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

Internet-based life have drastically changed the worldview of news utilization and their effect on data spreading and dispersion has been to a great extent tended to, Specifically, a few investigations concentrated on the expectation of social dynamics[19],[4]. with unique accentuation on data streams designs and the development of explicit network structures. Surely, internet-based life are in every case progressively associated with both the conveyance and utilization of news. As indicated by an ongoing report[29], around 51% of clients get news straight forwardly through web-based life, and such data experiences indistinguishable prevalence elements from different types of online substance, for example, selfies or kitty pictures. In the current, disinter-mediated condition clients effectively take an interest in substance creation and data is no longer intervened by columnists or specialists. Correspondence techniques and how messages are surrounded and shared over interpersonal organizations have changed. Without a doubt, online disinter-mediation evokes clients' inclination to select data that follows (and strengthens) their previous convictions, the alleged affirmation inclination. In such away, clients will in general structure gatherings of similarly invested individuals where they enrapture their assessment i.e., reverberation chambers[25],[34]. In addition, affirmation predisposition assumes a crucial job in enlightening falls. Exploratory proof shows that corroborative data gets acknowledged regardless of whether containing intentionally bogus cases [1], while disagreeing data is for the most part disregarded [14]. Bantering with similarly invested individuals has been appeared to impact clients' feelings adversely what's more, may even build bunch polarization [32].

It is of specific significance to comprehend the inborn components of such informal organizations, so as to productively distinguish potential changes in the clients' perspectives around the various issues they are discussing. By totaling literary data that characterizes clients with the elements of the dispersion they are a piece of, the objective is to foresee future positions of the people when they associate in the Social Network with their friends. This is of specific helpfulness since it gives a trace of how individuals change their positions in the wake of having cooperated with others, along these lines driving a to a superior comprehension of the systems.

1.2 Motivation

The primary objective of this work centers principally around recognizing and anticipating the conduct of individuals associated with Online Social Networks [23] when they are presented to particular sorts of data and cooperate with individuals previously supporting various thoughts. For achieving this objective we utilize a contextual investigation of Brexit. By deciphering the situation of members regarding the matter of the withdrawal of the United Kingdom from the European Union, we are keen on foreseeing conduct changes in the talk of people, who are already end up being a piece of various networks.

We are additionally keen on identifying if there are different elements that may cause shifts in the assessments of members in the online discussions, for example, the quantity of messages they trade, with what sort of people they do trade messages, the prevalence of their messages, and so forth. In any case, the social impact [33] and the homophily [11] are likewise angles we are attempting to exploit on during the time spent anticipating the direction of just members.

Despite the fact that there are a considerable amount of papers on points, for example, data dissemination in online interpersonal organizations [22], [24], [35], and specifically regarding the matter on Brexit [6], these works generally attempt to distinguish the networks (leavers and remainers) and discover the subjects that are hot among these two gatherings. In our examination, we are beginning principally from these works and attempt to propel them, by searching for designs among the various dispersions and we plan to foresee future changes in places of people who have been presented to particular sorts of data. For example, if two people X and Y share similar conclusions, if individual X is presented to a specific succession of messages of a specific kind, suppose master Brexit, and thusly changes the tone of the talk in the accompanying messages, what are the odds that individual Y will likewise do likewise whenever presented to this grouping of messages. X second distinction between our work and different examinations distributed so far comprises of the sort of data utilized: while different works utilize literary data for arranging clients, we depend just on data got from the social communication they have on online stages.

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 resulting areas are organized as follows: Section 2 portrays the best in class in the field of network discovery, data dispersion in Online Networks, Homophily, and Social Influence on the informal organizations.

In Section 3,we presents the dataset associated with the improvement of our proposed approach and the manner in which the information was preprocessed. We detail the stage utilized for gathering the information. We utilize the Reddit dataset to prepare our position recognition classifier(pro, against, or impartial around Brexit subject). Next, this classifier is to decide the two principle networks from this Online Social Network, and by amassing various highlights, we foresee future patterns in the system.

In Section 4,we subtleties the technique behind Homiphily and How social impact is affecting the client to change the practices. In this strategy, we can perceive how client polarization is characterized. We ascertain the Edge homogeneity between the hubs and picture how Edge Homogeneity assumes a job in data dispersion between the clients.

In Section 5, we present the methodological commitment proposed during this temporary position with respect to the political position forecast. Right off the bat, the primary Reddit study intends to foresee the position of a client considering its cooperation with different individuals from the network. This exploration way disclosed a subsequent way, because of the need of marking and dividing the clients. Segment 5.2 portrays this classifier and the manner in which it was prepared.

At last, Section 6 presents the outcomes that got in the wake of leading this investigation, while, in Section 7, we reach inferences and list the following stages which must be made so as to improve our outcomes.

2 State of the Art

2.1 Information Diffusion in Online Networks

Numerous examinations concentrated in transit interpersonal organizations data stream can be demonstrated in a manner that permits investigating, understanding and anticipating the data dispersion. A careful overview of the various techniques identified with these subjects is performed by [18]. Since the start, the creators are demonstrating the Online Social Networks utilizing components from chart hypothesis. In this manner, every vertex speaks to a client from the system, while a current bend between two clients implies that the source vertex is presented to data originating from the destination vertex.

