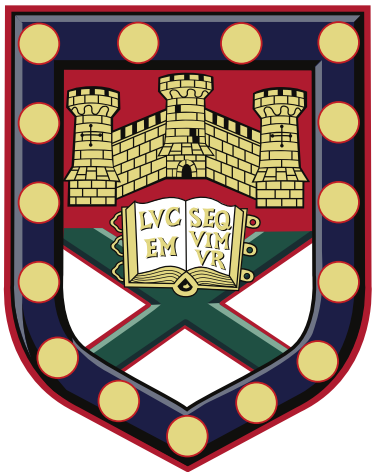
A BIRD'S EYE VIEW: TO WHAT EXTENT CAN WE SOCIALLY SENSE LOCAL ATTITUDES TO POLITICAL PARTIES USING TWITTER?



Christopher Tyson
Department of Computer Science
University of Exeter

Submitted for the degree of
Master of Data Science

September 2020

Abstract

With the modern ubiquity of networked devices, it's clear we're leaving more digital fingerprints online than ever. This work aims to examine a number of those fingerprints, found in a UK Location based collection of tweets taken from the run up to the 2019 UK General Election. Novel indicators were produced to infer the voting intention of Twitter users, which were mapped on to location tags to produce "tweet votes". A bias corrective method using simulated annealing was produced, which aimed to build a demographic bridging between the voting intentions of the Twittersphere and the UK electorate, and a number of interesting insights concerning the relationship of geospatial political orientation on Twitter and in real life were found. This task revealed a large number of non-trivialities in attempting to predict elections using Twitter, but provided hopeful results for future attempts. To the authors own knowledge, this is the only work which has combined Twitter location tagging with user retweets, to perform political and electoral analysis.

Contents

Introduction

With Twitter reporting 15 million tweets posted relating to the 2019 UK General Election (a 66% increase from 2017)[19], and 589 out of 650 UK Members of Parliament now on Twitter[1], it's apparent that social media is becoming more integral to the political process. It seems likely that online participation in political discourse is on the increase, and this has piqued the attention of researchers kicking off an endeavour to try and make sense of the masses of resulting data. A number of studies have sought to leverage such data for use in a predictive capacity, and have shown positive results in predicting and tracking trends in the stock market[6], movie sales[3] and pandemic likelihood[16]. Additionally researchers have found success in using Twitter data for "event detection" in which real time warnings can be derived from tweets, signaling phenomena such as floods[2], fires[7] and riots[1].

Over the last several years some of the same methodology has been applied with the aim of predicting elections[27, 11], a task which has found itself at the centre of a growing industry, with the performance of many tried and tested models faltering in recent elections. Numerous researchers have reported successful models built using social media data for the prediction of elections, however such studies have drawn criticism[18, 4] for some of their usage of data and lack of attention to sample biases.

The 2019 United Kingdom General Election took place on December 12th 2019 after legislation for another early election was approved, and was predominantly contested by the right wing Conservative Party and it's leader Boris Johnson, and the left wing Labour Party with it's leader Jeremy Corbyn, with a number of smaller minority parties also vying for seats. The election saw a number of "red wall" seats held by Labour for generations change hands, and resulted in a win (and 80 seat majority) for the Conservative Party. Despite most polling of the election coming close to this result, a lot of left leaning commentators were surprised by the extent of the Conservative victory, often citing that the political landscape they observed on social media did not reflect what they saw play out on election night. This claim, along with other observations has lead many to believe that Twitter's political environment is biased towards left wing views, which has even pushed some to jump ship to other social networks[9].

The aim of this work is threefold. Firstly using the 2019 United Kingdom General Election as a case study, the predictive capacity of Twitter is examined, by evaluating whether the localised distribution of political views in UK Parliamentary constituencies reflects reality, and by investigating the claim that Twitter's political

[1] https://www.mpsontwitter.co.uk

discourse is left-skewed. Viewpoints towards seven political parties are examined as a measure of these views, the parties being: The Conservative and Unionist Party, The Labour Party, The Liberal Democrats, The Green Party of England and Wales, The Scottish National Party, Plaid Cymru (The Party of Wales), and The Brexit Party. There are 650 Parliamentary constituencies in the United Kingdom, however a relaxation of electoral prediction will be used for this work which examines the 632 British constituencies as the inclusion of the main parties of Northern Ireland would add (at least) a further 5 distinct parties that would needed to be classified. Secondly we aim to validate results on a handful of constituencies which saw an unprecedented handover from the Labour Party to the Conservatives, testing whether decreased support for Labour can be especially felt in those "red wall" seats. And thirdly, the feasibility of a bias correcting method is studied, in order to ease the demographic differences from Twitter's usership to that of the electorate.

