Polarisation and Influence: Investigating Brexit Opinion Dynamics on Reddit

Duy Khuu

A thesis submitted for the degree of
Bachelor of Advanced Computing (Honours)
The Australian National University

November 2021

Except where otherwise indicated, this thesis is my own original work.

Duy Khuu
15 November 2021

I would like to dedicate this thesis to my parents.
Thank you for always supporting me.

Acknowledgments

I would like to express my gratitude to everyone who has made this work possible. Thank you to all the Behavioral Data Science group members - namely Tom, Frankie and Rohit - for all of your feedback and the fun memories over the past year. Special thanks to Andrew Law for his involvement in both this project and the results of his own complementary work. Finally, I want to thank my supervisor, Dr Marian-Andrei Rizoiu, for allowing me to be part of this project and the tremendous amount of support you have provided me this year - your guidance has been invaluable.

Abstract

The popularisation of social media has led to widespread occurrences of echo chambers, selective exposure and misinformation. This is particularly concerning with regard to contentious topics, where lack of interaction with opposing views can lead to complacence or stubbornness. We build on past work in an attempt to determine how exposure to differing opinions affects an individual's future opinion. We quickly discover that: 1. The problem goes far beyond a simple discrete classification task due to the subtleties of user sentiment and 2. Future stance information being conditional on users choosing to remain active in the discussion network.

We address the first issue by proposing a continuous polarity metric to quantify the attitudes of users and find that individuals who choose to remain are polarised users who are stubborn in their beliefs.

To resolve the second point we must first determine what makes users choose to leave. We find that future presence correlates with user interaction and social neighbourhood size.

Finally, we propose a sequence model that takes into account individual interactions to predict future user behaviour.

Contents

Chapter 1

Introduction

1.1 Motivation

Social media has become an increasingly prevalent form of discussion towards contentious topics. The anonymity and accessibility of online platforms such as Twitter and Reddit can lead to extreme opinions being expressed to a wider audience when compared to physical discussion networks. Well-known phenomena present in social network theory such as homophily - increased contact and interaction between individuals with similar characteristics [McPherson et al., 2001], along with echo chambers (a potential byproduct of homophily) - the reinforcement of existing opinions due to lack of interaction with alternative views, also exist in online discussion networks [Colleoni et al., 2014].

Athough the increased ability to communicate with like-minded individuals may seem beneficial, providing a safe space for harmonious entities to engage and interact, this may lead to blind faith in opinions shared by members of such a community resulting from such beliefs being seldom challenged. Exposure to differing opinions has been found to improve the ability to justify one's own political views along with conceive reasons why others may disagree with them [Price et al., 2002].

Social networks have also been found to be particularly dangerous with respect to the dissemination of misinformation. Personalised content algorithms that isolate the user to experience only agreeing viewpoints and ideas (filter bubbles) along with selective exposure - avoiding content that conflicts with personal opinions, have contributed to the wide spread of misinformation in cases such as the 2016 Presidential Election [Spohr, 2017].

While it is clear that lack of opposing discussion can lead to hive minds and naive ignorance, what is not clear is the dynamics between users with different opinions on contentious topics - more specifically **how does interacting with different views affect future stance?**

Being able to understand these dynamics and how they influence future opinion could be used to help prevent homophilic behaviour regarding these topics. However, they could also be used to influence democratic processes, such as elections. Democratic votes are generally decided by the "undecided voters", a group whose final decision is uncertain and can consist of as much as 15%-30% of voters - who are generally less informed and not as interested in politics [Schill and Kirk, 2014].

Political candidates are also aware of the importance of appealing to the undecided demographic. During his 2012 US Presidential Election campaign, Republican candidate Mitt Romney expressed to a private audience in a later leaked speech that, "what I have to convince are the five to ten percent in the center that are independents" [Throsby, 2013].

Uncovering the types of interactions that are able to shift future stance could potentially change the outcome of similar voting processes by influencing voters in the undecided group. The 2016 Brexit referendum was decided by a margin of less than 4% [BBC, 2020]. This implies that being able to sway just 2% of overall voters could have lead to a different result.

1.2 Research Questions and Contribution

Based on the above motivation we seek to answer the main question **How does interacting with different views and opinions affect future user behaviour?**

We also aim to improve on past work in this area that utilises only high-level summaries of the stances of users and types of discussions that individuals have engaged with to predict future stance [Largeron et al., 2021]. We evaluate the effectiveness of incorporating the user network (i.e. links between users who have interacted with each other) along with atomic-level interactions (i.e. individual replies and interactions) to answer the question **are network-level features and discussion structure influential in determining future behaviour?**

Finally, we wish to interpret our findings to go beyond the notion of predicting stance and discover specific interactions that contribute to stance chance. This leads to the question **what drives stance change?**

In this thesis we initially explore network-level features to assess their impact in the prediction of future stance, however we quickly discover that this problem goes far beyond a simple discrete classification task. We find that our **future stance information is conditional on the user choosing to participate again in a future time period**.

Additionally due to the nature of social media comments, we find that classifying users into three discrete stances (Pro, Against, Neutral) is too naive and introduce a **continuous polarity metric** to improve the quantification of sentiment - turning the stance prediction task into a regression problem. Utilising feature importance methods on models trained on this regression task we deduce that **users who remain in the system are already polarised and unlikely to change their stance**.

Due to future stance information being conditional on remaining active in the social network, we must first answer the question **what makes users leave?** We find that both **the number of interactions and unique users an individual converses with correlates with future presence**. We go one level deeper to examine whether the structure of these interactions is also important and propose a **sequence model classification model for predicting future user presence**.

In summary, our contributions are:

- Identifying the conditional nature of the future stance prediction task and its implications
- Continuous metric for quantifying user stance polarity and insights from feature importance methods in the future stance regression task
- Sequence model that considers atomic-level interactions between users for future presence prediction

Chapter 2

Background and Related Work

2.1 Relevant Literature

2.1.1 Opinion Dynamics

Past work on opinion dynamics in social media has generally assumed that knowledge of a user's "social neighbourhood" is known. One example of this is the list of individuals a particular user follows on Twitter, which provides a general idea of the types of tweets they are exposed to.

State-of-the-art performance opinion forecasting of Twitter users is achieved by [De et al., 2016] who model user opinion as a continuous-time stochastic process. The latent estimation of user opinion is affected by the range of sentiment expressed in the user's tweets, along with those made by their neighbours.

[Zhu et al., 2020] model stance dynamics over time using a Recurrent Neural Network (RNN) model. In addition to the textual features of the tweets made by an individual user in a given time period, the model also considers the features of recent tweets made by their neighbours as additional context. Tracking user opinion over time, they found that predictions made for periods where users did not tweet (and thus made solely from the neighbourhood context) generally aligned with their sentiment in future periods where they chose to participate.

[Das et al., 2014] find that while users are influenced by the opinions of their neighbours, we cannot simply aggregate the views of the users in their neighbourhood to determine their future opinion. They propose that future stance is also influenced by a user's level of stubbornness (fixation on own ideas) and conformity (tendency to adopt the opinions of others).

The above literature implies that knowing a user's connections and neighbourhood is important in deriving their future opinion. Additionally, different users who hold the same stance may react discordantly when presented with the same sentiments from their neighbourhoods due to individual tendency to stick to their own beliefs or conform to others - suggesting that additional context around a user's behaviour must also be taken into consideration.

2.1.2 Network Features for Discussion Forums

While the above work seems to necessitate knowledge of a user's neighbourhood network in predicting stance dynamics, this is difficult to determine for social networks that are primarily discussion forums - such as Reddit - where this type of information is hidden or impossible to determine due to the types of possible interactions.

This network is approximated by [De et al., 2014] who consider discourse between two individual users on Reddit as an undirected connection in their neighbourhood and uses this to predict user sentiment regarding political topics on Reddit. Surprisingly, this method performed better than the same models trained on Twitter data that created neighbourhoods from examining each user's follower data.

[Chua et al., 2007] find that measures of the overall network structure such as centralisation (influence of prominent individuals who have a high level of connection) and inclusiveness (number of connections in the network) are significant factors in predicting future participation on discussion forums.

The timing and ordering of atomic-level interactions on Reddit are also explored by [Horawalavithana et al., 2021] who show that sequence modelling can be used to incorporate these interactions and predict future thread engagement.

Despite the lack of more concrete network information compared to social media websites such as Twitter, these findings show we can still obtain meaningful network data from discussion forums such as Reddit from examining individual interactions between users and their ordering, along with the type of users and level of interaction present in the network.

2.2 Reddit Structure

In this project we focus on Reddit, an online social media platform with over 52 million active users each day and is used by 25% of adults in the United States [Dean, 2021]. Discussions take place in a "subreddit", which can be thought of as an overarching topic or category that categorises the discussions that occur inside. Any user may start a new discussion by starting a new post in a given subreddit - which creates a new thread. Users may then choose to comment a reply to one of these posts or to another comment made inside the thread, starting a new nested discussion. Figure 2.1 shows an example of such a nested discussion, along with a topological representation of the discourse.

Unlike other forms of social media such as Twitter where there are clear methods of endorsement such as "following" another user to subscribe to their tweets or "retweeting" another user's tweet, this information is not as clear on Reddit. While users may "upvote" and "downvote" posts, the purpose of this feature is to determine the relevance of posts to the topic, with more relevant threads and comments appearing higher based on their "karma" rating (although in practice many users use this to express their sentiment on the post [Graham and Rodriguez, 2021]). Additionally, information about which users have "upvoted" and "downvoted" particular posts is not shown to others. While the lack of endorsement methods and use of karma sug-

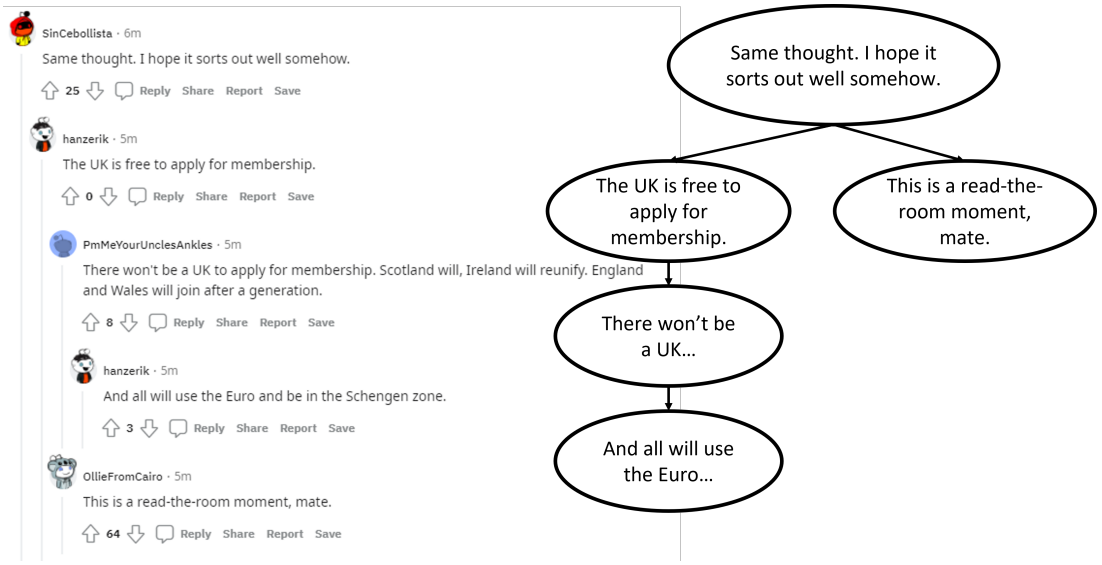

Figure 2.1: Structure of a Reddit discussion [Reddit, 2021]

gests that it is more difficult for filter bubbles to occur, this also makes it difficult to determine the sentiment of users due to having information on only a subset of their activity.

2.3 Dataset

The dataset used contains over 800,000 posts regarding the contentious topic of Brexit, a referendum that left many dumbfounded in 2016 [Clarke et al., 2017] as just under 52% of voters voted yes to pass the motion for the United Kingdom to leave the European Union [BBC, 2020]. While the referendum occurred over five years ago as of 2021, discussion in the Brexit subreddit remains vibrant due to redditors (Reddit users) dissatisfied with the result along with events such as the rejection of the first white paper and resignation of Theresa May reigniting discourse around the merits of Brexit.

In order to provide a distinction between current and future stance, the data has been split into 27 time periods based on certain events that have occurred. Table 2.1 shows high-level information about which events have determined the split in addition to the timeframe and number of comments each period spans.

The ground truth for the prediction task has been created by classifying the textual content of each post into one of three classes (Pro-Brexit, Against-Brexit, Neutral). While the classifier will be discussed further in Section 2.4.2, we then aggregate this to the user-level by labelling each user with the most frequently appearing class in the posts they have made in a given period.

