Relationship between Social Media Presence and Voting Behaviour

Candidate no. 195105

Department of Informatics and Engineering
Supervisor Prof David Weir

September 3, 2019

Abstract

This work undertook research into the relationship between social media presence and voting behaviour in Members of Parliament in the government in the United Kingdom. The majority of work in this area focuses on members of the general public, which brings difficulties due to a lack of ground truth surrounding their actual voting intentions or how their voting history. UK Members of Parliament provide an interesting use case as all of the votes in the House of Commons are publicly available. This study takes advantage of these two pieces of publicly available data to look the task of describing or predicting an MPs previous or future voting pattern. There is a lot of critique of work in this area that suggests that well performing models are only describing votes that have occurred during the time frame of their data set and not predicting the future. With this in mind this work will look to both 'describe' and 'predict' various votes to see if the critiques are just. A focus is placed on hashtags and follower graphs as these are both fairly simple models. In proving some form of descriptive or predictive power using simple models this works hope to inspire further work in this area that focuses on politicians and not just the general public. This work found that critiques to work in this area were just, producing a disparity between the descriptive and predictive power of certain models. Content-based models were effective at describing whereas network based models were effective at predicting.

Declaration

The author declares that the work of this report is their own.

Acknowledgements

The author gratefully acknowledges the Joe Holmberg scholarship for enabling them to undertake this course of study. Further they gratefully acknowledge Chris Inskip and supervisor Prof David Weir for their support and advice throughout the undertaking of this work. A special acknowledgement is due to Brandwatch Ltd for providing access to tooling, knowledge and data that allowed this project to take place. Thanks is due to Billiejoe Charlton, Phil Messenger and Oskar Holm of Brandwatch for their aid in accessing data. A special thanks is due to Hamish Morgan, Colin Sullivan and Paul Siegel of Brandwatch for their continued support and advice throughout the duration of this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Twitter stands as one of the most popular micro-blogging services with 261 million international users [1]. Having increased the character limit of a tweet from 140 to 280 characters expressing meaningful opinions on Twitter has become more achievable. With such a high volume of users and an average of 500 million Tweets a day [1] Twitter has become a popular source of data for data science and machine learning applications. Amassing such a wide user base the social media platform is not exclusively used by members of the general public. Companies utilise the platform's reach for advertising and customer-service claims while politicians use it as a platform to communicate with their constituents. Political discourse is particularly active on the platform and a number of users look to explicitly define their political preference using their Twitter bio. Containing both interactions between the general public and politicians as well as users self identifying their political preference, the platform has found popularity in attempting to predict the outcome of elections or the political preferences of an individual.

The content of a users Tweets has been used in an attempt to predict actual election outcomes [2–7] and the political preferences of an individual user or their voting intentions] [8–11] to a remarkably high accuracy. Further work in this area has called these findings in question, citing a misleading oversight as to the distinction between describing and predicting data [12, 13]. Research further dismissed high performing predictive models to be correct only by chance [14] suggesting that many results were not reproducible and that application to any other event other than the vote in question would not produce similar performance. A further concern in political predictions regarding the general public is a lack of ground truth to evaluate these models against. Although some Twitter users self-identify their political leaning on Twitter, this is not ubiquitous nor absolute when predicting how an individual actually voted. As a result the majority of work that looks to predict election outcomes focuses on aggregated data as its source of ground truth, be that simply predicting the winner or percentage share of votes for candidates or parties. In selective cases, work has instead analysed the Twitter presence of politicians themselves [15] as there is more publicly available information that can be used to form a more comprehensive ground truth.

This work looks to follow this trend in considering the relationship between Members of Parliament (MP) in the United Kingdom and their voting records on motions/acts in the House of Commons. The votes for every motion and act discussed in the House of Commons are publicly available through a government website [16] and a public facing API [17]. Combining this with the publicly available Twitter profiles of MPs who are active on the platform provided a dense data set with a reliable source of ground truth. The existence and availability of this source of ground truth makes politicians a particular appealing focus of research in the relationship between politics and Twitter. In considering the relationship between these two entities this work will look to address a number of research questions that look to both explore this relationship and address the concerns raised about work in this area. These research questions can be split into two groups: questions about the descriptive power of Twitter and questions about the predictive power of Twitter.

1.1 Descriptive Questions

R1. Is tweet content useful for describing how an MP has voted? In the context of this work to be descriptive is to accurately 'predict' the results of a vote that has occurred in the past or present. This definition is aligned with the definition used in work that looks to rebut much of the work in this area [12–14]e. Addressing this question will involve experimentation in predicting votes that have occurred within the same time frame as the collected data was produced. Resulting from this conclusion can be made as to the semantic value of hashtags.
R2. Is there a relationship between party and Twitter presence? In analysing the dataset, this work will look to draw conclusions as to the similarities and differences to how parties and the MPs that represent these parties use the platform. Are some parties more active than others? Are there some MPs who subvert the norm set by their party?

1.2 Predictive Questions

R3. Is tweet content useful for predicting how an MP will vote? In the context of this work a prediction must concern an event that occurs outside the time-frame of the data that is used to 'predict' it, be that in the past or the future. A primary focus will be placed on predicting the future, as this application provides more value in the real world.
R4. Are follower networks useful for predicting votes? Follower networks are continuously changing but the core of these networks, especially the follower relations between MPs remain relatively static. As a result this work looks to investigate the value held in these graphs as this area is less explored among present research into the area of politics. With reservations expressed as to the reliability of content as a means of predictions, network structure provides a possible alternative.
R5. Are all votes as predictable as others? The majority of work in this area focuses on single sporadic events such as general and Presidential elections. In taking a topic-centric instead of event centric approach, the descriptive and predictive power of different topics can be considered and conclusions made as to why some topics may be more predictable than others.

1.3 Contributions

The contributions of this work can be summarised as follows: firstly it introduces a means of investigating the connectivity of active MPs on the twitter platform by providing a dataset that contains two separate follower graphs. The graph of follower relationships among MPs and the graph of MPs that are '1 hop' apart, i.e they share a follower relationship with the same user. Second, provide evidence that hashtags far exceed their function as defined by Twitter and carry meaningful semantic value. Thirdly, provide lists of Twitter handles, mappings between names and handles and lists of active followers that provide the building blocks to construct a similar UK MP centric dataset that allows further experimentation with the Twitter content of MPs, their voting patterns and the content of their followers that go beyond the scope of this work.

1.4 Achievements

The achievements of this work manifest themselves in the ability to confidently answer the research questions listed above. The findings of this work are as follows.

R1. Is tweet content useful for describing how an MP has voted?

This work has found that hashtags are very descriptive of how an MP has voted providing consistent results of greater than 80% accuracy, far above the baseline measurement.

R2. Is there a relationship between party and Twitter presence?

There is some evident correlation between political party and activity on Twitter with the Scottish National Party utilising the platform more than any other party. Moreover, this work found that the percentage of share of seats in Parliament and the percentage share on Twitter are very similar, suggesting that Twitter provides scope as a test-bed for a lot of political research.

R3. Is tweet content useful for predicting how an MP will vote?

Tweet content proved to perform very poorly in predicting how an MP will vote in the future. This performance is in line with the findings of work in this area that critique the main body of work suggesting there is confusion between describing and truly predicting data.

R4. Are follower networks useful for predicting votes?

Follower networks have proved to be very effective at predicting votes, providing consistent performance above 90% accuracy across a number of different votes. This consistency is seen between topics and between time periods.

R5. Are all votes as predictable as others?

Votes that are expected to go along party lines proved to be more difficult to predict than those that are surrounding heavily divisive topics such as Brexit. More research could be done in this area by introducing more voting topics into the experiments listed as part of this dissertation.

1.5 Report Structure

The remainder of this report will be structured as follows. Related work will be discussed and followed by a background section that outlines the knowledge required to interact with this work. Professional considerations with be addressed before the theoretical approach is discussed in relation to the above research questions and related work to justify the direction of this work. With the presentation of a new dataset, analysis will be performed directed towards addressing **R2**. This will be followed by a brief discussion around the implementation and the experimental set up. The results of each experiment will then be addressed at an objective level. The discussion section will draw further conclusions from these results and look to link these to the above research questions. This section will conclude with a discussion of potential further work before this work is summarised and concluded.

Chapter 2

Related Work

2.1 Data

Twitter is a popular platform amongst the body of research into the relationship between social media and another entity. Increasingly, research in this area focuses on predicting an attribute about a user or a group of users outside of the realm of Twitter. Political leaning and electoral votes are an attractive prospect as there is plenty of publicly available data, albeit in aggregated form. As a result a lot of papers in this area look to predict the outcome of elections [2–5]. Given the sporadic nature of elections some research in this area has centred around predicting or describing the political leaning or voting intentions of the general public [8–10]. The USA and its politics arises as a particularly popular test-bed in this area of research with a lot of work in this area focusing on congressional and presidential US elections [3, 6, 9, 11, 18, 19]. This focus is not exclusive however, with work looking at Scotland [7], Switzerland [15], Canada [10], Singapore [20] , Germany [5], the UK [2,8] France and Italy [21]. Despite focusing on a number of different nations and elections, the majority of the aforementioned work shares a focus on the general public with only Rauchfleisch et al [15] taking a particular interest in the politicians themselves.

2.2 Criticisms

The applicability of social media platforms as a tool for predicting and reasoning about the political activity of the general public has been rebutted by a number of papers in this area. Mellon et al [22] found that Twitter and Facebook users 'differ substantially' from the general public in the areas of vote choice, turnout, gender, age and education. All areas that they describe to be 'politically relevant dimensions'. Moreover, Malik et al [23] and Mislove et al [24] found that Twitter users in the USA predominantly were a younger demographic who lived either in urban areas or in wealthier areas according to their geo-tagged data. In considering the political leanings of Twitter users Vaccari et al [25] found that Italian users were predominantly left wing and male. Barbera and Rivero [26] looking at Twitter users in Spain and the USA in 2011-2012 to be males who sit on the far left or right. Metaxas, Gayo-Avello et al criticised much of the work in this area by finding issue with the widespread use of the term 'prediction' [12]. They rebutted findings in this area that shared success stories with high accuracy models by suggesting that they were not predicting results but simply describing the past or present. With some work in this area not falling into their descriptive and not predictive narrative, they further suggest that genuine predictive success in this area is a coincidence and occurs by change, citing the work of Lui et al who found that the predictive power of Google search volumes for US congressional election outcomes were equally accurate for some votes as they were inaccurate for others concluding that any success in this area occurs purely by chance [14]. Gayo-Avello et al [13] tested this hypothesis by attempting to in their eyes truly predict the outcomes of US elections in 2010. They found that predicting the share of votes between the two candidates was no better than chance, suggesting that the mean average error (MAE) of professional polling services of 2-3% leaves these models redundant with a MAE of 17.1% in their experiments.

2.3 Approaches

2.3.1 Content-Based

A trend across the majority of this work is the use of content based approaches as a basis for making these predictions. With 500 million tweets a day, the amount of publicly available data when taking a content based approach can provide a challenge. Work in this area looks to address this by providing a means of discerning the relevance of a tweet. This relevance is defined by the inclusion of keywords within a tweet. These keywords sets are either generated manually by hand [5, 18, 27] or semi-automatically [19, 28, 29]. Burnap et al [2] considered a tweet to be relevant if it contained a reference to a political party or candidate by name whereas Connover et al take a more complex approach [19]. They consider a tweet to be relevant if it contains a politically relevant hashtags. These politically relevant hashtags are identified using hashtag co-occurrence. After filtering out ambiguous hashtags that carry too broad a meaning, they were left with a keyword set of 55 hashtags that they used to identify 252,000 relevant tweets. Makazhanov et al recognised the potential value in politically relevant hashtags but suggest that some are more relevant than others leading them to go further to define an 'interaction profile' [10]. They suggest the mentions, replies and retweets that concern political candidates and parties provide more value as to the political preference of a user. As a result they compile a per party ranked list of weighted terms that they coin as the 'interaction profile' of the party.

