UNIVERSITY OF BIRMINGHAM

Mining Twitter Data for Improving Lexicon-Based Election Predictions and Candidate Analysis on Political Issues: Hybrid Topic-Based Sentiment Analysis with Issue Filtering

Samuel Kopelowitz, BSc Computer Science w/ Year in Industry FT,

Under Supervision of Prof. Uday Reddy,

School of Computer Science, April 2020

ABSTRACT

Twitter data mining techniques have been used in the run-up to elections to predict their outcomes and perform analysis to explain results. Due to the popularity of the social media platform it is possible to collect large amounts of data with which often lexicon-based sentiment analysis has been used to accomplish these tasks, mostly because of its efficiency and simplicity. More recently, hybrid techniques, which in addition to calculating tweet sentiment also incorporate topic modelling methods to extract the main "topics" from a corpus of text, have been applied independently for both election prediction and analysis. It is possible to use hybrid methods to analyse different political issues (e.g. economic, social, etc) and the public opinion for candidates in respect to them; and other hybrid methods have been shown to outperform baseline sentiment analysis approaches for election prediction. A mining solution which can accomplish both of these tasks non-exhaustively is desirable for better predictions and a greater understanding of election outcomes. This report will present a novel approach to mining Twitter data, Hybrid Topic-Based Sentiment Analysis with Issue Filtering (HTBSA), which will not only pose as a potential improvement upon state-of-the-art techniques for election prediction; but can be abstracted to perform candidate analysis on any individual political issue, proposing a baseline methodology for doing this. This research approach has effectively outperformed all of the well-established methods in the realm of lexicon-based election prediction, giving a mean average error as low as 2.20% from true vote share. This technique was performed on data collected on the run up to the UK General Election 2019 and in an addition to this, it has successfully been black box tested on an unseen dataset. Based on the empirical evidence given by our results, HTBSA* can be relied upon to predict elections occurring in the future, but analysis results in respect to individual political issues may be inconsistent, suggesting further work is required. Lines of research that come as a result of this study have the potential to tackle election mining problems in new ways, which are more sophisticated than what has been done previously.*

Keywords: *Sentiment analysis; Topic modelling; Election prediction; Twitter data; Mining public opinion; Final year project report*

TABLE OF CONTENTS

1. INTRODUCTION

The increasing use of social media globally has seen a sharp uprise in the amount of data that is available to analyse various trends. Twitter in particular has become a popular communication channel for people to express their opinion and the social media giant boasts 16 million active users in the UK alone, as well as up to 328 million users worldwide [London School of Economics, 2017]. During political campaigns and elections, social media has become a useful tool for both contesting parties and their voters to express their opinions. With so much political discourse now taking place online, it should be no surprise that in 2017 political parties spent a record-breaking £3.2m on Facebook advertising alone [Guardian, 2018]. The influence of internet advertising, particularly in politics, has led to increasing pressure being put on social media firms to control their content and in October 2019 Twitter CEO Jack Dorsey announced an indefinite ban on all political advertising on the platform.

A popular area in which Twitter data has been used is in the analysis and forecasting of results in political elections. Traditionally, opinion polls and public surveys have been the bridge between public opinion and politicians. These have historically played an important role on the run-up to elections by trying to depict election results and various other statistics such as public opinion on candidates in respect to different issues or voting intention by a particular demographic. With such a large amount of data available and the election forecasting 'industry' growing more and more, techniques which rely on Twitter data to analyse elections and their outcomes are becoming increasingly more credible in tackling the problems in which opinion polls have traditionally been used. Criticism of the consistency and clarity of such methodologies have been voiced in the past, however, there is ample evidence of mining techniques which have given surprisingly high levels of accuracy – even exceeding that of opinion polls in recent years. The cost-effectiveness of such solutions makes them highly desirable from a commercial standpoint in replacing the well-established methods used in public surveying.

The standard approach for mining Twitter data for election prediction is to rely on lexicon-based sentiment analysis alone (LSA). More recently "hybrid" techniques have emerged which, as well as analysing tweet sentiment, also extract the main "topics" from a candidate's set of tweets in order to perform election analysis. Such techniques have been shown independently to outperform state-of-the art lexicon-based methods for predicting election outcomes and also to successfully perform candidate analysis in respect to different political issues. A mining solution which can accomplish both of these tasks non-exhaustively is desirable for improving election predictions and gaining a greater understanding of their outcomes in future studies.

In this project report, we propose a novel lexicon-based approach to mining Twitter data: Hybrid Topic-Based Sentiment Analysis with Issue Filtering (HTBSA*). The proposed approach builds upon current methods that use sentiment analysis and topic modelling to predict elections by improving the underlying principles used to calculate voter intention and applying the most up-to-date techniques in this area of research. This will be performed on Twitter data collected on the run-up to the UK General Election 2019 which will not only seek to improve upon state-of-the-art methods for prediction but can be abstracted to perform candidate analysis for the public opinion on any given number of political issues. The latter could propose a baseline methodology for this type of problem. In addition to this, we will apply an identical implementation of HTBSA* to an unseen dataset, to objectively evaluate its performance to analyse and forecast election outcomes in the future.

2. LITERATURE REVIEW

Twitter has become the most popular social media outlet for researchers, mostly due to the convenience of having such a large amount of data available for capturing key trends. A range of different studies have been developed that have used Twitter data to examine a wide variety of disciplines such as sociology, computer science, media and communication, political science, and engineering to name only a few. In this section we will review how data mining techniques have been applied to election analysis problems in the past. Most of these will be in regard to predicting vote share as this has been a dominant area for researchers in recent years. Standard Twitter datasets which can be used for election mining will label tweets into distinct groups for a given number of candidates.

The simplest form of election prediction that can be done using Twitter data is analysing the raw volume of tweets for each candidate. Typically, other methods of election prediction will be compared with Volume Analysis because it has been shown that volume of tweets alone can successfully represent election polls [Tumasjan et al, 2010]. Although some primary results given by analysing volume of tweets were promising, [Gayo-Avello, 2011] concluded in a comprehensive review of the literature that results were unpredictable and there was a lack of strong evidence to consider it a valid method of prediction.

Early research that used sentiment analysis to predict candidate vote share also gave mixed results. Sentiment analysis is commonly used as an umbrella term for methods that are able to disclose polarity and subjectivity from a given piece of text. Polarity can be summarised as how positive/negative the emotions expressed in a piece of text are whereas subjectivity distinguishes sentences that express factual information from those that are based on subjective views and opinions, i.e. it is a measurement of how subjective a piece of text is [Liu, 2012]. There is generally considered to be two main approaches to performing sentiment analysis. The first approach is known as the "learning-based" one, also known as the machine learning approach. This approach requires pre-defined knowledge about data categories and uses these to train a classifier. The second approach is known as the "lexicon-based" approach, which relies on a dictionary of pre-defined positive and negative terms to disclose the sentiment of text and requires no prior knowledge about its categories. [Gayo-Avello, 2011] showed that using sentiment analysis was better than raw tweet count for forecasting elections but still underperformed the prediction baseline. [Metaxas et al. 2011] demonstrated that sentiment analysis was able to outperform both raw volume count and the baseline, however, others including Gayo-Avello and Metaxas himself, concluded that contemporary methods that used lexicon-based sentiment analysis to predict elections were close to using random classifiers. Later studies also gave indecisive results [Bermingham & Smeaton, 2011; Tjon Kim Sang & Bos, 2012] and [Gayo-Avello, 2011] concluded that overall results were contradictory, but it was clear that even naïve sentiment analysis could outperform the baseline and further research was needed. Following on from these remarks, [Gayo-Avello, 2012] then outlined a set of recommendations for improving future election predictions and cited the flaws that past researchers had made. His top recommendations can be summarised: predict an election happening in the future, clearly define what constitutes as a "reliable indicator of a vote" and take into account the biases within the data in relation to the candidates being studied. Finally, he called upon researchers to devote more time to improving sentiment analysis methods rather than relying on simplistic assumptions.

These principles have been useful in predicting elections since, for example [DiGrazia et al, 2013] demonstrated a positive relationship between number of mentions in tweets and electoral vote share for both the 2010 and 2012 U.S. Congressional elections. A supervised sentiment analysis method was used by [Ceron et al, 2014] to show that such a method was capable of supplementing traditional off-line polls for Italian political leaders in the 2011 parliamentary elections and candidates in the French 2012 Presidential election. More supervised learning techniques have been employed in subsequent studies to calculate polarity of political texts. For instance [Marozzo & Bessi, 2017] used Random Forest algorithm to do this on Twitter and news texts related to political campaigns. Such a technique requires a large training dataset and significantly more computational power compared to lexicon-based methods. In 2015 [Burnap et al, 2015] proposed a baseline lexicon-based semantic analysis method based upon Gayo-Avello's principles. Burnap's methodology used positive tweets to calculate overall vote share. This was then applied to the true result of the preceding election in order to calculate a measure of national swing, which was then applied on a constituency-by-constituency basis to produce an estimate of which party would win a given seat. Of course, this step can only be applied to UK General Elections but with some abstraction it is likely that it could be applied to other electoral systems. This performed effectively on tweets collected on the run-up to the 2015 UK General Election. Little to no progress has ever been made in solving the problems posed by using Twitter data as a population sample as highlighted by [Gayo-Avello, 2012] and this barrier has caused somewhat of a halt in progressing the field of election prediction mining in recent times; all of Gayo-Avello's other recommendations have been explored to a much greater extent. The complications arising in the use of Twitter data have been well-documented in many election prediction studies: Twitter bots (spam accounts), underrepresenting particular demographics such as the older generation and those living in rural areas as well as the varying proportions of active Twitter users in each candidate's voting population are the main problems. Some have even claimed it is unclear whether we can improve our current mining techniques without further analysing these problems first. [Mellon & Prosser, 2017] argued that current methods are not useful in predicting the voting intentions of an entire population as they can only be representative of those who are active Twitter users, which would lead to mixed results without appropriate adjustment. Mellon evidenced this claim by highlighting various disproportions in social media data when compared with true population statistics such as

the overrepresentation of the left-leaning Labour Party in data collected from Facebook and Twitter. It is true that work has begun to address these problems and studies have made attempts to scale population samples (i.e. weighting data based on demographics so that total population is more accurately represented) [Filho et al, 2015; Wang et al, 2014] and this has led to some success, however, the techniques proposed require detailed data and lots of care to implement. Overall, there is a lack of reliable solutions as of yet.

One factor not considered in [Gayo-Avello, 2012] work was the possibility of creating more sophisticated methods by taking other features of the corpora generated by Twitter data into account (other than volume or sentiment-based statistics). [Bansal & Srivastava, 2018] recognised this and proposed a novel method which took word relations and co-occurrences in candidate datasets into account as well as tweet sentiment. This technique is known as topic modelling and Bansal showed that it could be applied successfully in conjunction with principles given by Burnap's LSA to calculate vote share, he named this method Hybrid Topic-Based Sentiment Analysis (HTBSA) and claimed it provided a better estimate of sentiment polarity and score of tweets than LSA methods for election prediction. Before we can review HTBSA as part of this section we must touch briefly on topic modelling. Topic modelling is a computational technique used to group words together that frequently co-occur in a corpus of texts, resulting in a list of distinct "topics" generated from the corpus. Topic modelling techniques have been applied to data mining problems in a number of different contexts from climate change discussion to tweets about Uber [Dahal et al, 2019; Alamsyah et al, 2018]. A standard generative technique used is latent Dirichlet allocation [Blei et al, 2003], which assumes each document in a corpus of text is made up of a mixture of topics. LDA is typically used on longer documents to train a topic model and there may be a loss in performance observed by treating smaller texts (such as tweets) as separate documents. Thus, methods that attempt to group together similar tweets into pseudo-documents, "pooling" methods, have been shown to improve performance when compared with regular LDA [Mehrotra et al, 2013]. HTBSA incorporates the same measurement of what constitutes as a reliable indicator of a vote proposed by [Burnap et al, 2015]'s baseline LSA – number of positive tweets. The difference is that HTBSA incorporates the sentiment of each candidate's generated topics and topic proportions within texts to calculate tweet sentiment. Bansal's implementation of HTBSA used a newer form of topic modelling known as biterm modelling to generate topics candidate topics [Yan et al, 2013]. BTM produces word co-occurrences by generating pairs of words called "biterms". Unlike LDA, BTM uses a rich corpus that models documents as a bag of biterms rather than a list of words. In his baseline LSA proposal, Burnap stated that using the total magnitude of positive tweets (Lexicon-Positive Magnitude Analysis (LPM)) would lead to more accurate predictions rather than total positive tweet count (Lexicon-Positive Volume Analysis (LPV)). This was based upon the fact that when two parties have equal positive tweet counts, total magnitude would differentiate them. However, even when taking this into account, these claims are yet to be proven and because of this Bansal ran HTBSA experiments using both techniques, for which LPV analysis results performed slightly better, contradictory to Burnap's earlier statements. HTBSA was shown to outperform the baseline given by Burnap's method for tweets collected on the run up to the 2017 U.P. State Election in India (for both LPV and LPM analysis results).

Generally speaking, methods that combine both sentiment analysis and topic modelling in order to solve data mining problems are far from completely new; however, few studies have been successful in establishing methods within the realm of election analysis. Although there is a need for replicating Burnap's HTBSA results for elections happening in the future, HTBSA is a rare example of such a hybrid method. One other hybrid method that has been shown to perform successfully in election analysis problems was proposed by [Karami et al, 2018]. Karami produced positive and negative topics for candidate tweet sets collected on the run-up to the 2012 US Presidential Election (Obama and Romney). He filtered these topics in respect to the top Economic issues and the result was a ranking of the two candidates for each issue based on net number of positive topics. A criticism of Karami's approach other than the lack of replication across other datasets is that it is too simplistic, in particular the inability to produce percentile results to show each candidate's share of public opinion makes comparison to issue surveys and other methods a challenge. Karami did perform a form of election prediction by observing the overall results for each issue and using this to state which candidate would have an advantage based on the findings, but this incomparable to other election prediction methods as no technique to predict vote share is outlined. Although this method lacks sophistication, it was shown to perform well on the dataset used in Karami's study and remains the only known solution to performing candidate analysis on individual issues using Twitter data.