In Section 4.1 we find that this level of detail is too shallow and reformulate

Period	Start Date	Comments	Important Event(s)
1	2015/11/16	3367	Referendum, David Cameron resigns
2	2016/06/26	6265	Theresa May accepts Queens's invitation to form government
3	2016/07/14	3084	UK House of Commons votes in favour of Article 50
4	2016/12/08	1466	Brexit is initiated
5	2017/01/27	2300	Two year process begins
6	2017/03/30	4102	Brexit negotations commence
7	2017/06/20	54505	White paper finalised, Secretary of State resigns
8	2018/07/09	23067	EU rejects white paper
9	2018/09/22	15385	Brexit withdrawal agreement published
10	2018/11/16	3718	Other 27 EU member states endorse withdrawal agreement
11	2018/11/26	25568	UK Government defeated in withdrawal vote
12	2019/01/16	54850	Second withdrawal vote is defeated; Extension vote passed
13	2019/03/15	9119	First request for Article 50 extension
14	2019/03/22	13414	Third defeat of UK Government
15	2019/03/30	9509	Second request for Article 50 extension
16	2019/05/25	27781	UK holds elections to European Parliament, Theresa May resigns
17	2019/08/29	78434	Boris Johnson becomes prime minister
18	2019/09/10	19662	MP's reject motion to call general election
19	2019/09/25	18872	Supreme court throws out PM's decision to prorogue parliament
20	2019/10/03	10608	White paper published outlining plan to replace Irish backstop
21	2019/10/18	17174	UK and European Commission revise agreement
22	2019/10/30	13546	Third extension of Brexit deadline
23	2019/12/14	30510	Conservatives win general election
24	2020/01/23	44958	Withdrawal Agreement Bill passes parliament
25	2020/02/01	12368	UK begins withdrawal from EU
26	2020/03/18	52131	EU publishes draft proposal for new partnership with UK
27	2021/01/01	215382	UK completes separation with EU

Table 2.1: Description of the timeframes the dataset is split into [Largeron et al., 2021]

Feature Set	Description	Features
F0	Textual features	- Textual features - Stance at current timeframe
F1	User activity	- Number of initiated diffusions - Number of submitted comments - Quantiles of number of received comments per post - Stance at current timeframe
F2	User activity per stance	- Number of comments submitted to posts from each stance - Quantiles of the number of received comments of each stance per post
F3	Diffusion overview	- Quantiles of the number of comments in diffusions the user participated in per stance - Stance at current timeframe

Table 2.2: Description of feature set used in previous work

stance as a continuous metric between -1 (polarised Against-Brexit) and 1 (polarised Pro-Brexit), with 0 indicating a user has posted only Neutral stance comments in a particular period.

2.4 Prior Work

2.4.1 Textual Content and Diffusion Overview Model

The past work this project aims to improve on utilised an earlier version of this dataset containing only the first 15 periods in Table 2.1, with ground truth generated using a stance classifier trained on Brexit Twitter data (achieving an F1 score of 0.88), employing transfer learning to classify the Brexit subreddit data. [Largeron et al., 2021] evaluated the effectiveness of textual features (F0 in Table 2.2) along with high-level descriptions of users (F1, F2) and the discussions they participate in (F3) on the future stance prediction task - given a particular user's stance and comments in period t, predict their stance in period $t + 1$.

While the feature sets include specific information about users such as the number of posts they have submitted directly to the Brexit subreddit (initiated diffusions), number of submitted comments (replies) and their current stance, there is also a high-level description of the types of threads a user engages in.

Feature set F3 computes the vector $[N_t^{sp_1}(u), ..., N_t^{sp_5}(u)]$ for each percentile $p_x \in \{0, 25, 50, 75, 100\}$. Letting y_s denote the vector containing the (normalised) number of posts classified as stance s $\in \{ProBrexit, AgainstBrexit, Neutral\}$ in each discussion user u participates in, $N_t^{sp_x}(u)$ represents the p_xth percentile of y_s for user u in period t.

Examining this vector can provide insight into the types of discussions a particular user prefers to participate in, for example a large value of $N_t^{ProBrexit25}(u_1)$ may indicate that user u_1 engages mainly in discussions with a high number of Pro-Brexit comments. On the other hand, a small value for $N_t^{Neutral100}(u_2)$ suggests that u_2 tends

to avoid posts with a large number of neutral comments.

[Largeron et al., 2021] find that feature set F3 is able to achieve the best performance, with a macro F1 score of 0.539 - outperforming models that include all four features (which will be denoted as the F0123 feature set from now on). Given that this is a 3-class classification problem, random guessing would produce a score of 0.33 (assuming an equally balanced dataset) suggesting that the classifier performs relatively well compared to random chance. In Section 3.2 we find that the dataset is in fact highly imbalanced. Random guessing results in an F1 score of 0.22 - implying that this result is more impressive than it initially suggests.

Given that the F3 feature set considers only the number of comments of each stance in each discussion, we seek to determine whether incorporating network-level features i.e. information about the authors of the comments that the user has interacted with can improve performance. In addition, we also wish to evaluate the effectiveness of considering atomic-level interactions i.e. the actual structure of each discussion thread as opposed to a high-level summary.

2.4.2 BERT Stance detector

While the results in Section 2.4.1 seem fruitful, the use of the Twitter classifier to determine the stance of Reddit comments imposes an additional layer of uncertainty that propagates throughout the rest of the process (Twitter classifier is used to classify Reddit comments $\rightarrow$ Reddit comments are aggregated to form user-level stance predictions $\rightarrow$ future stance predictor uses these predictions for both current and future stance ground truth).

A complementary project was undertaken by [Law, 2021] to improve the stance classifier and train it on (Brexit) Reddit comment data as opposed to Twitter data. To obtain actual ground truth for Reddit comments, they used Amazon Mechanical Turk to pay real people to examine over 5895 comments in the dataset and determine if they should be classified as Pro-Brexit, Against-Brexit or Neither (which we will assume is the same as Neutral in this project for consistency with previous work). Comparing the Mechnical Turk classes to the predictions made by the Twitter classifier yielded the shocking finding that roughly only 25% of comments contained the same class in both datasets. This was mainly due to the Twitter classifier predicting "Neutral" for a significant number of comments that Mechanical Turk users annotated with a non-"Neutral" stance - suggesting that the Twitter classifier is overly cautious in predicting stance.

This has major implications for the previous work in section 2.4.1, suggesting that the features must be reevaluated using a more accurate ground truth. In addition to extending the dataset by almost two years (adding 12 additional periods of comments), [Law, 2021] has trained a BERT (Bidirectional Encoder Representations from Transformers) language model on the text of the comments in the dataset, using the Mechanical Turk annotations as ground truth in order to produce a more accurate Brexit comment stance predictor which achieves an F1 score of 0.55 on the extended Brexit subreddit dataset. The work in this project utilises the stances from the BERT

model along with the extended dataset to ensure more precise current stance predictions are used in the resulting models and analysis.

2.5 Summary

In this chapter, we examine past work on opinion dynamics and the importance of a user's social neighbourhood in influencing their opinion. We discuss the difficulties in approximating this network for discussion forums and examine how features such as atomic-level interactions and overall network structure have been used to model interaction and opinion dynamics.

We then provide an overview of the Reddit social media platform and discuss the difficulty in determining user sentiment due to the lack of information available outside of a user's posts. We introduce the dataset along with current progress on the future stance prediction task. However, we find that the quality of the data used in previous work is quite poor compared to the real world ground truth. We address this by exploring the more accurate BERT stance predictor that will be used for the ground truth of individual comments throughout this project.

This poses the question of how well the feature sets used in previous work perform on this improved dataset. We will explore this in the next section, along with evaluating whether network-level features can enhance these results.

Chapter 3

Future User Stance Classification

In this chapter we will first examine the skewed nature of the distribution of users and comments in the dataset in Section 3.2. We also revisit the future stance classification task and compare network-level features and models that incorporate user-level interactions to the high-level descriptors used in past work in Section 3.3. In Section 3.4 We discover that this task goes well beyond a simple discrete classification problem due to the **current labels being inadequate in capturing the nuances of user sentiment** and the **conditional nature of the future stance prediction task**.

3.1 Assumptions

We will first introduce simplifying assumptions made throughout this project regarding the dataset. We assume that the stances for individual comments predicted by [Law, 2021]'s BERT model are correct and treat them as ground truth. We also treat each period defined in Table 2.1 as independent from external factors, with a given user's stance influenced only by their interactions in the previous period. Addressing these sources of uncertainty ensures that the project can be reasonably scoped, however avenues such as the impact of external events and outside context on a user's stance provide motivation and direction for future research.

3.2 Stance and Transition Distribution

In this section we will explore the extended dataset to show the lopsided nature of the stances of comments and types of users present in the dataset.

From examining Table 3.1 we can already see that comments posted in the Brexit subreddit are heavily skewed towards the neutral class, which makes up 78% of the comments posted. Additionally, the proportion of Against comments is almost

Comment Stance	**Against**	**Neutral**	**Pro**
Proportion	0.14	0.78	0.08

Table 3.1: Distribution of the predicted labels for each comment in the dataset

User Stance	Against	Neutral	Pro
Proportion	0.13	0.83	0.04

Table 3.2: Distribution of the predicted labels for each user in the dataset: authors who appear in multiple periods are treated as separate users for each period they are present

p \ p+1	**Against**	**Neutral**	**Pro**
Against	386	1790	58
Neutral	1608	18425	431
Pro	73	511	44

Table 3.3: Matrix of the user stance transitions between periods p and p + 1, $p \in [1, 26]$

double the proportion of Pro comments, suggesting that participants in this subreddit are generally more Against-leaning.

This is further exemplified when user-level stances are generated by aggregating the most frequent stance of each user in a given period as done in previous work by [Largeron et al., 2021]. Table 3.2 shows that while the number of Against users is relatively proportionate to the number of Against comments made, only 4% of users are Pro-leaning compared to the 8% of comments classified as Pro. While this implies that the dataset is heavily unbalanced and should be taken into account during evaluation, this is also partially due to the aggregation method used - which will be discussed further in Section 3.4.

Tabulating the types of transitions that occur (Table 3.3) shows that the majority of users transition to Neutral in period p + 1, regardless of their stance in period p. This goes against the notion of polarised users, as we would expect most users to remain in the same stance. While this could be regarded as insight, we will argue that this is a result of categorising users into only three, very rigid, classes along with the conditional nature of future stance ground truth in Section 3.4.

From the above observations, we see that both the stance of comments and types of users who participate in discussions on the Brexit subreddit are highly Neutral-leaning - with Against comments and users outnumbering those from the Pro class. We also observe counter-intuitive behaviour regarding the stance transitions of users, where most polarised users choose to switch to the Neutral stance rather than remain polarised.

3.3 Network Features for Future Stance Classification

In this section we will introduce two different forms of network modelling and apply them to the future stance classification problem. We then use the features from previous work (recomputed on the improved dataset) as a baseline to compare the performance of these models to.

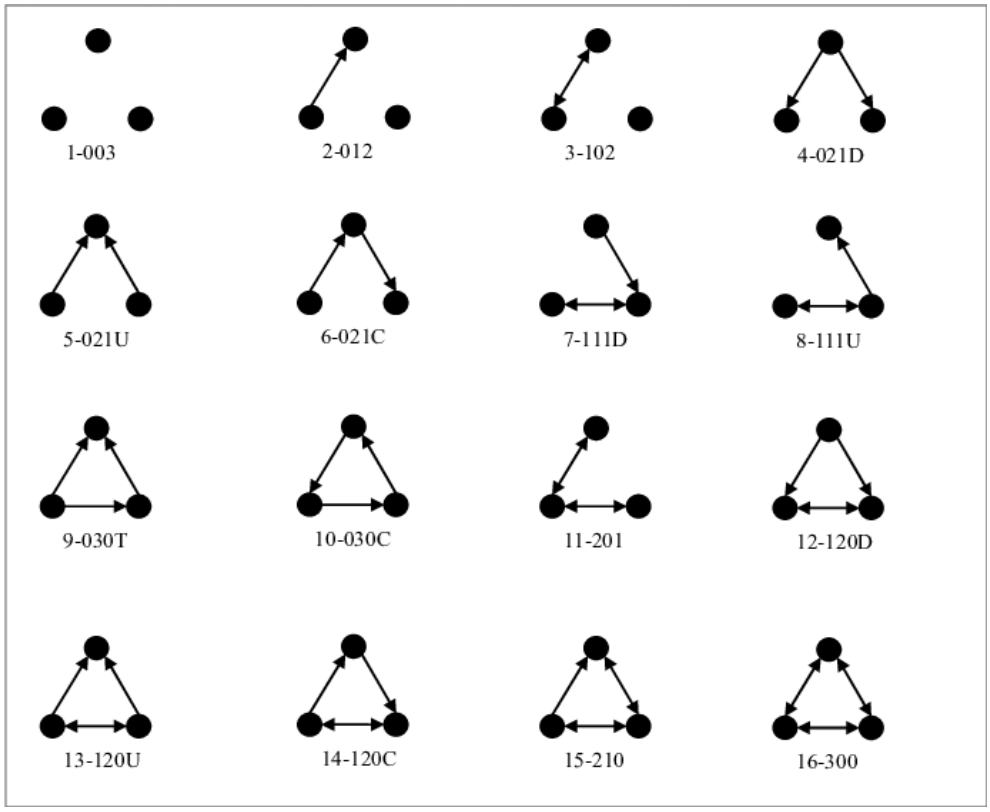

Figure 3.1: The 16 directed triads used in social network theory [Uddin and Hossain, 2013]

3.3.1 Triadic Closure Features

Triadic closure is a social network theory concept that revolves around the idea that if two individuals, A and B, each have ties with another individual C, there is an implicit connection between individuals A and B. Triads are the connections between three individuals in a network. Figure 3.1 shows the 16 possible combinations of triads in a directed graph.