The value of hashtags in Twitter centric literature in general has not gone unnoticed with Liu et al proposing the Hashtag2Vec model [30]. This builds on the initial work of Mikolov et al who proposed the Word2Vec model [31]. They proposed a means by which words and phrases can be embedded into a vector space while maintaining their semantic value. The motivation behind this was to produce an efficient means by which words can be vectorised, circumventing the dense matrix multiplications that make other methods slow. Resulting from this came a model that encodes linguistic regularities and patterns, many of which can be represented as linear translations [32]. Hashtag2Vec follows the trend of '2Vec' models that have come out of the Word2Vec model. Liu et al propose a joint embedding framework that allows hashtag, tweet and word representations to be learned simultaneously and reinforced mutually. This manifests itself in the hierarchical heterogeneous graph G where $G = (V^h \cup V^t \cup V^w, E^{hh} \cup E^{ht} \cup E^{tw} \cup E^{ww})$. G has three kinds of vertices: V^h (hashtags), V^t (tweets) and V^w (words); and four kinds of edges: E^{hh} (hashtag-hashtag), E^{ht} (hashtag-tweet) , E^{tw} (tweet-word) and E^{ww} (word-word). Each of these relationships can be represented as an adjacency matrix: M^{hh} , M^{ht} , M^{tw}, and M^{ww}. One could argue that the separation of word and hashtag is not necessary given that hashtags cluster together within a tweet when they do co-occur as hashtags conventionally suffix the content of a tweet. Therefore the hashtag-hashtag relationship can be considered as a word-word relationship and similarly for the hashtag-tweet relationship. This would result in graph $G = (Vt \cup Vw, E^{tw} \cup E^{ww})$. A model is this form is not dissimilar to the Doc2Vec model [33]. Le et al proposed a joint learning model in which they defined a Paragraph Vector, an unsupervised framework that learns continuous distributed vector representations for pieces of texts. This can be applied to any length of text from a sentence or paragraph to a large document. Comparing this to the Hashtag2Vec model, the tweet-word relationship is encapsulated by considering each tweet to be a document and the occurrence of words within each document encapsulates the word-word relationship. Such a model is worthy of application in the research area of political prediction.

2.3.2 Network-Based

Other work in this area focuses less on a content based approach and places its focus on network based approaches. Many of the interactions on Twitter can be represented in graph form. Examples of these networks include a followers network, the following network, a mentions network, a retweet network, a reply network and an amalgamation of the previous three to form an interaction network. These networks can be combined to produce a social graph [34–37]. Volkova et al used a social graph alongside a number of networks they defined to infer the political preference of Twitter users leading up to the 2012 US Presidential election [38]. They defined three additional networks: a candidate-centric graph representing politically active users; a geo-centric graph representing less politically active users and a ZLR graph. The candidate-centric graph was constructed by labelling users as either Democrat or Republican dependent on their following relationship with exclusively two Democrat or Republican candidates. The candidate-centric graph for each user is then generated by taking the immediate k-neighbours from follower, friend, user mention, reply, retweet and hashtag graphs. The geometric-centric graph is generated by randomly sampling from the social graph of users with their political preference in their Twitter bio from the states of Maryland, Virginia and Delaware. The ZLR graph is constructed from the dataset proposed by Zamal et al for political affiliation classification [39]. There is far less relevant work in this area of research that considers network structure when compared to the body of work that focuses on content and sentiment analysis.

Chapter 3

Background

This section outlines the prior knowledge required surrounding the underlying technologies and concepts to engage with this work. The addressed technologies and concepts will be explored in enough depth to ensure this work is digestible and as a result some complexities will be omitted or abstracted over.

3.1 Term Frequency Inverse-Document Frequency

3.1.1 Term Frequency

Term frequency is the statistical count of the number of times a term appears in a given document. Therefore, term frequency can be defined as $tf(t,d) = ft,d$ where ft,d refers to the raw count of term t in document d.

Variant definitions of term frequency do exist however the raw count provides the simplest definition and therefore any reference to term frequency in this paper proceeding from this point reference this definition as defined above.

The usefulness of this statistic is reliant on the assumption that a term that appears frequently in a document is highly descriptive of the document itself. This assumption does not hold on raw documents due to the frequency of stop words like 'the' so some form of preprocessing it required if term frequency is to be useful as a standalone statistic.

3.1.2 Inverse-Document Frequency

Inverse-document frequency is a statistical measure of how much information/descriptive power a given term holds, i.e. how unique or common a term is across all documents. It is a logarithmically scaled inverse fraction that is obtained by dividing the total number of documents by the number of documents containing that term.

This inverse fraction is designed to downweight the importance of common terms like stop words while placing greater importance on terms that appear in only a few documents. It is not concerned about the count of a term in a given document just that it exists within that document. Therefore given term i and term j if both terms appear in n documents both terms would be considered to be equally descriptive of that document despite i appearing 5 times in each document and j appearing only once in each document. The intuition here is that i is likely to be more descriptive than j but the idf weighting will not reflect this.

3.1.3 Term Frequency Inverse-Document Frequency

As the product of term frequency and inverse-document frequency, *tfidf* can be defined as *tfidf(t, d, D) = tf(t, d) · idf (t, D)* or more precisely by the equation:

$$\text{tf-idf}_{t,d} = (1 + \log \text{tf}_{t,d}) \cdot \log \frac{N}{\text{df}_t}$$

The motivation behind taking the product of these two statistics is to gain the merits of both while negating the drawbacks that have been previously discussed. The inverse-document frequency downweights the terms that appear incredibly frequently, negating the issue the term frequency has with stop words and over common terms across the language. The term frequency gives a means by which terms that appear in the same number of documents can be distinguished, placing greater importance on terms that appear more frequently within the document itself. The combined statistic therefore weights a terms descriptive importance based on its document wise importance and corpus wise importance.

3.2 Regression

3.2.1 Linear Regression

Linear regression is a linear approach to modeling the relationship between one or many independent variables and a single dependent variable. The model looks to predict the numerical value of the dependent variable Y given the input set of independent variables X. The linear transformation performed on X is often the conditional mean, conditional median or some other quantile.

3.2.2 Binary Logistic Regression

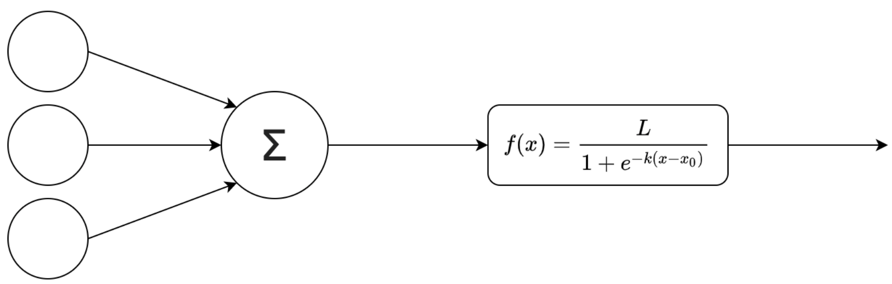


Figure 3.1: Logistic Regression

Unlike linear regression, a binary logistic regression model looks to predict the probability of the input set of independent variables X belonging to class 0 or class 1 in the single dependent variable Y [41]. It operates off of the assumption that the input space can be separated by a linear boundary, i.e. the input can be separated into two distinct classes. As demonstrated in Figure 3.1 a linear transformation on X is performed as in linear regression. The output of this transformation is then passed through the logistic function which transforms the numerical value into a continuous value between 0 and 1. This is considered to model the probability of a given X sitting in class 0 or class 1. If this probability sits closer to 0 than 1 then Y value is predicted to be 0 and vice versa.

3.3 Embedding

Embeddings provide a mapping between discrete categorical variables and vectors of continuous numbers. The motivation behind producing these mappings stems from the mathematical underpinning of machine learning models. With a lot of these models relying on matrix multiplication and other transformations, a vector representation of a categorical variables opens up scope for utilising more complex machine learning methods. There are three primary applications for embeddings:

1. Visualisation of relations between categorical variables

2. Clustering to find nearest neighbours

3. Input to a machine learning model

Neural networks are often used in producing embeddings as they are capable of reducing the dimensionality of the discrete variables without a significant loss of meaning in the representation. A popular application of this in the Natural Language Processing space is word embeddings.

3.3.1 Neural Network

Neural networks are computing models that are somewhat inspired by the biological neuron and how a network of these functions as part of an animal's brain. These neurons are arranged in layers that perform 3 different functions as shown in Figure 3.2.

Layers

The input layer's purpose is to bring the input data into the network, it may perform some form of transformation to this data to ensure it is in the correct form for the network to use. The hidden layer or layers are where the 'learning' takes place. Weights are passed as input into each node in the hidden layer, transformed using the activation function of the node and outputted as a single weight for that node. The output layer produces the given output in the form of a label or probability.

Weights

Weights in the model represent the strength of the connection between two connected nodes. The weights are used to make initial nodes in the input more or less influential in the network. Optimising these weights against a loss function defines the 'learning' process that is performed during the training of a neural network.

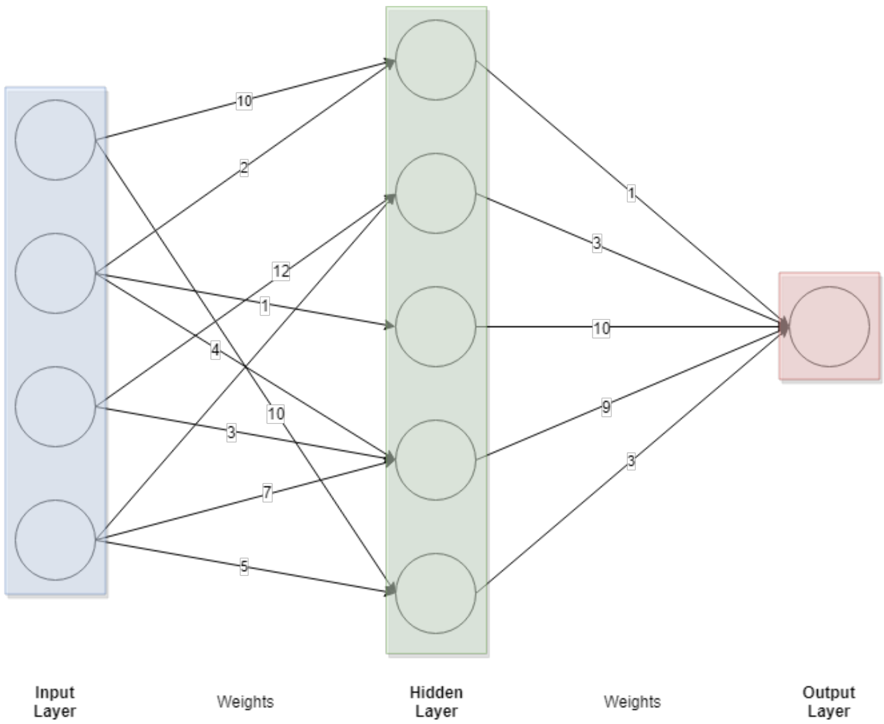


Figure 3.2: An example neural network

3.3.2 Word Embedding

Text data can often result in large, sparse data sets due to the complex grammatical structures present in most languages and the existence of synonyms. Embeddings are useful in this application as they can model both the relationship between words in terms of their functions as a verb, noun etc... and the relationship between synonyms by modelling them using similar vectors. On of the most widely used models for this is Word2Vec [31].