Taking all of the present literature on into account, more work needs to be done in order to establish hybrid methods as reliable techniques for analysing elections in future studies. With this being said, it is clear that better election predictions and performing

candidate analysis in respect to individual issues are two very useful functions in which future election mining studies should look to include. In this research we aim to build upon the work done by both [Bansal & Srivastava, 2018] and [Burnap et al, 2015], and we will develop our own novel election prediction technique. We will explore the implications of applying issue filtering proposed by [Karami et al, 2018] to the HTBSA technique in regard to improving the reliability and accuracy of lexicon-based approaches. The main contribution of our research approach directly follows on from one of [Gayo-Avello, 2012]'s main principles: clearly defining what constitutes as a "reliable indicator of a vote". Lexicon-based methods do not share the added utility given by hybrid methods of being able to do issue filtering, which enables us to define new measurements of public opinion in regard to specific subjects such as political issues. Furthermore, our research aims to extend the work done by [Karami et al, 2018] to potentially propose a baseline technique for performing candidate analysis on any political issue. We will (partially) address the common flaw made by past researchers by applying our approach to unseen data after it has been pre-configured for one dataset. The step that would follow this — to make a "real" prediction for an election occurring in the future [Gayo-Avello, 2012] — will be discussed in more detail during the final sections of the report. Although solving the substantial issues surrounding the use of Twitter data as a population sample is out of scope for this research, they are still very much valid problems within election prediction mining and their solutions may lead to major advancements within the field. Nevertheless, this study still has the potential to yield new lines of research that tackle election mining problems in different ways to what has been done previously and addressing the difficulties posed by Twitter data would only increase the fruitfulness of these lines of work, but such ideas need developing separately.

3. SPECIFICATION AND DESIGN

This section will outline the steps that will be taken to implement Hybrid Topic-Based Sentiment Analysis with Issue Filtering (HTBSA*). This will be performed on Twitter data collected on the run-up to the UK General Election 2019 to predict the vote share each of the main contesting candidates. Following this, we will apply an identical HTBSA* implementation to an unseen dataset, in order to objectively evaluate its performance to predict outcomes and analyse issues in elections occurring in the future. HTBSA* is an extension to [Bansal & Srivastava, 2018]'s HTBSA, which calculates vote share by counting the number of positive tweets in each candidate dataset – an established baseline approach first proposed by [Burnap et al, 2015] (as previously reviewed). HTBSA* calculates tweet sentiment in the same way as regular HTBSA, by calculating sentiment of candidate topics and then carrying this forward to calculate sentiment of tweets. The difference is that HTBSA* only carries forward topics that are related to key political issues, i.e. those that have the potential to depict election outcomes. We can also generate public opinion of candidates in respect to these issues individually using HTBSA*. This is achieved by counting the number of positive tweets for each candidate when we isolate only the topics that relate to a single one of these issues in our application of issue filtering (rather than all of them at once).

In order to evaluate the performance of HTBSA* we will perform comparisons to a number of standard and state-of-the-art approaches for election prediction: Volume Analysis, LSA [Burnap et al, 2015] and regular HTBSA [Bansal & Srivastava, 2018]. These will all be applied to our datasets to predict candidate vote share alongside HTBSA*. We will compute mean average error from true vote share to be used as a benchmark for comparison (MAE). We can justify the inclusion of Volume Analysis in our study as it simple to implement and provides a minimum comparison metric for our technique. Burnap's approach is included because it is the current state-of-the-art lexicon-based method and its inclusion is imperative in evaluating any novel approach such as ours. Furthermore, it is important to remember that HTBSA has not been applied to enough elections to be considered a baseline method for comparison like Burnap's LSA. Although it is still important to compare results given by regular HTBSA with our method, it also needs to be included as part of the complete evaluation of its ability to predict future elections — this study includes further verification of Burnap and Bansal's work as a side effect of their involvement in the project. This leads us on to an important assertion that must be made about a pseudo-limitation of our project design considerations: this research project does not aim to complete a *full* objective evaluation of HTBSA*'s performance to predict election outcomes in the future, nor does it aim to establish it as a new baseline method in the field (yet). The only way to achieve this would be to apply HTBSA* to multiple elections in the future and subsequently produce consistent results alongside this [Gayo-Avello, 2012]. Among our main objectives is to outline our novel approach so that it can be understood well and replicated in future studies. Alongside this, it is important that we evaluate its performance against current contemporary methods for an election it has been pre-configured for, as well as on unseen data. Although it would be extremely useful to demonstrate multiple applications of our approach across elections of the past, present and future as part of this project, we cannot predict an election occurring in the future until the aims of this research are met. Moreover, obtaining datasets is extremely difficult

(we had to write our own tweet collector) and a sizeable amount of the work required in this research has been in the selection and the application of the most up-to-date techniques for each step in our method (i.e. how to do sentiment analysis, topic modelling, etc). We see this as the next step which will be covered in full detail at the end of our project report, along with recommendations. We will also be evaluating HTBSA*'s effectiveness to perform candidate analysis of individual issues by comparing it with results obtained using the only other known method of this kind [Karami et al, 2018]. This will be the only known application of Karami's method to take place outside of his original study and will therefore contribute to the objective evaluation of this approach. In addition to this, we will compare the shares of public opinion generated for each candidate by HTBSA* with public surveys on different political issues, observing the MAE for each one.

We have touched on how HTBSA* is used to perform election prediction and candidate analysis of individual issues but there remains much more to examine on this subject. As well as needing to discuss how HTBSA* is an improvement upon previous lexicon-based techniques (in other words, explaining why it works effectively), we need to explain what steps need to be undertaken to implement it and therefore, regular HTBSA must be understood to a sufficient level. The rest of this section will include implementation details on HTBSA* and all of the other election analysis experiments we are performing in this project. Before we proceed, it must be noted that the purpose of this section of the report is to (comprehensively) explain what these techniques do to solve a given election analysis problem. The reader should not necessarily be concerned with how these can or will be implemented in our study as this information will be conveyed in the following section.

3.1. Volume Analysis:

Performing Volume Analysis is trivial. We calculate vote share by dividing the number of tweets for each candidate by the total number of tweets collected. This formula is outlined below:

$$VSx = \frac{TTVx}{\sum_{i=1}^{n} TTVi} \quad (1)$$

Where VSx is the vote share for candidate x, TTVx is the number of tweets collected for candidate x and n is the total number of candidates.

3.2. Lexicon-Based Sentiment Analysis [Burnap et al, 2015] – LSA:

As previously mentioned, we must understand regular HTBSA before we can begin to understand HTBSA*. Similarly, regular HTBSA is built upon baseline LSA principles first set out by [Burnap et al, 2015] and understanding this form of LSA is a prerequisite to knowing regular HTBSA. Burnap's main principle is the assumption that tweets containing a particular party or party leader's name with positive sentiment can be considered a reliable indicator of a vote.

The idea is to employ a sentiment analysis technique to calculate the "score" of a tweet. This score is the polarity component of the tweet's sentiment. Burnap also refers to the polarity as the "magnitude" of the tweet's sentiment. To calculate vote share using this methodology, you first need to calculate the combined scores of all of the tweets with positive magnitudes in each candidate dataset. Once this has been obtained, simply carry forward each candidate's total score and divide it by the combined score of all candidates to calculate vote share. This formula is denoted below:

$$VSx = \frac{TPMx}{\sum_{i=1}^{n} TPMi} \quad (2)$$

Where VSx is the vote share of candidate x, TPMx is the total positive magnitude of candidate x and n is the number of candidates. If we wanted to use the total positive volume of tweets rather than total positive magnitude to calculate candidate vote share, we could use the following formula:

$$VSx = \frac{TPVx}{\sum_{i=1}^{n} TPVi} \quad (3)$$

Where VSx is the vote share of candidate x, TPVx is the total volume of tweets related to candidate x with positive scores and n is the number of candidates. A final step can be performed to calculate the winner of each seat on a constituency-by-constituency level as explained in our review of the literature and it is worth noting that we will be ignoring this step in our research project — it would be very easy to extend our study to do this. Both HTBSA methods could also implement this step just as easily as any LSA method.

3.3. Hybrid Topic-Based Sentiment Analysis [Bansal & Srivastava, 2018] – HTBSA:

HTBSA relies on the aforementioned assumption used in [Burnap et al, 2015]'s baseline LSA that a positive tweet containing a party or party leader's name can be used as a reliable indicator of a vote. The difference is that positive tweets are classified differently to the sentence level sentiment classification used in LSA techniques. Tweets are collected for a given number of contesting parties then, if we want to classify positive texts for each candidate using HTBSA, we must first of all generate a set of topics for each dataset. Following on from our description of topic modelling given in our literature review, we may have topics for a given candidate that resemble a similar form to this:

Topic no.	Example Words
Topic 1	"leave", "remain", "brexit", …
Topic 2	"care", "nhs", "health", …
Topic 3	"vote", "labour", "election", …
Topic 4	"corbyn", "leader", "party", …
Topic 5	"referendum", "result", "respect", …

***Figure 1**: Example sample topics for a candidate dataset.*

Clearly, this could be the top three words for a set of five topics generated from tweets relating to the Labour Party leading into a recent UK General Election. In reality, the number of topics generated could be anything and will depend on the type of topic modelling technique used and decisions made by the implementor, which may be based upon requirements of the project or some heuristic(s) applied to the dataset. In addition to this, each topic is comprised of many words and in truth it is unlikely the top words will be so interpretable; however, it would not be unreasonable for a good topic model to produce topics with these terms in their top 10 or even top 5 list of words. Perhaps a more realistic example would look more like this:

Topic no.	Example Words
Topic 1	…, "leave" * 0.02, …, "remain" * 0.01, …, "brexit" * 0.01, …
Topic 2	…, "care" * 0.03, "nhs" * 0.02, …, "health" * 0.015, …
Topic 3	…, "vote" * 0.05, …, "labour" * 0.04, …, "election" * 0.03, …
Topic 4	…, "corbyn" * 0.05, …, "leader" * 0.05, …, "party" * 0.04, …
Topic 5	…, "referendum" * 0.01, …, "result" * 0.01, …, "respect" * 0.01, …

***Figure 2**: More realistic example sample topics for a candidate dataset.*

Where the "…"'s indicate a number of words which occur between the example terms in each topic and the "* w"'s indicate the amount of significance "w" a word has within a topic – it's weight. This is another feature of topic models and word weightings are used to determine the top words in each topic. Following on from this, once we have generated topics for a candidate, we would calculate the sentiment of a tweet in two steps: first calculate sentiment of topics, then using these and the weighted proportions of each topic in a tweet, sentiment of tweets. This process is outlined more clearly below, to calculate the sentiment of a topic:

$$STx = \sum_{i}^{n} Wi * SWi \quad (4)$$

Where STx is topic sentiment of topic x, Wi is the weight of word i and SWi is the polarity of word i. This formula takes the top n words from each topic. [Bansal & Srivastava, 2018] set this value to 20. Sentiment of words are calculated using a lexicon which contains sentiment mappings for many different words. The final result is a k-dimensional vector we will refer to as "K", where k is the number

of topics in a dataset and each dimension in K corresponds to a topic sentiment e.g. {ST1, ST2, … STn}. It is possible to run topic models on individual tweets to get the distribution of topic proportions within them. The result of this will be a k dimensional vector P(z|t) where each dimension corresponds to topic z posterior for tweet t e.g. {P(1|t), P(2|t), … P(n|t)}. Hence, to calculate the sentiment of a tweet we sum the products of each dimension i of K and P(z|t), for all topics k:

$$TweetSt = \sum_{i}^{k} Ki * P(i|t) \quad (5)$$

By calculating sentiment of tweets, you can proceed to calculate vote share via counting the total number of positive tweets for each candidate and dividing this by the overall number of positive tweets (equation 3). Otherwise you can choose to use total magnitude of positive tweets instead (equation 2).

After observing the example sample topics, it may seem that topics contain a lot of terms which have neutral sentiment polarities — such as nouns. It may often be the case that the top words in a topic contain words that have low polarities in either direction (positive/negative), therefore it is essential to use an appropriate lexicon. Lexicons are pre-trained on thousands of documents to give precise measurements of the overall sentiment of individual words occurring in a piece of text. The more similar the training set is to the type documents you want to examine in the future, the more accurate the lexicon will be at determining sentiment of the words occurring within those documents [Taboada et al, 2011]. Among this, a lot of other types of words will occur within topics which will have stronger sentiment polarity components, such as adjectives. These words will typically contribute more to the overall topic sentiment but it important to remember that this method takes the weightings of each word into account as well as their sentiment, this way the significance of words towards their topic is also incorporated. Furthermore, it is true that some topics may only be slightly positive or negative, but this is not necessarily a “bad” thing; similar to using weightings of words to calculate topic sentiments, the proportion of these topics within tweets are the most significant factor in determining the sentiment of a tweet. These proportions are calculated by taking all of the words within a tweet into account and the most prevalent topics across a candidate dataset will contribute more to determining overall vote share.

3.4 Sentiment Analysis with Topic Modelling and Issue Filtering [Karami et al, 2018]:

Before we can outline our novel approach, HTBSA*, there is one more technique which needs to be discussed, [Karami et al, 2018]’s hybrid issue filtering method for performing candidate analysis of individual issues. Of course, it is important for the reader of this project report to understand the steps required to implement this method because we will be performing it ourselves later in this research. The same also applies to the other methods explained up to this point; however, along with most of those methods, understanding this technique is also a fundamental pre-requisite to understanding HTBSA*. The notion of issue filtering and how it can be used to solve the problem of mining public opinion on individual issues is the final piece of knowledge that must be understood before we can discuss HTBSA*. [Karami et al, 2018]’s method is very simple to implement. The first step required is to perform sentiment analysis on each candidate dataset and split the tweets into positive and negative sets, discarding neutral tweets along on the way. Typically, this would be done by classifying tweets as positive or negative based on the polarity component of sentiment analysis results. After this, you should generate topics in both of these datasets for each candidate and choose a number of political issues to analyse. Karami decided to analyse the public opinion for contesting candidates in his study in respect to different economic issues. For reference, these were: Economy in General; Budget Deficit; Healthcare and Tax but could have been any arbitrary number of political issues. Once these have been decided you can manually inspect the topics produced in each candidate’s positive and negative datasets and filter them based on the chosen issues.