We leverage this concept to produce features that aim to capture connections between different users. Rather than utilising the 16 directed triads, we instead propose an undirected model (as done in the work by [De et al., 2014]) that considers the stances of users in the triad definitions. In addition to the links between the users in a triad, we incorporate a further distinction between the stances of each user. To avoid confusion with the traditional definition of triads, we will refer to these as *stance triads*. Figure 3.2 shows an example of three different stance triads - with nodes colour-coded by stance (Blue: Pro, Grey: Neutral, Red: Against). While the stance triads containing users ABC and DEF would be considered the same under traditional triadic closure theory, this notion of stances provides additional information about the types of users an individual has interacted with.

Due to the explosion of feature space size when incorporating stance into triadic

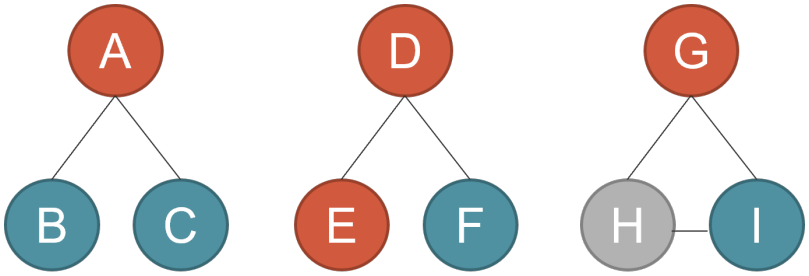

Figure 3.2: Examples of stance triads

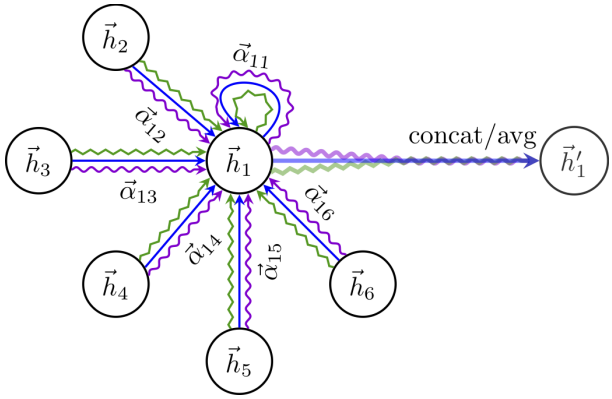

Figure 3.3: Illustration of the multi-head attention mechanism in the GAT model

closure theory, we assume that there is no significant distinction between the direction of interaction and consider only undirected triads. We generate these stance triad features by first computing a (undirected) network graph of all interactions between users in a given period. For each unique stance triad, we assign a feature value to each user by tallying the number of matching triads they are a part of. These features were evaluated on the future stance classification task. The performance of these models will be discussed in Section 3.3.3.

3.3.2 GAT Model

In addition to triadic closure inspired features, we also evaluate the effectiveness of the Graph Attention (GAT) model which leverages the attention mechanism commonly found in NLP models. Predictions for a specific node in the graph are generated based on its own features along with features of neighbouring nodes. Attention is used to weight neighbouring features based on their importance to the current context.

An example of this can be seen in Figure 3.3 where a new representation for vector $\vec{h}_1$ is calculated as the weighted composition of its original representation along with its neighbours $\vec{h}_2 ... \vec{h}_6$. α_{1n} denotes the weight assigned to node n, i.e. how much "attention" should be paid to n, in the new representation $\vec{h}'_1$. The different coloured

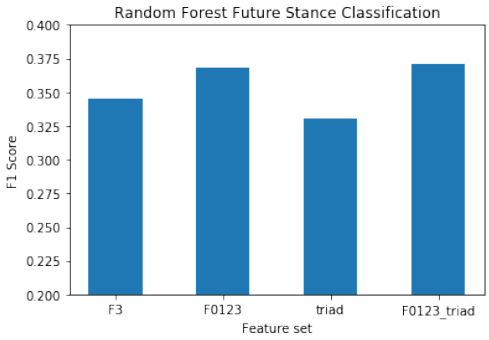

Figure 3.4: Random forest results

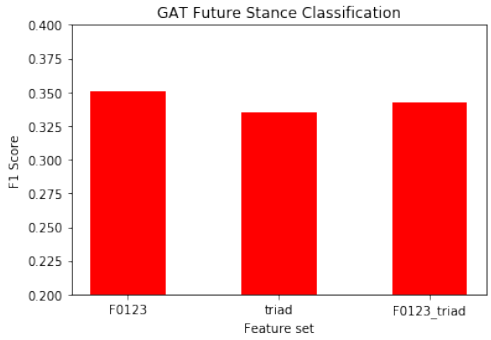

Figure 3.5: GAT results

lines in the figure represent the multiple heads which are aggregated to form the final representation - which can be attached to a classification head to produce predictions.

This provides a finer level of detail compared to stance triads as more than two neighbours can be considered at a time. We leverage this approach by denoting the users as nodes and assigning them features from both the F0123 and stance triad feature sets. We then train the model on the future stance classification task.

3.3.3 Methodology and Results

In Figure 3.4 we present the results of random forest models trained on the F3 and F0123 datasets from previous work and stance triad features along with evaluating them both as the combined feature set F0123_triad. Due to the class imbalance, macro F1-score was used throughout this project. The training methodology used in the previous work was adopted for the random forest models, where stratified 3 fold nested cross-validation was used to select hyperparameters and evaluate the selected models. The GAT models were trained by splitting the dataset into train, validation and testing sets, and performing grid search to find the best performing hyperparameters in the validation set. Final performance scores were obtained by evaluating the best performing model on the testing data.

We can see a clear impact of the improved and extended dataset, with none of the models coming close to the F1 score of 0.539 obtained by [Largeron et al., 2021] on the Twitter-trained dataset. While all features clearly outperform the random chance baseline of 0.22 macro F1-score mentioned in Section 2.4.1, the F0123 feature set performs significantly better than using only the diffusion overview feature set (F3) - producing contrary results to previous work.

The stance triad features perform very poorly compared to the other features, achieving a score of only 0.330. Comparing the results of F0123 and F0123_triad suggests little improvement from adding the stance triad features, improving test F1-score from 0.368 to just 0.371.

Furthermore, the results from the GAT models (Figure 3.5) suggest that incorporating network features in the form of triads actually leads to worse performance.

F0123 outperforms both triad and F0123_triad, and the random forest models score better the GAT models trained on the features (with the exception of triad). We hypothesis that the former is due to stances triads contributing little-to-no additional information as the GAT model already takes into account relations between users through performing attention between a user and all it's neighbours (as opposed to only two neighbours, as with triads). The GAT model likely outperforms the random forest model when both are trained only on stance triad features due to having additional information about the exact users that a given individual interacts with (as opposed to only the types of the triads they are part of).

3.4 Issues Regarding User Stance and Presence

Against	Pro	Neutral	Freq
0	0	1	0.31
1	0	0	0.10
0	0	2	0.09
0	0	3	0.04
1	0	1	0.04
1	0	2	0.02
0	0	4	0.02
1	0	3	0.02
0	0	5	0.01
2	0	0	0.01
0	1	0	0.01
1	0	4	0.01

Table 3.4: Frequency of different combinations of comment stances per user in a given period with frequency ≥ 0.01

Introducing network features in the form of stance triads, along with the GAT model in Section 3.3 did not appear to produce any meaningful improvement in the future stance predictions. Additionally, in Section 3.2 we find odd behaviour where the majority of polarised users become Neutral in the next time period. In this section we will attempt to explain this by addressing issues and areas of uncertainty present in the current approach due to the discrete stance aggregation method, along with trends regarding user participation in the Brexit subreddit.

Although the current approach to classifying the stance of users by taking the majority class of their comments is simple and intuitive, this leads to only extremely polarised users being classified as either Against or Pro Brexit due to the skewed nature of the comments in the dataset. This is likely one of the reasons for the surprisingly low proportion of users classed as Pro Brexit in Table 3.2 when compared to the overall distribution of comment stances.

In Table 3.4 we see that four percent of users posted one Against comment and either two or three Neutral comments in the same period. Despite expressing some

level of Against-Brexit sentiment, these users would be considered the same as an individual who posted ten comments which were all Neutral. Some of the "Neutral" comments made by these "semi-polarised" users were simply clarifications or words such as "yes" or "agreed" which require more context to interpret. This suggests that the current stance aggregation method is too naive and aggregates only based on the extremes. More leniency in the user-level stances would allow us to better quantify the stance and polarisation of a user.

Pattern	**Frequency**
1	0.705
11	0.050
101	0.013
1001	0.010
111	0.009
10001	0.008
100001	0.005
1111	0.005
10000000001	0.005
1011	0.005

Table 3.5: Frequency of different user stay/leave patterns where 1 denotes a user being present in the network for a particular period and zero denotes that a user is not present

Table 3.5 shows the distribution of user presence in the network. We see that the majority of users (over 70%) only ever appear in one period. Additionally, we note that users who take a break from posting in the subreddit before returning (such as patterns 101 and 1001) are not considered in the future stance prediction task due to not appearing in two consecutive periods.

This indicates that our predictions are severely biased as the model is trained only on the small number of users who choose to post in two consecutive time periods. We must take care in our interpretations to consider the fact that the future stance prediction task is actually **conditional on the user choosing to participate in the next period**. This makes the original question of answering how interacting with different opinions and views influences future stance much harder than initially anticipated as we do not have information about the future sentiment of over 70% of the users in the dataset.

3.5 Summary

In this chapter, we find that the poor performance of the Twitter comment stance classifier used in the original work significantly affects the reliability of the presented results. We examine the effectiveness of introducing network information in the form of stance triads and considering all neighbours using the GAT model, however these

do not appear to produce any meaningful improvement to the features used in the previous work.

We find that this may be in part due to the naive assumption made that user stance can be aggregated into just three discrete classes - resulting many users being classified as "Neutral" when their actual stance may be more nuanced. Additionally, we find that the majority of users choose not to remain active beyond a single time period, which has major implications on the interpretation of our future stance prediction models and their ability to answer our main research question. These insights cause the original project to be significantly more complicated than the discrete classification task initially outlined due to requiring a new way to quantify user stance, along with taking into account the conditional nature of our predictions.

To address the first issue, we will introduce a continuous notion of user stance in Chapter 4. We will then address the users not accounted for in the stance prediction task by considering a new question in Chapter 5 - what makes people leave?

Chapter 4

Future User Stance Regression

In this chapter we will introduce and explore a continuous metric for quantifying the stance of users. We will then use this metric to create high-level network features that provide insight into the types of interactions that occur on the Brexit subreddit. After this, we will evaluate the effectiveness of these features on the future stance regression task and apply the SHAP feature importance framework to interpret how the best-performing model makes its predictions.

4.1 Continuous Polarity Metric

We define polarity for a user in a given period $\in [-1, 1]$ as

$$\frac{1}{2}(\frac{P-N}{P+N}+1) - \frac{1}{2}(\frac{N-A}{N+A}+1) \tag{4.1}$$

where P: number of Pro posts, N: number of Neutral posts, A: number of Against posts by that user in the period. This weights both the number of Pro and Against comments made by users against the number of Neutral posts made. This provides a more nuanced overview of an individual's stance and their level of polarisation. A user with polarity 1 is very clearly Pro-Brexit, while a user with polarity 0.10 may be slightly Pro-Brexit leaning but still uncertain - the latter user would be one of the "undecided" users that would be targeted if we wanted to sway user opinion.