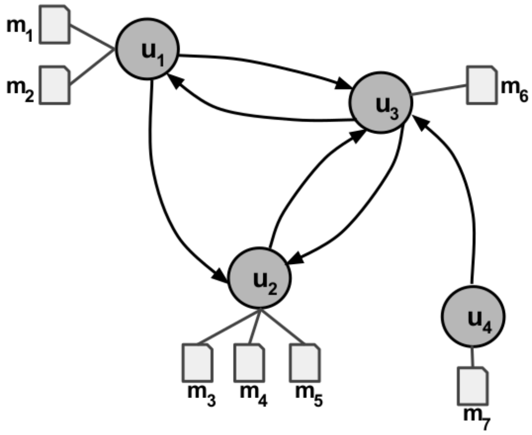


Figure 1: Online Social Network modelled as graph. Each vertex u_i is a user, while a directed edge (u_i, u_j) means user j exposes user i to a certain information, via message k, denoted as m_k Source: *Adrien Guille, Hakim Hacid, C. Favre, Djamel Abdelkader Zighed. Information Diffusion in Online So- 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 social orders homophily, the inclination of comparable people getting related and fortified with one another is known to be a prime data factor between a couple of individuals[26]. This affiliation and holding can be identified with at least one highlights including sexual orientation, race, age, training level, monetary and economic well being, and numerous more. Thus, homophily largely affects various on a very basic level significant social wonders like isolation, disparity, recognition inclinations, and the transmission of data between gatherings of people. The component of central conclusion, i.e., the way toward framing connects to others with shared qualities yet without normal associates can be considered as a run of the mill sign of homophily[31].

2.3 Brexit Position Classification of Twitter Users

Another significant bit of work that speaks to the beginning stage for our exploration is spoken by [3]. They break down messages sent on Twitter, particularly sent in the time of the Brexit choice. Their primary objective is to have the option to precisely anticipate by the content sent in the message if the creator is ace or against Brexit. In the subsequent part, they consider the two gatherings and examine the most significant subjects that they had drawn into.

The first part of their work is a valuable instrument for our exploration: they give a formula to ordering Remain versus Leave accounts, by utilizing a managed Machine Learning calculation, to be specific Naive Bayes Classifier. The errand is especially troublesome in light of the fact that the set up of the issue doesn't give names or ground truth to the gathered answers.

Be that as it may, their principle commitment is in building such a ground truth. They locate the main 200 most referenced records in that period and physically survey their participation in one of the previously mentioned classes. At that point, for the two discovered gatherings, they investigate the hashtags and make two arrangements of hashtags: ProBrexit and Against Brexit. Next, they total the tweets of each record and by investigating all records messages, they pick the ones that are utilizing those hashtags, register a leave list, which is the number of leave hashtags less the number of remain hashtags and sorts the records as indicated by the file. At last, they pick the primary 10% as professional Brexit and the main 10% as against Brexit. At that point, they train an NB classifier and mark every single other record.

2.4 Reddit Studies

In spite of the fact that initially, research on Reddit datasets has caught up in the last years most recent years. For example, [8] carried out an investigation on the discussions facilitated on the Reddit online system, concentrating on the quantitative and subjective parts of the messages. The primary objective of the investigation is arranging the discussions and discovering properties as far as volume, the responsiveness of the clients, and the propensity of getting famous and spreading rapidly. In this manner, the creators reasoned that a viral dispersion will in general have increasingly troublesome content, though falls that will remain not all that obscure to the huge open have less complex, shorter messages. In addition, the creators detail how a huge discussion is really comprised of a between twinning of countless messages, the greater part of which are sent by a little arrangement of one of a kind clients. Maybe as anyone might expect, each sub-network has various qualities, while subreddits containing media content like photographs and recordings, news and conversations sub-networks tend to prompt viral discussions.

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

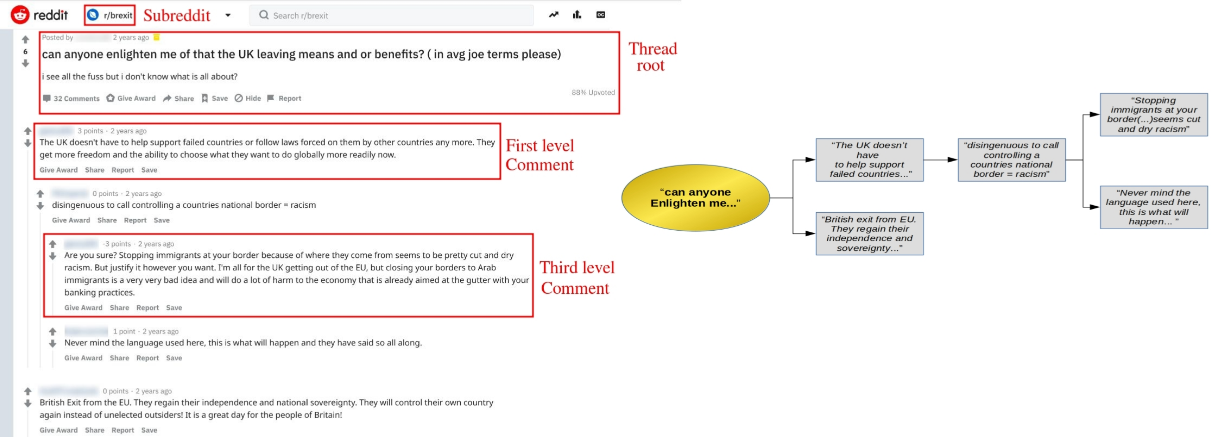
Reddit is a social media platform for online conversations to clients who share news, articles, and suppositions with one another on the specific regions of client intrigue. The territories of effective interests in Reddit are classified ”subreddits”, every one of which fills in as a free network. A subreddit can be made by any client who is keen on a specific subject, e.g., game, legislative issues, or sports. Each subreddit is overseen by several”moderators” who are answerable for directing and policing discussions among individuals. 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 gathered utilizing the Pushshift API [30], which is created inside the Pushshift Project. This is a major information stockpiling and investigation venture which permits downloading countless reddits, from a specific period, regarding diverse forced



6

Posted by [redacted] 2 years ago

can anyone enlighten me of that the UK leaving means and or benefits? (in avg joe terms please)

i see all the fuss but i don't know what is all about?