If a topic model is good enough, the majority of the top words occurring in each topic can easily be associated with a single theme or subject. In the case of such a politically oriented dataset, a significant number of topics will be about different subjects across the political spectrum and many of these subjects will be related to various issues in politics. The process of issue filtering involves associating topics that have been produced with different political issues. If you refer to our original example sample topics in figures 1 and 2, you can see that topics 1 and 5 can be associated with Brexit and topic 2 can be related to the National Health Service. We can attach a subject to each topic as shown:

Topic no.	Example Words	Subject
Topic 1	…, “leave” * 0.02, …, “remain” * 0.01, …, “brexit” * 0.01, …	Brexit
Topic 2	…, “care” * 0.03, “nhs” * 0.02, …, “health” * 0.015, …	Health Care/NHS
Topic 3	…, “vote” * 0.05, …, “labour” * 0.04, …, “election” * 0.03, …	Vote Labour in GE
Topic 4	…, “corbyn” * 0.05, …, “leader” * 0.05, …, “party” * 0.04, …	Labour Leader
Topic 5	…, “referendum” * 0.01, …, “result” * 0.01, …, “respect” * 0.01, …	Brexit

Figure 3*: Labelling example sample topics with a single subject.*

The final step to generating public opinion in respect to the chosen issues is performed by totalling up the number of positive and negative topics with labels for each issue and carrying forward the net positive number of topics for them, DPNT (difference between positive and negative topics). The final result is a ranking of the candidates by ascending DPNT, for each issue. The full pipeline can be observed below:

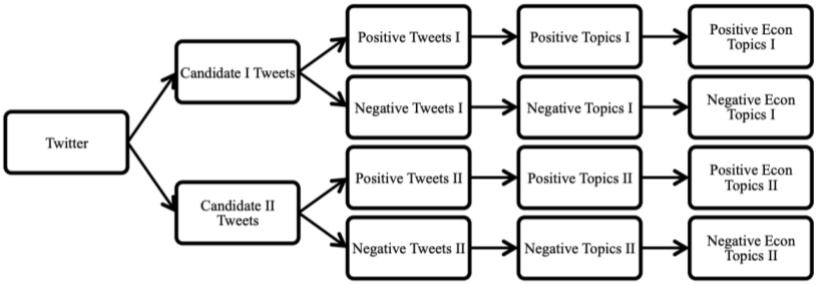


Figure 4: *Full pipeline of [Karami et al, 2018]’s approach to performing election analysis on economic issues.*

We have discussed mining methods that predict vote share and typically these techniques are compared with opinion polls before true vote share. Likewise, [Karami et al, 2018] compares the candidate rankings for each issue with public issue surveys and this should be recommended.

3.5. Hybrid Topic-Based Sentiment Analysis with Issue Filtering — HTBSA*:

Now that sufficient knowledge of regular HTBSA and issue filtering has been established, we can present the novel approach which this research proposes, Hybrid Topic-Based Sentiment Analysis with Issue Filtering. As well as outlining how regular HTBSA works, we have discussed how an innovative approach which also uses topic modelling and sentiment analysis to analyse public opinion on an individual issue level can be applied to Twitter data [Karami et al, 2018]. By filtering topics in HTBSA that only relate to a single issue, it is possible to perform HTBSA on one issue exclusively. The result of this would be the public opinion on each candidate's policies towards a given issue, according to the Twitter data. This is the basis of Hybrid Topic-Based Sentiment Analysis with Issue Filtering, HTBSA*. The steps can be seen below:

- *Step 1:* Select one or more political issues to filter
- *Step 2:* Generate topics for each candidate dataset and for each set of topics:
  - *Step 2.1:* Calculate topic sentiment vector using standard method given in regular HTBSA
  - *Step 2.2:* Manually inspect and filter topics by discarding all topics that are not associated with chosen issues
- *Step 3:* Calculate tweet sentiment for each candidate as follows:
  - *Step 3.1:* Calculate topic distribution for each tweet and filter topics as shown in previous step
  - *Step 3.2:* Normalise weighted distribution of topics so they sum to 1 (optional)
- *Step 3.3:* Calculate tweet sentiment by using standard method given in regular HTBSA but for filtered topics only

HTBSA* can be used to perform candidate analysis on any given number of issues. To allow a better understanding of how this process differs from regular HTBSA, we can refer back to our labelled example sample topics in figure 3:

Figure 5 (Below)*: Labelling example sample topics with a single subject*

Topic no.	Example Words	Subject
Topic 1	**…, "leave" * 0.02, …, "remain" * 0.01, …, "brexit" * 0.01, …**	**Brexit**
Topic 2	…, "care" * 0.03, "nhs" * 0.02, …, "health" * 0.015, …	Health Care/NHS
Topic 3	…, "vote" * 0.05, …, "labour" * 0.04, …, "election" * 0.03, …	Vote Labour
Topic 4	…, "corbyn" * 0.05, …, "leader" * 0.05, …, "party" * 0.04, …	Labour Leader
Topic 5	**…, "referendum" * 0.01, …, "result" * 0.01, …, "respect" * 0.01, …**	**Brexit**

Normally, regular HTBSA would calculate tweet sentiment in a dataset which generated such a list of topics by carrying forward a topic sentiment vector of the form {ST1, ST2, ST3, ST4, ST5} to calculate tweet sentiment — where all of the topics' sentiment values were included. Using HTBSA*, we could perform candidate analysis on the issue of Brexit alone by calculating the tweet sentiment vector based upon just the topics that were labelled by this issue. If we wanted to generate public opinion on Brexit for the example given above, we would only carry forward topics 1 and 5 when calculating tweet sentiment and our topic sentiment vector would {ST1, ST5}, where the dimensions are the topics in the dataset filtered by the sole issue of Brexit.

This alone may already be a valuable contribution to future studies that want to predict elections using an emerging lexicon-based hybrid technique such as HTBSA, but also want to generate public opinion in respect to individual issues to support prediction results and explain election outcomes in their analysis. Evaluation of results given by running HTBSA* on a single issue similar to the example given above can be made by direct comparison with public issue surveys because they are of a percentile form that is identical to vote share predictions given by regular HTBSA. In our study we will use this method of evaluation to get MAE of HTBSA*'s ability to generate public opinion in respect to various political issues on the run-up to an election and compare results with those given by the only other known method of this kind [Karami et al, 2018].

Taking this forward, no election prediction technique(s) to our knowledge have taken into account key political issues and used them to calculate vote share. Removing topics that have little relevance to major political issues (that would otherwise be included in regular HTBSA) is the extra step that we propose will further enhance the HTBSA prediction methodology and lead to more accurate results. Essentially, the idea is to perform HTBSA* in respect to a number of key issues which will assumingly strongly depict the result of an election. One may argue that this technique is exactly the same as trying to model a traditional issue survey which asks participants to select which party would handle a range of different issues best and calculating vote share by averaging the results for all of the issues (but using Twitter data instead). Although it wouldn't be unreasonable to compare this result with MAE of traditional opinion polls (see appendix), HTBSA* operates slightly differently. If we were trying to model the above, we would calculate a prediction by running separate HTBSA* experiments for each issue individually and averaging those results, rather than running a single experiment over all the topics that cover the issues. HTSBA* should be thought of as trying to model the results of a survey which asks participants to select which party would handle ALL of the issues best, which is significantly different.

This election prediction method has an advantage over regular HTBSA, rather than treating every topic generated in a dataset equally to calculate vote share, it recognises that topics that resemble key political issues contribute more in regard to voting intention. Furthermore, due to the nature of HTBSA, a topic that contains a lot of positive words (or negative) will often have a much larger sentiment score than others, these outliers are shown in an example below:

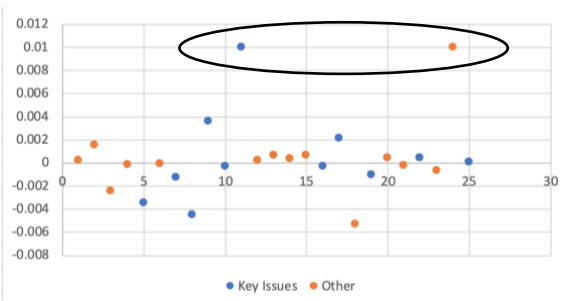


Figure 6 (Left)**: Example Topic Sentiment distribution for example dataset with 30 topics **(X-Axis = Topic; Y-Axis = Topic Sentiment)

Let's say in this example dataset we have chosen three key issues to filter our topics with. For the sake of clarity these issues are Economy, Healthcare and Brexit – these topics are blue dots in the scatter graph and all others are orange. Take the two topics circled in the graph, the sentiment scores of these topics are both unbalanced and will have a significantly larger effect on tweet sentiment scores than other topics. The blue topic includes positive terms about Economic policy

and the other topic can be grouped under the subject of a recent debate. Tweets that are heavily made up of either of these two topics will most likely be classified as positive and therefore count as a "vote" in HTBSA prediction; however, we strongly believe a positive opinion of a candidate's economic policies will be always (on average) be a more *reliable indicator of a vote* [Gayo-Avello, 2012]. Our explanation for this does not suggest that a tweet needs to be related to a specific political issue to be used as voting intention — it is based upon the idea that positive tweets about a candidate's policies towards a key issue would have a lower false positive rate. This hypothesis is grounded in the knowledge that key issues are those that are statistically proven to be important factors for members of the public when deciding on which party to vote for. By filtering our topics by as many of the top ones as possible, we lose any outlying topics that are unrelated to the most important political issues and add unwanted noise. This subtle difference in measuring intention to vote is what we propose will lead to a more accurate election prediction.

Of course, some issues will be more important for different groups of people, such as their demographics: geolocation, age, etc. The last section of this project report explores this idea further, but the purpose of our research is to concentrate on key issues which have the potential to depict an entire election outcome i.e. important issues for an entire population rather than different sub-groups. This study will be the first ever election prediction made using HTBSA* and as mentioned, we will later put forward more interesting ideas which can be developed from it and make use of HTBSA* to tackle other Twitter-based election analysis challenges.

4. APPLICATION OF APPROACH

Now that we have discussed what steps need to be taken to perform HTBSA* as well as other election mining techniques, it is time to outline how we will implement them during this research. The structure of the rest of this section will provide a comprehensive coverage of how numerous natural languages processing techniques have been executed in this study. Not only will we present our application of different data mining methods themselves (i.e. LSA, HTBSA, etc) but we will cover each of the discrete steps required to perform these methods, justifying the choices that were made at each stage in the process. Lastly, LPM analysis must be mentioned before we can proceed. In our literature review, we discussed mixed results to support [Burnap et al, 2015]'s hypothesis that Lexicon-Positive Volume Analysis (equation 2) was superior to Lexicon-Positive Magnitude Analysis (equation 3) and in light of this, we will run LPM as well as LPV analysis experiments to give two sets of results for all lexicon-based methods used in our project. Further evaluation of Burnap's statements about magnitude-versus-volume analysis can be considered a minor objective of this piece of work. Our configurations for various techniques will be built for data collected on the run-up to the UK General Election 2019. Following this section, we will display the results given by this set of experiments and apply an identical black-box implementation of the approach to an unseen dataset, which will serve as an objective evaluation of our methodology.

4.1. Data Collection:

We began our study by collecting tweets relating to the four main contesting UK political parties according to YouGov opinion polls as of 8th October [YouGov, 2019]. The Twitter Developer platform is a powerful tool which offers a number of different options for harvesting data from the social media site.[1] There are two APIs which allow us to collect tweets: the Twitter Streaming API and the Twitter Search API. The former allows us to select tweets occurring in real-time based on a given query whereas the latter allows us to retrieve a sample of tweets relating to the query that are most relevant. Along with this, the Search API allows us to assemble more bespoke queries that can be built upon numerous parameters such as geo-location, language and many more. Due to wanting to keep our dataset clear of irrelevant tweets, we decided to select this API for our task. We collected tweets relating to the Conservative, Labour, Liberal Democrat and Brexit parties using the Twitter Search API [Oikonomou and Tjortjis, 2018] and tweets were retrieved using a set of queries for each party. These included terms such as party names, party leader names, Twitter handles, etc. Below is the set of queries used for each party:

Party	Conservative	Labour	Liberal Democrat	Brexit Party
Queries	Conservatives, Conservative Party Tories Boris Johnson Johnson	Labour LabourParty Labour Party Jeremy Corbyn Corbyn	LibDem Liberal Democrats Lib Dems Jo Swinson Swinson	BrexitParty brexitparty_uk Brexit Party Nigel Farage Farage

Figure 7: *Queries used for candidates during data collection phase.*

[1] http://www.developer.twitter.com

The query sets were consistent for each party; the same number of queries were used as were the types of search terms that were included. We wrote a program which swept the tweets from the Twitter Search API every 15 minutes as per Twitter's rate limits and stored them in a relational database. The Twitter Search API returns a large amount of metadata for every tweet and only some of these were of interest to our project. Therefore, we stored the full text of each tweet and kept additional fields such as Tweet id, Twitter user id and the date and time the tweet was posted. Along with this we stored the name of the party from whose list of queries contained the search term that was used to find the tweet, in order to label candidate datasets. Only Tweets that were geo-tagged and posted from a location within the UK were added to the collection [Bansal & Srivastava, 2018]. We started to collect data on the 21st October 2019 and the election was officially confirmed on the 29th of the same month. Our strategy was to collect tweets on the run-up to the election up until the end of election day at midnight. We can justify this time period based on the fact that similar time periods were used in [Burnap et al, 2015] and Bansal's studies, and nearly all of the election prediction work covered in our literature review stopped data collection at the end of election day by similar convention or otherwise. On top of this, the time period enables us to capture public opinion for the entire duration of the candidate's election campaign.

4.2. Data Cleansing:

Before we can get any results or think about applying aforementioned techniques to our candidate datasets it is important to purify the data. This is done by removing elements from our corpora which are unrelated or have to potential to corrupt results. This is a separate step to what is typically known as data pre-processing or preparation which is discussed later. After harvesting 501,471 tweets over a total period of 53 days we followed a data cleansing process which resembled the baseline method for election prediction proposed by [Burnap et al, 2015]. Firstly, we removed Tweets from the collection of each party which included terms from any of the other party's query sets. This removed any duplicate tweets and also reduced the list of tweets to only those which were related to one party. This is important for producing accurate topic models and sentiment analysis results, specifically the latter because there is a danger of misidentifying sentiment in tweets that contain multiple parties (e.g. "I'm voting Conservative because I think Labour have leadership problems"). The total number of tweets after this step was 300,722.