One note of caution for this metric would be that if a user posts an equal number of Pro-Brexit and Against-Brexit posts, they would be assigned a polarity of 0 - identical to users who post only Neutral comments. While this may abstract away any distinction between the groups (perhaps users in the former group should be categorised as "indecisive" rather than Neutral), in practice this does not affect many users. Less than 4% users posted an equal (non-zero) number of Pro-Brexit and Against-Brexit comments, with 70% of them posting a comment with exactly one of each stance. Thus we will ignore this nuance, however accept it as a valid concern for datasets with a larger proportion of such "indecisive" users.

Looking at the distribution of polarity in Figure 4.1, we can immediately see the impact of switching to a continuous metric. While there are pockets of extremely polarised users at either end of the spectrum, we also see varying levels of polarity

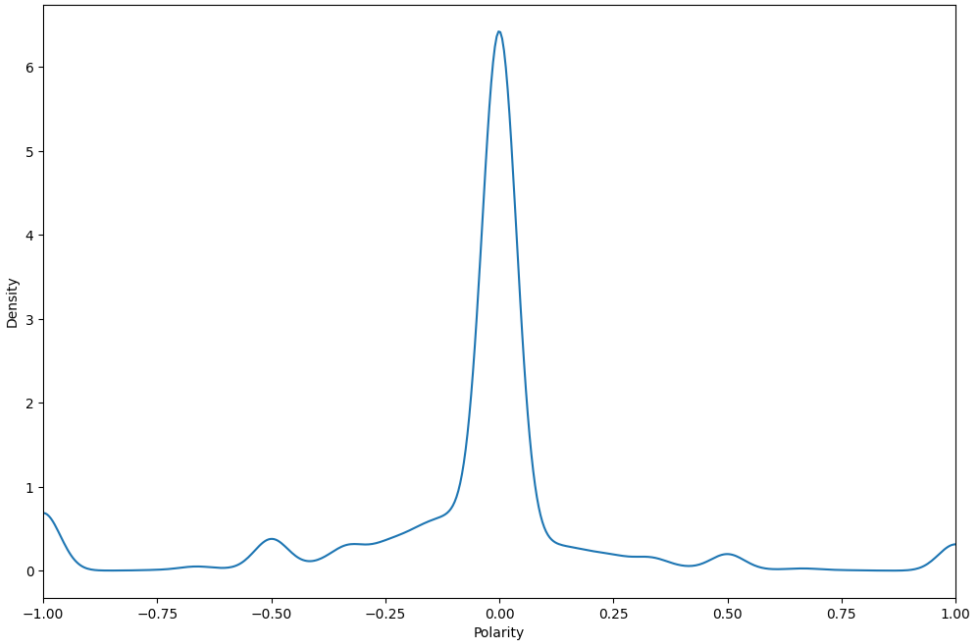

Figure 4.1: PDF of polarity distribution for all users across all periods

between the neutral stance and completely polarised. Much like when using discrete labels, the majority of users are considered "true Neutral" (polarity = 0) - however the proportion of users with a polarity of 0 is 58%, compared to the 0.83 calculated in Table 3.2 for discrete classes, implying that this metric provides a higher level of granularity of how "Neutral" these users are.

We also examine the distribution of users who are considered "Pro-" (polarity > 0) or "Against-" (polarity < 0) leaning in Figure 4.2. To account for the relative size of each group we plot the cumulative density (CDF) rather than the PDF. We see no major differences in the shapes of these curves, suggesting that the distribution of individuals within each group is relatively similar. However, it is worth noting that the cdf of Pro-leaning users grows more quickly than Against-leaning users. This suggests that Against-leaning users are generally more polarised, however this effect appears to be very slight and may be influenced by the previously mentioned fact that the dataset contains substantially more comments labelled as Against than Pro (Table 3.1). This may also suggest the effect of selective exposure, as the relative dominance of Against-Brexit sentiment compared to Pro-Brexit may attract more polarised Against-Brexit users to the subreddit.

4.2 Edge Polarity Features

Due to the stance triad features requiring a discrete notion of stance to "colour" each node, they cannot be directly be used as features when stance is quantified as a continuous metric. We will evaluate two features that take advantage of the

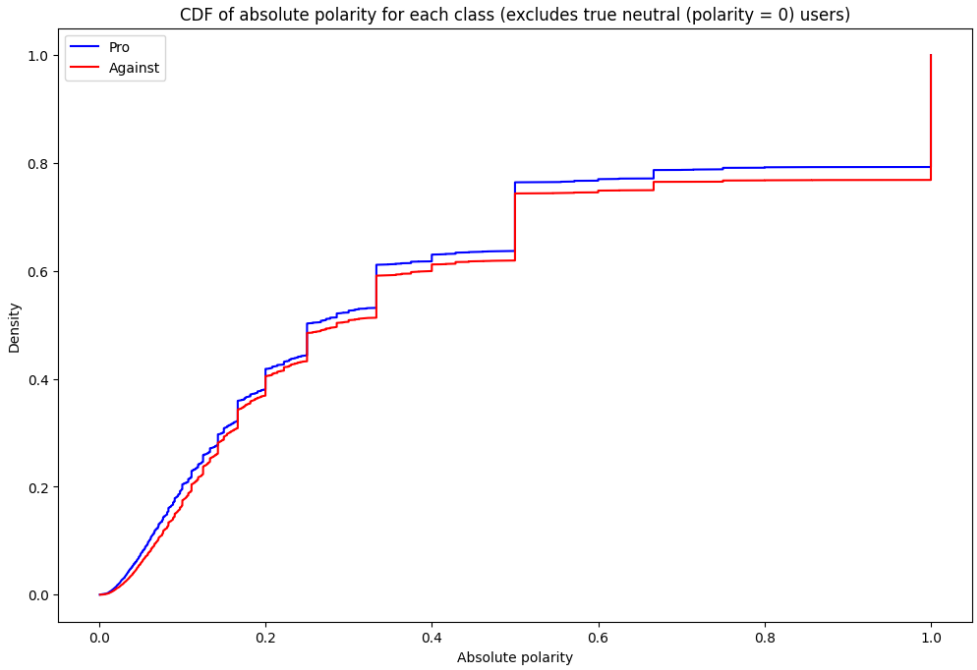

Figure 4.2: cdf distribution for absolute polarities of "Pro-leaning" (polarity > 0) and "Against-leaning" users (polarity < 0)

continuous polarity metric: interaction polarity and edge homogeneity.

For any individual i, we denote the $\vec{n}_i$ as the vector containing all users that i has interacted with (replied to or received a reply from) in a given period. We also define the function $polarity(u)$ as the continuous polarity of user u - so $polarity(i)$ would express the polarity of user i and $polarity(\vec{n}_i)$ is the vector containing the polarity of each user in $\vec{n}_i$.

$$interaction\_polarity(i) = polarity(\vec{n}_i) \tag{4.2}$$

$$edge\_homogeneity(i) = polarity(i) * polarity(\vec{n}_i) \tag{4.3}$$

We define interaction polarity in Equation 4.2 as the polarity of each individual that user i has interacted with. This provides a high level summary of the types of individuals that i has engaged with - perhaps they choose to talk only to users that lean towards a particular stance, or those with low levels of polarisation. However, this does not incorporate the context of the interaction.

To remedy this we utilise edge homogeneity - which has previously been used in literature to investigate homophily and cascade dynamics [Del Vicario et al., 2016]. Given that the continuous polarity metric is bounded between $[-1, 1]$, taking the product of the polarity of two interacting users provides insight on the circumstances. A positive value indicates discourse between two agreeing users, while a negative value indicates the opposite. The value of this metric also quantifies the intensity of

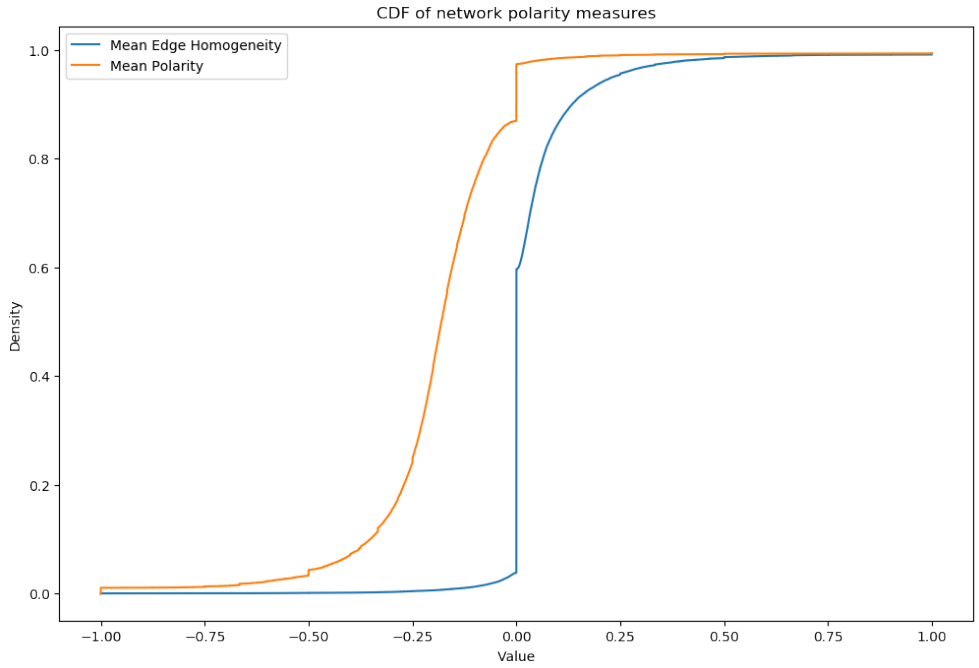

Figure 4.3: CDF of network features

this interaction e.g. a value of -1 implies discussion between two extremely polarised users that disagree with each other.

Thus we denote edge homogeneity as the product of the polarity between a user and all individuals they interact with in Equation 4.3. This provides additional information about the types of interactions that a user participates in when compared to the actual polarity of each user.

In Figure 4.3 we examine the interaction polarity and edge homogeneity of each user by taking the mean of each vector and plotting the distribution. We notice that the majority of mean interaction polarities lie below zero. This is not surprising due to the Against/Neutral-Brexit leaning composition of post and user stance labels that we have previously observed.

Additionally, the proportion of users with a negative mean edge homogeneity is very low and almost 40% of the user-level interactions are "positive". This suggests evidence of an echo chamber effect, as users rarely engage with individuals that have opposing beliefs.

4.3 Future Stance Regression

Due to changing from a discrete to continuous metric for polarity, we now have a regression task rather than a classification task for predicting the future stance of remaining users. We incorporate the edge polarity features in the previous section by calculating the five number summary {min, 1st quantile, median, 3rd quantile, max}

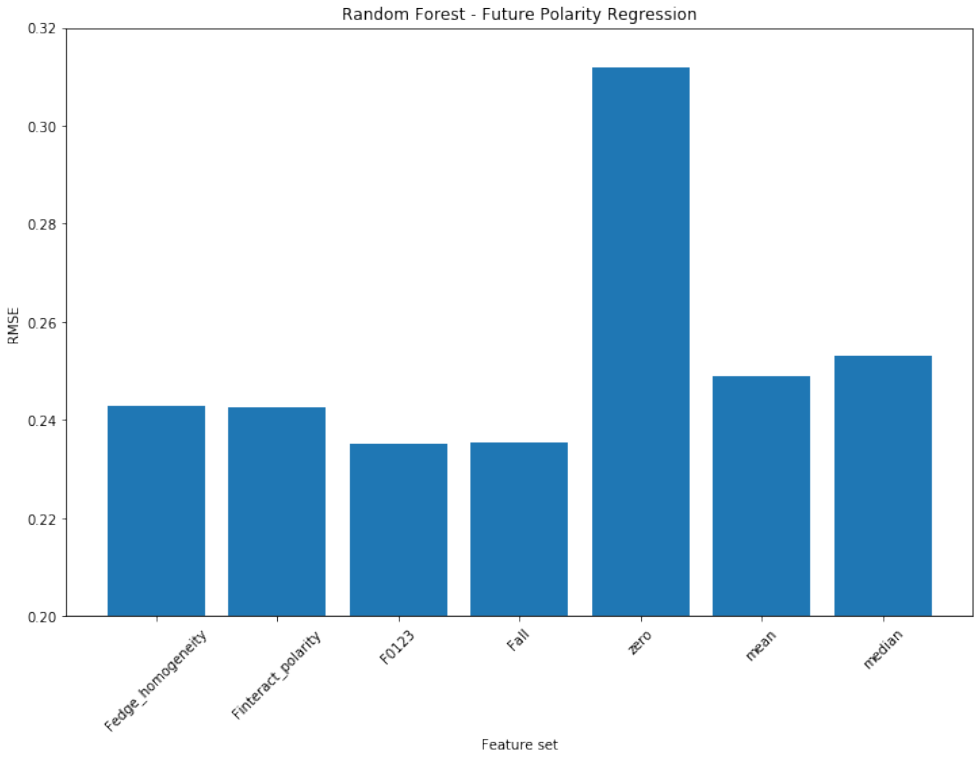

Figure 4.4: Random forest regression results

for each user's interaction polarity and edge homogeneity vectors and using these as features along with user's current polarity and mean interaction polarity/edge homogeneity.