32 Comments Give Award Share Save Hide Report 88% Upvoted

[redacted] 3 points · 2 years ago

The UK doesn't have to help support failed countries or follow laws forced on them by other countries any more. They get more freedom and the ability to choose what they want to do globally more readily now.

Give Award Share Report Save

[redacted] 0 points · 2 years ago

disingenuous to call controlling a countries national border = racism

Give Award Share Report Save

[redacted] -3 points · 2 years ago

Are you sure? Stopping immigrants at your border because of where they come from seems to be pretty cut and dry racism. But justify it however you want. I'm all for the UK getting out of the EU, but closing your borders to Arab immigrants is a very very bad idea and will do a lot of harm to the economy that is already aimed at the gutter with your banking practices.

Give Award Share Report Save

[redacted] 1 point · 2 years ago

Never mind the language used here, this is what will happen and they have said so all along.

Give Award Share Report Save

[redacted] 0 points · 2 years ago

British Exit from the EU. They regain their independence and national sovereignty. They will control their own country again instead of unelected outsiders! It is a great day for the people of Britain!

Give Award Share Report Save

"can anyone Enlighten me..."
"The UK doesn't have to help support failed countries..."
"British exit from EU. They regain their independence and sovereignty..."
"disingenuous to call controlling a countries national border = racism"
"Stopping immigrants at your border(...)seems cut and dry racism"
"Never mind the language used here, this is what will happen... "

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.

rules, similar to the subreddit. The information is absolutely liberated from utilize and can be gotten to effectively, through a Python content utilizing a pre-characterized API. Consequently, we gathered all reddits from **November, 2015** to **April, 2019**, which were a piece 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), score and number of comments (specific for the root of the threads)**[9]. From submission id, we can relate to finding 22010 different threads with each thread that has a different number of participates in the discussion.

We have time dissemination of the gathered message which is spoken to in Figure 3. This Figure represents a climbing pattern, which focuses on the developing significance of this Brexit subject as it is seen by regular individuals utilizing online networking stages. Despite the fact that the monotonicity is expanding, we can watch a spike in June 2016. This is created by the way that on the 13th of June 2016, British individuals were relied upon to decide in favor of the national submission. This occasion delighted insignificant consideration from British media and news channels, which was additionally reflected in the movement of online informal organizations users[5].However, On the other hand, the top as far as the number of submitted messages is in February - March 2019. This is a result of the official timetable of the Brexit procedure, which ought to have finished in March 2019.

Most of the messages are remarked as delineated in Figure **??** (a), rather than an underlying, string beginning messages. As far as remarkable writers, Figure **??** (b) shows that 20% of all the exceptional writers are **only** string initiators. This implies they just send a solitary message, beginning a conversation string, in which they never post again. Then again, 19% of the writers, are both string starters and analysts, implying that they start strings and partake effectively in the conversations, posting answers in their own

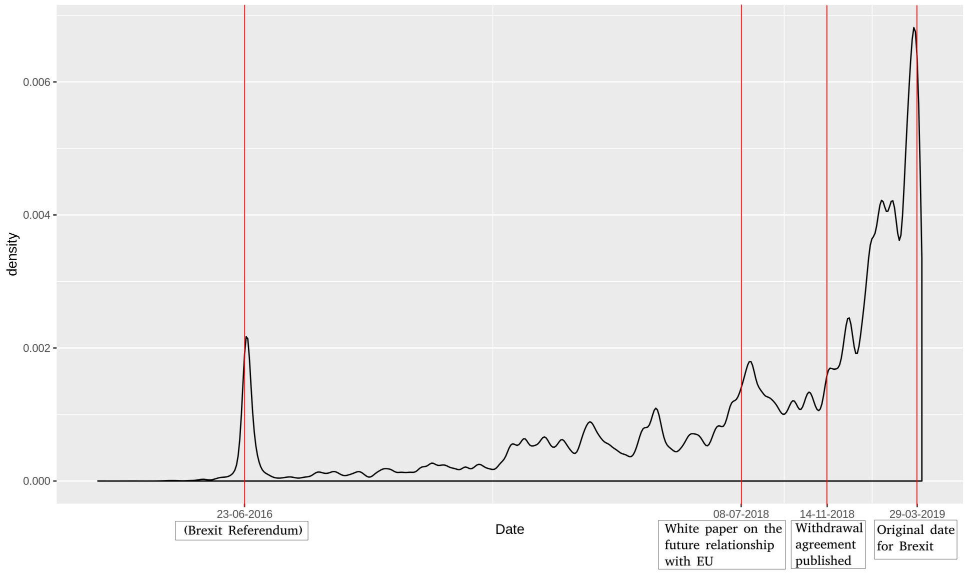


Figure 3: Time distribution of the submissions collected from Reddit (subreddit brexit), between November, 2015 and April, 2019.

began string or engaging in other discussions[9]. Most of the one of a kind clients are **only** analysts, implying that they never start conversations, yet generally participate in them.