The final step in the cleansing process involved calculating the proportion of tweets in candidate sets that were actually related to that candidate's party. An example of irrelevant tweets that may exist in our datasets would be those collected for Labour that were about the word relating to work or the final stages of pregnancy rather than the political party. This case was common in [Burnap et al, 2015]'s study and had to be taken into account by sampling tweets for each party and labelling ones which were false positives. The proportion of tweets that were related to the candidate was then carried forward and used to scale results later on. We followed the same approach and randomly sampled 1000 tweets from each candidate dataset. We then manually inspected each sample of tweets and observed that only a small number were unrelated to the political candidates. The maximum number of false positives was 25 (2.5%) for the Brexit Party and the rest of the parties fell around 1-2%. We can conclude that the decision to take geographic location into account (and using a low number of queries) was largely responsible for this low percentile (Burnap's study ignored this step) and considering the variation of false positives between each party was fairly negligible, we decided to ignore these results for the rest of the study.

4.3. Volume Analysis:

With the data collection and cleansing phases complete, we now have our final dataset. This equates to four distinct sets of tweets, labelled for each of our candidates. We have reached the stage where we can start to apply different data mining techniques in order to make election predictions and perform individual issue analysis. The first and most simple of our election prediction techniques which will be compared with the performance of HTBSA* is Volume Analysis. Performing Volume Analysis to predict vote share is trivial. We take the number of tweets for each candidate and divide this by the total number of tweets to calculate vote share (equation 1).

4.4. Sentiment Analysis:

With Volume Analysis out of the way, we still need to apply a few more data mining techniques to our dataset as a part of our research. Along with needing to execute HTBSA* to predict vote share and analyse individual issues, we still need to perform baseline LSA, regular HTBSA and [Karami et al, 2018]'s hybrid issue filtering method to be compared with the results of our novel method.

Having read the project report up until this point, the reader should be familiar with the role that these different methods will play in our research and also the implementation steps they require. Nevertheless, for the purposes of simplifying these terms that will be repeated throughout this section, we will refer to these methods as HTBSA*, LSA, HTBSA and Hybrid Issue Filtering (HIF) from now on.

One feature that all of these techniques have in common is that they all rely on sentiment analysis at some point, albeit on different levels (and in different ways), in order to be performed. Although a machine learning approach is considered to be the most accurate way of performing sentiment analysis, we decided to apply a lexicon-based approach to find positive, negative and neutral tweets. This was mainly due to the practicalities of the project. For example, acquiring a dataset that can be used to train a machine learning classifier requires a lot of effort and we do not have such prior knowledge readily available on our dataset. Furthermore, building a pipeline based on a machine learning function that incorporates topic models to classify tweet sentiment could stand as an entirely independent project and is therefore not in scope. We can break down the usage(s) of sentiment analysis into two categories for our experiments: the first way we will need to implement it will be to classify tweets as positive, neutral or negative. This is done in LSA and HIF. In LSA, all positive tweets will be used in calculating candidate vote share (equations 2 & 3) and in HIF we need to split our candidate datasets into positive and negative tweet sets (discarding neutrals). Based on the identicality of the way sentiment analysis is executed in these two methods, we can refer to this function of it as "sentiment classification". The following passage will explain how we intend to implement this in our research. This leaves us with remaining methods HTBSA and HTBSA* that fall out of this category. This is because these two methods classify tweets as positive, negative and neutral based upon topic proportions within tweets and topic sentiments, rather than the sentiment of tweets alone. This can only be achieved after topic sets have been generated for each candidate and in light of this, we will discuss how this is done after the topic modelling chunk of this section.

4.4.1. Sentiment Classification:

We will be mostly be following [Burnap et al, 2015]'s original approach to classifying tweet sentiment. Burnap's LSA uses SentiStrength[1] [Thelwall et al, 2010] to analyse tweet sentiment, a lexicon-based classifier designed for analysing short pieces of text. We can justify the use of this classifier in our experiment because in a state-of-the-art TSA comparison done by [Zimbra et al, 2018] it was shown that SentiStrength was among the top general-purpose classifiers as well as one of the best for academic research with a focus on Twitter-Based Sentiment Analysis. As mentioned above, the sentiment classification component of HIF is to split the tweets into positive and negative sets (figure 4). The original implementor of HIF used Linguistic Inquiry and Word Count (LIWC) to calculate sentiment [Karami et al, 2018; Pennebaker et al, 2007]. LIWC does not understand irony, sarcasm or metaphors [LIWC, 2015] which can lead to misclassifications and similarly SentiStrength has been shown to underperform for text taken from news and politics against other data sets due to the expressive nature of the language used in these sources [Thelwall, Unknown]. We can enrich our course of action to classify tweet sentiment by adding an additional step taken by a paper that analysed climate change tweets in 2019 [Dahal et al, 2019]: using the SentiStrength 'scale' classifier which gives a score between -4 and 4, we can adjust the threshold used to classify tweets. The default approach is to classify tweets with scores greater than zero as positive and vice-versa for negative, however, adjusting this threshold will affect precision and recall values, which are metrics that represent a classification model's ability to minimise false positives and false negatives respectively [Manning et al, 2008]. We first of all took a random sample of 500 tweets and manually labelled them as positive, negative and neutral. We then ran the scale classifier over each tweet in our sample set and scaled scores down to fall into the range of -1 and 1. In other words our scores fell into absolute values from 0 with increments of 0.25, up until 1. By comparing sentiment classifications with our labels for thresholds 0, 0.25, 0.5 and 0.75 we observed that, based on precision and recall values (and not wanting to discard too many tweets classified as neutral), 0 gave us the best results. See the appendix section for a full set of results for these experiments and also how precision and recall values were calculated.

After deciding on how we would classify tweets as positive and negative, we were ready to perform LSA by carrying forward tweets with polarities greater than zero to calculate vote share (equations 2 & 3). We were also able to split candidate tweets into positive and negative datasets as per the first step in HIF by using the same threshold (figure 4).

4.5. Topic Modelling:

Having already performed Volume Analysis and LSA successfully to calculate vote share on our dataset (results to be shown in following section), the remaining experiments that are a part of our research, HTBSA, HTBSA* and HIF, all rely on topic modelling at

---

1 http://www.sentistrength.wlv.ac.uk

some point. As we will be implementing all of these approaches in our study, we will be using topic modelling in two ways: to generate a) the overall topics in each candidate's tweets to perform HTBSA/HTBSA* and calculate vote share/public opinion in respect to political issues and b) the positive and negative topics in each candidate's positive and negative datasets after sentiment classification, to be filtered based on political issues in HIF. Before we can do either of those operations, we need to select an appropriate topic modelling methodology.

To determine which topic modelling technique would be best to use throughout our study we will compare four different implementations, three of these are variations of latent Dirichlet allocation (LDA) [Blei et al, 2003] and the other is Biterm Topic Modelling (BTM) [Yan et al, 2013]. The steps required to perform these methods are very similar and all require the number of topics k to be provided in advance. It is worth noting at this stage that there is a type of topic modelling known as Hierarchical Dirichlet Process (HDP) [Teh et al, 2006] which does not require the number of topics to be specified in advance; however, it is more complicated to implement. In our case we decided to run multiple models for a different number of topics to select a best fit.

We compared each model's performance against one another based on a metric known as CV coherence, which detects the quality of the topic models produced by evaluating the semantic similarity of the words in each topic. See [Röder et al, 2015] for a definition and explanation of how this value is calculated.

4.5.1. Data Pre-Processing:

To generate accurate topic models, we need to remove undesirable features from our data. There are some features that are native to our dataset, in other words, Twitter specific — such as handles and hashtags. Previous researchers have opted to either get rid of these features all together [Karami et. al 2018] or include them both in topic models [Dahal et. al 2019, Basal et. al 2019]. Although Twitter handles can be used to reference a subject that can provide useful information to our models (party leaders, MPs, party accounts etc), the majority of these handles are found at the start of tweets which are responses to other users. For this reason, handles are removed entirely from the dataset and coming up with sophisticated methods which attempt to selectively remove undesired handles is out of scope. Although the original implementation of HTBSA does not remove handles or hashtags, it can be argued that this is a big mistake due to the fact that the overwhelming majority of these strings (especially those without underscores) cannot be parsed as single words using trivial methods. Results of HTBSA significantly rely on calculating sentiment of the top words in each topic using a lexicon (dictionaries which map single words to sentiment values), hence most strings used in handles/hashtags would not be mapped to anything other than zero and should be replaced in topics. Hashtags always provide information to our models because they are a part of the speech which is deliberately added by the user (unlike handles which are added automatically to replies). However, splitting these compounds perfectly is a challenge and also out of scope, for this reason we split the hashtags that use title case (e.g. #LikeThisExample) into individual words by the location of capital letters but leave strings that have less than 4 characters as these are usually acronyms (e.g. UK, USA, etc). Initially the text component of all hashtags is kept, and our first pre-processing steps involve removing handles, punctuation, emojis, URLs, processing hashtags by means described above and converting all text to lowercase.

Lastly, we tokenise the tweets, remove "stopwords" (words that appear very frequently in a language and add no real value to models) and apply lemmatisation after part-of-speech tagging (the vast majority of undesirable hashtags that remain are lost here, as only words with noun/verb/adjective/adverb POS tags and lemmas are kept). We decided to use lemmatisation over stemming — two NLP techniques which reduce words to their "root" form — because stemming (which was used in the original HTBSA) can produce roots that aren't real words, therefore leading to the exact same aforementioned problem with handles and hashtags when applying lexicon-based methods later. These steps leave us with a significantly reduced set of words to increase the accuracy of our topic models.

4.5.2. Model Selection:

We applied the Mallet implementation of LDA [McCallum, 2002] with default parameters on each candidate dataset for three different LDA implementations (regular, hashtag-pooling and author-pooling). For each of these methods we ran LDA on the datasets for different numbers of topics k, starting with 5 up until 25 with incremental steps of 5 in-between and also for k = 50 and k = 100. In the end we ran this process three times for each LDA method and for each candidate dataset we were left with average CV coherence values for the different values of k. Regular LDA can be performed relatively straightforwardly on our dataset. We simply produce a

corpus from the original set of documents and run LDA for different values of k. This leaves us with the number of documents being the same as the number of tweets we collected (300,722). Author-Pooled LDA groups together tweets that were posted by the same user into single documents. The number of documents in the resulting corpus was reduced to 137,922. The final LDA method we compared was Hashtag-Pooled LDA. This involves examining the n hashtags present in each of the tweets and adding the tweet to n different documents for each hashtag. The result is a set of documents which individually represent tweets that share the same hashtag. Tweets that do not contain any hashtags are added as separate documents. The number of documents in the set decreased to 275,620.

For BTM, we ran an implementation written by the author of BTM[1] to run our topic modelling experiment across each of our datasets. Like LDA we needed to supply the number of topics in advance and also some hyperparameters. Following [Yan et al, 2013]'s suggestions, we supplied an α value of 50/k and after running some tests we found that 0.001 was the optimum value for β. We ran 1000 iterations with a step size of 2 and below are the average coherences for the different topic modelling techniques:

Topic Modelling Technique	Average CV Coherence
Regular LDA	0.381
Author-Pooled LDA	**0.473**
Hashtag-Pooled LDA	0.468
BTM	0.395

Figure 8: *Performance of different topic modelling techniques on our dataset measured by average CV coherence for different values of k.*

From our analysis we observed that the best performing topic modelling techniques on our dataset were Author-Pooled and Hashtag-Pooled LDA. However, considering only a small proportion of the tweets in our dataset actually contained hashtags (~50,000 or 17%) and therefore, most of the corpus produced by Hashtag-Pooled LDA would be identical to regular LDA (abandoning all other tweets is not an option), we decided to select Author-Pooled LDA as our preferred topic modelling technique. After this stage was complete, we were ready to start producing topic models for each candidate dataset in order to apply HTBSA and HTBSA* using Author-Pooled LDA. We were also ready to do the same in positive and negative candidate corpora to execute HIF. To calculate the number of topics 'k' for a candidate dataset or positive/negative candidate dataset, we generated topic models for many different values of k and took the CV coherence values of each model. We then plotted a graph showing topic coherence against k and selected the best value according to this trade-off [Prabhakaran, 2018] (see appendix). For the latter parts of our experiment (where we perform issue filtering) it suited us to have a large number of topics, so we also took this into consideration when selecting k for our different datasets (note that HTBSA/HTBSA* results are not directly affected by our choice(s) of k, but HIF results are).

4.6. Topic Sentiment:

For the sake of clarity, we will recap what has been implemented so far. We have been able to fully apply Volume Analysis and LSA to our dataset in order to predict vote share for the main four contesting parties in the 2019 UK General Election. Now, we have reached the stage where we have topic models for our candidate datasets which will allow us to perform HTBSA to predict vote share. Along with this, we have produced positive and negative topics for our candidates, and we will be able to perform HIF to analyse public opinion on candidates towards key political issues. The results of all of these various techniques will be compared with the performance of HTBSA*, which can be abstracted away from election predictions to analyse individual issues as well. In this passage we will denote how we were able to calculate topic sentiments in candidate datasets, a required step in implementing HTBSA and HTBSA*. By the end of this section we will have fully applied HTBSA to our dataset and in the next section, we will select a number of key issues to apply to the remaining steps of HTBSA* and HIF.

To calculate the sentiment of a topic, we firstly extract the top 20 words [Bansal & Srivastava, 2018]. From these words, we calculate the sum of the sentiment of each word in the topic multiplied by its weight (equation 4). After doing this for all topics we have a k-dimensional vector which corresponds to the sentiment of each topic. Word sentiments are calculated using the lexicon Sentiwordnet [Baccianella, Esuli & Sebastiani, 2010], which was found to be the best performing lexicon not just in Bansal's paper but also for microblog posts in general during a comparison study by [Musto et al, 2014] in 2014. We took the polarity component of the sentiment mappings held in the lexicon for each of the words in a topic (if multiple mappings existed for a word then we took the average polarity of all mappings (alternative solution would be to embed POS tags in topic models)) and these were used as their sentiment.