We train random forests model on the future stance regression task with each of these feature sets along with the F0123 and Fall (combination of all features) using the same 3 fold nested cross-validation methology used for the future stance classification task in Section 3.3.3. Due to the continuous nature of predictions, we cannot calculate an F1-score and instead use Root Mean Squared Error (RMSE) for evaluation. We use the loss when trivial predictions are made for all users (i.e. zero, the mean and median of all polarities in the training set) as a baseline for comparison.

4.3.1 Results

We examine the results in Figure 4.4, noting that lower is better when comparing loss. While all feature sets perform better than the trivial baselines, the edge polarity features do not appear to provide any additional useful information when compared to the F0123 feature set - in fact, combining the edge polarity features with F0123 results in a (slightly) higher loss.

This suggests that the features are too high-level in the information about the user network they provide - we only calculate the polarity of immediate connections in each user's social neighbourhood.

4.3.2 SHAP Feature Analysis

To interpret the best performing model (F0123) we use the SHAP (SHapley Additive exPlanations) framework to determine feature importance and quantify the impact of feature values on predictions [Lundberg and Lee, 2017]. This is a game-theoretic approach to feature importance that aims to quantify the contribution of each feature to the final "payout" (prediction).

For each input to the fitted model, SHAP values are calculated for each one of it's features. We define the *base value* as the mean of all predictions. The SHAP value of a feature for an input is how much it influences the prediction from the base value. Adding together all the SHAP values for each feature along with the base value produces the final prediction for the input. Due to the potentially competing nature of these values they can be thought of as "forces", pushing and pulling the base value in different directions.

An illustration of these forces in action can be see in the force plot in Figure 4.5. We can see that the base value for the regression is -0.1889 and the final prediction of this particular user's future polarity is -0.22 (bolded). The plot shows the impact of different features on the regression result. We see that the features with the most impact on the final prediction are Polarity (the user's polarity in the current time period) - which was -0.1175 in the current time period, pA-75% and pA-50% (the meaning of these feature names will be explained shortly). Note that the equated values are the values of each feature for the user, not the SHAP values. From the size of the bar, we can see that the "force" (SHAP value) of the pA-75% feature was around -0.04 - pushing the model to lower the predicted value (i.e. more left-leaning). After these values have been calculated for all features and all users, we determine the overall impact of each feature by averaging the absolute values of its SHAP values for all inputs.

Features px-y% where $x \in A, N, P$, $y \in 0, 25, 50, 75, 100$ are the diffusion overview model features (F3) introduced in Section 2.4.1. To briefly recap, if vector $\vec{d}_x$ contains the proportion of posts of stance x (A, N, P represent Against, Neutral and Pro respectively), px-y% is the value of the yth percentile of this vector.

Figure 4.6 shows the SHAP value plot for the 5 most significant features (in descending order of significance). This is a three-dimensional plot (the third dimension being colour) of each feature's influence on the predictions of input (i.e. user). The x-axis denotes the SHAP value of each feature on the given user's prediction. The density of these values is represented by the number of dots (individual users) in the area. Finally, colour denotes the relative value of the feature for a particular user.

It is clear that Current Polarity has a major influence on predictions. The big red cluster for "Polarity" implies that users with a high polarity value (e.g. Pro-Brexit) encourage the model to predict a higher future polarity value (compared to the baseline). The sea of blue dots on the left of the zero line suggests a similar effect for users with lower polarity values (i.e. Against-Brexit) who encourage a lower future polarity prediction, however the effect is less pronounced (likely due to the baseline already being a negative polarity). This indicates that users who choose to

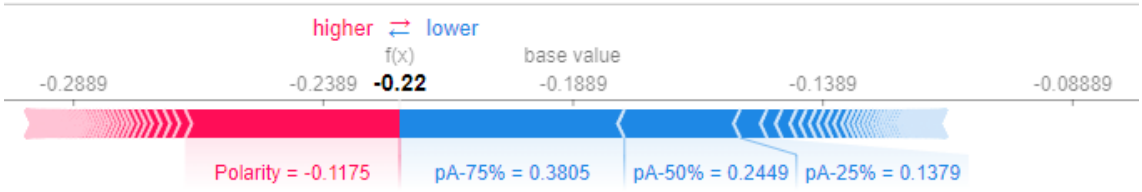

Figure 4.5: Example of a SHAP force plot

remain active (recall that this is a conditional problem) generally do not change their opinion.

The SHAP value plot also implies that the level of Against-Brexit sentiment in the discussions the user engages in also has a strong impact on the future stance prediction. Participation in mainly Against-Brexit leaning discussions (i.e. high values of $pA-50\%$ and $pA-25\%$) correlates to lower future polarity predictions. This is also the case for $pA-75\%$, which implies that participating in even a small number of highly Anti-Brexit dominated discussions encourages the future polarity prediction to lean towards Against-Brexit. A similar (but smaller effect) exists for the Pro-Brexit sentiment, which can be seen by the plot of $pB-75\%$.

While this appears to suggest that users who have participated in discussions with high levels of Against-Brexit comments and choose to participate again in the next period are more likely to form a negative sentiment towards Brexit, this is more likely a result of the echo-chamber effect. Recall that the majority of mean edge homogeneities in Figure 4.3 were greater than or equal to zero.

Instead, we posit that users who remain in the system are generally rigid in their beliefs, and choose to keep participating in echo-chambers that reaffirm their own ideals.

4.4 Summary

In this chapter we introduce a continuous notion of polarity and demonstrate how it introduces further context about a user's stance along with their interactions by considering the polarity and edge homogeneity of the individuals they have interacted with.

While useful for profiling the network, these features do not appear to be powerful enough to provide additional information for the regression of future stance when compared to features from previous work.

Applying feature importance methods to the regression model suggests that polarity in the current time period along with the level of Anti-Brexit sentiment in the discussions that users participate have a significant impact in the model's predictions.

We interpret the SHAP value plots and deduce that **users who choose to remain in the network are polarised and unlikely to be swayed**.

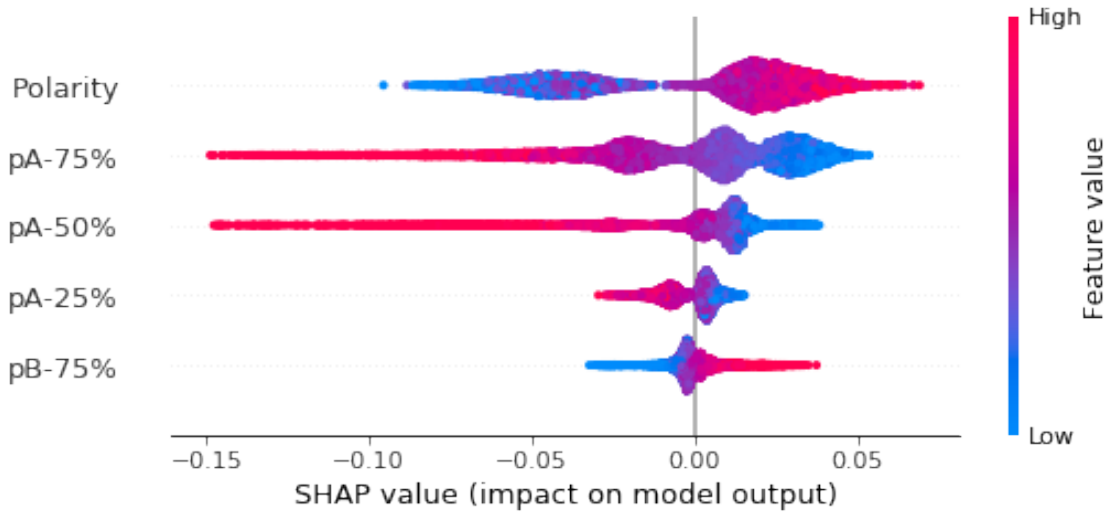

Figure 4.6: SHAP values for the five most significant features in the F0123 random forest model for future presence

Chapter 5

Future Presence Analysis

In the previous chapter, our predictions and analysis were conditional on users remaining in the system. However, this makes interpretation difficult along with leaving out the majority of users in the dataset. In this chapter, rather than focusing on the users that remain, we ask ourselves: what makes users leave? We tackle the future presence classification task - i.e. given a user's behaviour and interactions in this time period, will they also remain active in the next period? Unlike the future stance prediction task, we have complete knowledge regarding the presence of each user in a given time period so our interpretations are not conditional on the user remaining. The ground truth is also not influenced by the uncertainty of a stance predictor. In Section 5.1 we first explore two features that when combined provide a high-level overview of the user's interaction and participation habits. We will fit and interpret future stance classification models to evaluate these features in Section 5.3. Section 5.4 then introduces a pipeline that uses sequence modelling to model atomic-level interactions between users.

5.1 Activity and Degree Features

To introduce additional network information that complements the edge polarity features explored in Chapter 4, we attempt to quantify the size and diversity of interactions a user is involved in.

We first define *degree* as the number of unique users that an individual has interacted (replied to or received a reply from) with in a particular period. Note that this is an undirected measure of connection as it does not consider the direction of interaction, but provides a general overview of the size of a user's neighbourhood.

We then define *activity* as the number of comments that a user makes in a given period. This is a measure of how often an individual chooses to interact with the network.

Looking at both features allows us to gain insight about the posting habits of a user and the types of content they post. For example, high activity and high degree suggests that the user is very active and chooses to engage in discourse with many individuals (even if they do not receive many replies). On the other hand, low activity and high degree may indicate that a user posts particularly interesting or provoking

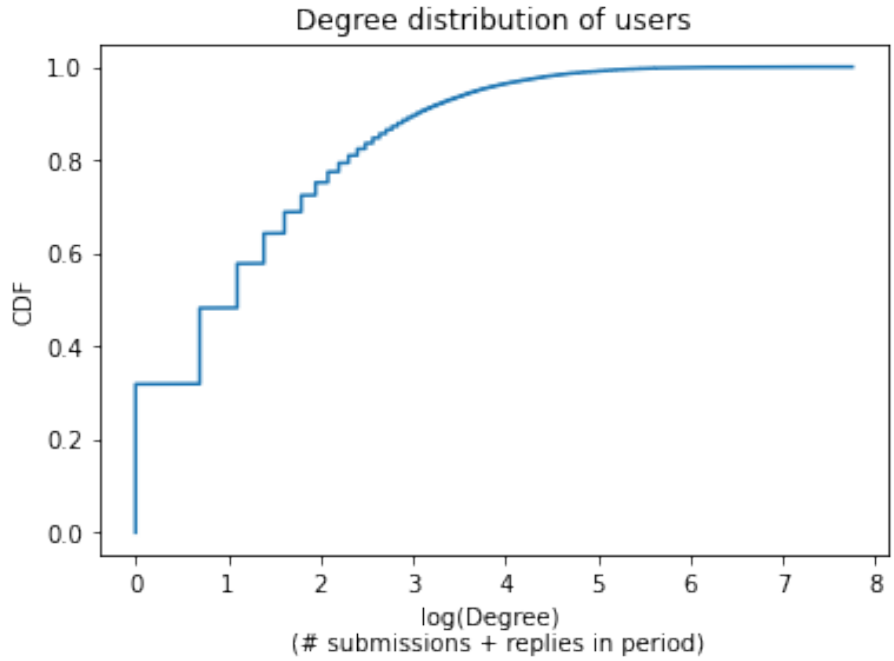

Figure 5.1: log distribution of User degree

content that sparks many replies.

Figures 5.1 and 5.2 suggest that both these features follow long-tailed distributions, where there are many users with low values for these features and very few with extremely high values. This is consistent with past observations regarding social network behaviour [Du et al., 2012].

5.1.1 Leave heat map

To examine the correlation between activity and degree on future presence, we plot these features against each other in Figure 5.3. For each "bucket" $b_{ad} : a, d \in [0, 24]$, we calculate the proportion of users with activity $\in [p_{activity}(a), p_{activity}(a+4))$ and degree $\in [p_{degree}(d), p_{degree}(d+4))$ who choose to participate again in the next period - where $p_x(y)$ is the percentile function for feature x.