Word2Vec

The Word2Vec model is a two-layer neural network which can reconstruct the linguistic context of words. It is trained on a large corpus of sentences, mapping each unique word to a vector in a vector space. With this vector space words used in similar contexts, where a context is defined as a set of surrounding words, sit closely together. This results in synonyms like good and great sitting closely together and also allows for surprising linear transformations such as *France - Paris + England = London*. There are two flavours of Word2Vec, the Continuous Bag-Of-Words(CBOW) model that looks to predict a word given its context and the Skip-Gram model, as shown in Figure 3.3, that looks to predict a context given a word.

Architecture

In training the model the input layer takes a word as input and the output layer has a neuron for every unique word in the corpus. A given output in the skip-gram model therefore represents the probability that a word in a randomly selected position in a context containing the input word is the word represented by the neuron. After

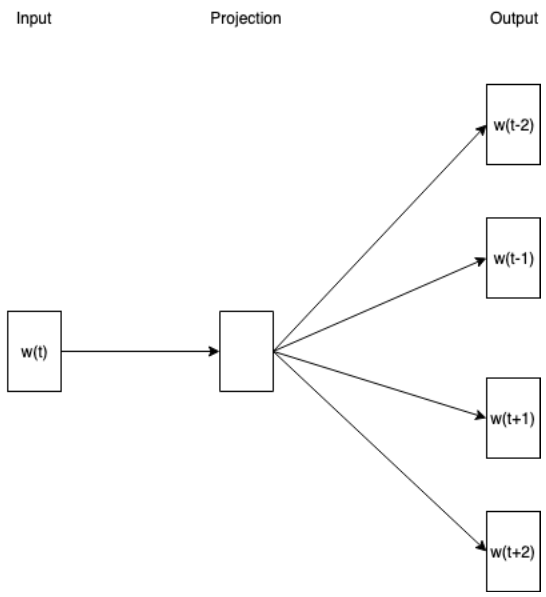


Figure 3.3: Skip-Gram Model

several iterations of training the weights of the hidden layer, or projection as in Figure 3.3, are actually word embeddings of the size of the number of neurons in the hidden layer. These weights can therefore be pulled out of the model and these embeddings used to train and predict and in other models.

3.3.3 Doc2Vec

The Doc2Vec model looks to apply the theory behind Word2Vec to generate an embedding for documents of varying length, described as a Paragraph Vector [33]. The model introduces the idea of a paragraph or document id which is used to tag a sentence or group of sentences. In training therefore, not just word embeddings are learnt but embeddings for the set of unique tags are also generated. This results in two different flavours of the Doc2Vec model, the Distribution Memory(DM) model that predicts a word given a context and paragraph id and the Distributed Bag of Words(DBOW) that predicts a set of words it expects to appear within the document. In looking at Figure 3.3 and Figure 3.4 it is evident that the Skip-Gram Word2Vec model and DBOW Doc2Vec models are very similar with Skip-Gram predicting an expected context given a single word and DBOW predicting an expected context given the context of a document. Their architecture is the same, the differences exist in the input and concept of a context. This model therefore allows one to say, given Document A, what words do I expect to exist within this document.

3.4 Label Propagation

3.4.1 Graphs

A graph G can be defined as G=(V,E) where V is a set of nodes and E is the set of edges between these nodes. Graphs can therefore be used to represent lots of different relationships and interactions between entities. Additional information can

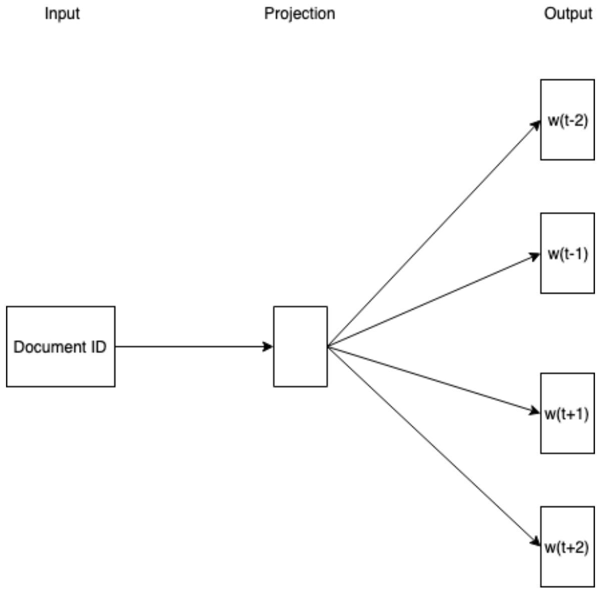


Figure 3.4: DBOW Model

be encoded into a graph by adding attributes to the nodes and edges. An undirected graph puts no restrictions on the direction an edge can be traversed, resulting in the edge of any adjacent node being traversable.

3.4.2 Propagation

Traditionally, label propagation algorithms are used to define communities within a graph, where a community is defined as a connected subgraph that appear to be distinct from other nodes in the graph. They find popular application in machine learning settings as labelled data is often sparse but can still be valuable in reasoning about large sets of unlabelled data. In its simplest form, labels can be propagated by passing the label of all labelled nodes to each of its neighbours on each iteration. In the case where the data is completely unlabelled, random labels can be generated and propagated to reveal communities. Figure 3.5 demonstrates the initial graph when considering a graph that contains 4 randomly labelled nodes.

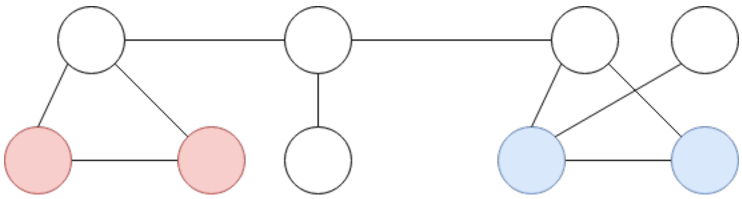


Figure 3.5: Label Propagation First Iteration

After the first iteration of the algorithm, these labels are propagated to the immediate neighbours as shown in Figure 3.6.

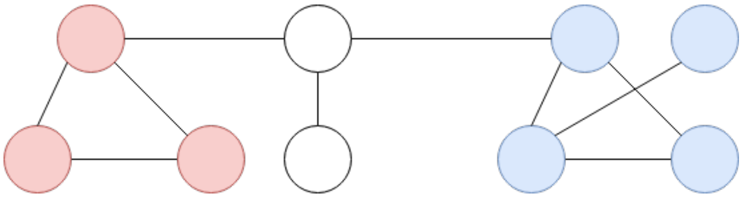


Figure 3.6: Label Propagation Second Iteration

During the third iteration of the algorithm two different labels are propagated to the same node. As demonstrated in Figure 3.7 this results in a new community label being introduced that is a combination of the two labels.

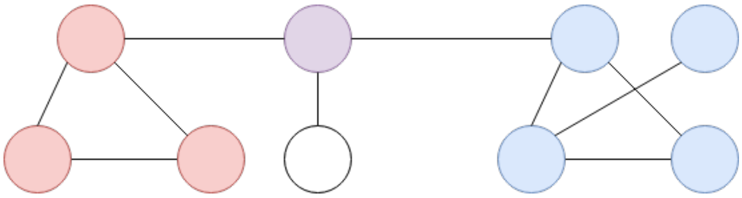


Figure 3.7: Label Propagation Third Iteration

Finally, after the fourth iteration, all nodes have a label and there are 3 distinct communities as shown in Figure 3.8.

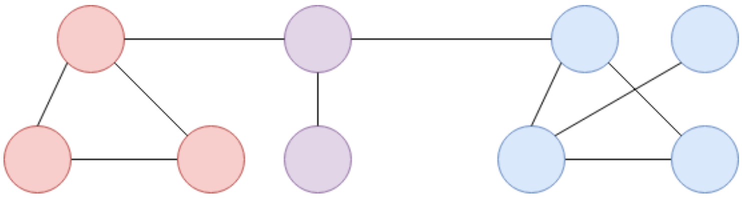


Figure 3.8: Label Propagation Fourth Iteration

In considering a binary classification problem the above can be applied using ground-truth labels. The purple labels in this case represent the probability of sitting exactly between the two classes. Some form of a threshold may be utilised to decide if these nodes sit in the red or the blue class.

In this section document and term based statistics, binary logistic regression, word and document embeddings and label propagation have all been discussed in enough depth to ensure the reader is adequately informed in navigating the rest of this dissertation. The following section will outline the professional considerations that were adhered to throughout this work before the theoretical approach and implementation are discussed.

Chapter 4

Professional Considerations

The professional considerations detailed in this chapter have been adhered to throughout the entirety of this project. Moreover, a conservative effort has been made to use only publicly available data and to discuss only people who are deemed to be public figures. The British Computer Society (known as BCS, the Chartered Institute for IT) [40] outline a Code of Conduct [41] that any work within the IT sector must adhere to. The following sections from the BCS Code of Conduct are addressed and detailed below: Public Interest, Professional Competence and Integrity, Duty to Relevant Authority and Duty to Profession.

4.1 Public Interest

This project may at times use code libraries or tooling that have been produced by a third party. In these instances, the relevant third party or parties involved will be given proper credit and be explicitly referenced. Given that this work addresses well known public figures in the form of Members of Parliament, its results and conclusions may benefit the wider public. In light of this, this work will adhere to Section 1(c) of the BCS Code of Conducts in that it will be conducted 'without discrimination on the grounds of sex, sexual orientation, marital status, nationality, colour, race, ethinic origin, religion, age or disability, or of any other condition or requirement'.

4.2 Professional Competence and Integrity

With this work being undertaken as part of an Advanced Computer Science MSc masters degree, the project will predominantly focus on knowledge areas obtained through the study of this course as delivered by the University of Sussex. Any knowledge that has been gained from background research or further reading will be acknowledged via appropriate reference. In collaborating with Brandwatch Ltd the author will not claim any level of competence that they do not possess and will provide credit through reference or acknowledgement where any contributions have been made.

4.3 Duty To Relevant Authority

This work will be conducted with due care and diligence in accordance with the requirements as set out by the University of Sussex. Permission has been granted to carry out this project by supervisor Prof David Weir and the University of Sussex having sought advice about the relevance of this project. In working in collaboration with Brandwatch, any decisions made in the use of their data were not made without prior consultation and granted permission. In conducting this work, any situation that may result in a conflict of interest between Brandwatch Ltd or the University of Sussex will be avoided. Moreover, the author of this work accepts full responsibility for all work carried out within its confines. This project will not disclose or attempt to disclose any confidential information without abiding by the relevant legislation or without the permission of both the University of Sussex and Brandwatch Ltd.

4.4 Duty to Profession

The author will accept their duty to 'uphold the reputation of the profession and not take any action which could bring the profession into disrepute' as described by Section 4(a) of the BCS Code of Conduct. The author will support fellow members in encouraging them in their future development and shall act with integrity and respect in their professional relationships with all members of the BCS and members of other professions with whom they work with in a professional capacity.