If Twitter were found to be a successful instrument for predicting political elections, the process of electoral prediction could be made cheaper and less resource intensive, and the methodologies could be utilised by political parties in order to inform more intelligent campaigning strategies.

The following work will be sectioned into seven main parts, a background section which will provide a brief literature review on a number of topics related to this work, a data section which will discuss the process of data collection as well as how data was formatted and preprocessed, a methods section will provide insight into each of the methods developed throughout this work, a results section will detail a number of evaluation metrics computed from the found results, as well as presenting a number of plots, a discussion section will talk about the results found in this study, and finally a conclusion section will discuss some of the author's recommendations for further study into this area of research.

Background

The main principals and ideas in this work draw on successes from two domains which have utilised large amounts of Twitter data: Social Sensing / Event Detection and Latent Attribute Inference.

Social Sensing / Event Detection

Social sensing refers to a set of techniques which treat human beings as sensors, capturing data about the real world and posting this online, with the aims of extracting reliable information from a mass of this unreliable and noisy data[29]. With the modern ubiquity of networked technology, people are posting more about their everyday lives online and so applications can be built to sift through masses of this data to find content relevant to a myriad of applications. Event Detection builds on these methodologies, augmenting them with location data to track and monitor spatially embedded events, often natural disasters and weather. Due to this, Event Detection often relies on a mixture of location inference methods, as well as techniques that are able to derive insight from the content of posts. Social Sensing and Event Detection pipelines can be broadly be distilled into 4 broad steps: Data Collection, Filtering/Classification, Location Inference, and Summarising Results.

With Twitter based topics, most studies opt to use the Twitter Streaming API to collect tweets, as this gives the option to search for tweets by keyword and/or location. For some applications, such as in natural disaster event detection the scope of the content needed may be quite small and so a keyword search is adequate[2], whereas others may wish to take a larger sample by getting all tweets and applying filtering later on[6]. It may also be deemed useful to apply additional filtering at this level, to get rid of easily identified erroneous content - for instance retweets, quote tweets, or tweets from an irrelevant timezone[2].

For detecting spatially embedded phenomena it's important to have a measure of where the tweet was authored, however occurrences of tweets with location tags can be quite sparse. A 2013 study[25] found that only around 0.85% of tweets contained exact geographic metadata, which means that for more niche applications after content filtering is applied the sample of workable data can be rather small. To get around this a number of location inference methods have been developed. A multi-indicator approach to location inference is often observed in the literature, one such being the method proposed by Schulz, Hadjakos, Paulheim, Nachtwey and Mühlhäuser[24] which pulls spatial indicators from the tweet's body and user's

account, that are then passed to lookup services. DBPedia, a project which has compiled data from Wikipedia into a structured format was used to look up candidate location features, and if a feature corresponded to a location then GADM - a database of local administrative areas - was used to grab a polygon for this location. This collection of polygons can then be distilled to a final location for a user/tweet by stacking polygons into a height map. Smaller polygons are given larger heights as they correspond to more granular locations. It follows that the highest point on the map then corresponds to the inferred location of the user. The authors found that they could locate 92% of tweets in a set with a median accuracy below 30km, which was a workable number for many applications.

As Social Sensing and Event Detection applications focus on tracking tweets with specific content, some method of filtering or classification needs to be applied to hone in on the tweets which fit into the required topic category. Arthur, Boulton, Shotton, and Williams[2] looked at the social sensing of flooding across the UK, and developed a Naive Bayes relevance filter to knock out erroneous tweets leaving a subset relevant to immediate flooding that could be studied. Multinomial Naive Bayes was used, rather than a simpler binary classifier, so that multiple categories of irrelevancy could be specified, which were: warnings (no flood observed, but "chance of flooding" stated), historical, and otherwise irrelevant (e.g. "floods of tears"). Sakaki, Okazaki and Matsuo's[23] work on the social sensing of earthquakes utilised additional features, and instead opted for a Linear Kernel SVM as their initial method of filtering. To train this model the tweet data was split into three feature groups: statistical, keyword and word context, with later results showing that perhaps surprisingly the statistical features (the number of words contained in a tweet as well as position of the keyword) were the most helpful for classification.