The bottom left and top left quadrants suggest that a higher level of degree and activity correlates with a higher level of future participation, which is intuitive as users who are more engaged with the subreddit would likely want to come back. However, there is also a very prominent section of users who choose to remain that contains users above the 96th percentile of degree and between the 44th and 64th percentiles of activity. These users have a very high level of degree but comparatively lower activity, suggesting that they post sparingly compared to the number of replies they receive.

It turns out that due to the extreme combination of values, only two users are present in this area. Both users are very Anti-Brexit polarised (polarity -0.5, -0.667)

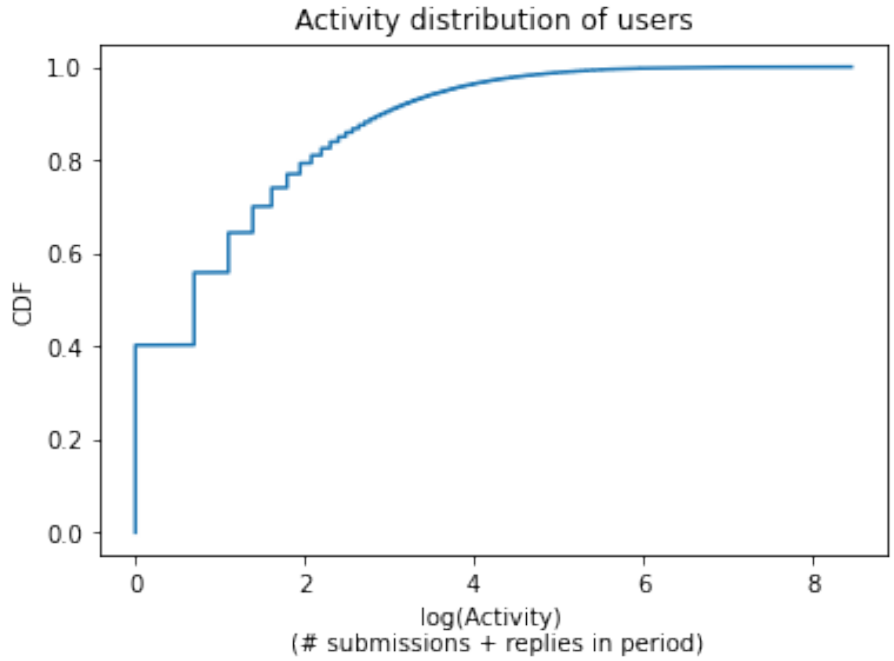

Figure 5.2: log distribution of User activity

and contributed two and three times respectively. While the second user appears to have incited attention for submitting a failed attempt at comedy, the first user is present in nine out of the twelve periods since they first started posting on the subreddit. In all but one of these periods (in which they were classed as neutral), they were labelled as Anti-Brexit leaning and exhibited a similar trend of posting a small amount compared to their level of degree.

Looking at the posts made by the first user, they generally appear to be more humour-based or extremely polarised opinions rather than attempts to start discussion. Examples of their content include: "But but fish, blue passports blah blah", "Stop the madness. Stop Brexit.", and "You underestimate the power of the blue passport!"

This suggests that the user chooses to keep returning due to positive reception to their jokes (which aim to make fun of those on the Pro-Brexit side) and/or to continue sharing their polarised opinions.

5.2 Influential Users

While we have investigated the effect of a user's own neighbourhood on future presence, this begs the question: how does interacting with *extreme* users - individuals with a significantly high level of activity and/or degree - affect future behaviour?

We consider "extreme" users to be those at or above the 99th percentile of degree and/or activity. We create two contingency tables, Tables 5.1 and 5.2, depicting the

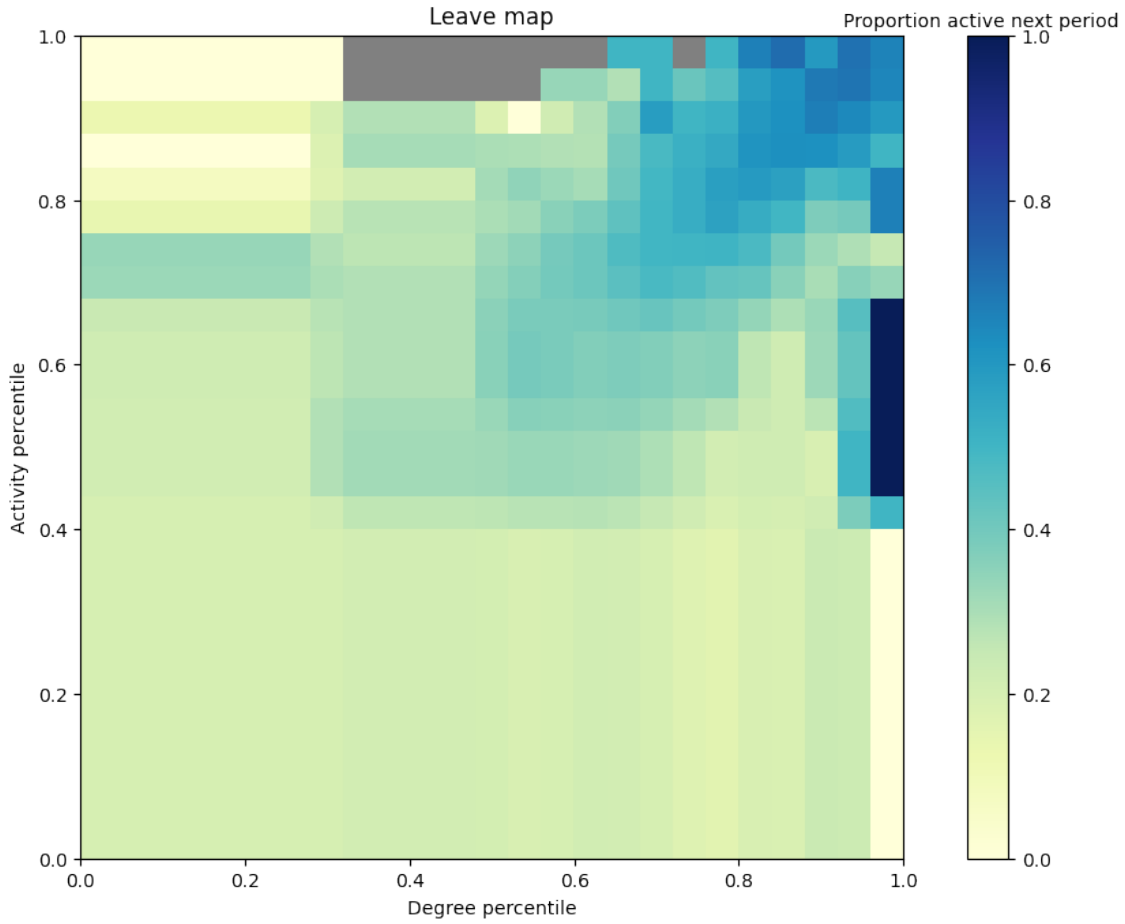

Figure 5.3: Activity-polarity percentile leave map

	Remain	**Leave**
Interact with extreme activity user	14259	14184
Does not interact with extreme activity user	9743	20963

Table 5.1: Contingency table of users who remain/leave given their interaction/lack of interaction with users with degree above the 99th percentile

	Remain	**Leave**
Interact with extreme activity user	13234	13190
Does not interact with extreme activity user	10768	21957

Table 5.2: Contingency table of users who remain/leave given their interaction/lack of interaction with users with activity above the 99th percentile

number of users who remain or live, separated by whether or not they have interaction with a high degree/activity user. Note that (unsurprisingly) 73% of extreme users belong in both categories - so there is considerable overlap.

Assuming the null hypothesis that the proportion of users who remain active in the next period given that they have interacted with a user with extreme degree/activity is no different to the proportion of users who remain active who have not interacted with such users, we conduct Chi-squared contigency tests on both tables.

This produces Chi-squared statistics of 2072 and 1788 for degree and activity respectively, corresponding to p-values of practically 0. Under a confidence level of $\alpha = 0.05$ we reject the null hypothesis and conclude that there is a significant difference between these groups of users. We can infer the direction of this relationship by looking at the difference in remain and leave rates between each category in the aforementioned tables to see that this implies interacting with extreme users correlates significantly with remaining active in the next time period.

This supports the ongoing hypothesis throughout this thesis that taking into account network level interactions provides additional information in modelling future user behaviour.

5.3 Random Forest Classification and Analysis

We generate a random forest baseline for the future classification task using the same methodology as the previous future stance classification and future polarity regression tasks (stratified 3 fold nested cross-validation), however this time we predict the presence of each user in the next period. We denote leaving the network with class 0 and remaining with class 1. We combine the previously discussed interaction polarity and edge homogeneity feature sets with each user's degree and activity to produce the Fnet feature set. Like before, we also evaluate the F0123 feature set and Fall - the combination of Fnet and F0123.

The F1-scores in Figure 5.4 suggest that while all models perform much better than random chance (0.5 F1-score for a two class classification problem), there is

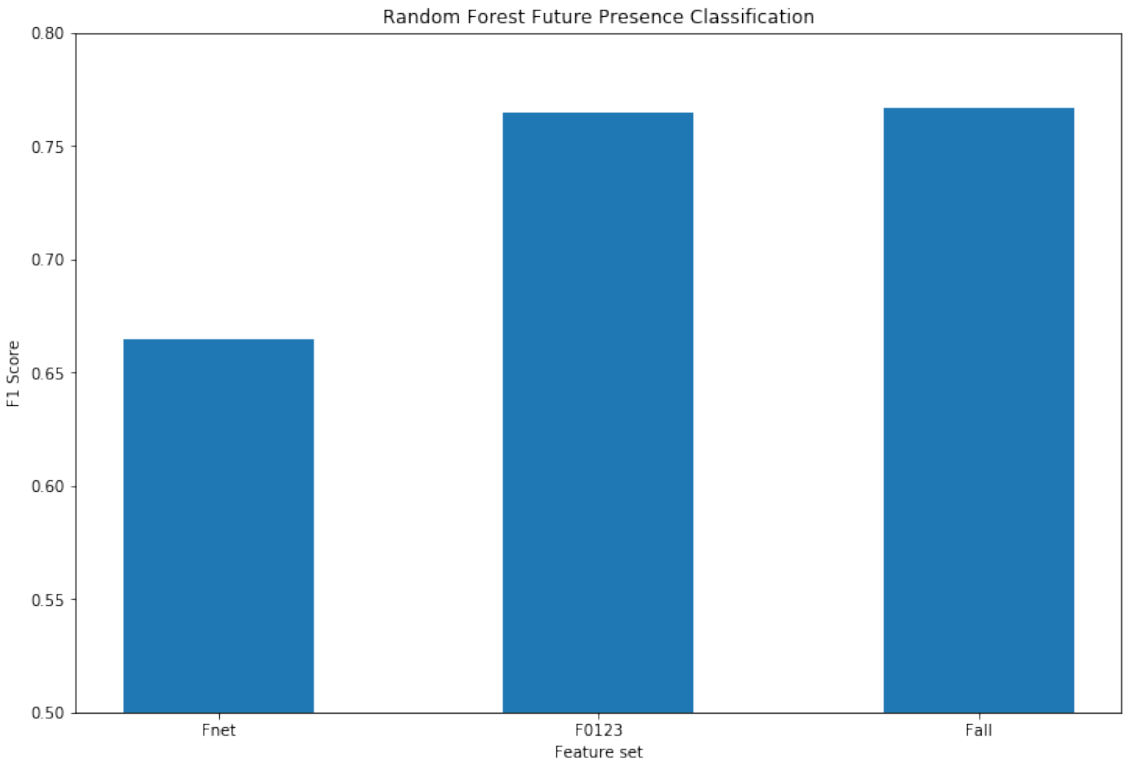

Figure 5.4: Random forest results on future presence classification task

little additional information added by Fnet. While there is a performance increase from adding the Fnet features to F0123, it appears to be negligible.

As remaining present in the next time period is class 1, a positive SHAP value in Figure 5.5 implies the feature value encouraged the model to predict the user to remain. We can see from the red clusters for $pB - 100\%$ and $pA - 100\%$ that participating in at least one highly polarised discussion greatly influences the model into predicting remain. We also observe interesting behaviour that a high value for $pA - 0\%$ (implying that the user engages only in very polarised Against-Brexit discussions) actually decreases the likelihood of remaining. We observe a similar phenomena for feature $pN - 0\%$, where participating only in predominantly neutral threads is associated with leaving - however this agrees more with previous insights regarding mainly polarised users remaining in the system.