Chapter 5

Approach

This section outlines the theoretical approach that has been taken in addressing this area of research. Decisions made in terms of the approach are justified with reference to the research questions laid out in the introduction and to other work in this area as laid out in the related work section. The discussion around the theoretical approach is centred around the choices made regarding the chosen task and dataset. Implementation decisions are laid out in the Implementation section.

5.1 Location

With the majority of work in this area of research focusing on the politics of the USA, the UK is a relatively unexplored space in comparison. The political space in the UK has been an increasing desirable area for investigation in the last few years given the controversy surrounding 'Brexit' and several Prime Ministers stepping down. This provides some interesting dynamics with some of the political discourse being very party independent, i.e. Brexit and other discourse more party aligned. The party system in the UK is somewhat similar to the US in that it is dominated by two main parties with the Labour party and the Conservative party dominating UK and the Republicans and the Democrats dominating the USA. However, unlike the US, and increasingly so in more recent years, the voices of smaller parties have become louder with UKIP, the Brexit Party and the likes of the SNP and Liberal Democrats gaining a greater share of the public vote. As a result, party wise discussion is not as binary as it is in the data that is generally used in this area of research. This motivated a focus on UK politics between 2017 and 2019. Considering data over a large time frame provides scope for considering a number of different votes as well as providing the opportunity to 'predict' votes in the past, present and future, helping to address **R5**.

5.2 Classification Task

The task that a lot of the work in this area looks to address is predicting the outcome of an election or some form of political leaning or preference amongst the general public. Voting patterns of the general public and their political preferences are not publicly available information and therefore any form of ground truth can only be obtained through aggregated figures or some form of survey. Members of Parliament are an interesting case as they regularly vote on issues within the House of Commons and their voting records are publically available. These records are available through a publicly exposed API [17]. As a result of focusing on MPs and not the general public the resultant dataset is easier to evaluate when training a model as there is scope for in depth analysis as to why false positives and false negatives are predicting on a per party basis. Moreover, having a genuine source of ground truth allows for more compelling arguments to be made about the relationship between Twitter presence and politics as one can speak with certainty. Moreover, having a genuine source of ground truth helps in addressing the criticisms that have been made regarding bodies of work in this area. This focus provides the foundations for addressing research questions **R1**, **R3** and **R4** with confidence.

Social media presence has become an increasingly important part of any MPs public persona and profile. Unlike a member of the general public it is therefore important for them to remain active. As a result this leads to a data set that is less sparse in comparison than one that focuses on the general public as their Twitter activity is likely to be far more sporadic. Moreover, unlike the average user an MP is likely to be discussing political content leaving the resultant dataset cleaner and generally less noisy. With the majority of work in this area focusing on elections, it is difficult to predict more than a single event given the sporadic nature of elections in general. This has lead some work to predict across different countries which relies on them having very similar political systems. With MPs voting on the House of Commons so regularly, there is scope for considering a number of different issues and the votes regarding these. This allows one to speculate about the robustness of a given model as well as making comment on the difficulties of addressing different topics in relation to **R5**.

5.3 Timeframe

Focusing around elections, the majority of work looks to predict events that fall just before or just after the time frame in which their content was derived. This results in a lot of the predictive tasks or descriptions of the dataset occuring in what would be considered the present. The advantage of considering the MP dataset with the frequency of voting events is that one can try and predict data from the past, present and future providing a deeper insight into the value of social media presence when considering politics. Having scope for investigating both the descriptive power of past and present and the predictive value for the future allows for more compelling arguments to be made about this relationship through **R1**, **R3** and **R4**. In addition, exploring both the descriptive and predictive power in these methods aids in addressing the criticisms raised about misleading claims regarding true predictive power.

5.4 Pre-Processing

It is a trend amongst research in this field to filter out 'noisy' data that comes with examining the general public by selecting tweets that contain words from hand crafted sets of keywords. These keywords may include political party names, the names of politicians, political events etc... However, in filtering on a handcrafted set of keywords there is some bias that is introduced into the data and perhaps some value lost. With a dataset orientated around MPs and a content focus on the value of hashtags, there is little preprocessing to be done in terms of removing noise from the dataset. In generating this dataset therefore the only consideration made was to omit tweets that did not contain a hashtag. With a dense dataset, the relationship between party and Twitter can be analysed in detail, addressing **R2**.

5.5 Methodology

The potential power of hashtags has already been recognised in research within this field. Hashtags have been used to filter content, define interaction profiles and construct parts of social graphs. However, work in this area is yet to consider the simpler approach of using hashtags in their raw form. Considering only hashtags and not the rest of the content in these tweets helps to address **R2**. Moreover, in proving the value in something as small as a hashtag, one can motivate confidence in this area of research that may have diminished as a result of critique by other work in this field.

In this area a number of different graphs have been considered as input data for solving this task. The simplest of these are arguably the follower or friends graph that encapsulate the follower or following relationship. With MPs using Twitter as a political tool one could hypothesis that politicians that share followers are likely to share values. As a result this dataset also considers a followers graph that encapsulates the follower relationship between a Twitter user and each MP to address **R4**.

This section has justified the decision to focus on and create a UK centric dataset including MP votes, their tweets that use hashtags and their follower networks. The process of generating this dataset and an exploration of its contents will be discussed in the next section.

Chapter 6

Data

This section will outline the process of constructing the dataset that was used as part of this work. The remainder of this section will explore this data, looking at how different parties and MPs utilise the Twitter platform. Sections of the dataset were curated in collaboration with Brandwatch Ltd [42] and they and their products will be referenced where used.

6.1 Data Retrieval

The data retrieval process can be split into a number of key steps: MP handle curation, vote curation, MP follower network curation, follower graph generation and tweet curation.

MP Handle Curation

The initial list of MP Twitter handles was curated by accessing the House of Commons Members API [43] and filtering based on MPs who had a non-empty Twitter Handle field. Some of the handles on this list were found to be out of date and were manually updated by hand. The data returned from this API call also contained information about each MPs party, which was used for the exploration of the data in this section. In curating this list of handles, a mapping between MP full name and handle was also generated as the House of Commons Divisions API [17] that is used for curating the vote information, does not include Twitter handles in its data but instead addresses each MP via their full name.

Vote Curation

The votes for each of the 4 selected topics were curated programmatically using the House of Commons Divisions API [17]. Each call to the API returned a list of MPs names mapped to their vote. A mapping between MP name and Twitter handle was used to convert each of these to a mapping between Twitter handle and vote. Votes returned from the API calls that regards MPs who do not use Twitter were disregarded.

MP Follower Network Curation

Instead of using the public facing Twitter API [44] to generate a list of followers for each MP, API calls were made to Brandwatch's Audiences product [?] to generate a followers list for each of the MPs handles. This could be replicated using the Twitter followers API [44] but is rate limited so was considered to be slower than using the Audiences product. These followers lists were then exported and filtered to remove followers who were not very active users of Twitter within the timeframe 2017-2019 where activity is based off of the number of tweets from an account in a year, i.e tweeting more than once a month. This left a list of active followers for each of the MPs.

Follower Graph Generation

The follower graph for the set of MPs was generated using the NetworkX Python package [45]. This was achieved by iterating over each follower list and generating a list of nodes and edges that make up this subgraph. These nodes and edges were then added to the overall follower graph before the next iteration began. In a case where an MP already existed within the graph as a follower of another MP, there node was removed from the subgraph before adding it to the overall followers graph.

Tweet Curation

The tweets for each of the MPs were curated using Brandwatch's ForSight product [46]. This allowed for tweets to be curated from the last few years and not just a stream of tweets from the last week as the public users of the Twitter Search API [47] are restricted to. The ForSight product allowed for the curation of tweets to be filtered so only tweets that contained a hashtag were collected for the list of MP twitter handles. These tweets were then accessed by day in a programmatic fashion using the hexpy Python package [48]. The returned results from each day were concatenated to produce a list of all of the MPs tweets that contained a hashtag from within the timeframe.

6.2 Data Composition

The dataset itself contains 4.5m Twitter handles, 570 of which are MPs and 250,000 tweets. These 250,000 tweets originate from a time period spanning from 1st January 2017 - 12th July 2019 and come from the accounts of the 570 MPs. The remainder of this section will explore how the data is composed, looking at both parties and MPs.

Parties on Twitter

At the time of writing, of the 650 MPs that sit in the House of Commons 584 of those have active Twitter accounts. Table 6.1 shows the by party percentage of current MPs that are active on Twitter. The majority of the smaller parties are relatively well represented on Twitter with 94.3% of MPs who do not represent the Conservative Party or Labour Party in the House of Commons active on Twitter. In comparison to this The Conservative Party are not as well represented with only

Party	On Twitter
Conservative	85% (263/311)
Labour	95% (235/247)
Scottish National Party	100% (35/35)
Independent (No Party)	88% (14/16)
Liberal Democrat	100% (12/12)
Democratic Unionist Party	80% (8/10)
Sinn Fein	100% (7/7)
The Independent Group	100% (5/5)
Plaid Cymru	100% (4/4)
Green Party	100% (1/1)

Table 6.1: Parties On Twitter

Party	Percentage of Seats (1 d.p)	Percentage of Dataset (1 d.p)
Conservative	47.90%	45%
Labour	38.00%	40.20%
Scottish National Party	5.40%	6.00%
Independent (No Party)	2.50%	2.40%
Liberal Democrat	1.90%	2.10%
Democratic Unionist Party	1.50%	1.40%
Sinn Fein	1.10%	1.20%
The Independent Group	0.80%	0.90%
Plaid Cymru	0.60%	0.80%
Green Party	0.20%	0.20%

Table 6.2: House of Commons and Twitter Percentage Shares

85% of their MPs active on Twitter. Using this grouping, the Labour Party are the best represented party with 95% of their MPs active on Twitter. The dataset itself, containing 14 less MPs, maintains this trend. The dataset contains 569 of these active MPs as it does not account for MPs that have joined Twitter in the last 3 months. Given his influence within British politics Nigel Farage has also been included as he is an interesting case, especially around the topic of Brexit.

In comparing representation in the House of Commons to representation on Twitter, the Conservative party remains the best represented party with the Labour party second best. However, as can be seen in Table 6.2 the share of representation between the Conservative Party and the Labour party is fairer on Twitter than it is within the House of Commons. Unlike the Conservative Party and the Labour party there is very little drop off between the representation in the House of Commons and the representation on Twitter for the remaining parties with only 4 of the 90 MPs in these parties represented in the House of Commons but not on Twitter. Despite this disparity, the share of representation amongst the parties remains the same in terms of ordering so one can consider the active MPs on Twitter to be a reasonably fair representation of the House of Commons.

Name (handle)	Number of Followers	Number of Active Followers	Party
Jeremy Corbyn	2,024,186	1,256,551	Labour
Nigel Farage	1,386,426	829,721	Brexit Party
Boris Johnson	927,031	361,083	Conservative
Theresa May	893,930	658,149	Conservative
Edward Miliband	757,023	545,038	Labour
David Lammy	541,653	306,628	Labour
Caroline Lucas	405,923	272,859	Green Party
Chuka Umunna	380,868	282,513	Liberal Democrat
Jacob Rees-Mogg	320,229	106,764	Conservative
Tom Watson	307,744	181,507	Labour

Table 6.3: Most Popular MPs on Twitter

6.2.1 MPs on Twitter

Popularity

In this context popularity on Twitter is defined in terms of followers and not interactions. As demonstrated in Table 6.3, Jeremy Corbyn is the most popular MP on Twitter with 118.4% more followers than Boris Johnson, the current Prime Minister. The second most popular MP Nigel Farage is no longer an active MP. It is worthy of note that the least represented party within both the House of Commons and this dataset, the Green party, are represented by the 7th most popular MP with Caroline Lucas having 405,923.