[1] http://www.github.com/xiaohuiyan/BTM

Using this method, we were able to calculate the topic sentiments of all candidate topics. As explained in the previous section, tweet sentiments for a candidate were calculated by running the candidate's topic model against the tweet to calculate topic proportions within it, and then summing the products of topic sentiments and topic proportions (equation 5). All positive tweets were used to calculate HTBSA vote share in the same way previously described in the last section of this report (equations 2 & 3).

4.7. Issue Filtering:

We have now applied Volume Analysis, LSA and HTBSA on our dataset to predict vote share of each candidate. We will include the results of these predictions in the following section along with their MAEs from true vote share, which will be a means of comparison with our novel technique HTBSA*. In order to be able to apply HTBSA* on candidate tweets to predict vote share for the 2019 UK General Election, we need to select a number of key issues in UK politics which we will use to filter candidate topics. The reader of this project report should of course know by now, that if we execute HTBSA* on candidate tweets and isolate topics that are filtered based on a single issue alone, we can generate public opinion in respect to this issue according to our data. On top of our election prediction evaluation, we will compare results given for the individual issue analysis with public surveys and results given by HIF, the only other known method that can perform this type of election analysis.

To complete our application of HIF on our dataset, the only remaining step is to filter the positive and negative topic sets for each candidate in respect to our chosen issues (figure 4). We will be performing separate HTBSA* experiments for five of the top seven issues in UK politics according to [YouGov, 2019]. These statistics were taken a week prior to the election and the issues were NHS, Asylum/Immigration, Law & Order, Economy in General and Brexit.

All of the issues chosen in our experiment have been statistically proven to be important factors for members of the public when deciding on which party to vote for [Ipsos MORI, 2019]. We will also filter topics based on these issues in our application of HIF and last of all, we will perform one more HTBSA* experiment which will filter all of the topics at once rather than individually to predict vote share. Another reason these issues were chosen in particular was that [YouGov, 2019] also ran a survey to find out the best parties facing a range of issues, according to public opinion, and these issues were the most significant ones (according to YouGov) that were also available on the survey. This survey will be used as a direct comparison to evaluate HTBSA*'s performance for each issue. After generating topics using Author-Pooled LDA in our execution of HTBSA* for each of our candidate datasets, we discovered a total of 243 issue-related topics across all of our candidate datasets. The full breakdown of these topics can be found below, along with some real examples of words occurring in topics filtered by our issues:

Issue	Conservatives	Labour	Liberal Democrats	Brexit Party
NHS	13	8	3	2
Asylum/Immigration	2	4	0	2
Law & Order	7	6	5	3
Economy in General	31	53	8	14
Brexit	22	13	13	34
Total	**75**	**84**	**29**	**55**

Figure 9a: *Number of topics filtered by key issues for each candidate.*

Example Issue	Conservatives	Labour	Liberal Democrats	Brexit Party
NHS	sell nhs buy trump	doctor nurse nhs staff	protect abstain nhs sell	doctor train work nhs
Asylum/ Immigration	immigrant country citizen migrant	cheap wage worker migrant	N/A	country immigrant free immigration
Law & Order	crime police criminal knife	force police charge cut	electoral law reform join	law set pass legal
Economy in General	financial global crash austerity	economic economy policy investment	policy economic spend economy	fish fishing industry water
Brexit	leave remainer remain vote	results referendum respect leave	majority libdem remain remainer	brexit shamble leave pro

Figure 9b: *Real examples of words occurring in a single topic filtered by an issue, for each candidate.*

Topics were filtered by carefully inspecting the top 5-10 words within them in order to associate them with a single subject; if this subject clearly and indisputably fell underneath an issue, it was labelled by it. Figure 9b demonstrates words from topics that were labelled with our chosen issues for each candidate dataset: e.g. Conservative NHS topic about the NHS being on the table during post-Brexit trade negotiations with the USA. After we had filtered the topics in candidate datasets, we were able to execute HTBSA* to predict vote share and perform individual issue analysis using the implementation details outlined in the previous section. One last measure must be taken before we have finished applying all of our methods to our dataset, the issue filtering required to execute HIF in order to perform candidate analysis on our issues. We can do this in an identical fashion to how we applied issue filtering in HTBSA* but for positive/negative tweet sets as opposed to our original candidate datasets (figure 4). See appendix for a breakdown of the topics across each issue, their values are directly integrated within the issue analysis results given by HIF (net positive topics), which are included in the next section.

We have now explained the exact steps which were undertaken in this research to perform our novel approach to mining Twitter data for election prediction and individual issue analysis, Hybrid Topic-Based Sentiment Analysis with Issue Filtering. As well as this, we have carried out the same process for other methods that are used in this study for the evaluation of its performance on the 2019 UK General Election, these are Volume Analysis, baseline LSA, regular HTBSA and Hybrid Issue Filtering. Throughout this section we have considered the different options that were available to us, based on the most recent research techniques and the requirements of the project, and we have had to justify all of the choices we have made throughout. Without further or do, we will now progress to the next section of this report which will include the results of our application of the aforementioned techniques, as well as discussion of their significance.

5. RESULTS AND DISCUSSION

In this study we performed two separate experiments to evaluate our novel approach. Firstly, one which evaluates HTBSA* 's ability to analyse public opinion in respect to key political issues, by method of comparison to issue surveys and the only other known method of this kind. The second experiment compares election prediction results given by HTBSA* with true vote share and other results given by baseline and emerging prediction techniques. We configured the approaches during our running of the experiments on data collected for contesting candidates on the run-up to the UK General Election 2019. In addition to this, we performed the experiments with an identical configuration on an unseen dataset to objectively evaluate the methodology used in our report. This dataset was a set of tweets for the UK General Election 2017 — results for this set of experiments will be discussed at the very end of this section.

5.1. Lexicon-Positive Magnitude Analysis vs Lexicon-Positive Volume Analysis (2019 Election):

Before we discuss the results for both candidate analysis of individual issues and election prediction made by HTBSA*, it is worth noting that other than raw volume of tweets, all of our experiments have been run with both LPV and LPM analysis methods of calculating vote share (equations 2 & 3). LPM analysis was better at ranking candidates when performing HTBSA* to analyse public opinion on key issues; however, by all other metrics, LPV analysis outperformed LPM analysis so significantly such that MAE results given by LPM analysis were incomparable with traditional issue surveys and opinion polls. For this reason, we will display single results for LSA, HTBSA and HTBSA* which will represent the results for LPV analysis only, and all lexicon-based results mentioned in subsequent sections should be assumed to refer to their LPV implementations unless otherwise stated. A comprehensive list of all results in full can be found in the appendix section of this project report, which also includes the performance of LPM analysis when applied within the different methods on our dataset. The significance of this observation will be discussed as part of the evaluation section.

5.2. Candidate Analysis for Individual Issues (2019 Election):

We compared the results of HTBSA* directly to issue surveys and two metrics were used to evaluate its performance. These metrics were MAE for each run of HSBTA* with the public opinion survey for a specific issue and the overall rankings of the candidates in both of these. We also performed the only other known method for analysing public opinion on individual issues, HIF [Karami et al, 2018], which gave candidate rankings for each of the issues based on net positive topics found in positive and negative sets of candidate tweets. We were then able to compare these orderings with the surveys and HTBSA* results.

It must be noted that Karami's research worked with two candidate datasets that were Obama and Romney, whereas here we have four. The significance of this is that predicting the correct order of four candidates is more difficult than two and therefore analysing these results can only tell you so much. Furthermore, going off candidate ordering results alone can be misleading if two parties are closely matched in opinion polls and the party rankings should only be used to supplement evaluation done by observing MAE from traditional surveys. This is ultimately the biggest downfall of Karami's technique because it becomes a challenge to objectively evaluate an election prediction method which is not able to generate percentile results — especially when the outcomes are non-binary. In regard to predicting the top party for each issue in our study, Karami's method correctly predicted 1/5 outcomes (Brexit) when compared to the YouGov survey. HTBSA* correctly predicted 2/5 of these outcomes (NHS and Brexit).

Overall, Karami's method predicted 5/20 rankings (25%) across all the issue surveys and didn't flawlessly predict the public opinion rankings of any issues. HTBSA* predicted 10/20 (50%) rankings and flawlessly predicted the public opinion rankings of candidates in respect to the issue of Brexit. These results along with the MAE from the YouGov survey for public opinion on each issue generated by HTBSA* can be found below:

Issue	Advantage, HIF [Karami et al, 2018]	Advantage, HTBSA*	Advantage, Public Survey [YouGov, 2019]
NHS	Brexit, Labour, LibDem	Labour	Labour
Asylum/Immigration	Labour	Labour	Conservative
Law & Order	Labour, LibDem	Labour	Conservative
Economy in General	LibDem	Labour	Conservative
Brexit	Conservative	Conservative	Conservative

Figure 10a (Above): *Top performing candidates for each issue shown for [Karami et al, 2018]'s HIF method and HTBSA*, compared with [YouGov, 2019] survey results (2019 election).*

Issue	Correct out of 4, HIF [Karami et al, 2018]	Correct out of 4, HTBSA*
NHS	0	2
Asylum/Immigration	0	2
Law & Order	2	2
Economy in General	1	0
Brexit	2	4

Figure 10b (Above): *Number of correct rankings generated for each issue for HIF and HTBSA* when compared with [YouGov, 2019] Survey (2019 election).*

Figure 10c (Below): *MAE between HTBSA* results and the YouGov issue survey for our chosen issues. Average MAE for all issues and also for just NHS, Economy and Brexit only also shown (2019 election).*

Issue	HTBSA* MAE
NHS	7.97%
Asylum/Immigration	17.86%
Law & Order	13.46%
Economy in General	9.29%
Brexit	10.43%
Average	11.80%
Average (NHS, Econ, Brexit)	**9.23%**

As you can see from figure 10c, the MAEs for Asylum/Immigration and Law & Order issues were significantly worse than the others. This can be attributed to the fact that only 6 topics were found relating to the issue of Assylum/Immigration across the whole dataset (this was the worst performing issue) and the number for Law & Order was also low at 21. If we refer to figure 9a, we can also observe that the number of topics found for the other issues was significantly larger (especially for the largest datasets — Labour and Conservative). We will evaluate this discovery along with all of the results in more detail later.

5.3. Election Prediction (2019 Election):

Due to findings in the above section, we have included two results columns for HTBSA* when comparing it to other election prediction techniques: one which incorporates the topics filtered for all key issues and another which doesn't take topics associated with

Asylum/Immigration and Law & Order into account.

In both of these cases HTBSA* was the best performing election prediction methodology, with a slight improvement observed when all of the key issues were included rather than the reduced set. None of the election prediction methods ranked the candidates in the correct order. Before we evaluate the results, they can be found in full in the figure below:

Party	True Vote Share	Volume Analysis	Baseline LSA [Burnap et al, 2015]	HTBSA [Bansal & Srivastava, 2015]	HTBSA* (All Issues)	HTBSA* (NHS, Econ, Brxt)
Conservatives	43.60%	38.40%	35.99%	29.85%	38.51%	37.40%
Labour	32.20%	42.50%	40.81%	50.79%	44.13%	44.59%
Liberal Democrats	11.50%	4.60%	5.52%	6.17%	5.39%	5.70%
Brexit Party	2.00%	14.50%	17.68%	13.19%	11.97%	12.30%
MAE	**N/A**	**8.73%**	**8.86%**	**12.21%**	**8.27%**	**8.67%**

Figure 11: *MAE for exiting baseline, state-of-the-art and emerging lexicon-based election prediction methods, as well as those for HTSBA* when compared to true vote share (2019 election).*

5.4. Unseen Dataset (2017 Election):

This dataset was selected as it was the only available alternative collection of election tweets that we were able to find online[1], the fact that it is the preceding UK General Election is a mere coincidence. We applied an identical configuration of all methods used in our original experiments after the data collection phase, as a means of objectively evaluating our approach. The implementation of HTBSA* in particular served as a true black-box test where we could feed in an unseen dataset as an input and observe the results that were given. Details on the exact configuration(s) used on this dataset can be found in the appendix (with a full set of results) but for all intents and purposes, the steps taken to undergo our experiments on this dataset were identical to that of the original ones (only the data collection strategy differed (obviously)). For clarification, the candidates we analysed for the 2017 election were identical to 2019 with one exception being that the Brexit Party was replaced by UKIP. The most important results of this test are included in this report.

5.4.1. Lexicon-Positive Magnitude Analysis vs Lexicon-Positive Volume Analysis:

In our demonstration of the results for the UK General Election 2019 given by lexicon-based techniques, LPM analysis performed poorly in all but every aspect when compared to LPV analysis. The same outcome was observed when comparing LPV and LPM results given by lexicon-based methods on the unseen dataset; with the exception that this time around the candidate orderings for issue analysis results given by LPM HTBSA* were not superior to those given by LPV HTBSA*. We will evaluate/conclude the Lexicon-Positive magnitude-versus-volume component of this research in the final few sections but again, all results given by lexicon-based methods that follow in this section should be assumed to be referring to their LPV analysis counterpart.