5.4 Sequence Modelling

While the models discussed in this thesis consider high-level summaries of discourse and network features, they do not consider atomic-level interactions (e.g. a coherent stream of individual responses and replies throughout a discussion thread). We propose a bi-directional Long Short-Term Memory (LSTM) sequence model to incor-

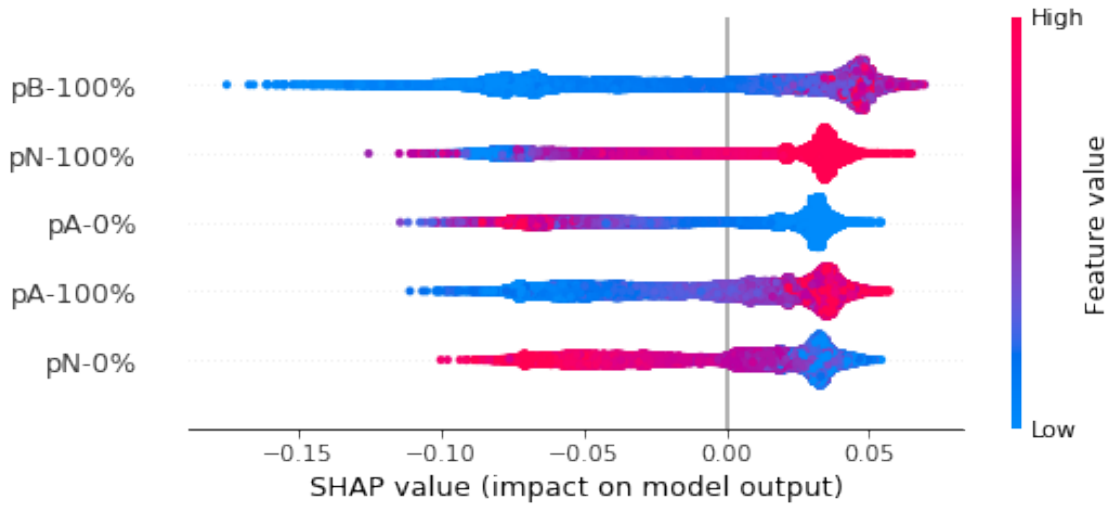

Figure 5.5: SHAP values for five most significant features in Fall random forest model

porate the structure of a thread into the decision making process by treating each comment as part of a sequence and processing them one at a time.

To help understand the process better we first examine Figure 5.6, which is a tree representation of a discussion thread to be processed. Each circle (node) represents a comment in the discussion, with the top-most node being the root comment that starts the thread. Nodes are colour-coded based on the author of the comment. Finally, each comment is annotated with letters representing the order in which each comment was posted (lexicographical order starting with A).

We now turn our attention to Figure 5.7, which provides a high-level overview of the overall pipeline. We propose two methods for the order in which comments are processed by the model: chronological, where comments are ordered by the time they are posted (would result in the nodes in the diagram being ordered alphabetically) and discussion-based (pictured) which is inspired by the depth-first search technique. Discussion-based ordering begins with the root node and proceeds with the earliest posted child of the root node. We continue this process, traversing down the discussion tree until we reach a "leaf" comment. We then traverse back one level to the parent and select the next oldest child until we hit another "leaf" comment or all children have been explored in which we would traverse back another level and the process would repeat until all nodes are exhausted.

We hypothesise that the discussion-based method more accurately mimics the way in which discussion threads are read. When reading a forum, users typically traverse down a chain of replies before moving onto another set of discourse rather than reading through all the top-level comments, then the children of these comments etc. - using this to way flatten threads may provide a more accurate context around the interactions occur between different users.

Once we have determined the sequence in which comments will be processed, we

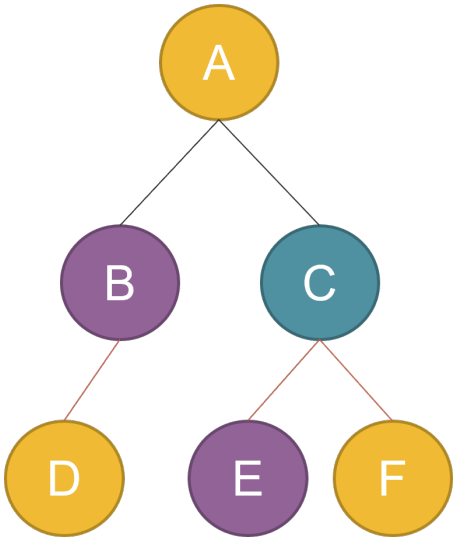

Figure 5.6: Discussion tree processed in the pipeline in Figure 5.7

obtain features from the author of each comment and feed them to the model. Since replies may also influence the behaviour of the user being replied to, we model the process in both directions and sum the representations. We then output a prediction about whether the author of the comment will remain in the next period or leave the system. Finally, we pool the predictions made for each user and take the most frequently predicted class to produce a user-level prediction of future presence.

5.4.1 Methodology

The internal LSTM model is trained on comment-level predictions of future presence. We determine the optimal hyperparameters using a 5/6 train and 1/6 validation split and train using an Adam optimiser with Cross-Entropy Loss. We evaluate prediction performance at the user level to maintain comparability with other models. However, simply computing predictions for all the authors in the dataset trivialises the task as the model already has access to all the labels during training.

Instead, we split the *users* into a train and test dataset (75/25 split). In the comment-level train and validation sets we hide the label of comments made by all users present in the test set. Loss calculations also ignore the predictions made for comments by these these users. After the final model has been selected, we evaluate the aggregated stance predictions of only the users in the test set. This approach solves the issue of not exposing the model to any labels in the test-set, while allowing the comments made by these authors to remain part of the training process - which is extremely important given that we are trying to model atomic-level interactions.

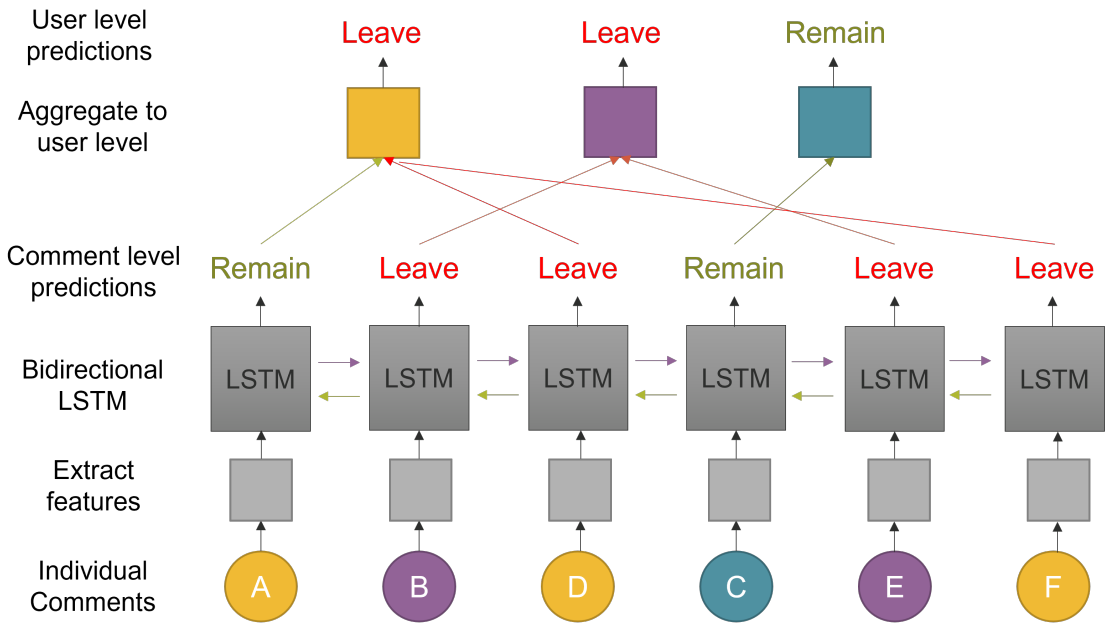

Figure 5.7: Sequence model pipeline

5.4.2 Preliminary Evaluation

Due to the amount of time it takes to train and fine-tune this complex model, we are able to perform a preliminary evaluation only on a subset of the available data. We use only comment data from up to and including period 14 (Recall Table 2.1) and train on the Fnet feature set due to it containing significantly less features than the F0123 set. For a baseline, we retrain the random forest classifier in the previous section using only Fnet features from the same timeframe to ensure a fair comparison.

The huge gap between the chronological ordering and other features in Figure 5.8 immediately stands out. Despite training on many different hyperparameters, the performance of this model barely improved. While this may be due to poorer performance of the ordering, given that random chance would produce a score of 0.5, the difference suggests that this needs to be looked into further to determine if this was a result of an implementation error.

Comparing the discussion-based ordering and random forest baseline suggests that these models perform very similarly on the data. Given that the limited knowledge provided by the feature set (current polarity and the types of interactions that occur) is likely what would already be inferred by the model (due to knowing the current polarity of each user along with a general idea of who they interact with), this result is not very surprising.

However, this does bring up the question of what types of features would be useful. One such idea would be to incorporate textual features such as representations from the BERT stance predictor by [Law, 2021]). This is a promising avenue

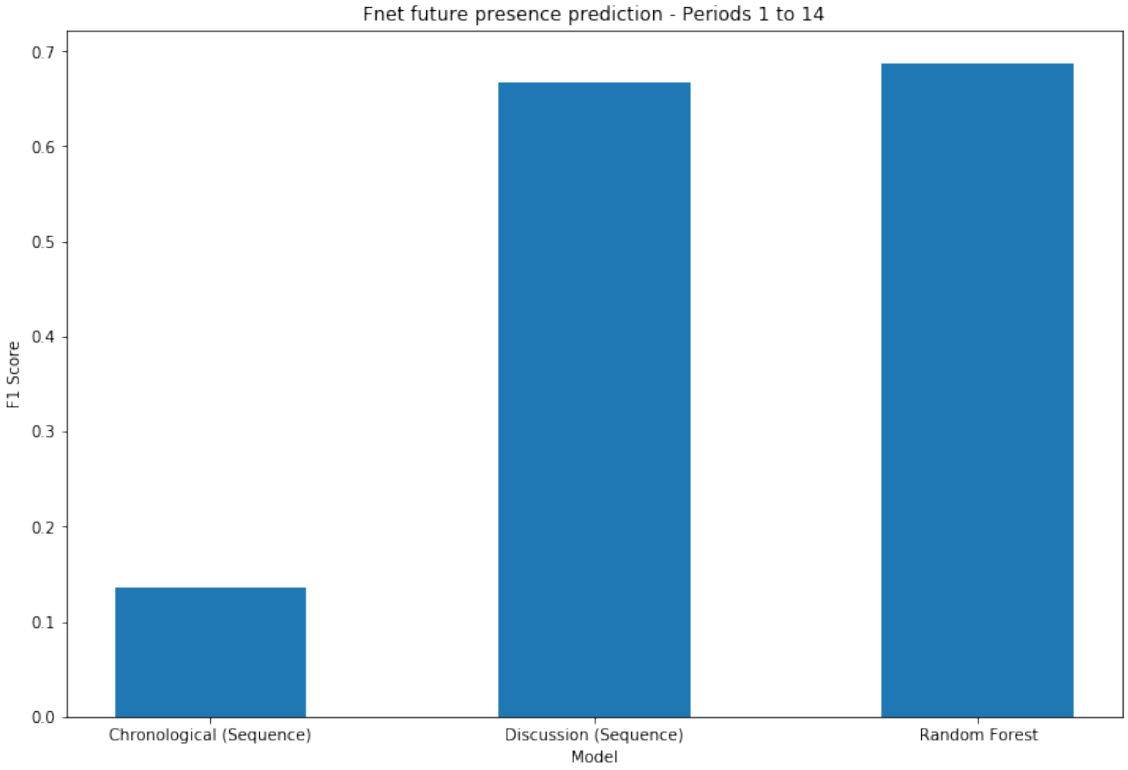

Figure 5.8: Results of sequence model (and sequence ordering) compared to random forest baseline

as it would provide additional context for the interactions that occur. If textual information was available comments such as, "agreed", could become more powerful predictors of stance dynamics as we could likely infer that they sympathise with (and perhaps even endorse) the sentiment of the user they are replying to.

5.5 Summary

In this section we explore the future presence task in an attempt to answer the question why do users leave? We find that measures of participation and interaction such as activity and degree positively correlate with a user's desire to remain active. We also find that users with extreme values of activity and degree have a significant level influence over the individuals they interact with in regards to future presence. Examining feature importance on the future presence task suggests that users who remain tend to have participated in a polarised discussion. Finally, we propose a sequence based model to take into account individual interactions between users and provide suggestions for future direction in this area.

Chapter 6

Conclusion

Throughout this thesis, we have investigated various avenues for modelling and understanding stance dynamics for users on the Brexit subreddit. We have used these features and models to analyse the types of interactions that occur, along with observing the impact of specific features on future user behaviour. Finally, we provide a solid foundation for conducting future work in this area.