From Figure 4 We can watch the long tail of the realistic introducing the quantity of messages per client. In this figure, the Complementary Cumulative Distribution Function of the quantity of messages shows that an exceptionally huge number of clients send just a couple of messages, though there are a couple of clients sending numerous messages in the watched stretch. These clients can either be feeling formers or essentially 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.

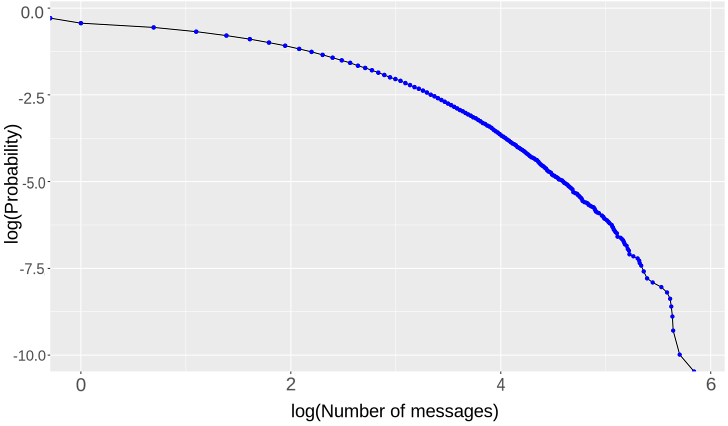


Figure 4: Complementary Cumulative Distribution Function of the number of messages sent by each user.

4 Information Spread around Brexit on Reddit

These days, web-based social networking stages comprise a significant segment of a person's social cooperation. These stages are viewed as a powerful data scattering instrument to communicate suppositions and offer perspectives. Individuals depend on these instruments as the fundamental wellspring of news to interface with the world and to get moment updates[28]. They have a significant gainful site that permits people to investigate different parts of developing subjects, express their own perspectives, get moment criticism, and investigating open perspectives. The gigantic reliance of clients on these stages as the fundamental wellspring of correspondence permitted scientists to consider various parts of online human conduct, including an open position towards different social and political aspects[10]. The position is characterized as the outflow of the speaker's point of view and judgment toward a given suggestion Stance identification assumes a significant job in diagnostic investigations estimating general conclusion via web-based networking media, especially on political and social issues[2]. The idea of these issues is typically dubious, where individuals express restricting assessments 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 5 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.

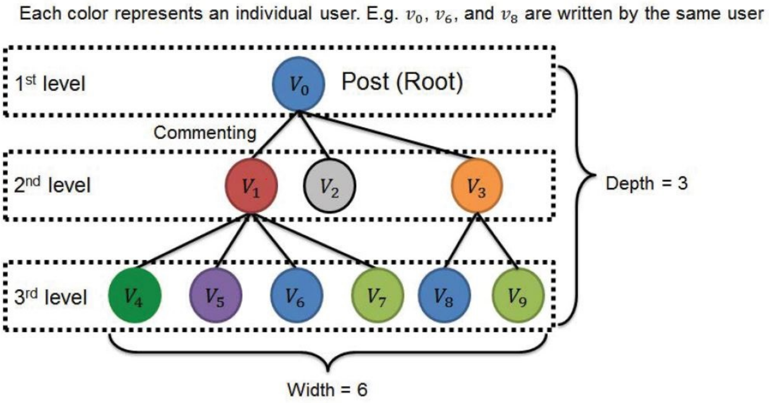


Figure 5: A comment tree is illustrated for a post that has 9 comments

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.

- $-1 \leq \sigma \leq -0.2 \implies$ **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 6. 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.

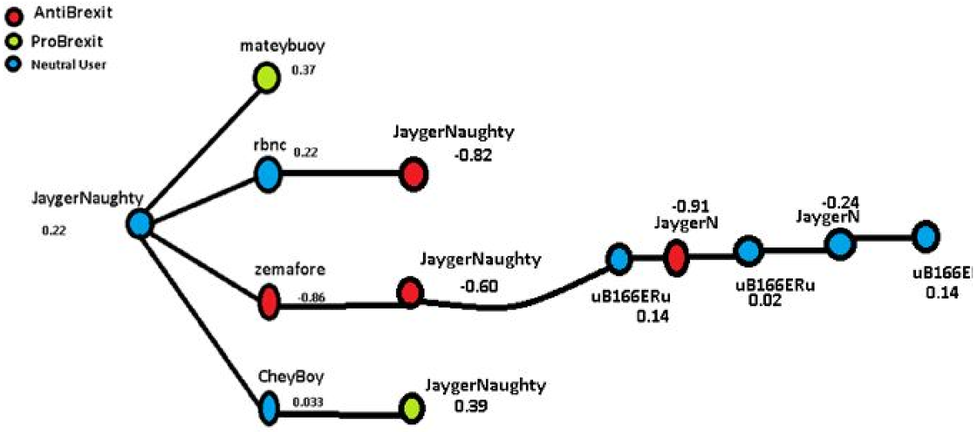


Figure 6: Smallest Discussion Thread in Reddit Dataset.

