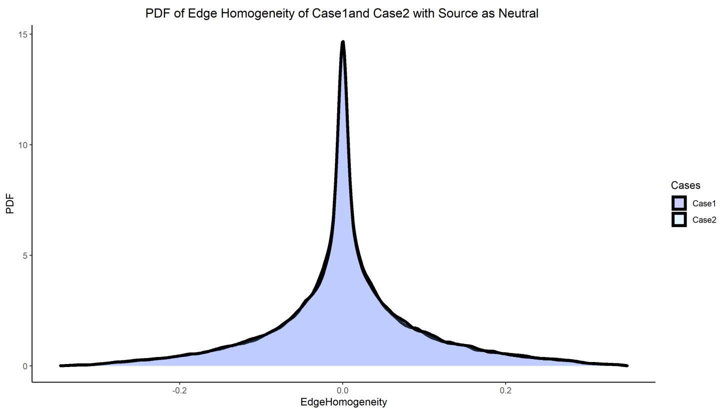

Figure 19: PDF of Edge Homogeneity of Neutral Stance when Case1 and Case2

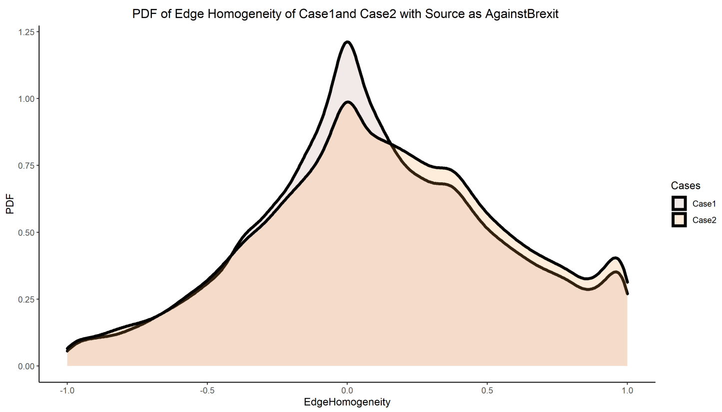

Figure 20: PDF of Edge Homogeneity of Against Stance in Case1 and Case2

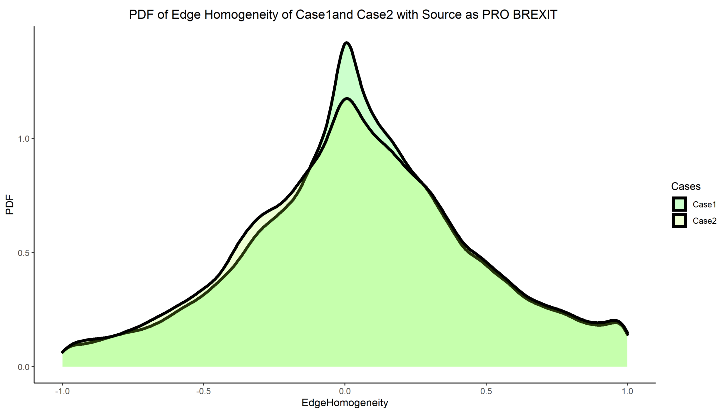

Figure 21: PDF of Edge Homogeneity of ProBrexit Stance in Case1 and Case2

to see the behaviour of ProBrexit users which tends to interacts more with the different user.From Figure 21 ProBrexit user interacts with AgainstBrexit or Neutral user rather than ProBrexit Users. However the information transmission occurs inside the homogeneous clusters and as well the mixed neighborhoods. From various results we can say the edge homogeneity play a vital role in information diffusion.

Model Evaluation : Model evaluation is done on the F1 score and Accuracy for LSTM model without the Textual features.In Table 1 we calculate the confustion matrix of Actual Stance and Predicted Stance from LSTM model, with the help of this we can obtain the F1 Score and Accuracy in Table2. Howvever results are not promising at this moment. We can improve the results by hypertuning the parameters in LSTM model and motivated to get the good prediction of user stances. Moreover the computation constraints for training the Neutral network requires the very high GPU. Lastly, This results are not including the Textual features, as we already proposed LSTM model with textual features we can see the results of LSTM model with textual features. From the Tabel 2 and Table 3 we can see when the textual features are added to the network the model give better accuracy compared to the other.Even though task of prediction of stance is difficult our predictor managed to predict the user future stance.

Table 1: Confusion Matrix of Actual Stance and Predicted Stance from LSTM model

Actual \ *Predicted*	**Against**	**Neutral**	**Brexit**
Against	3790	5544	3354
Neutral	5817	8174	4818
Brexit	3857	5238	3210

Table 2: Accuracy, precision, recall and F1 score for the LSTM Model without Textual Features used for predicting user's stance based on their submitted posts.

Set	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Test	0.350	0.3496	0.3496	0.3496

Table 3: Accuracy, precision, recall and F1 score for the LSTM Model with Textual Features used for predicting user's stance based on their submitted posts.