Not all Twitter users are as active as others, some sit dormant and do not tweet for an extended period of time. As a result, not all of the followers referenced in Table 6.3 are of use in this dataset. Table 6.3 demonstrates the number of followers lost when only the active followers are considered. An active follower is defined as a follower that has tweeted more than once a month in the last 12 months. There is a significant drop off in terms of followers for all of the most popular MPs on Twitter.

Activity

None of the MPs accounts are considered to be inactive with all of them showing reasonable activity over the last few years. The most active MPs on Twitter come from three main parties: the Scottish National Party, the Labour Party and the Labour and Co-operative party who stand for both the Labour Party and the Co-operative Party. The only exception to this, a member of the Independent Group, is formerly of the Labour and Co-operative Party. With the tweets collected in this dataset containing only hashtags, the most active MPs on Twitter in general and the most active MPs in this dataset differ as a result of the usage habits of different MPs on the platform. The only MP who features in both lists is Angus MacNeil MP a member of the Scottish National Party. As shown by Table 6.4, the most active users of hashtags among MPs is Luke Pollard MP a member of the Labour and Cooperative Party.

Author	All Posts	Hashtag Posts	Party
@PeterGrantMP	37,810	440	Scottish National Party
@MikeGapes	35,888	324	The Independent Group
@AngusMacNeilSNP	28,247	2,853	Scottish National Party
@jessphillips	24,704	933	Labour
@GwynneMP	21,899	1,246	Labour
@StewartMcDonald	18,909	968	Scottish National Party
@PaulJSweeney	18,793	639	Labour (Co-op)
@alisonthewliss	18,519	1,549	Scottish National Party
@BarrySheerman	18,129	1,327	Labour (Co-op)
@annaturley	17870	765	Labour (Co-op)

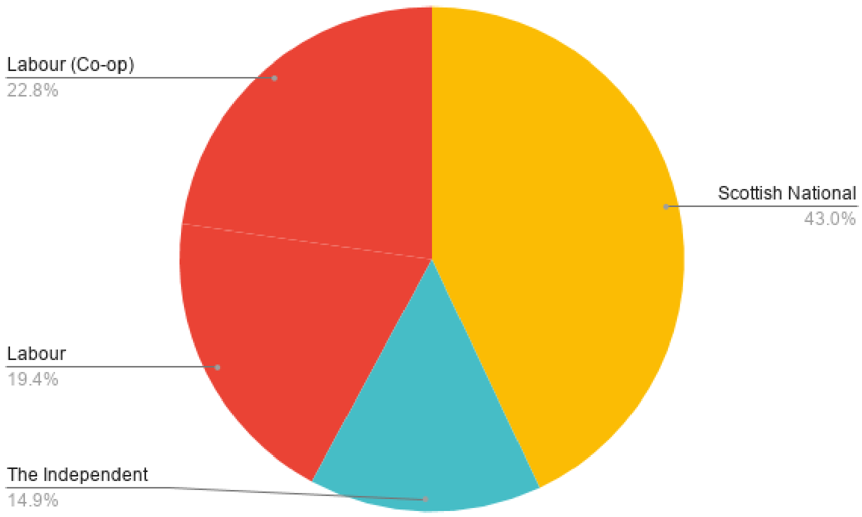
Table 6.4: Most Active MPs on Twitter

The top 10 most active users of hashtags represent a fairer share of the parties than the top 10 most active tweeters. Demonstrated by Figure 6.1(a) and Figure 6.1(b) 6 parties are represented by the 10 top users of hashtags whereas only 4 are represented by the most active users.

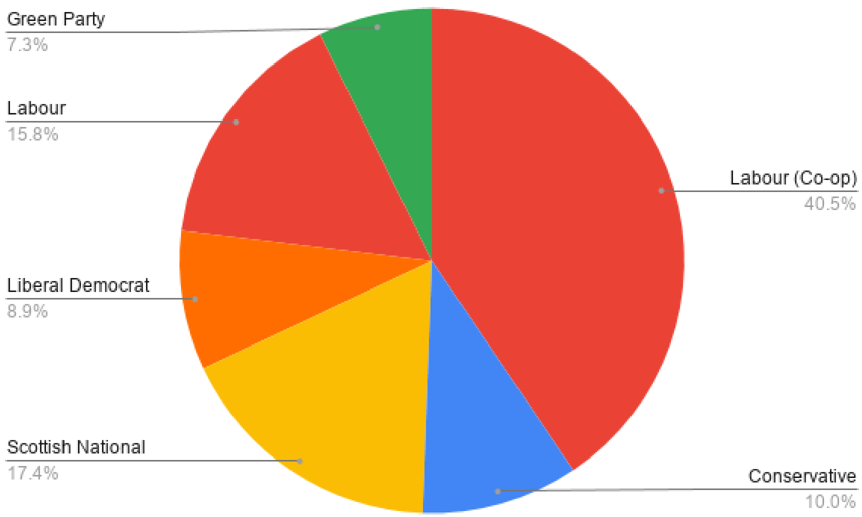
6.2.2 Content

Most Used Hashtags

Figure 6.2 demonstrates a word cloud of the most used hashtags across the entire dataset. The topic of Brexit clearly dominates the conversation in terms of the number of hashtags that reference it in some capacity: *#brexit*, *#peoplesvote*, *#brexitvote*, *#remainers*, *#hardbrexit* etc... There are also references to *#universalcredit*, *#socialcare*, *#theresamay*, *#tories* and *#labour* amongst many others. This variety demonstrates that hashtags are used to discuss a wide range of issues and are not centred around one or two use cases.



(a) Most Active Tweeters



(b) Most Active Users of Hashtags

Figure 6.1: Top Activity Share By Party

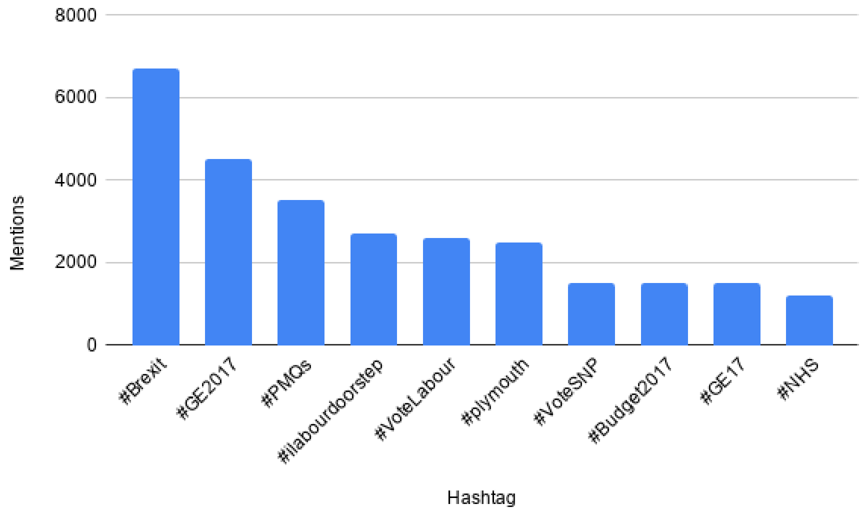
Figure 6.2: Most Frequent Hashtags

As shown by Figure 6.3(a) and Figure 6.3(b) the top 10 hashtags used between 12th July 2017 - 12th July 2018 and the 12th July 2018 - 12th July 2019 are quite different. The most used hashtag does not change with *#Brexit* dominating in both timeframes. *#PMQs* and *#plymouth* are the only other hashtags to appear across both timeframes. Although the mass of *#plymouth* is quite surprising, neither *#Brexit* or *#PMQs* are particularly surprising given the position Brexit holds in the forefront of British political dialogue and the regularity of Prime Minister's Questions.

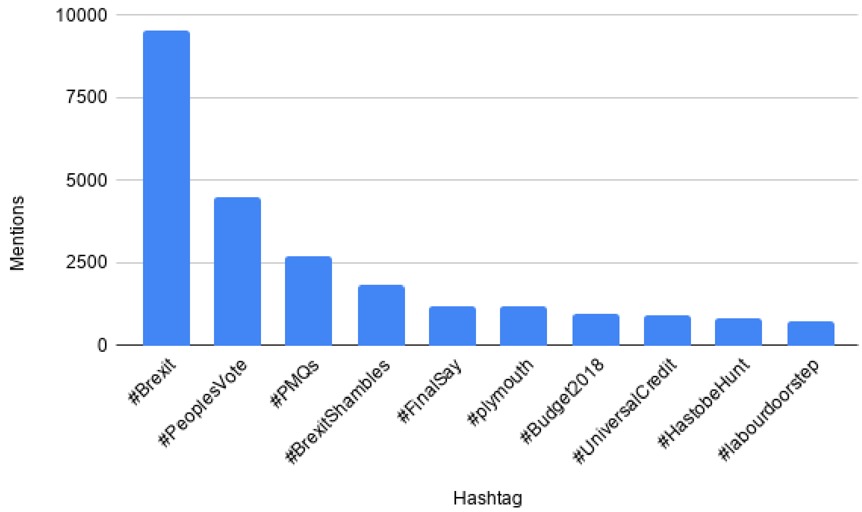
#NHS appears in the first timeframe and not the second whereas *#UniversalCredit* appears in the top 10 of the second timeframe but not the first.

6.2.3 Networks

Table 6.5 demonstrates the top 10 most connected MPs in the subgraph of the generated followers graph that contains only MPs. None of this group of MPs also sit in the list of most popular MPs on Twitter as shown in Table 6.3.



(a) Most Used Hashtags 12/07/2017-12/07/2018



(b) Most Used Hashtags 12/07/2018-12/07/2019

Figure 6.3: Most Used Hashtags

Handle	Edges
@mattwarman	540
@lindsayhoyle_mp	535
@lukepollard	527
@andrewrosindell	511
@jamescleverly	495
@wesstreeting	482
@thomasbrake	461
@nadhimzahawi	414
@rhonddabryant	386
@eleanor4epping	377

Table 6.5: Highest Degree MPs

Chapter 7

Implementation

This section outlines the implementation decisions that were made in producing the models. It will be split into a discussion around content-based models and a discussion around network based models. Raw hashtag, Word2Vec [31] and Doc2Vec [33] based models will be addressed in the content-based section and a label-propagation model will be addressed in the network-based section.

7.1 Raw Hashtag Model

A logistic regression model was used from the scikit learn linear-model library [49]. The c hyperparameter of this model was optimised utilising the scikit learn model_selection library [0] to perform a random search across a range of 0-2 using increments of 0.1. These optimisations were performed using k-fold cross validation [0] and evaluated using the accuracy_score metric [0]. The dimensions of the hashtags given as input were reduced by calculating the document frequency of each hashtag and removing those that were not used by more than 10 MPs where an MP was considered to be a document.

7.2 Word2Vec Model

The training data used to train the Word2Vec model was the list of all tweets each MP had produced in the last two years that contained a hashtag. Sentences from this list were taken and randomly shuffled and added as new sentences to provide further sentences for sampling. The Word2Vec model in the Gensim Python package [0] was used to generate this model. The average embedding for each MP was calculated and fed into the same logistic regression model used by the raw hashtag model.

7.3 Doc2Vec Model

The training data used to train the Doc2Vec model was the same as the Word2Vec model with each sentence tagged with the Twitter handle of the tweeting MP. The Doc2Vec model in the Gensim Python package [0] was used to generate this model. The document embedding for each MP was then taken and fed into the same logistic regression model used by the raw hashtag model.