From Figure 6 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

4.2 Homiphily: Engagement and friends

In this area, we need to comprehend if the commitment in a particular sort of comment is a decent intermediary to recognize a gathering of clients with comparable perspectives. Homophily – i.e., the propensity of clients to total around basic interests – has been as of now called attention to as a factor in changing the client conduct/disposition swing[14]. In fact, online substance blends the subjective, conduct, and social parts of a client network. The resulting ecosystem system permits us to research the different procedures at play in the cooperation of people and to consider the manners by which clients relate with the sort of data they connect with. In this way, users' preferring movement across substance of the various classes [34] might be planned as the special disposition towards the other user behavior

4.3 Social Influence on changing client behavior

Social impact power that individual A (i.e., the influencer) applies on individual B to present a difference in the conduct, as well as an assessment of B Influence, is a causal procedure. where the activity of a client is activated by different suppositions. A case of this situation is the point at which a client purchases an item since one of his/her companions has as of late purchased a similar item This alludes to the marvel that the activity of people can prompt their companions to act along these lines. This can be through setting a model for their companions (as for the situation

of style), illuminating them about the activity (as on account of viral promoting), or expanding the estimation of activity for them (as on account of the appropriation of innovation). In our specific circumstance, we would be intriguing to know whether client practices are impacts to change the client position from one position to another people.

In order to study the Homophily and Influence, we take Reddit dataset and define two cases as Case1 and Case2. In Case1 we fix one stance as the source node and the target node is combined with all other stances. By this, we can see which Target node is attracted to the source node. Similarly, we do vice-versa 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 which source node is attracted most by the target node.

To analyze 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 the Target node is combined with all the other user stances(AgaintBrexit, Neutral, and ProBrexit). This defines as 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 is interested to comment on the Neutral stance node through Pie-diagram in Figure 7

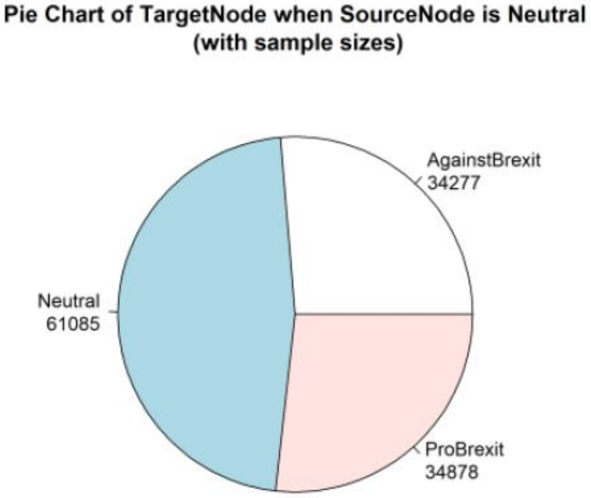


Figure 7: Pie-Chart of the Target Node Stance when Source Node is Neutral Stance

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

Considering the Case1- Scenario2, In this scenario Source node, is AgainstBrexit stance node and the Target node is combined with all the other stance nodes. Technically this states the target node user who is interested and attracted to the AgainstBrexit stance node. However, we can observe the number of Target nodes who are attracted to the AgainstBrexit stance in Figure 8. We can see that the when source node is Against majority of the users are attracted to the Source Node are Neutral stance and followed by AgainstBrexit Stance and ProBrexit stance. Users do interact with other stances when they are on the discussion.

Moving forward to the Case1 - Scenario3, wherein this scenario Source node is ProBrexit stance node and the 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 9. However the user interaction results to the similarity as in scenario2 and scenario3 as ProBrexit stance node attracted by all the other stance users

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. The interesting part to be to see the behavior of the which source node stance got attracted by Target node when it is neutral.We can see the results in Figure10. From the results, we can say Neutral stance nodes are the majority and it states that the neutral

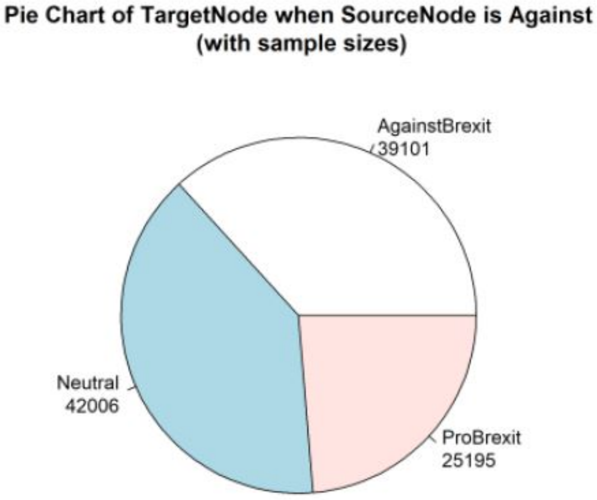


Figure 8: Pie-Chart of the Target Node Stance when Source Node is AgainstBrexit Stance

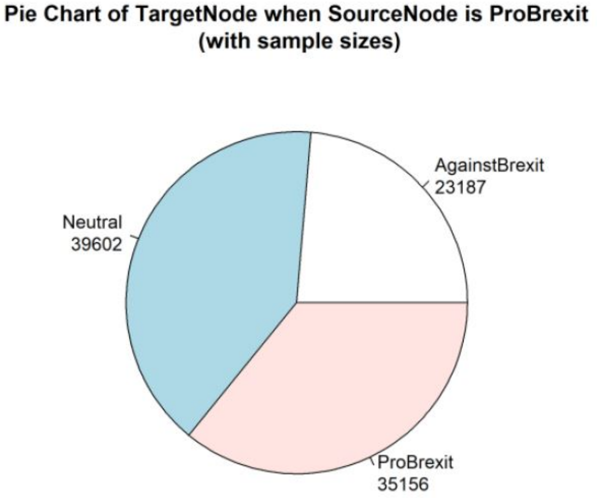


Figure 9: Pie-Chart of the Target Node Stance when Source Node is ProBrexit Stance

node attracted are mostly in this scenario. Also, the other two stances got attracted too.