For summarisation Arthur et al.[2] split the area of study into a grid, giving a score to each cell based on the tweets found in that area using their novel multi-indicator method. Point scores were normalised by real world factors: i.e. population density taken from the last census, and then tuned by a parameter α via a validation process. The resulting score was dubbed the "floodiness" of the grid cell, and was used to produce maps that detailed real time flood observations which could be used to validate flood warnings produced from quantitative meteorological data. Sakaki et al.[23] built on the idea that Twitter users can work like often defective/unreliable sensors observing events, and developed a spatiotemporal probabilistic model. Methods from pervasive computing were employed such as Kalman Filtering and Particle Filtering, and used the locations of the human sensors to better determine the origin point of earthquakes. The pipeline produced by Sakaki et al.[23] was used as the backend for an early warning system, and was able to notify people in the vicinity

of an earthquake 5 minutes faster than the Japanese Meteorological Agency.

Latent Attribute Inference

Latent Attribute Inference is the notion of using available unstructured data generated by online individuals to infer demographic attributes that don't tend to be information held or accessible on some online services, i.e. age, gender, political orientation[30] and even coffee preference[21]. This can be useful for a broad range of research applications, and many methods developed are robust enough to be transferable to capture more niche attributes, these methodologies also being applicable to tasks such as user profiling or user recommendation.

Solutions to Latent Attribute Inference tasks often utilise a broad range of features for classification. Pennacchiotti and Popescu[21] leveraged 4 categories of user features in their implementation of a robust Latent Attribute Inference pipeline that was used for several classification tasks. These features included: **profile features** such as username, bio and location, **tweeting behaviour** which captured statistics of when and how often a user tweeted, as well as the number of replies a user authored, **linguistic content features** which included prototypical words and hashtags used by various user classes, as well as LDA features and text sentiment, and **social network features** which included user followership as well as accounts replied to and retweeted. The authors used Gradient Boosted Decision Trees as an initial classification method, which was then refined using a graph based label updating function and found that their initial classification method achieved a good performance which was difficult to refine using social graph features. The authors looked at the inference of three demographic variables: political orientation, ethnicity and whether the user was a "starbucks fan". Their classifier observed a good accuracy reporting that reported linguistic features seemed to give the most robust performance across the varied tasks, and found it was difficult to improve the initial classifier by adding social graph information.

The seminal work by Conover, Ratkiewicz, Francisco, Goncalves, Flammini and Menczer[12] shows how the polarized nature of political discourse online has the potential to be used in classification tasks by determining that the topology of a retweet graph exhibits a partisan structure. The authors collected politically relevant messages authored during the 2010 US Midterm elections, and built two networks, one in which edges correspond to retweets between user nodes, the other in which edges represent mentions. Networks were clustered using label propagation seeded with an initial partition of node labels computed using an eigenvector modularity maximisation method. Distinct community structure was observed in the retweet graph,

with little salient structure seen in the mention graph, and results were validated by examining content homogeneity in the clusters. It can be concluded from this work that retweets are a strong method of endorsement, as there was little information sharing in the form of retweets between communities.

A number of other studies have looked at various methods for political text/user classification. Rao and Spasojevic[22] used an LSTM with word embeddings and were able to get a 87% accuracy in the classification of tweets marked as Democrat or Republican, with their classifier returning a continuous value in $[0, 1]$ conveying the intensity of the message. The authors claimed that whilst a lot of work has went into sentiment analysis of text, contextual analysis often proves to help with classification tasks such as identifying political leaning as messages supporting a political party may carry varied sentiment. Zamal, Liu and Ruths[30] also tackled Latent Attribute Inference by considering text features, and built feature vectors from the content a user posted, looking at the top k-differentiating words, stems, digrams, trigrams, co-stems and hashtags. The authors exploited the principle of homophily - the notion that online individuals tend to seek out and associate with similar users that they agree with - by constructing feature vectors for the individuals a user follows, and merging these into the users own feature vector by either averaging or concatenation. An SVM was used to classify feature vectors, and the authors found a significant increase in accuracy against conventional methods, as well as good performance using the neighbourhood features alone, giving credence to the idea that substantial amounts of information about a user can be inferred from their online neighbours.

Cohen and Ruths[10] warn that the success observed by many Latent Attribute Inference papers in the literature focusing on political orientation may be optimistic in the reporting of their results, citing the way in which validation sets for such studies are constructed. The authors prepared sets of political figures, politically vocal individuals and politically modest individuals to validate their claim that many inference methods enjoy such high degrees of accuracy as they only focus on users that are easiest to classify. The authors found that by using novel methodologies such as those proposed by Conover et al.[12] and Zamal, Liu and Ruths[30] on data sets of "normal" Twitter users were only able to achieve accuracies as high as 65%.