7.4 Label Propagation

The label propagation model was implemented in Python utilising the NetworkX Python package [45] and the generated follower graph. Labels were propagated by taking the average of the labels of all labelled-neighbouring nodes and performing several iterations until all nodes has a label.

The implementations of the various models that were used have been discussed in this section. The next section will outline the experiments that were performed using these models to address the research questions that were laid out in the Introduction.

Chapter 8

Experiments

This section outlines the experiments that were performed to address the research questions this work is considering. An exhaustive description of the set up for each experiment can be seen in Appendix 1 where the inputs for each step of each experiment are listed.

8.1 Validity and Reliability

To ensure that each experiment is valid, reliable and reproducible each experiment will be performed in the same fashion. Unless specified otherwise in the experiment each case will be run using a seeded 80 20 training testing split. This will be achieved using k-fold cross validation to ensure that any results are robust. The generation of the seed nodes in each network-based experiment will be performed at random using a seed to ensure they are reproducible. The same random seed will be used regardless of the percentage of nodes that are being seeded in the experiment.

8.2 Experiment 1

This experiment will look to address **R1** by attempting to describe how an MP has voted using all 3 of the content based models.

8.3 Experiments 2, 3, 4

These experiments will look to address **R3** by attempting to predict votes in the immediate and distant past and future using the raw hashtag based model.

8.4 Experiments 5, 6

These experiments will look to address **R4** by attempting to predict votes in the immediate and distant past using the label propagation model.

This section has laid out the experiments that were performed using the models described in the Implementation section. The next section will contain an objective discussion surrounding the results of these experiments.

Chapter 9

Results

This section will present the results of the experiments laid out in the previous section. In keeping with the structure set out in the above section the results will be considered in relation to the research question that they address.

9.1 Experiment 1: Predict 2019 vote from data spanning the previous two years

9.1.1 Raw Hashtag Model

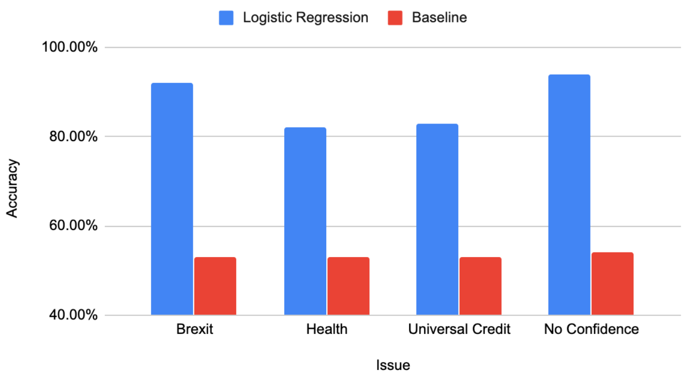


Figure 9.1: Describing Using a Raw Hashtag Model

This performed well in predicting all four of the different votes. The baseline performance was far exceeded in all cases where the baseline measurement used was predicting in favour of the majority class. The average baseline performance was 50% with the model showing an improvement of 34% on average. In considering the performance against the four different votes the votes naturally split into two different groups from the results. Predicting the Brexit vote and vote of no confidence demonstrated a similar accuracy of 90% and 90%. Whereas predicting the NHS vote and Universal Credit vote demonstrated accuracy of 82% and 82%.

9.1.2 Word Embedding Model

This model did not perform well. The performance was in line with the baseline with the average score of the model 55% and the baseline average 55%. Predicting

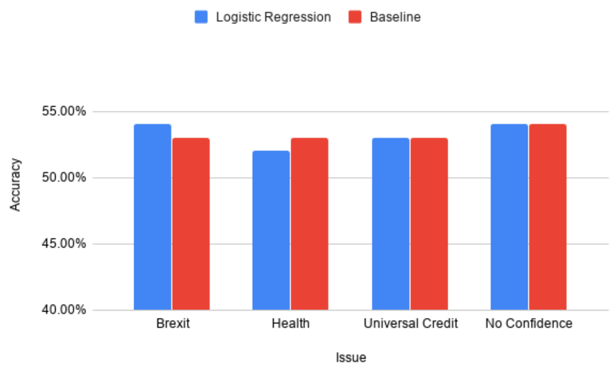


Figure 9.2: Describing Using a Word Embedding Model

the Brexit vote performed marginally above the baseline with a 3% improvement. In contrast, predicting the health vote performed marginally below the baseline with a decrease of 2% from the baseline of 54%. Both predicting the Universal Credit vote and vote of no confidence the performance was exactly in line with the baseline. Unlike the first experiment, there is no obvious separation between the votes in terms of performance.

9.1.3 Document Embedding Model

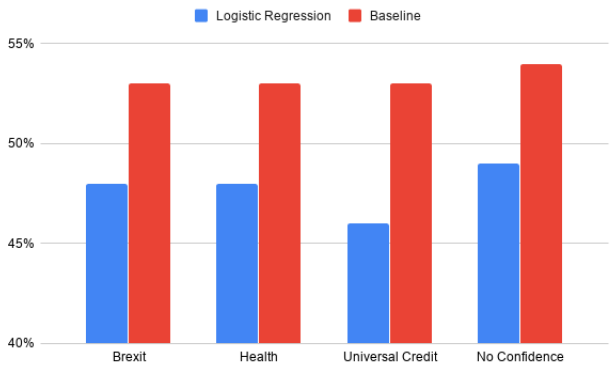


Figure 9.3: Describing Using a Document Embedding Model

This model did not perform well. The performance was consistently below the baseline with the average score of the model 48% and the baseline average 55%.

9.2 Experiment 2: Predicting future votes from Twitter data of previous years

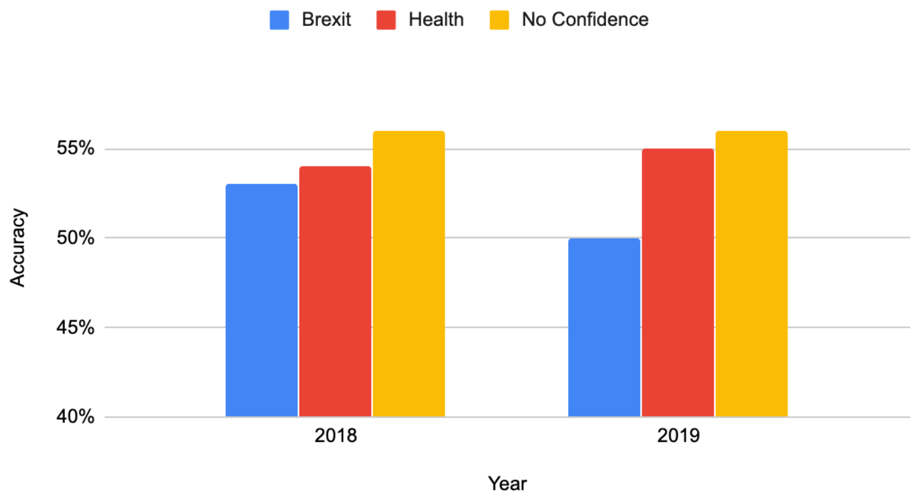


Figure 9.4: Predicting the Future

The logistic regression model using raw hashtag count did not perform well in predicting future votes. In predicting a Brexit, Health and vote of no confidence in 2018 and 2019 the model did not outperform the baseline average of 54%. Brexit was the most difficult vote to predict, achieving an average of 52% whereas the vote of no confidence was the easiest to predict with a slight increase in accuracy with an average of 56%.

9.3 Experiment 3: Predicting past votes from Twitter data of previous years

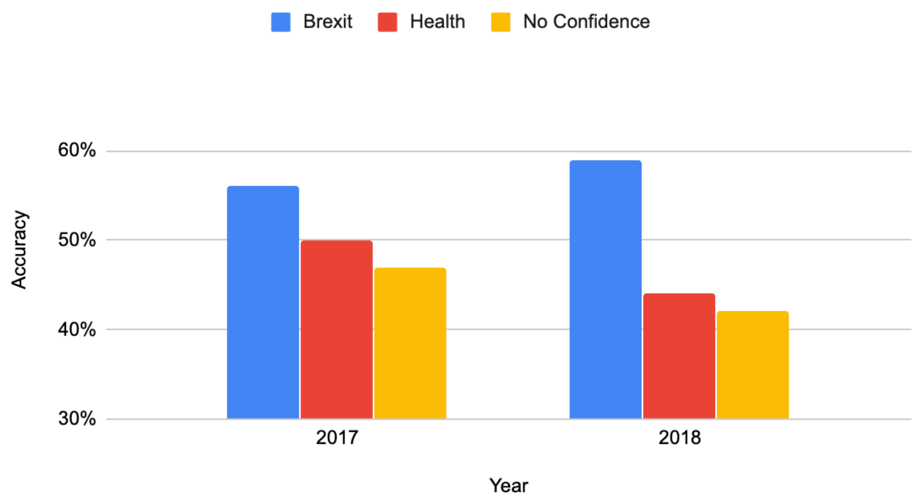


Figure 9.5: Predicting the Past

The logistic regression model using raw hashtags did not perform well in predicting votes from the past. Predicting the Brexit vote produced the best results in both years with 56% and 59% accuracy for 2017 and 2018 respectively. The worst performing vote was the vote of no confidence performing well below the baseline with 47% and 42% accuracy.

9.4 Experiment 4: Predicting next day votes from Twitter data of previous years

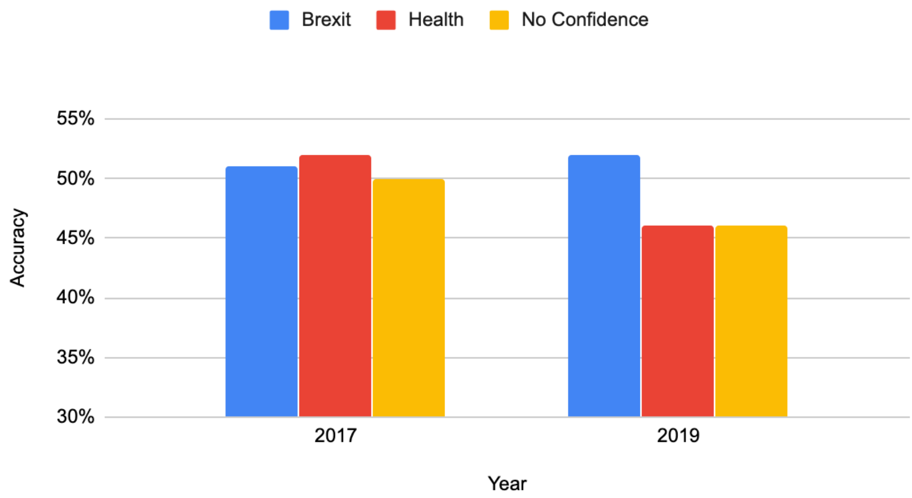


Figure 9.6: Predicting the near past and future

The logistic regression model using raw hashtags performed below the baseline in predicting both votes in the preceding and following year. The average performance for predicting the past year was slightly higher sitting at 51% compared to predicting the future at 47%. Both of these figures sit below the baseline of 53% and 54% respectively.

9.5 Experiment 5: Predict 2019 vote using different seed percentages of initially labelled nodes

The label propagation model performed well in predicting all four of the votes with a high performance achieved across a range of starting seed percentages. Best results were achieved for all votes with a 90% starting seed of the labeled nodes. In a similar fashion to the first experiment there is a clear separation in the votes in terms of performance. The Brexit and vote of no confidence votes showed consistent performance regardless of starting seed percentage sitting in the range between 94-99%. Both the health and Universal Credit votes achieved 90% accuracy with a 90% starting seed, performing between 80-91% across the different seed percentages.