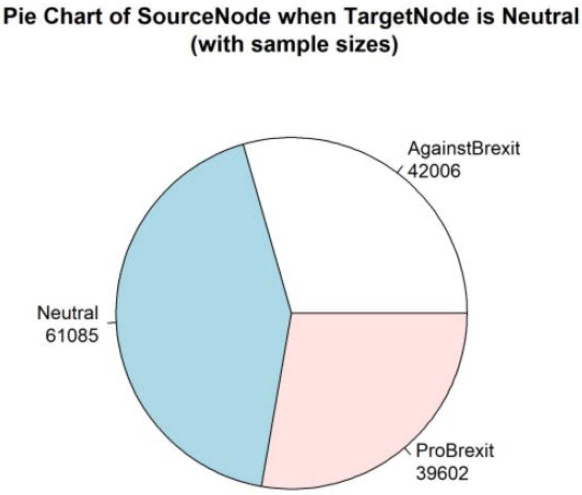


Figure 10: Pie-Chart of the Source Node Stance when Target Node is Neutral Stance

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 Figure11. 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.

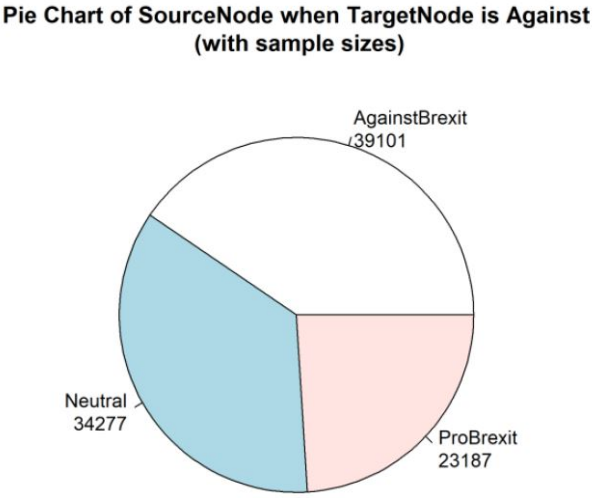


Figure 11: Pie-Chart of the Source Node Stance when Target Node is Against Stance

Lastly, we have Case2 - Scenario3 were Target Node stance is ProBrexit stance nodes and Source node are combined by all other stances.Moreover, we can see from Figure12 which node is attracted to 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 Figure 12

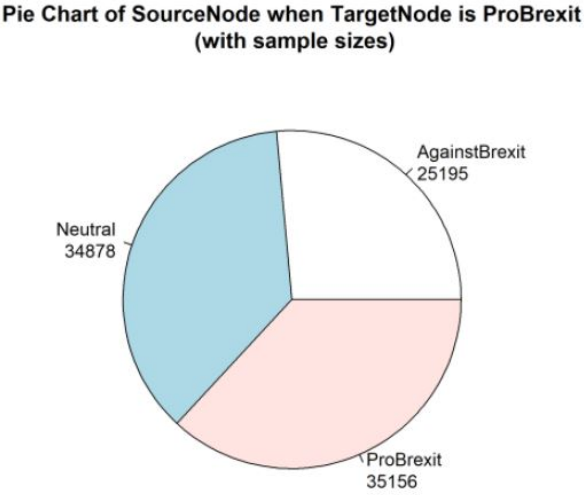


Figure 12: Pie-Chart of the Source Node Stance when Target Node is ProBrexit Stance

As of now, we can relate results in the different Cases with several scenarios and analyze 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 the 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 Figure13 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.

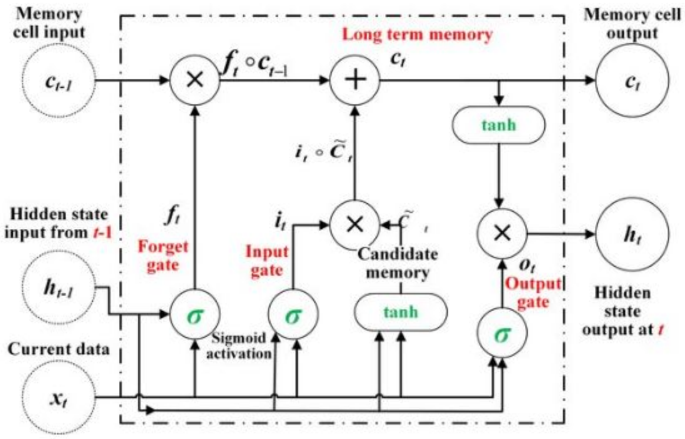


Figure 13: Architecture of a Long-Short Term Memory (LSTM) unit.

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, the LSTM model is good at learning when input features are more and it also requires large data to process to result in good output. Our problem is a 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 the 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, the data set is imbalanced with the majority of against Brexit. To overcome the problem of data imbalance were the prediction will effect due to the imbalanced classed data is normalized to get better prediction. As we are using the LSTM model, we need to reshape our data into the 3D format expected by LSTMs, namely (samples, timestamps, features). After reshaping of data we use this data as input data to the network. We define number activation function as "categorical cross-entropy as this suits best for the classification problem.