5.4.2. Candidate Analysis of Individual Issues:

The top issues for the UK General Election 2017 were identical to the previous dataset with the only change being that the issue of Defence/Security was more important than Law & Order [YouGov, 2019]. For reference, the breakdown of topics across candidate datasets are given in the following figure:

Issue	Conservative	Labour	Liberal Democrat	UKIP
Brexit	17	6	16	6
NHS	3	3	1	0
Economy in General	21	28	8	2
Asylum/Immigration	2	3	0	1
Defence & Security	20	8	3	4
Total	**63**	**48**	**28**	**13**

Figure 12: *Candidate topic/issue breakdown for top parties/issues in 2017 election.*

[1] https://data.world/wwymak/uk-election-tweets-2017-june-1

Poor results were observed universally for HTBSA* performed in respect to all of the issues in terms of MAE from surveying data given by [YouGov, 2019], and a decrease in the ability of HBTSA* to rank candidates for these issues was also observed. [Karami et al, 2018]'s technique performed slightly worse than HTBSA* on this dataset. Results for candidate analysis of individual issues given by both methods are included below:

Issue	Advantage, HIF [Karami et al, 2018]	Advantage (HTBSA*)	Advantage (YouGov Survey)
Brexit	Conservative	Labour	Conservative
NHS	Conservative	Labour	Labour
Economy in General	Conservative	Labour	Conservative
Asylum/Immigration	UKIP	Conservative	Conservative
Defence & Security	UKIP	Conservative	Conservative

Figure 13a: *Top performing candidates for each issue shown for [Karami et al, 2018]'s method and HTBSA*, compared with [YouGov, 2019] survey results (2017 election).*

Issue	Correct out of 4, HIF [Karami et al, 2018]	Correct out of 4, HTBSA*
Brexit	1	2
NHS	0	2
Economy in General	2	0
Asylum/Immigration	2	2
Defence & Security	0	2

Figure 13b: *Number of correct rankings generated for each issue, for HIF and HTBSA*, compared with [YouGov, 2019] survey (2017 election).*

Average MAE	21.08%
Average MAE (Brxt, Econ, Def)	20.32%

Figure 13c: *Average MAE for all issues given by HTBSA* and also for just Brexit, Economy in General and Defence issues only (2017 election).*

Karami's method ranked the candidates correctly only 25% of the time (identical to 2019 result) and predicted the top candidate for 2/5 issues (slight improvement from 1/5 in 2019 result). HTBSA* correctly generated 45% of candidate rankings (slight decrease from 50% in 2019 result) and predicted the top candidate for 3/5 issues (improvement from 2/5 in 2019 result). As explained previously, MAE from issue surveying data is the most useful metric of comparison; therefore, these outcomes cannot be considered a strong result. For the issue analysis results in the previous dataset, we were able to justify issues for which HTBSA* gave a poor MAE by attributing this to an insignificant number of topics being available; however, even for issues when this wasn't the case the average MAE given in the 2017 dataset was weak (see appendix). For these issues, it was often the case that one of the top parties (Tory/Labour) had an unusually low share of public opinion, which can be attributed to a suspected high prevalence of negative topics amongst candidate tweets.

5.4.3. Election Prediction:

Party	True Vote Share	Volume Analysis	Baseline LSA [Burnap et al, 2015]	HTBSA [Bansal & Srivastava, 2015]	HTBSA* (All Issues)	HTBSA* (Brxt, Econ, Defence)
Conservatives	42.30%	45.75%	36.57%	61.53%	41.03%	41.41%
Labour	40.00%	43.98%	52.59%	23.67%	43.39%	38.07%
Liberal Democrats	7.40%	5.72%	6.76%	2.88%	8.55%	8.60%
UKIP	1.80%	4.55%	4.08%	11.92%	6.55%	6.58%
MAE	**N/A**	**2.96%**	**5.31%**	**12.55%**	**2.64%**	**2.20%**

Figure 14: *MAE for exiting baseline, state-of-the-art and emerging lexicon-based election prediction methods, as well as those for HTSBA*, when compared to true vote share (2017 election).*

HTBSA* produced the strongest MAE from true vote share for the 2017 election Twitter dataset. Yet again, results given by Volume Analysis and [Burnap et al, 2015]'s LSA were also strong and candidate predictions given by regular HTBSA were incomparable to opinion polls. Out of all the techniques, only LSA and HTBSA* over the full issue set did not predict the correct candidate rankings, suggesting Volume Analysis and HTBSA* over the reduced set gave the best results.

6. EVALUATION

6.1. Lexicon-Positive Magnitude Analysis vs Lexicon-Positive Volume Analysis:

[Burnap et al, 2015]'s hypothesis was that LPM analysis should be preferred when making predictions as it can differentiate two candidates with an equal Lexicon-Positive Volume of tweets. When this scenario was closest to occurring – for example a few times in 2019 individual issue analysis results – LPM analysis sometimes did do a solid job of differentiating the candidates with similar vote shares so that they ended up in the correct order. However, it must also be noted that this was not a consistent observation across all of these cases (see appendix) and Burnap's claims were rather optimistic given that the scenario of having two candidates with identical vote shares is extremely unlikely when working with large datasets. Furthermore, (virtually) no lexicon-based techniques gave a better MAE for their LPM implementations in this study and [Bansal & Srivastava, 2018]'s paper when compared with the results given by LPV analysis. We can attempt to explain the cause of this: LPV analysis always uses a clear indicator of a vote (a positive tweet), whereas LPM adds more significant quantities towards a candidate's vote share from tweets which are "more positive" than others. This is not a clear nor a reliable indicator of a vote [Gayo-Avello, 2012]. In the future, LPV analysis should be preferred to calculate vote share and LPM analysis can provide an additional insight into the political discussion on Twitter. In the extremely rare scenario two candidate's vote shares are tied using Lexicon-Positive Volume Analysis, LPM analysis can be useful to differentiate between them.

6.2. Candidate Analysis for individual Issues:

With such a polarising set of outcomes for the two datasets, it is hard to draw any conclusions as to how well HTBSA* should be expected to perform candidate analysis on individual issues. Clearly, it would be a better solution than the only other known approach [Karami et al, 2018] as it performed much better in both election datasets for the only comparable metrics (candidate rankings and top candidates) and possesses the unique property of being able to produce percentile results, providing ample utility for comparisons with issue surveys. Yet this would only apply to our implementation of Karami's method, which may not be optimal and therefore an unfair evaluation of its ability to analyse political issues. Take the number of topics for example: unlike HTBSA*, the value of k has a direct effect on the results given by HIF — since a larger k creates the potential for a larger number of topics to be associated with different issues — and Karami did not give any details on how to determine this value. Furthermore, as mentioned previously, Karami's original study only analysed two candidates and judging from our results, it is not very effective at analysing more minor candidates (such as Liberal Democrats), so we cannot say for certain that HTBSA* is more effective than it.

For the unseen dataset, individual issue analysis results given by HTBSA* were extremely poor, but since the results from the 2019 election were so strong and we were not involved during the data collection (and most of the data cleansing) phase(s) for the unseen dataset results, the solution should not be written off so easily. Additionally, the unseen dataset did not take geo-tagging into account and the real-time Twitter Streaming API was used to collect tweets, which means it was likely to have included a much larger proportion of redundant tweets which can skew results.

6.3. Election Prediction:

The main takeaways from this research in regard to election prediction mining will be that first of all, regular HTBSA gave consistently poor results in comparison to baseline LSA (which is considered the state-of-the-art for lexicon-based approaches) and Volume Analysis. The implication of these observations is that they directly contradict the empirical evidence presented by Bansal's paper in [2018], which supported his claims that using HTBSA would consistently lead to reliable election predictions by providing better estimations of tweet sentiment. Out of all of the already-established approaches, analysing raw volume of tweets was the best performing method (Volume Analysis). Given that [Burnap et al, 2015]'s LSA only performed slightly worse, we can say that as expected both of these approaches performed effectively on our datasets. HTBSA was the least comparable benchmark technique to true vote share.

Conversely, HTBSA* was shown to outperform all of the chosen benchmark approaches and work very effectively on collected and unseen datasets. These results support our hypothesis that HTBSA* uses a more reliable indicator of a vote than that of HTBSA/LSA by prioritising key political issues. Given this, it is likely that HTBSA* could outperform these methods if applied to future election studies.

7. CONCLUSION

In this piece of research, we put forward a novel Twitter data mining method, Hybrid Topic-Based Sentiment Analysis with Issue Filtering, which could potentially perform candidate analysis of individual political issues at a baseline level and improve upon state-of-the-art lexicon-based election prediction techniques. We hypothesised this result based on the fact that our approach is currently the only known method that could perform candidate analysis in respect to individual issues and give results that can be compared directly with public surveys, and also because our election prediction technique is based upon newly defined principles which transcend the measurements that have been used in previous techniques to constitute reliable voter indication, a fundamental part of improving future studies that predict elections proposed by [Gayo-Avello, 2012]. On top of this, our solution encompasses two independent hybrid methods that use topic modelling and tweet sentiment and have been successfully used to mine elections in the past, so that future election analysis research can improve their predictions and be better equipped to explain their outcomes. We applied our technique on Twitter data collected on the run-up to the 2019 UK General Election and afterwards to an unseen dataset from the preceding election.

Performing candidate analysis in respect to individual issues proved to be a challenge and further work is needed to improve this aspect of the solution. HTBSA* results given by analysing public opinion on different issues were inconsistent and low number of topics as well as a high prevalence of negative topics amongst candidate tweets seemed to be the biggest contributing factors to poor results. Ideally, using a low number of issues when predicting elections with HTBSA* should be avoided to ensure the total aggregation of topics over all issues will not inherit any of the above properties that are observed when issues are analysed individually; therefore, we recommend using our chosen number as a minimum (five). Overall, we cannot say for sure that HTBSA* should be considered a baseline for the problem over [Karami et al, 2018]'s method despite performing better, since results were so mixed. Future work incorporating a data collection phase for an election occurring in the future is required in this area.

For election prediction, HTBSA* attempts to update the benchmark definition of what constitutes as a "reliable indicator of a vote" [Gayo-Avello, 2012] by reducing the number of topics used in HTBSA to calculate tweet sentiment to only those that relate to a key political issue. This principle was shown to work more effectively than [Burnap et al, 2015]'s original definition which was also used in baseline LSA and HTBSA. Although our election prediction results given by HTBSA* were extremely strong and gave a MAE as low as 2.20% from true vote share (on an unseen dataset), we are yet to successfully predict an election occurring in the future [Gayo-Avello 2012], which prevents us from making any real claims over the potential of our novel method to become the new state-of-the-art. This is the case with hybrid methods more generally, but this research project has nevertheless laid out a replicable novel method for election prediction mining; which performed effectively on tweets collected for the UK General Election 2019, as well as during a black-box test where it was applied to an unseen dataset.

Overall, this research has proven that it is possible to explore alternative ways of improving election mining techniques outside of the more conventional lines of research, such as trying to better account for the biases that occur in social media datasets or devoting ample time and/or resources to developing complex machine learning classifiers. It is also plausible that with further evaluation by method of application to future elections, HTBSA* could be an improvement upon state-of-the-art lexicon-based techniques.

Recommended upcoming political elections and polls in which HTBSA* could be applied to are the 2020 U.S. Presidential Election, the 2020 New Zealand General Election and the 2020 New Zealand Cannabis & Euthanasia Referendums. Based on our results, there still remains a lack of consistent empirical evidence to support [Bansal & Srivastava, 2018]'s claim that regular HTBSA can predict elections more accurately than standard LSA techniques in the future. As well as this, our findings do not support [Burnap et al, 2015]'s hypothesis that LPM analysis is better at measuring candidate vote share than LPV analysis. Lastly, based on empirical evidence gained by applying HTBSA* to self-collected and unseen datasets alone, this technique can be reliably replicated for future election prediction to give results that are comparable with traditional opinion polls.

7.1. Limitations & Future Work:

Addressing the many problems that can arise with using Twitter data as a population sample was out of scope for this research, however, it is clear that these issues do introduce some limitations. For example, in our 2019 election dataset, the proportion of active Twitter users was not equal across all of the parties; a candidate that highlights this well is the Brexit Party, as they had a much larger total tweet count than that of the Liberal Democrats, but a much smaller vote share in the final result. The results that we obtained in this study would support previous remarks made by [Mellon & Prosser, 2017] that using Twitter data to mine election outcomes without suitable adjustment would lead to overrepresentation of the left-leaning Labour Party. For example, nearly all of the vote share forecasting given by different election prediction methods executed on both datasets (figure 11/14) can be observed to demonstrate such a bias towards Labour. The same can be said for the individual issue analysis methods (figure 13a/10a), albeit to a lesser extent. It is also very likely that such biases change over time – for example Conservative Party Volume Analysis results between 2017 and 2019 – which would suggest any attempt to resolve them would require a dynamic approach. Such observations highlight the broader disproportions that social media platforms are shown to exhibit within their user bases when compared to true populational figures and until significant progress is made in addressing these problems, future election mining methods will continue to suffer. Search terms and time frames used to collect data also inadvertently introduce further biases. To conclude on these issues, it is important to stress that a project of this nature can never be done “perfectly” and will always be limited by such factors; however, alongside this, we are by no means suggesting that either of these final remarks are not true: firstly, that the impact of these factors can be reduced (up to a point) by ensuring diligent care is taken when making certain choices that can *systematically* exacerbate them (consistency of query selections, time period used, etc); and that in conjunction with ensuring such damage limitation is carried out correctly, broader methodological improvements should still be sought after in order to reduce errors that are introduced by *elements of randomness*, rather than systematically (such as Twitter sampling biases) — see our review of the literature for evidence of existing methods that can account for Twitter population sampling biases. Researchers should consider both of these items separately — systematic error reduction (damage limitation) and accounting for “random” biases — when trying to improve future studies.

When we collected our own data for the 2019 election, the decision to only use geo-tagged tweets as well as not using the real-time Twitter streaming API meant that our dataset was massively reduced. These restrictions were added largely due to reasons of wanting to preserve the integrity of our dataset; removing them would require a level of care which was out of scope and seemingly not so straightforward, as briefly discussed in the evaluation of our 2017 dataset. Nevertheless, an attempt to remove such restrictions should be made to increase the amount of data available in the future, which would make results more reliable. In the 2019 UK General Election, the SNP and the Green Party both secured more votes than the Brexit Party. This was hard to predict when the candidates were decided, due to the circumstances changing (Brexit Party stepped down in seats previously held by Tories). This was not possible to factor into our study but as a result of this a lot of the Twitter users who had counted as "votes" for the Brexit Party were likely to vote Tory in constituencies previously held by the Conservatives. More sophisticated techniques should be flexible enough to accommodate such electoral alliances in the future (see appendix). Furthermore, analysing different time frames in regard to forecasting vote shares and analysing key issues could be another future direction for researchers to examine; particularly in respect to significant political events (such as the example above) or even with how these can affect key issues over time. Researchers can improve Twitter-based sentiment analysis by using lexicons specifically designed for the political domain. An even better solution to this would be to develop a machine learning classifier or rather, a hybrid one which is able to classify candidate tweets based on topic sentiments. Amongst this, there are other ways to develop more sophisticated election prediction techniques with HTBSA*. In the United Kingdom for example, the voting system is done on a constituency basis, rather than by total vote share. If it were possible to get all of the tweets in each constituency and the key issues in them, HTBSA* could be used to calculate every individual seat in the UK. This is an example of mining voter intention by geo-location but in reality, such an idea could easily be abstracted to predict voting intentions by age, gender or any other given data point, providing the Twitter data and information on key issues is available. Such insightful statistics have been collected via surveying in the past and so far, no mining techniques have been developed to supplement them. Extending HTBSA* to calculate a measure of national swing and predict individual seats as done by [Burnap et al, 2015] could be undertaken to enhance future predictions. There a few manual steps required in HTBSA* such as issue filtering, solutions which are able to automate such steps would enable future studies to save time implementing HTBSA*. Retaining POS tags in topic models may have increased the accuracy of topic sentiments. One other constraint of our approach is that it can only be used to derive sentiment from tweets written in the English language. Further examination of HTBSA*’s ability to predict elections happening in the future is

imperative to completing a full objective evaluation of it as methodology, such studies should involve a data collection phase as a necessity, which was the only aspect of the original methodology that we were unable to replicate during our black-box test.