Set	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Test	0.430	0.4236	0.4216	0.4236

7 Conclusion and Future work

In this work, we mainly aimed to get the analysis of information diffusion affects the participants in social media platform.We choose Reddit social media platform other than others because it does it offer structured information and 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 discussed how the edge homogeneity plays vital role in the information diffusion and how users are influence by the others to change their opinion.

Secondly we build a classifier for predicting the future stance of user. Though we illustrate the performace of LSTM model with two different inputs to get the comparision between the models.At this moment, the results are intermediate and can be improved in Future.

In the following months , we would like to see how user behaviours are in the different time frames and we can see how the model can be enhanced to get improved results. We are also interested in comparing our model with different convolution networks which we can examine the model performance. As the data is timely distributed we can divided the whole time periods into individual time period and check the performance of the user behaviour and examine in which time period user changes it behavior.

Acknowledgements

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank Ms. Christine Largeron , for providing me an opportunity to do the project work in Research Laboratory Hubert Curien and giving us all support and guidance which made me complete the project duly. I am extremely thankful to her for providing such a nice support and guidance, although he had busy schedule managing the corporate affairs.

I owe my deep gratitude to our advisor Andrei Marian Rizoiu, who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.

I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of Master MLDM - University Jean Monnet, which helped us in successfully completing our project work.

Bibliography

[1] A. Bessi, M. Coletto, G. A. D. A. S. G. C. and Quattrociocchi, W. (2015). Science vs conspiracy: collective narratives in the age of misinformation. In *PloS one*, volume 10, page 02.

[2] Adamic, L. A. and Glance, N. (2015). The political blogosphere and the 2004 u.s. election: Divided they blog. In *3rd International Workshop on Link Discovery,LinkKDD*, volume 1, page 36–43. ACM Press.

[3] Amador Diaz Lopez, J. C., Collignon-Delmar, S., Benoit, K., and Matsuo, A. (2017). Predicting the Brexit Vote by Tracking and Classifying Public Opinion Using Twitter Data. *Statistics, Politics and Policy*, 8(1):85–104.

[4] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WIIAT),2010 IEEE/WIC/ACM International Conference*, volume 1, pages 492—-499. IEEE.

[5] Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. journal of conflict resolution. *Journal*, 8:203–226.

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

[7] Biber, D. and Finegan., E. (1988). "adverbial stance types in english". *Discourse processes*, page 1–34.

[8] 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.

[9] Christine, L., Andrei, M., and Andrei-Marian, R. (2019). Information diffusion in online communities. *Research Work*.

[10] Christopher A Bail, Lisa P Argyle, T. W. B. J. P. B. H. C. M. B. F. H. J. L. M. M. F. M. and Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Journal*, 37:9216–9221.

[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] Ding, D., Zhang, M., Pan, X., Wu, D., and Pu, P. (2018). Geographical Feature Extraction for Entities in Location-based Social Networks. In *Proceedings of the 2018 World Wide Web Conference.*, pages 833–842. Association for Computing Machinery (ACM).

[13] Eren Aldis1, E. B. and Malani2, J. (2015). A study on lstm and cnn models for news classification using word2vec embedding. In *Reserach Study*, pages 338–348, 2Department of Computer Science, Johns Hopkins University.
[14] F. Zollo, A. Bessi, M. D. V. A. S. G. C. L. S. S. H. and Quattrociocchi, W. (2017). Debunking in a world of tribes. In *PloS one*, volume 12, page 07.
[15] Geoffrey E Hinton, Nitish Srivastava, A. K. I. S. and Salakhutdinov., R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors.
[16] 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.
[17] Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.
[18] 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.
[19] H. Becker, M. N. and Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *third ACM international conference on Web search and data mining*, pages 291—-300. ACM.
[20] Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.
[21] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. 143:1735–1780.
[22] 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.
[23] 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.
[24] 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.
[25] M. Del Vicario, A. Bessi, F. Z. F. P. A. S. G. C. H. E. S. and Quattrociocchi, W. (2016). The spreading of misinformation online. In *Proceedings of the National Academy of Sciences*, volume 3, page 554–559. PNAS.
[26] McPherson, M., S.-L. L. . C. J. M. (2001). Birds of a feather: Homophily in social networks. *Journal*, 27:415–444.
[27] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.
[28] Newman, N. (2011). "mainstream media and the distribution of news in the age of social discovery,". *Reuters Institute for the Study of Journalism, University of Oxford*.
[29] Nic Newman with Richard Fletcher, A. K. and Nielsen, R. K. (2019). Digital news report 2019. *Reuters Institute*.
[30] Pushshift (2019). Pushshift. `https://pushshift.io/`.
[31] Schelling, T. C. (1969). Models of segregation. *Journal*, 59:488–493.
[32] Sunstein, C. R. (2002). The law of group polarization. *political philosophy*, 10:175 – 195.

[33] 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.

[34] W. Quattrociocchi, A. S. and Sunstein, C. R. (2016). Echo chambers on facebook. In *Available at SSRN*. SSRN.

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