Before training the model, we split into two subsets for training, and testing was performed, with a ratio of 80% training and 20% testing data. We train over the model with a cross-validation methodology 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 the LSTM model without Textual features added to it, we also build another model with adding the textual features and check the different outcomes and select which model gives us better results.

Datapreparation for LSTM model with Textual features

To start off, 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 unnecessary characters before stripping it into a vector indexing individual words[13].

Traditionally labeled 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.

6 Results

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

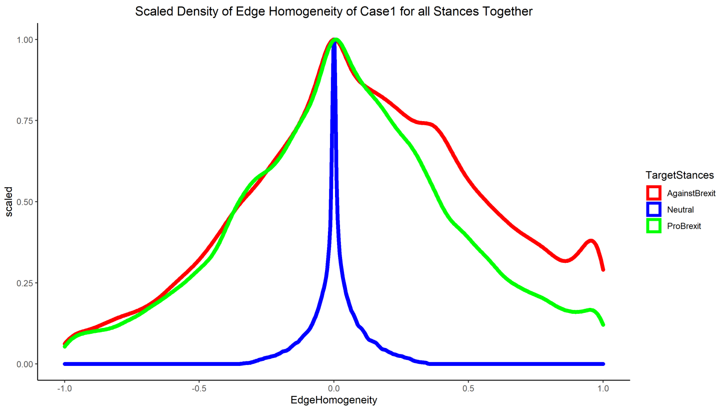


Figure 14: PDF of Edge Homogeneity of Case1 combined with all Source Node

In Figure14 We can see the AgainstBrexit and ProBrexit users are having similar notions of interaction. However, we are interested in seeing how AgainstBrexit and ProBrexit users are attracted to the source node. In Figure15 we have PDF of Edge Homogeneity between the AgainstBrexit and ProBrexit user stance.Most interesting analysis from Figures 8, 9 we know the majority of the user behaviours individually.But when we see the PDF of Edge Homogeneity, We can say that the 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 interact with other users as well. Similarly, ProBrexit users interact mostly with the ProBrexit users and they do involve in discussion with other user stances. From the Peak of the density plot, we can say in this Figure15 majority of the user are ProBrexit users. The majority of the user are ProBrexit but the density towards a similar user is in AgaintBrexit user.

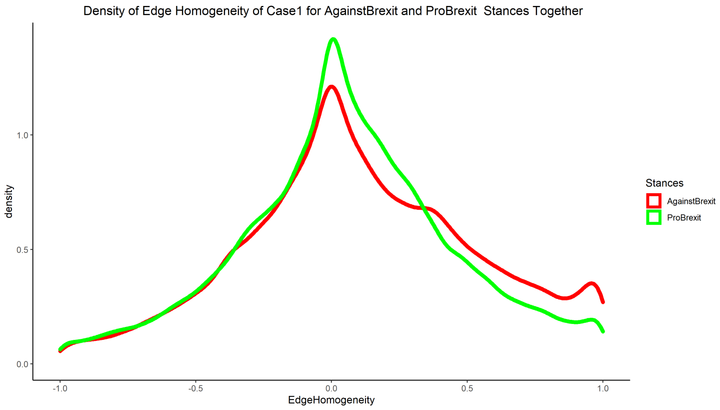


Figure 15: PDF of Edge Homogeneity of Case1 combined with AgainstBrexit and ProBrexit Source Node

It would be interesting to see the Case2 results when all the Target Node users combined together, moreover how the edge homogeneity plays the role of information diffusion and

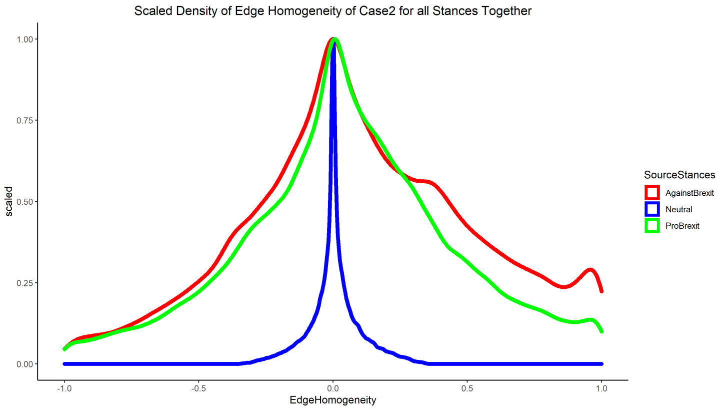


Figure 16: PDF of Edge Homogeneity of Case2 combined with all Target Node

analyze any change in the user behavior. Firstly we see PDF of Edge Homogeneity of all the Target Node combined in Figure16. Even though results seem to be similar to Case1 but they are not equal to each other. When in Figure13 we can find the user interaction with the same echo chambers is more but in Figure16 when the Target Node is combined with all other stance. Users interact with other users here. Moreover, we can identify ProBrexit and AgainstBrexit are having similarities. To dig deeper into the comparison we plot PDF of Edge Homogeneity of other than Neutral Stance to check the behavior of the users.