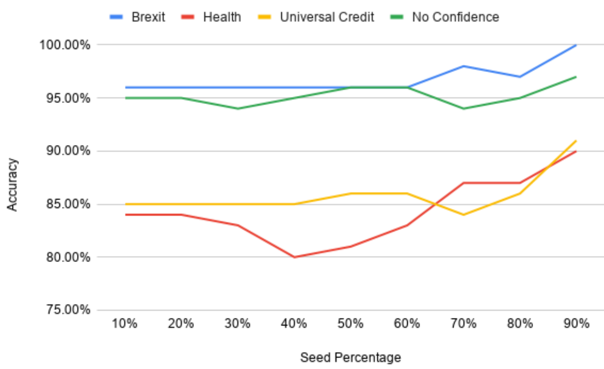


Figure 9.7: Predicts Across Different Starting Seed Percentages

9.6 Experiment 6: Predict 2017, 2018 and 2019 votes using 80% and 90% seeded nodes

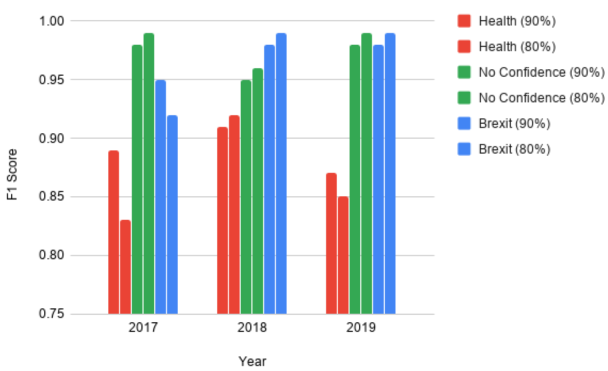


Figure 9.8: Predicting Using Label Propagation

The label propagation model performed well in predicting votes in 2017, 2018 and 2019. Both the Brexit votes and the votes of no confidence were consistently predicted at an accuracy greater than 90%. Although consistently performing below the other two votes predicting the health votes consistently far out performed the baseline. The votes in 2018 were the easiest to predict whereas 2017 was the worst performing year.

Chapter 10

Discussion

This section will further discuss the results of the experiments in the previous section and look to draw conclusions that explain these results. It will start with a general discussion before addressing each research question in light of these results.

Generally, the models were able to predict the outcome of different votes with varying successes with the more simple models, the raw hashtag model and the simple label propagation providing the better results. In line with the criticisms put forward by Metaxas et al [12] tasks that involved describing events that had occurred during the time frame of the training data performed with surprisingly high accuracy whereas the majority of tasks that involved predicting events outside of this time frame, for the content-based approach at least could not outperform the baseline of predicting in favour of the majority class. The results of the label propagation model suggests that there is genuine scope for utilising the Twitter platform to predict the voting behaviour of MPs. Given that MPs do not use Twitter like an average user, in the majority of cases, further experimentation would be required to conclude if the techniques laid out in this dissertation are applicable to the general public.

10.1 Is tweet content useful for describing how an MP has voted?

Tweet content has proven to be a very useful means of describing how an MP has voted. The notion that an MPs Twitter feed is very descriptive of how they might vote is not particularly surprising given that MPs utilise Twitter as a means of talking to their constituents, to campaign and to generally discuss the issues that they will be voting for. Moreover, in many cases their Twitter may be run by a member of their staff with each tweet specifically targeted to portray their political agenda. What is surprising here, however, is the accuracy one can achieve using only the hashtags the MP has used in the given timeframe. The use of hashtags is by no means consistent across MPs, as demonstrated in the Data section of this report, and therefore the consistency in which votes can be described across MPs is not immediately obvious. With this in mind, Twitters definition of the function of a hashtag as being used 'to index keywords or topics on Twitter' [50] is somewhat misrepresentative of the value that hashtags hold. The use of hashtags, by MPs at the very least, results in some semantic value regarding the topics that the hashtag refers to being carried by the hashtag.

The model that used hashtag embedding generated by a Word2Vec model instead of a raw hashtag count performed very poorly in describing how an MP had previously voted. A number of factors contributed to this performance with the simple choice of model being one of them. In taking the average of the hashtag embeddings for a given MP a lot of the information that is encoded by these hashtags is lost. Moreover, in generating these embeddings using context, Word2Vec models have the unwanted side effect of embedding antonyms to be very similar. In considering Brexit, the *#Leave* and *#Remain* hashtags should be considered to be at opposite ends of the spectrum in discussing the subject matter. However, the Word2Vec model used modelled the *#Remain* and *#Leave* hashtags as being each others second most similar hashtag. This is a result of both hashtags being used in the context of a sentence discussing Brexit. Similarly the Doc2Vec model performed poorly and can somewhat be attributed to characteristics that it inherits from the Word2Vec model as well as the naivety of the model used.

10.2 Is there a relationship between party and Twitter presence?

It is clear from the data that there is a relationship between how an MP utilises the Twitter platform and the party that they belong to. Members of the Scottish National Party and the Labour and Cooperative party are far more active on the platform than members of the Conservative party and the Democratic Unionist Party. In considering the political spectrum the platform appears to be utilised more by the so called 'left' than the 'right'. This is not particularly surprising given the work of Malik et al [23], Vaccari et al [25], amongst others [12, 24] concluding that users on the platform are more often younger and more 'left wing' in their political stance. There is also an evident disparity in the use of hashtags by different MPs on Twitter with some using the feature heavily and some utilising it sparingly. However, there does not appear to be an obvious correlation between a favoured use of hashtags and party as the top hashtag users all represent 6 different parties across the political spectrum. Although it is noted that of the top 10 hashtag users, 9 of them represent what would be considered more 'left wing' parties this is not all that surprising given that the two parties that would be considered most 'right wing', the Conservative party and the Democratic Unionist party, are both the worst represented on Twitter in terms of the percentage of their parties that use the platform. Interestingly, the percentage share of the seats of MPs on Twitter by party is well representative of the actual percentage share in the House of Commons suggesting that Twitter may provide a useful platform for pursuing further research into UK politics in general, not just in a machine learning setting.

10.3 Is tweet content useful for predicting how an MP will vote?

Poor predictive performance was observed when using a content-based model that falls in line with literature [12–14]. This large drop-off in performance when considering a predictive task opposed to a descriptive task can be attributed to the use of

	2017	2018	2019
2017	226	67	35
2018	67	362	53
2019	35	53	196

Table 10.1: Intersection of Hashtags

hashtags in the model. Hashtags are very specific and over time the keywords and therefore hashtags used to describe certain events change subtly. The divergence or short shelf life of hashtags in terms of politics is demonstrated in Table 10.1 showing the number of unique hashtags that exist in the intersection between the set of unique hashtags used in different years.

To combat this divergence the embedding space produced by the Word2Vec and Doc2Vec models could be used. A limiting factor in their application in this case is a focus on context places value of the ordering of hashtags whereas the permutation of the hashtags that an MP has used does not hold any obvious value when considered as input to a machine learning model. These permutation issues could result in two MPs who have used the same hashtags just in a different order of tweets being considered to be quite different. Although there could be some value in considering these MPs as different, the value is not immediately obvious.

10.4 Are follower networks useful for predicting votes?

Unlike the content-based models, the network-based label propagation model performed very well in predicting voting behaviour outside of the time frame that the follower network was generated in. Even when seeding a small percentage of the labelled nodes results well above the baseline were achieved across all votes, outperforming the content-based models in every experiment. In preparing this report, visualisations of the graph of MPs who follow other MPs could not be achieved to a reasonable visual standard due to how densely connected UK MPs are on Twitter. The high performance of the label propagation model, even after only 2 iterations, suggests that these MPs are connected into communities that define their voting patterns to a reasonable accuracy. These communities contain MPs that directly follow each other as well as MPs who are connected by a shared follower. Unlike hashtags, the core of a follower network is far more static and is unlikely to diverge dramatically from year to year. This allows the simple label propagation model to perform consistently across different years.

10.5 Are all votes as predictable as others?

In the poorly performing models there was no obvious correlation or pattern between the topic that the vote in question was addressing and performance. However, when looking at the high performance achieved in the descriptive task using hashtags and the predictive task using label propagation there is a clear separation between voting topics. Surprisingly, the votes that one would expect to sit along party lines, addressing Universal Credit and the NHS were harder to predict than those that one would not expect to sit along party lines. Despite being the most divisive and cross-party of all the issues, votes surrounding Brexit demonstrated a high performance consistently. With the clear separation between the NHS and Universal Credit centric votes and the Brexit and vote of no confidence centric votes, introducing further topics into the performed experiments, one might attempt to predict the performance for a given topic based off of how along party lines or divisive the topic is considered to be. In attempting to predict votes in the past and future both immediate and distant future, there was no correlation between the proximity of a vote to the trained time frame and its performance. This supports the findings Gayo-Avello et al [13] in suggesting that genuinely good performance in this area may just be down to chance.

10.6 Further Work

This work has shown that there is strong descriptive and predictive power in two very simple models. Given the success of these simplistic approaches, an extension of this work should build on this proof of concept by applying more complex models to the dataset. The results of the label propagation are especially promising and further work in this area, implementing more sophisticated algorithms and investigating the effectiveness of community detection algorithms on the completely unlabeled graph. An alternative direction may involve investigating the application of neighbourhood information given that success was found using both a content based and network based approach. A suggested first step in this direction would be to implement a Graph Convolution Network model [0]. However, more value is likely to be found in further exploring the network-based approaches in an attempt to address the criticisms of work in this area.

Chapter 11

Conclusion

In conclusion this work has explored the relationship between Twitter presence and voting patterns and has found that there is scope for interesting further work in this area in utilising follower network structure to predict the voting pattern of MPs. Further, in investigating this relationship this work has found that critique of work in this area is not unjust with their conclusions regarding the unsuitability of Twitter content as a means for predicting future votes can also be drawn from the results of this work's experiments.

Bibliography

[1] Omnicore, "Twitter statistics." https://www.omnicoreagency.com/twitter-statistics/. accessed: 01.09.2019.

[2] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: Using twitter to predict the uk 2015 general election," *Electoral Studies*, 05 2015.

[3] K. McKelvey, J. DiGrazia, and F. Rojas, *Twitter publics: How online political communities signaled electoral outcomes in the 2010 US house election*, pp. 46–60. 03 2016.

[4] P. Salunkhe and S. J. Deshmukh, "Twitter based election prediction and analysis," 2017.

[5] A. Tumasjan, T. Oliver Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," vol. 10, 01 2010.

[6] J. Digrazia, K. Mckelvey, J. Bollen, and F. Rojas, "More tweets, more votes: Social media as a quantitative indicator of political behavior," *PloS one*, vol. 8, p. e79449, 11 2013.

[7] A. Fang, I. Ounis, P. Habel, C. Macdonald, and N. Limsopatham, "Topic-centric classification of twitter user's political orientation," pp. 791–794, 08 2015.

[8] V. Lampos, D. Preotiuc-Pietro, and T. Cohn, "A user-centric model of voting intention from social media," 08 2013.

[9] P. Grover, A. Kar, Y. Dwivedi, and M. Janssen, "Polarization and acculturation in us election 2016 outcomes – can twitter analytics predict changes in voting preferences," *Technological Forecasting and Social Change*, 09 2018.