6.1 Summary

To summarise, we have presented:

- Relevant literature on opinion dynamics and the importance of a user's social neighbourhood in addition to previous work on the dataset.
- Issues with the discrete classification task and the implications of future stance being contingent on future presence.
- Continuous polarity metric for user stance and insight about the polarity of users who choose to remain active
- Insight on measures of individual user participation and interaction.
- Sequence modelling pipeline that incorporates individual interaction transactions between users.

6.2 Future Work

Due to the complex nature of this task, many simplifying assumptions were made in the modelling process that could be lifted to explore new avenues that involve additional information such as:

- Incorporating additional Reddit features such as the ratings of each post (upvotes, downvotes), a user's karma rating (total received upvotes - total received downvotes)

- Modelling the outside context based on real-world events that occured and/or discovering trends or seasonal patterns in user behaviour
- Improving the balance of the dataset by incorporating comments from other Brexit-related subreddits

There are also a myriad of future directions that this work could be taken, notably:

- Improving the sequence model pipeline and incorporating textual content features.
- Refining the analysis to uncover specific interaction patterns that influence user behaviour.
- Revisiting the notion of user neighbourhood (and perhaps the GAT model) for the future presence task.

As iterated throughout this work, the problem of how interacting with different views affects future stance is a lot more complicated than initially expected and presents a wide range of potential opportunities.

Bibliography

BBC, 2020. https://www.bbc.co.uk/news/politics/eu_referendum/results. (cited on pages 2 and 7)

Chua, Z.; Goh, K.-Y.; Kankanhalli, A.; and Phang, C., 2007. Investigating participation in online policy discussion forums over time: Does network structure matter? 117. (cited on page 6)

Clarke, H. D.; Goodwin, M.; Goodwin, M. J.; and Whiteley, P., 2017. *Brexit*. Cambridge University Press. (cited on page 7)

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 (03 2014), 317–332. doi:10.1111/jcom.12084. https://doi.org/10.1111/jcom.12084. (cited on page 1)

Das, A.; Gollapudi, S.; and Munagala, K., 2014. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14 (New York, New York, USA, 2014), 403–412. Association for Computing Machinery, New York, NY, USA. doi:10.1145/2556195.2559896. https://doi.org/10.1145/2556195.2559896. (cited on page 5)

Dawson, N.; Rizoiu, M.-A.; Johnston, B.; and Williams, M. A., 2019. Adaptively selecting occupations to detect skill shortages from online job ads. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 1637–1643. IEEE, Los Angeles, CA, USA. doi:10.1109/BigData47090.2019.9005967. http://arxiv.org/abs/1911.02302https://ieeexplore.ieee.org/document/9005967/.

De, A.; Bhattacharya, S.; Bhattacharya, P.; Ganguly, N.; and Chakrabarti, S., 2014. Learning a linear influence model from transient opinion dynamics. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 401–410. (cited on pages 6 and 15)

De, A.; Valera, I.; Ganguly, N.; Bhattacharya, S.; and Gomez-Rodriguez, M., 2016. Learning and forecasting opinion dynamics in social networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (Barcelona, Spain, 2016), 397–405. Curran Associates Inc., Red Hook, NY, USA. (cited on page 5)

Dean, B., 2021. Reddit user and growth stats (updated oct 2021). https://backlinko.com/reddit-users. (cited on page 6)

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W., 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113, 3 (2016), 554–559. doi:10.1073/pnas.1517441113. https://www.pnas.org/content/113/3/554. (cited on page 23)

Du, X.; Wang, Y.; Du, W.; and Feng, A., 2012. Discussion on social network learning from the long tail. *IERI Procedia*, 2 (2012), 492–497. (cited on page 30)

Graham, T. and Rodriguez, A., 2021. The sociomateriality of rating and ranking devices on social media: A case study of reddit's voting practices. *Social Media + Society*, 7, 3 (2021), 20563051211047667. doi:10.1177/20563051211047667. https://doi.org/10.1177/20563051211047667. (cited on page 6)

Horawalavithana, S.; Choudhury, N.; Skvoretz, J.; and Iamnitchi, A., 2021. Online discussion threads as conversation pools: predicting the growth of discussion threads on reddit. *Computational and Mathematical Organization Theory*, (2021), 1–29. (cited on page 6)

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

Kong, Q.; Rizoiu, M.-A.; Wu, S.; and Xie, L., 2018. Will This Video Go Viral: Explaining and Predicting the Popularity of Youtube Videos. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, 175–178. ACM Press, Lyon, France. doi:10.1145/3184558.3186972. https://arxiv.org/abs/1801.04117http://dl.acm.org/citation.cfm?doid=3184558.3186972.

Kong, Q.; Rizoiu, M. A.; and Xie, L., 2020. Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes. In *International Conference on Information and Knowledge Management, Proceedings*, 645–654. ACM, New York, NY, USA. doi:10.1145/3340531.3411861. https://arxiv.org/pdf/2001.11132.pdfhttps://dl.acm.org/doi/10.1145/3340531.3411861.

Largeron, C.; Mardale, A.; and Rizoiu, M., 2021. Linking user opinion dynamics and online discussions. *CoRR*, abs/2101.09852 (2021). https://arxiv.org/abs/2101.09852. (cited on pages 2, 8, 9, 10, 14, and 17)

Law, A., 2021. *Exposing the Stance of Reddit Users on Brexit*. Bachelor's thesis. (cited on pages 10, 13, and 37)

Lundberg, S. M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777. (cited on page 26)

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

McPherson, M.; Smith-Lovin, L.; and Cook, J. M., 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 1 (2001), 415–444. doi:10.1146/annurev.soc.27.1.415. https://doi.org/10.1146/annurev.soc.27.1.415. (cited on page 1)

Mihaita, A.; LI, H.; and Rizoiu, M., 2020. Traffic congestion anomaly detection and prediction using deep learning. doi:arXiv:2006.13215.

Mihaita, A. S.; Benavides, M.; Camargo, C.; and Cai, C., 2019a. Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models. *International Journal of Intelligent Transportation System Research (IJITSR)*, 17, 2 (2019), 125–141. doi:DOI:10.1007/s13177-018-0160-z. https://link.springer.com/article/10.1007/s13177-018-0160-z.

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

Mihaita, A.-S.; Li, H.; He, Z.; and Rizoiu, M.-A., 2019b. Motorway Traffic Flow Prediction using Advanced Deep Learning. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 1683–1690. IEEE, Auckland, New Zealand. doi:10.1109/ITSC.2019.8916852. https://ieeexplore.ieee.org/document/8916852/.

Mihaita, A.-S.; Liu, Z.; Cai, C.; and Rizoiu, M.-A., 2019c. Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting. In *Proceedings of the 26th ITS World Congress*, 1–12. Singapore. http://arxiv.org/abs/1905.12254.

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

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

Mihăiţă, S. and Mocanu, S., 2011. An energy model for event-based control of a switched integrator. *IFAC Proceedings Volumes*, 44, 1 (2011), 2413–2418. doi:https://doi.org/10.3182/20110828-6-IT-1002.02082. https://www.sciencedirect.com/science/article/pii/S1474667016439741. 18th IFAC World Congress.

Mishra, S.; Rizoiu, M.-A.; and Xie, L., 2018. Modeling Popularity in Asynchronous Social Media Streams with Recurrent Neural Networks. In *International AAAI Conference on Web and Social Media (ICWSM '18)*, 1–10. Stanford, CA, USA. https://arxiv.org/pdf/1804.02101.pdf.

Monticolo, D. and Mihăiţă, A., 2014. A multi agent system to manage ideas during collaborative creativity workshops. *International Journal of Future Computer and Communication (IJFCC)*, 3, 1 (Feb. 2014), 66–70. doi:10.7763/IJFCC.2014.V3.269.

Price, V.; Cappella, J. N.; and Nir, L., 2002. Does disagreement contribute to more deliberative opinion? *Political Communication*, 19, 1 (2002), 95–112. doi:10.1080/10584602317246506. https://doi.org/10.1080/10584602317246506. (cited on page 1)

Reddit, 2021. R/brexit - today my wife was 'let go'. then the person telling her that was. https://www.reddit.com/r/brexit/comments/nlptpa/today_my_wife_was_let_go_then_the_person_telling/. (cited on page 7)

Rizoiu, M. A. and Velcin, J., 2011. Topic extraction for ontology learning. In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* (Eds. W. Wong; W. Liu; and M. Bennamoun), 38–60. IGI Global. ISBN 9781609606251. doi:10.4018/978-1-60960-625-1.ch003. http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60960-625-1.ch003.

Rizoiu, M.-A. and Xie, L., 2017. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *International AAAI Conference on Web and Social Media (ICWSM '17)*, 182–191. Montréal, Québec, Canada. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15553https://arxiv.org/pdf/1703.01012.pdf.

Rizoiu, M. A.; Xie, L.; Caetano, T.; and Cebrian, M., 2016. Evolution of privacy loss in wikipedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 215–224. ACM, ACM Press, New York, New York, USA. doi:10.1145/2835776.2835798. http://dl.acm.org/citation.cfm?doid=2835776.2835798http://arxiv.org/abs/1512.03523http://dx.doi.org/10.1145/2835776.2835798.

Schill, D. and Kirk, R., 2014. Courting the swing voter: "real time" insights into the 2008 and 2012 u.s. presidential debates. *American Behavioral Scientist*, 58, 4 (2014), 536–555. doi:10.1177/0002764213506204. https://doi.org/10.1177/0002764213506204. (cited on page 1)

Shafiei, S.; Mihaita, A.; Nguyen, H.; Bentley, C. D. B.; and Cai, C., 2020. Short-term traffic prediction under non-recurrent incident conditions integrating data-driven models and traffic simulation. In *Transportation Research Board (TRB) 99th Annual Meeting, Washington D.C.* doi:http://hdl.handle.net/10453/138721.

Shafiei, S.; Mihăiţă, A.-S.; Nguyen, H.; and Cai, C., 2022. Integrating data-driven and simulation models to predict traffic state affected by road incidents. *Transportation Letters*, 14, 6 (2022), 629–639. doi:10.1080/19427867.2021.1916284. https://doi.org/10.1080/19427867.2021.1916284.

Spohr, D., 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34, 3 (2017), 150–160. doi:10.1177/0266382117722446. https://doi.org/10.1177/0266382117722446. (cited on page 1)

Throsby, E., 2013. Engaging the disengaged: Swinging voters, political participation and media in australia. *Platform*, 5 (10 2013), 97–106. (cited on page 2)

Uddin, S. and Hossain, L., 2013. Dyad and triad census analysis of crisis communication network. *Social Networking*, 2 (01 2013), 32–41. doi:10.4236/sn.2013.21004. (cited on page 15)

Unwin, J. T.; Routledge, I.; Flaxman, S.; Rizoiu, M. A.; Lai, S.; Cohen, J.; Weiss, D. J.; Mishra, S.; and Bhatt, S., 2021. Using hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17, 4 (apr 2021), e1008830. doi:10.1371/JOURNAL.PCBI.1008830. http://medrxiv.org/content/early/2020/07/17/2020.07.17.20156174.abstracthttps://dx.plos.org/10.1371/journal.pcbi.1008830.

Wen, T.; Mihăiţă, A.-S.; Nguyen, H.; Cai, C.; and Chen, F., 2018. Integrated incident decision-support using traffic simulation and data-driven models. *Transportation Research Record*, 2672, 42 (2018), 247–256. doi:10.1177/0361198118782270. https://doi.org/10.1177/0361198118782270.

Wu, S.; Rizoiu, M.-A.; and Xie, L., 2019. Estimating Attention Flow in Online Video Networks. *Proceedings of the ACM on Human-Computer Interaction*, 3, CSCW (nov 2019), 1–25. doi:10.1145/3359285. http://dl.acm.org/citation.cfm?doid=3371885.3359285.

Wu, S.; Rizoiu, M. A.; and Xie, L., 2020. Variation across scales: Measurement fidelity under Twitter data sampling. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 715–725. https://arxiv.org/abs/2003.09557.

Zhang, R.; Walder, C.; and Rizoiu, M.-A., 2020. Variational Inference for Sparse Gaussian Process Modulated Hawkes Process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 04 (apr 2020), 6803–6810. doi:10.1609/aaai.v34i04.6160. http://arxiv.org/abs/1905.10496https://aaai.org/ojs/index.php/AAAI/article/view/6160.

Zhu, L.; He, Y.; and Zhou, D., 2020. Neural opinion dynamics model for the prediction of user-level stance dynamics. *Information Processing & Management*, 57, 2 (2020), 102031. doi:https://doi.org/10.1016/j.ipm.2019.03.010. https://www.sciencedirect.com/science/article/pii/S0306457318308604. (cited on page 5)