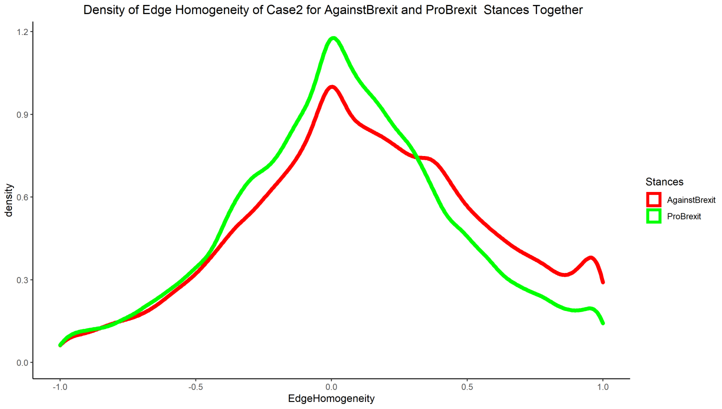


Figure 17: PDF of Edge Homogeneity of Case2 combined with AgainstBrexit and ProBrexit Source Node

In Figure17 we can see the AgainstBrexit is having more density indicating it interacts with alike minded people. However, ProBrexit users are involved in discussion with other stance users which defines edge homogeneity plays a major role in the user changing their behaviors and interacting with 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

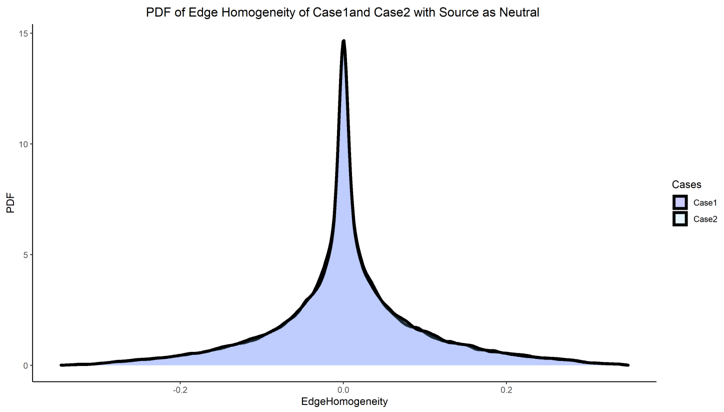


Figure 18: PDF of Edge Homogeneity of Neutral Stance when Case1 and Case2

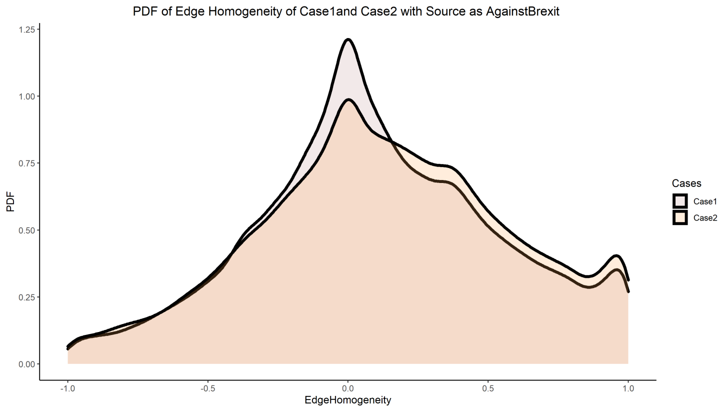


Figure 19: PDF of Edge Homogeneity of Against Stance in Case1 and Case2

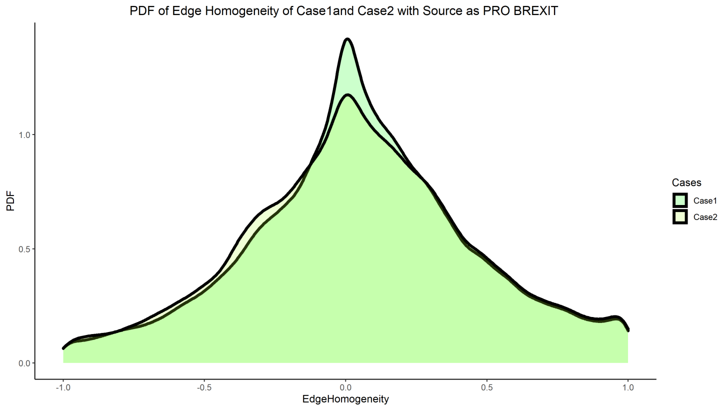


Figure 20: PDF of Edge Homogeneity of ProBrexit Stance in Case1 and Case2

stance, we can see in Figure 18 Neutral Stance users is identical in both the cases. Moreover, Neutral users do not involve much in discussion with other users. In Figure 19 we can clearly observe the AgainstBrexit users always tend to interact with similar opinion users. It states that the information is diffused in the echo chambers. The interesting part to see the behavior of ProBrexit users which tends to interacts more with the different users.From Figure 20 ProBrexit user interacts with AgainstBrexit or Neutral user rather than ProBrexit Users. However, the information transmission occurs inside the homogeneous clusters and as well as the mixed neighborhoods. From various results, we can say the edge homogeneity plays a vital role in information diffusion.

Model Evaluation: Model evaluation is done on the F1 score and Accuracy for the LSTM model without the Textual features. In Table 1 we calculate the confusion matrix of Actual Stance and Predicted Stance from the LSTM model, with the help of this we can obtain the F1 Score and Accuracy in Table2. However, results are not promising at this moment. We can improve the results by hyper tuning the parameters in the LSTM model and motivated to get a good prediction of user stances. Moreover, the computation constraints for training the Neutral network requires a very high GPU. Lastly, These results are not including the Textual features, as we already proposed the LSTM model with textual features we can see the results of the 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 gives better accuracy compared to the other. Even though the 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

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

Table 2: Performance Metrics for 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: Performance Metrics for 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.