[10] A. Makazhanov, D. Rafiei, and M. Waqar, "Predicting political preference of twitter users," *Social Network Analysis and Mining*, vol. 4, pp. 1–15, 05 2014.

[11] P. Grover, A. Kar, Y. Dwivedi, and M. Janssen, "The untold story of usa presidential elections in 2016 - insights from twitter analytics," pp. 339–350, 10 2017.

[12] P. Metaxas, E. Mustafaraj, and D. Gayo-Avello, "How (not) to predict elections," pp. 165–171, 10 2011.

[13] D. Gayo-Avello, P. Metaxas, and E. Mustafaraj, "Limits of electoral predictions using twitter," 01 2011.

[14] C. Lui, P. Metaxas, and E. Mustafaraj, "On the predictability of the us elections through search volume activity," 05 2012.

[15] A. Rauchfleisch and J. Metag, "The special case of switzerland: Swiss politicians on twitter," vol. 18, pp. 2413–2431, 11 2016.

[16] H. of Parliament, "data.parliament.uk." `http://www.data.parliament.uk/`. accessed: 01.09.2019.

[17] H. of Parliament, "Commons divisions." `http://explore.data.parliament.uk/?learnmore=Commons\%20Divisions`. accessed: 01.09.2019.

[18] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," pp. 115–120, 07 2012.

[19] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," pp. 192–199, 10 2011.

[20] M. Choy, M. Cheong, M. Nang Laik, and K. Ping Shung, "A sentiment analysis of singapore presidential election 2011 using twitter data with census correction," 08 2011.

[21] A. Ceron, L. Curini, S. Iacus, and G. Porro, "Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to italy and france," *New Media amp Society*, vol. 16, pp. 340–358, 02 2014.

[22] J. Mellon and C. Prosser, "Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users," *Research Politics*, vol. 4, p. 205316801772000, 07 2017.

[23] M. Malik, H. Lamba, C. Nakos, and J. Pfeffer, "Population bias in geotagged tweets," *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-scale Social Media Research*, pp. 18–27, 01 2015.

[24] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of twitter users," vol. 11, 01 2011.

[25] C. Vaccari, A. Valeriani, P. Barberá, R. Bonneau, J. Jost, J. Nagler, and J. Tucker, "Social media and political communication: A survey of twitter users during the 2013 italian general election," *Rivista Italiana di Scienza Politica*, vol. 43, pp. 381–410, 12 2013.

[26] P. Barbera and G. Rivero, "Understanding the political representativeness of twitter users," *Social Science Computer Review*, vol. 33, 12 2014.

[27] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," vol. 11, 01 2010.

[28] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," 01 2011.

[29] M. Marchetti-Bowick and N. Chambers, "Learning for microblogs with distant supervision: political forecasting with twitter," pp. 603–612, 04 2012.

[30] J. Liu, Z. He, and Y. Huang, "Hashtag2vec: Learning hashtag representation with relational hierarchical embedding model," pp. 3456–3462, 07 2018.

[31] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 10 2013.

[32] T. Mikolov, W.-T. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," *Proceedings of NAACL-HLT*, pp. 746–751, 01 2013.

[33] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st International Conference on Machine Learning, ICML 2014*, vol. 4, 05 2014.

[34] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: The structure of the twitter follow graph," pp. 493–498, 04 2014.

[35] M. Gabielkov and A. Legout, "The complete picture of the twitter social graph," pp. 19–20, 12 2012.

[36] L. Hong, O. Dan, and B. Davison, "Predicting popular messages in twitter," pp. 57–58, 01 2011.

[37] M. Gabielkov, A. Rao, and A. Legout, "Studying social networks at scale: Macroscopic anatomy of the twitter social graph," vol. 42, 04 2014.

[38] S. Volkova, G. Coppersmith, and B. Van Durme, "Inferring user political preferences from streaming communications," vol. 1, pp. 186–196, 06 2014.

[39] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," 01 2012.

[40] T. C. I. f. I. BCS, "Bcs, the chartered institute for it." `https://www.bcs.org/`. accessed: 01.09.2019.

[41] T. C. I. f. I. BCS, "Code of conduct for bcs members." `https://www.bcs.org/upload/pdf/conduct.pdf`. accessed: 01.09.2019.

[42] Brandwatch, "Brandwatch." `https://www.brandwatch.com/`. accessed: 01.09.2019.

[43] H. of Parliament, "Members." `http://explore.data.parliament.uk/?learnmore=Members/`. accessed: 01.09.2019.

[44] Twitter, "get-followers-ids." `https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids.html`. accessed: 01.09.2019.

[45] N. Developers, "Networkx." `https://networkx.github.io/documentation/networkx-1.9.1/overview.html`. accessed: 01.09.2019.

[46] Brandwatch, "Brandwatch forsight." `https://www.brandwatch.com/products/forsight/`. accessed: 01.09.2019.

[47] Twitter, "get-search-tweets." `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html`. accessed: 01.09.2019.

[48] C. Sullivan, "hexpy." `https://github.com/sullivancolin/hexpy`. accessed: 01.09.2019.

[49] scikit learn, "sklearn.linear$_m$*odel.logisticregression.*″. accessed: 01.09.2019.

scikit learn, "sklearn.model$_s$*election.randomizedsearchc.*″. accessed: 01.09.2019.

scikit learn, "sklearn.sklearn.model$_s$*election.kfold.*″. accessed: 01.09.2019.

scikit learn, "sklearn.metrics.accuracy$_s$*core.*″. accessed: 01.09.2019.

gensim, "models.word2vec – word2vec embedding." `https://radimrehurek.com/gensim/models/word2vec.html`. accessed: 01.09.2019.

gensim, "models.doc2vec – doc2vec paragraph embeddings." `https://radimrehurek.com/gensim/models/doc2vec.html`. accessed: 01.09.2019.

T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *CoRR*, vol. abs/1609.02907, 2016.

Appendix A

Appendix Title

A.1 Appendix 1

Experimental Setup

Baseline: Predicting the majority class

Experiment 1:

Cases:
Brexit:
X: MPs hashtags from 12 June 2017 - 12 June 2019
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: MPs hashtags from 12 June 2017 - 12 June 2019
y: MPs votes Vote of No Confidence 2019
Health:
X: MPs hashtags from 12 June 2017 - 12 June 2019
y: MPs votes NHS vote in 2019
Universal Credit:
X: MPs hashtags from 12 June 2017 - 12 June 2019
y: MPs votes Universal Credit vote 2019
Brexit:
X: MPs hashtag embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: MPs hashtag embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Vote of No Confidence 2019
Health:
X: MPs hashtag embeddings from 12 June 2017 - 12 June 2019
y: MPs votes NHS vote in 2019
Universal Credit:
X: MPs hashtag embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Universal Credit vote 2019

Brexit:
X: MP embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Brexit vote in 2019
Vote of No Confidence:

X: MP embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Vote of No Confidence 2019
Health:
X: MP embeddings from 12 June 2017 - 12 June 2019
y: MPs votes NHS vote in 2019
Universal Credit:
X: MP embeddings from 12 June 2017 - 12 June 2019
y: MPs votes Universal Credit vote 2019

Experiment 2:
Cases:
Brexit:
X: MPs hashtags from 1 January 2017 - 1 January 2018
X: MPs hashtags from 1 January 2018 - 1 January 2019
y: MPs votes Brexit vote in 2018
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: MPs hashtags from 1 January 2017 - 1 January 2018
X: MPs hashtags from 1 January 2018 - 1 January 2019
y: MPs votes Vote of No Confidence in 2018
y: MPs votes Vote of No Confidence in 2019
Health:
X: MPs hashtags from 1 January 2017 - 1 January 2018
X: MPs hashtags from 1 January 2018 - 1 January 2019
y: MPs votes NHS vote in 2018
y: MPs votes NHS vote in 2019
Universal Credit:
X: MPs hashtags from 1 January 2017 - 1 January 2018
X: MPs hashtags from 1 January 2018 - 1 January 2019
y: MPs votes Universal Credit vote in 2018
y: MPs votes Universal Credit vote in 2019

Experiment 3:
Cases:
Brexit:
X: MPs hashtags from 1 January 2018 - 1 January 2019
X: MPs hashtags from 1 January 2019 - 12 June 2019
y: MPs votes Brexit vote in 2017
y: MPs votes Brexit vote in 2018
Vote of No Confidence:
X: MPs hashtags from 1 January 2018 - 1 January 2019
X: MPs hashtags from 1 January 2019 - 12 June 2019
y: MPs votes Vote of No Confidence in 2017
y: MPs votes Vote of No Confidence in 2018
Health:
X: MPs hashtags from 1 January 2018 - 1 January 2019
X: MPs hashtags from 1 January 2019 - 12 June 2019
y: MPs votes NHS vote in 2017

y: MPs votes NHS vote in 2018
Universal Credit:
X: MPs hashtags from 1 January 2018 - 1 January 2019
X: MPs hashtags from 1 January 2019 - 12 June 2019
y: MPs votes Universal Credit vote in 2017
y: MPs votes Universal Credit vote in 2018

Experiment 4:
Cases:
Brexit:
X: MPs hashtags from 1 January 2018 - 2019 VOTE
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: MPs hashtags from 1 January 2018 - 2019 VOTE
y: MPs votes Vote of No Confidence in 2019
Health:
X: MPs hashtags from 1 January 2018 - 2019 VOTE
X: MPs hashtag embeddings from 1 January 2018 - 2019 VOTE
X: MP embeddings from 1 January 2018 - 2019 VOTE
y: MPs votes NHS vote in 2019
Universal Credit:
X: MPs hashtags from 1 January 2018 - 2019 VOTE
X: MPs hashtag embeddings from 1 January 2018 - 2019 VOTE
X: MP embeddings from 1 January 2018 - 2019 VOTE
y: MPs votes Universal Credit vote in 2019

Experiment 5:
Cases:
Brexit:
X: 10-90% labelled MP nodes with their true 2019 Brexit vote
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: 10-90% labelled MP nodes with their true 2019 Vote of no Confidence vote
y: MPs votes Vote of No Confidence 2019
Health:
X: 10-90% labelled MP nodes with their true 2019 NHS vote
y: MPs votes NHS vote in 2019
Universal Credit:
X: 10-90% labelled MP nodes with their true 2019 Universal Credit vote
y: MPs votes Universal Credit vote 2019

Experiment 6:
Cases:
Brexit:
X: 90% labelled MP nodes with their true 2019 Brexit vote
X: 80% labelled MP nodes with their true 2019 Brexit vote
y: MPs votes Brexit vote in 2017

y: MPs votes Brexit vote in 2018
y: MPs votes Brexit vote in 2019
Vote of No Confidence:
X: 90% labelled MP nodes with their true 2019 Vote of no Confidence vote
X: 80% labelled MP nodes with their true 2019 Vote of no Confidence vote
y: MPs votes Vote of No Confidence 2017
y: MPs votes Vote of No Confidence 2018
y: MPs votes Vote of No Confidence 2019
Health:
X: 90% labelled MP nodes with their true 2019 NHS vote
X: 80% labelled MP nodes with their true 2019 NHS vote
y: MPs votes NHS vote in 2017
y: MPs votes NHS vote in 2018
y: MPs votes NHS vote in 2019
Universal Credit:
X: 90% labelled MP nodes with their true 2019 Universal Credit vote
X: 80% labelled MP nodes with their true 2019 Universal Credit vote
y: MPs votes Universal Credit vote 2017
y: MPs votes Universal Credit vote 2018
y: MPs votes Universal Credit vote 2019