8. REFERENCES

Alamsyah, A., Rizkika, W., Nugroho, D., Renaldi, F. & Saadah, S. (2018). Dynamic Large Scale Data on Twitter Using Sentiment Analysis and Topic Modeling, *2018 6th International Conference on Information and Communication Technology (ICoICT),* Bandung, 2018, pp. 254-258.

Baccianella S, Esuli A & Sebastiani F (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: *LREC. vol. 10*; 2010. p. 2200–2204

Bansal, B. & Srivastava, S.(2018), On predicting elections with hybrid topic based sentiment analysis of tweets,*Procedia Computer Science*,Volume 135 , p 346-353.

Bermingham, A. & Smeaton, A. (2011), On Using Twitter to Monitor Political Sentiment and Predict Election Results, *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011),* 13.

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, 993-1022.

Burnap, P., Gibson, R., Sloan, L., Southern, R. & Williams, M. (2015), 140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election, [online] *Electoral Studies*, Available at: doi.org/ [Accessed 23/03/20]

Ceron, A., Curini, L. & Iacus, S. M. (2015), Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters—Evidence From the United States and Italy, *Social Science Computer Review*, 33(1), pp. 3–20.

Dahal, B., Kumar, S. & Li, Z. (2019), *Spatiotemporal Topic Modeling and Sentiment Analysis of Global Climate Change Tweets*.

DiGrazia, J., McKelvey, K., Bollen J. & Rojas, F. (2013) More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PloS one. 8.*

Gayo-Avello, D. (2011), A meta-analysis of state-of-the-art electoral prediction from Twitter data, *Social Science Computer Review*. 31. p13

Gayo-Avello, D. (2011), Don't turn social media into another 'Literary Digest' poll, *Communications of the ACM*, vol. 54, no. 10, p121–128.

Gayo-Avello, D. (2012), *"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data.*

Filho, RM., Almeida, JM., & Pappa, GL. (2015), Twitter population sample bias and its impact on predictive outcomes. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 25–28 August 2015, pp.1254–1261.

Ipsos.com, (2019), Ipsos MORI General Election Campaign Tracker, [online] *Ipsos MORI,* Available at: ipsos.com/ [Accessed 18/03/20], p1

Karami, A., Bennett, L. & He, X. (2018), Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election, *International Journal of Strategic Decision Sciences (IJSDS)*, 9(1), 18– 28.

Liu, B. (2012), Sentiment Analysis and Opinion Mining, *Morgan & Claypool Publishers*

Manning, C., Raghavan, R. & Schütze, H. (2008), [online] *Introduction to information retrieval*, Available at: nlp.stanford.edu [Accessed 17/30/20].

Marozzo, F. & Bessi, A. (2017), *Analyzing* Polarization of Social Media Users and News Sites during Political Campaigns, *Social Network Analysis and Mining.* 8.

McCallum, A. K. (2002), [online] *MALLET: A Machine Learning for Language Toolkit*, Available at: mallet.edu [Accessed 23/03/20]

Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. (2013). *Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling*. 889-892.

Mellon, J. & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media user, *Research & Politics*

Metaxas, P., Mustafaraj, E. & Gayo-Avello, D. (2011), How (Not) to Predict Elections, *Proceedings of PASSAT/SocialCom 2011*, IEEE Computer Society, 165–171.

Musto, C., Semeraro, G. & Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog posts, *CEUR Workshop Proceedings*. 1314. 59-68.

Oikonomou, L. & Tjortjis, C. (2018). *A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter*. 1-8.

Pennebaker, J. W., Booth, R. J. & Francis, M. E. (2007), *Linguistic inquiry and word count:* [online] *Liwc [computer soft- ware]*, Available at: liwc.net [Accessed 08/04/20]

Roder, M., Both, A. & Hinneb, A. (2015*).* Exploring the Space of Topic Coherence Measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408

Prabhakaran, S. (2018), Topic Modeling with Gensim (Python), [online] *Machine Learning Plus*, Available at: mlp.com [Accessed 17/03/20]

Sabbagh, D. (2018), Rise of digital politics: why UK parties spend big on Facebook, [online] *The Guardian,* Available at: theguardian.com/ [Accessed 04/03/20]

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*. 37. 267-307.

Teh, Y., Jordan, M., Beal, M & Blei, D. (2006). Hierarchical Dirichlet Processes. *Machine Learning*. 1-30.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. & Kappas, A. 2010. Sentiment strength detection in short informal text. *J Am Soc Inform Sci Technol 61*

Thelwall, M. (Unknown), [online] *The Emotionality of Discourse*, Available at: http://wlv.ac.uk/ [Accessed 17/03/20], slide 7

Tjong Kim Sang, E. & Bos, J. (2012), Predicting the 2011 Dutch Senate Election Results with Twitter, *13th Conference of the European Chapter of the Association for Computational Linguistics*, April 23-27, 2012. Avignon, France.

Tumasjan, A., Sprenger, TO., Sandner, PG., & Welpe, IM. (2010), Predicting elections with twitter: What 140 characters reveal about political sentiment, [online] *Icwsm*, Available at: aaai.org/ [Accessed 05/03/20]

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2014), Forecasting elections with non-representative polls, *International Journal of Forecasting*, 31(3): 980–91.

Yan, X., Guo, J., Lan, Y. & Cheng, X. (2013), A biterm topic model for short texts, *Proceedings of the 22nd international conference on World Wide Web*, ACM; 2013. p. 1445–1456.

YouGov.co.uk. (2019), Best Party on Issues (GB), [online] *YouGov,* Available at: cloudfront.net [Accessed 16/03/20]

YouGov.co.uk. (2019), Top Issues Tracker (GB), [online] *YouGov,* Available at: cloudfront.net/[Accessed 16/03/20]

YouGov.co.uk. (2019), Voting Intention Tracker (GB), [online] *YouGov,* Available at: cloudfront.net [Accessed 16/03/20]

Zimbra, D., Abbasi, A. & Zeng, D. (2018). The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation, *ACM Transactions on Management Information System*

APPENDIX A — ADDITIONAL DATA FROM 2019 ELECTION:

1) Method of Prediction by Average Issue Survey Candidate Shares:

Taking a step back and using surveying data provided by [YouGov, 2019], if we were to make an election prediction using this exact method, we would get a MAE of 7.05% and 7.13% for results closest to the last two elections. it wouldn't be unreasonable to compare this result with MAE of traditional opinion polls (2.53% and 2.05%) [YouGov, 2019]. These results can be found below:

2019	Conservative	Labour	Liberal Democrat	Brexit		
NHS	26%	35%	6%	2%		
Brexit	31%	13%	11%	12%		
Law & Order	36%	19%	6%	3%		
Immigration	30%	18%	8%	12%		
Education	26%	30%	9%	1%		
Taxation	34%	24%	7%	2%		
Unemployment	28%	27%	5%	2%		
Economy	37%	19%	7%	2%		
Defence	39%	15%	5%	3%		
Housing	22%	32%	6%	1%		
Average	30.90%	23.20%	7.00%	4.00%	**MAE**	**7.05%**
Real	43.60%	32.20%	11.50%	2.00%		
Opinion Poll	44.00%	28.00%	16.00%	3.00%	**MAE**	**2.53%**

2017	Conservative	Labour	Liberal Democrat	UKIP		
NHS	22%	41%	4%	2%		
Brexit	37%	19%	7%	6%		
Law & Order	36%	26%	4%	4%		
Immigration	30%	20%	5%	12%		
Education	25%	36%	5%	2%		
Taxation	30%	30%	5%	2%		
Unemployment	29%	30%	2%	2%		
Economy	39%	25%	4%	2%		
Defence	37%	22%	3%	4%		
Housing	21%	35%	4%	2%		
Average	30.60%	28.40%	4.30%	3.80%	**MAE**	**7.13%**
Real	42.40%	40.00%	7.40%	1.80%		
Opinion Poll	44.00%	36.00%	7.00%	4.00%	**MAE**	**2.05%**

Issue surveying data from last two elections showing MAE of vote share predictions made by method of averaging top four candidates score over all the issues. Compared with opinion polls.

2) Precision and Recall Values:

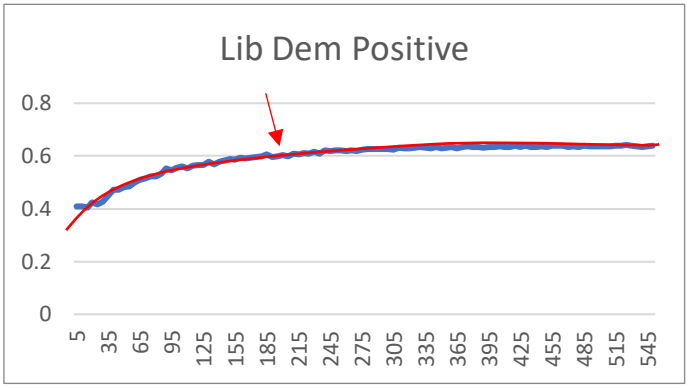
Precision is the ratio between true classifications and total classifications for either positive or negative texts. Positive recall is calculated by dividing the number of true positives by the number of true positives combined with the number of false negatives; and vice-versa for negative recall. We first of all took a random sample of 500 tweets and manually labelled them as positive, negative and neutral. We then ran SentiStrength over each tweet in our sample set and calculated sentiment scores using the 'scale' classifier. This is a linear scale from -4 to 4 and we scaled scores down to fall into the range of -1 and 1. In other words our scores fell into absolute values from 0 with increments of 0.25, up until 1. By running sentiment classification for thresholds 0, 0.25, 0.5 and 0.75 we observed the following results:

	Classified as Neutral	+ Precision TP/P	- Precision TN/N	+ Recall TP/TP+FN	- Recall TN/TN+FP	**Avg Precision**	**Avg Recall**
0	38.2%	78.4%	90.67%	83.5%	87.5%	84.6%	85.5%
0.25	72.8%	78.0%	94.74%	86.5%	90.9%	86.4%	88.70%
0.5	90.8%	83.3%	92.5%	62.5%	97.4%	87.9%	79.93%
0.75	98.8%	N/A	100.0%	N/A	N/A	100.0%	100.0%

Different thresholds used to categorise tweets using SentiStrength.

3) Determining Number of Topics 'k' – Example:

Below is a graph showing different values of k against cv coherence. This was taken from the positive Liberal Democrat tweets' dataset during our implementation of HIF. As you can see there is a slight curve with decreasing gradient. The blue line shows the results we obtained, and the red line indicates a rough outline of the trend observed. We decided to select 200 topics for this dataset (see arrow) because after this point the curve begins to completely flatten out. We also made sure the ratio of the number of topics between candidate datasets was relative to the ratio between number of tweets in them.

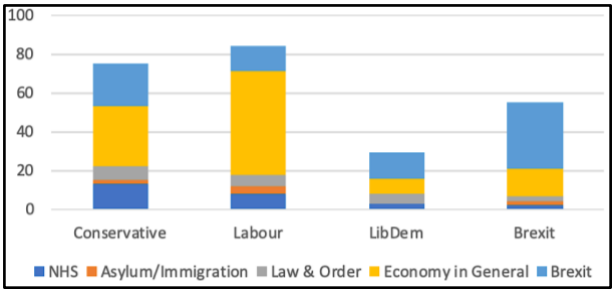


Number of topics vs cv coherence for the Liberal Democrats' positive tweets.

4) Number of Topics Produced for Different Topic Models:

After the topic modelling process was complete, for our four candidate datasets which we were going to perform HTBSA/HTBSA* upon, we were left with 350 topics for the Conservatives, 400 for Labour and 250 each for the Liberal Democrats and Brexit Party and for the eight positive/negative dataset pairs we were going to perform HIF upon, our positive topic numbers in the same order as above were 235, 255, 200 and 190, and our negative topic numbers were 175, 200, 195 and 185. These numbers are not essential in any way, but we have included them for reference.

5) Further Breakdown of Issues in Candidate Datasets:



Breakdown of the filtered topics in each candidate dataset, in respect to key political issues.

6) Full Set of Results*:

*Doesn't include some results that were shown in main body of report.

Party	True Vote Share	Baseline LSA [Burnap et al, 2015]	HTBSA [Bansal & Srivastava, 2015]	HTBSA* (All Issues)	HTBSA* (NHS, Econ, Brxt)
Conservative	43.60%	35.77%	7.35%	38.21%	27.83%
§	32.20%	40.43%	60.59%	39.69%	40.06%
Lib Dems	11.50%	18.08%	22.65%	19.20%	26.83%
Brexit	2.00%	5.73%	9.40%	2.90%	5.28%
MAE	**N/A**	**9.48%**	**21.85%**	**9.67%**	**13.67%**

ABOVE: *Lexicon-Positive Magnitude analysis results for lexicon-based election prediction methods.*

NHS	Party	Net	Ranking	Real	Correct
	Conservative	0	4	2	N
	Brexit	-1	3	4	N
	Labour	-1	3	1	N
	Lib Dem	-1	3	3	N

Asylum/Immigration	Party	Net	Ranking	Real	Correct
	Conservative	2	2	1	N
	Brexit	-1	4	3	N
	Labour	3	1	2	N
	Lib Dem	0	3	4	N

Law & Order	Party	Net	Ranking	Real	Correct
	Conservative	-4	3	1	N
	Brexit	-6	4	4	Y
	Labour	-1	2	2	Y
	Lib Dem	-1	2	3	N

Economy in General	Party	Net	Ranking	Real	Correct
	Conservative	1	3	1	N
	Brexit	1	3	3	Y
	Labour	-7	4	2	N
	Lib Dem	2	1	4	N

Brexit	Party		Ranking	Real	Correct
	Conservative	8	1	1	Y
	Brexit	6	2	3	N
	Labour	0	3	2	N
	Lib Dem	-1	4	4	Y

ABOVE: *HIF results for analysing chosen key issues.*

BELOW: *HTBSA* results for analysing chosen key issues*

(true candidate orderings not shown because they are above).

NHS	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	26.00%	37.38%	Y	20.51%	Y
Labour	35.00%	45.21%	Y	68.48%	Y
Brexit	2.00%	11.85%	N	7.55%	N
Lib Dems	6.00%	5.56%	N	3.45%	N
MAE	**N/A**	**7.97%**	**50.0%**	**11.77%**	**50.0%**

Asylum /Immigration	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	30.00%	47.56%	N	79.63%	Y
Labour	18.00%	52.16%	N	20.05%	Y
Brexit	12.00%	0.28%	Y	0.32%	Y
Lib Dem	8.00%	0.00%	Y	0.00%	Y
MAE	**N/A**	**17.86%**	**50.0%**	**17.84%**	**100.0%**

Law & Order	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	36.00%	46.44%	N	63.53%	Y
Labour	19.00%	53.48%	N	36.42%	Y
Brexit	3.00%	0.03%	Y	0.01%	Y
Lib Dem	6.00%	0.05%	Y	0.03%	Y
MAE	**N/A**	**13.46%**	**50.0%**	**13.48%**	**100.0%**

Economy in General	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	37.00%	38.89%	N	38.77%	N
Labour	19.00%	45.33%	N	55.45%	N
Brexit	2.00%	9.86%	N	2.73%	Y
Lib Dem	7.00%	5.92%	N	3.05%	Y
MAE	**N/A**	**9.29%**	**0.0%**	**10.73%**	**50.0%**

Brexit	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	31.00%	41.36%	Y	23.03%	N
Labour	13.00%	37.78%	Y	2.20%	N
Brexit	12.00%	14.22%	Y	65.80%	N
Lib Dem	11.00%	6.64%	Y	8.97%	N
MAE	**N/A**	**10.43%**	**21.85%**	**18.65%**	**0.0%**

Averages	All Issues	Only NHS, Econ, Bxt
LPV HTSBA* MAE	11.80%	9.23%
LPV HTBSA* Ordering	50.0%	50.0%
LPV HTBSA* #1 Candidate	40.0%	66.6.%
LPM HTBSA* MAE	18.12%	13.71%
LPM HTBSA* Ordering	60.0%	33.3%
LPM HTBSA* #1 Candidate	80.0%	33.3%
HIF Ordering	25.0%	33.3%
HIF #1 Candidate	25.0%	33.3%

ABOVE: *Average results given by LPV HTBSA*, LPM HTBSA* and HIF for the different issues based on three metrics, average MAE, average correct ordering and average correct #1 candidate.*

7) Adjusting Election Prediction Results for Conservative-Brexit Party Alliance:

If we were to subtract the Brexit Parties true vote share from the forecasted one and add it to the Conservative vote share for all techniques, this would be how the results would have looked:

Party	True Vote Share	Volume Analysis	Baseline LSA [Burnap et al, 2015]	HTBSA [Bansal et al, 2015]	HTBSA* (All Issues)	HTBSA* (NHS, Econ, Brxt)
Conservative	43.60%	50.90%	51.67%	41.04%	48.48%	47.70%
Labour	32.20%	42.50%	40.81%	50.79%	44.13%	44.59%
Lib Dems	11.50%	4.60%	5.52%	6.17%	5.39%	5.70%
Brexit Party	2.00%	2.00%	2.00%	2.00%	2.00%	2.00%
MAE	**N/A**	**6.13%**	**5.67%**	**6.62%**	**5.73%**	**5.57%**

Forecasted vote shares given by different techniques after the above adjustment was made.

Please note that these figures are purely based on speculative ideas and have no bearings on evaluating our research methods or the results given by them, they are mentioned here as a mere demonstration of how the changes in circumstance had an adverse effect on our results such that all election prediction techniques that were ran in our study over-predicted the Brexit Party. Our results went from no techniques predicting the correct candidate ordering to all of them doing so; the MAE for each technique was also reduced.

APPENDIX B — ADDITIONAL DATA FROM 2017 ELECTION:

1) Configuration for 2017 Election:

Data Collection:

Dataset found online. It used a very simple set of search terms (ge2017, generalelection, labour, conservatives, tories, ukip, libdems, greens, snp, brexit) to collect data starting a week prior to the election, including the election day itself. Dataset wasn't labelled by candidates; therefore, we selected the top four candidates based on same survey used in 2019 election provided by [YouGov, 2019], but for data most recent to the 2017 election – Conservative, Labour, Lib Dem & UKIP. Only additional step was to produce query sets for each party to extract and label tweets from the single corpus that we had downloaded, these can be found below:

Party	Conservative	Labour	Liberal Democrat	UKIP
Queries	conservative conservativeparty conservatives tories theresa may tory torys	labour labourparty labour party jeremy corbyn	libdem liberal democrats liberal democrat liberaldemocrats liberal democrat lib dem libdems lib dems tim farron	paul nuttall ukip uk independence party ukindependenceparty united kingdom independence party unitedkingdomindependenceparty

Queries used for candidates during data collection phase.

NB: It may seem that the query sets do not follow the same format used for the 2019 election, as there are a different number of queries for each candidate; however, the Twitter Streaming API used to collect the 2019 dataset in a finite number of cases would return tweets matching a given string query, even if the occurrence of the string in the tweet wasn't exact – these cases were if the string existed in the tweet without whitespaces e.g. lib dem/libdem, or if the plural (or non-plural) version of a string existed e.g. tory/tories – so we added in all these different cases to emulate this feature of the streaming API, and to ensure our configuration for the unseen dataset was as close to the 2017 one as possible.

Data Cleansing:

Unfortunately, we did not have the luxury of having pre-processed geo-tagged tweets, so we were bound to have a significantly larger proportion of irrelevant tweets in candidate datasets. We decided not to adjust the vote shares to account for these proportions in order to keep the implementations identical, yet even if we did so, it wouldn't have prevented redundant tweets form being used to generate topics in HTBSA, HTBSA* and HIF results.

The rest of the 2017 election tweets' configuration was identical the 2019 one:

Volume Analysis:

Identical Implementation: trivially used same method.

Sentiment Analysis:

Identical implementation: SentiStrength used with 0 threshold to classify tweets as positive, negative or neutral in HIF. LSA performed using SentiStrength.

Topic Modelling:

Identical Implementation: Author-Pooled LDA used and values for 'k' were determined by same method as the 2019 election for HIF and HTBSA/HTBSA*.

Topic Sentiment:

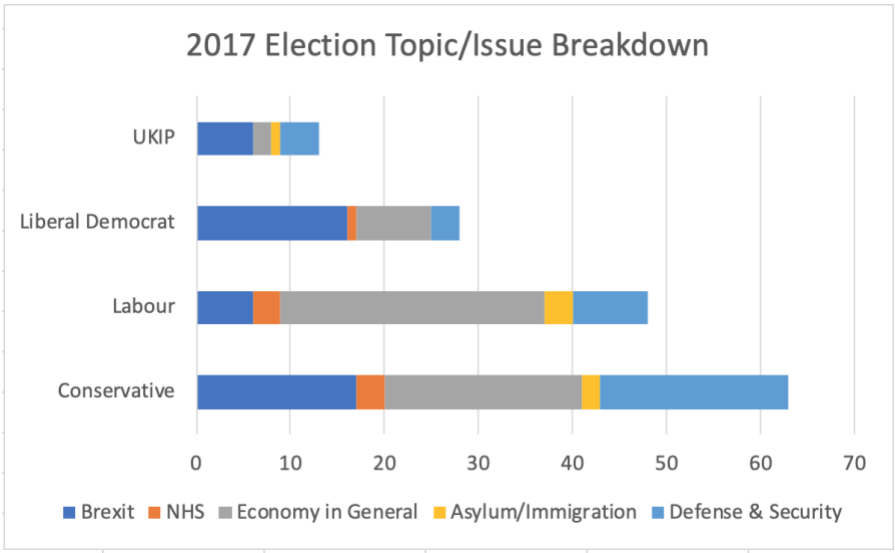
Identical implementation: Sentiwordnet was used in same way and top 20 words were used for calculating topic sentiment in HTBSA/HTBSA*.

Issue Filtering:

Identical implementation: top issues were determined from same survey data used for 2019 election; however, we used the data most recent to the 2017 election [YouGov, 2019] and selected the top five issues that were available to us on both the top issue survey, as well as the public opinion on various issues survey – Brexit, Health, Economy, Asylum/Immigration & Defence/Security. These issues used for issue filtering done in HIF and HTBSA*.

2) Example and Breakdown of Issues from 2017 Election:

Example Issue	Conservative	Labour	Lib Dems	UKIP
Brexit	back brexit future bright	corbyn pro remain position	negotiation start negotiate brexit	turn brexit table mandate
NHS	privatise nhs pay private	start nhs privatise privatisation	vote save nhs people	N/A
Economy in General	nhs education privatise destroy	industry job skill skilled	economy plan economic liberal	foreign aid money cut
Asylum/ Immigration	immigration target promise reduce	immigration policy mass control	N/A	immigration policy immigrant control
Defence & Security	nuclear war weapon bomb	police cut officer force	attack victim terrorist terror	terrorism islamic terrorist problem



Breakdown of number of topics filtered for each issue across candidate topics for the unseen dataset.

3) Full Results from 2017 Election:

*Doesn't include some results that were shown in main body of report

Party	True Vote Share	Baseline LSA [Burnap et al, 2015]	HTBSA [Bansal & Srivastava, 2015]	HTBSA* (All Issues)	HTBSA* (Brxt, Econ, Defence)
Conservative	42.30%	35.52%	13.12%	27.33%	27.25%
Labour	40.00%	53.65%	63.61%	38.29%	37.52%
Lib Dems	7.40%	6.81%	9.27%	27.52%	29.05%
UKIP	1.80%	4.02%	14.00%	6.86%	6.18%
MAE	**N/A**	**5.81%**	**16.71%**	**10.46%**	**10.89%**

ABOVE: *Lexicon-Positive Magnitude analysis results for lexicon-based election prediction methods.*

Brexit	Party	Net	Ranking	Real	Correct
	Conservative	4	1	1	Y
	Labour	-2	3	2	N
	Lib Dem	-5	4	3	N
	UKIP	2	2	4	N

NHS	Party	Net	Ranking	Real	Correct
	Conservative	2	1	2	N
	Labour	1	3	1	N
	Lib Dem	0	4	3	N
	UKIP	1	3	4	N

Economy in General	Party	Net	Ranking	Real	Correct
	Conservative	0	1	1	Y
	Labour	-9	4	2	N
	Lib Dem	-8	3	3	Y
	UKIP	-2	2	4	N

Asylum/Immigration	Party	Net	Ranking	Real	Correct
	Conservative	-1	3	1	N
	Labour	0	2	2	Y
	Lib Dem	-2	4	4	Y
	UKIP	1	1	3	N

Defence & Security	Party		Ranking	Real	Correct
	Conservative	-8	4	1	N
	Labour	-7	3	2	N
	Lib Dem	-6	2	4	N
	UKIP	-5	1	3	N

ABOVE: *HIF results for analysing chosen key issues.*

BELOW: *HTBSA results for analysing chosen key issues*

(true candidate orderings not shown because they are above).

Brexit	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	37.00%	15.43%	N	4.76%	N
Labour	19.00%	63.15%	N	28.20%	N
Lib Dem	7.00%	12.11%	Y	57.51%	N
UKIP	6.00%	9.31%	Y	9.52%	N
MAE	**N/A**	**18.54%**	50.00%	**23.87%**	0.00%

NHS	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	22.00%	0.00%	N	0.00%	N
Labour	41.00%	100.00%	Y	100.00%	Y
Lib Dem	4.00%	0.00%	N	0.00%	N
UKIP	2.00%	0.00%	Y	0.00%	Y
MAE	**N/A**	**21.75%**	50.00%	**21.75%**	50.00%

Economy in General	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	39.00%	6.52%	N	3.50%	N
Labour	25.00%	66.38%	N	89.15%	N
Lib Dem	4.00%	12.54%	N	7.35%	N
UKIP	2.00%	14.55%	N	0.00%	Y
MAE	**N/A**	**23.74%**	0.00%	**26.25%**	25.00%

Asylum /Immigration	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	30.00%	85.56%	Y	83.19%	Y
Labour	20.00%	1.62%	N	2.33%	N
Lib Dem	5.00%	0.00%	Y	0.00%	Y
UKIP	15.00%	12.82%	N	14.49%	N
MAE	**N/A**	**20.28%**	50.00%	**19.09%**	50.00%

Defence & Security	Survey Results	HTBSA* Lexicon-Positive Volume Analysis	Correct Order?	HTBSA* Lexicon-Positive Magnitude Analysis	Correct Order?
Conservative	37.00%	83.14%	Y	89.85%	Y
Labour	22.00%	4.59%	N	1.29%	N
Lib Dem	3.00%	0.05%	N	0.02%	Y
UKIP	4.00%	12.22%	Y	8.85%	N
MAE	**N/A**	**18.68%**	50.00%	**20.35%**	50.00%

Averages	All Issues	Only Brxt, Econ, Def
LPV HTSBA* MAE	21.08%	20.32%
LPV HTBSA* Ordering	45.0%	33.3%
LPV HTBSA* #1 Candidate	60.0%	33.3.%
LPM HTBSA* MAE	22.26%	23.49%
LPM HTBSA* Ordering	35.5%	25.0%
LPM HTBSA* #1 Candidate	60.0%	33.3%
HIF Ordering	25.0%	25.0%
HIF #1 Candidate	20.0%	66.6%

ABOVE: *Average results given by LPV HTBSA*, LPM HTBSA* and HIF for the different issues based on three metrics, average MAE, average correct ordering and average correct #1 candidate.*