Twitter Based Electoral Prediction

Perhaps the most cited example of electoral prediction using Twitter is the work of Tumasjan, Sprenger, Sandner and Welpe[27], who looked to study if Twitter can serve as a predictor for electoral success. The work examined 104,003 political tweets prior to the German National Election, with tweets being collected using a

keyword search filtering for the names of the 6 parties represented in the German Parliament, as well as the names of prominent politicians from these parties. LIWC (Linguistic Inquiry and Word Count) text analysis software which is able to assess the emotional, cognitive and structural components of text was used to analyse the content of messages shared during the election period. The authors claimed that Twitter could be used as a predictor of electoral success, citing that even the share of party mentions found in data reflected the share of the election result for each party of study. The Mean Absolute Error (MAE) was calculated for the tweet share against the real vote share to be 1.65% which the authors claimed to be comparable to major pollsters.

Burnap, Gibson, Sloan, Southern and Williams[8] looked at the 2015 UK General Election as a case study for prediction with Twitter, and collected almost 14 million tweets which mentioned political parties/leaders over a large period spanning from November 2014 to March 2015. The authors filtered the tweets to those that only included a single political target, and computed sentiment analysis scores for each tweet with the rationale that tweets with positive sentiment containing a party/leader implied voting intention for that party. Sentiment scores were summed across tweets to get a sentiment polarity with respect to the parties, and the shares of positive sentiment were normalised to obtain "twitter vote shares". The vote shares corresponded to an election result which gave a hung parliament with Labour as the majority party, which was not the correct result of the election. The authors also noted that without geolocation there was a significant under-estimation of regional parties such as the SNP, and stated that future methods of electoral prediction utilising Twitter would need to factor in location based information.

Coletto, Lucchese, Orlando and Perego[11] authored the only paper found in the literature which stressed the non trivial issue of sample biases in using Twitter data, addressing this with a novel machine learning approach. The study was also one of very few which utilised location information, with the authors constructing a set of 95,627 geo-located tweets matched to 20 Italian regions to study the primary elections of the Italian political party "Partito Democratico". The authors evaluated the performance of a number of "predictors", which functions that produced an estimate of the share of votes that a candidate received. The authors considered a tweet to mention a candidate if the tweet contained a hashtag of the candidates name, and built probabilistic estimators of vote share based on this measure of candidate mentions. Bias correcting weights were learned for each candidate using linear regression in order to minimise:

$$(y_c - w_c \cdot x_c)^2$$

where y_c was the actual share of votes for candidate c, x_c was the predicted share from Twitter, and w_c was the bias correcting weight. The authors found a 15% improvement MAE by using bias correcting factors and found that the correct ranking was able to be predicted in 15 out of 20 regions. An acknowledgement in the use of training data pertaining to the results or polling of the election was made, and the authors conceded that it wouldn't have been possible to present such results before the election took place. Although a claim was made that the weight vectors could be stable, and it may be the case that weight vectors could be learned from previous events or in two-round voting systems the weights could be learned in the first round.

Criticism of Twitter Based Electoral Prediction

The existence of papers proposing that elections can be predicted with Twitter has attracted criticism from a number of researchers. Perhaps the most vocal critic has been Daniel Gayo-Avello who laid out a balanced survey[4] on claims of electoral prediction using twitter with the rather stark conclusion being:

"*No, you cannot predict elections with Twitter.*"

In this survey Gayo-Avello details 8 flaws present in current research into electoral prediction using Twitter, principally among them: there hasn't been a single paper observed which actually "predicts" with all work being post-hoc analysis, sentiment analysis is applied as a black box, all tweets are assumed trustworthy, and perhaps most importantly Twitter is a biased sample, both in it's demographic differences to the electorate, and in the self selection bias present when a minority of politically active users control the majority of political conversation on Twitter. He remarks that any future work into electoral prediction should apply the same methods and attempt to predict an election that hasn't yet occurred, look at adjusting results based on demographic differences, note at the effect of spam/bot accounts on predictive methods, and focus on producing robust and accurate methods of political sentiment analysis before applying them to predictions, remarking that sarcasm and humour often get in the way in state of the art methods.

Gayo-Avello also notes that the "silent majority" are often ignored in papers that have looked at electoral prediction, an effect that was also studied by Mustafaraj, Finn, Whitlock, and Metaxas[20]. The authors compared two groups found in conversations regarding a 2010 Massachusetts Senate Election, the **vocal minority** being users that tweeted very often, and the **silent majority** being users that

tweeted only once. The researchers concluded that content generated by the two groups was significantly different, finding that the vocal minority behaved as a resonance chamber and spread content which aligned with their own opinions.

Data

In this section the datasets used in this work are described, as well as the collection and processing methods used to construct each dataset.

Location Based UK Collection

A rolling collection of tweets was used as the primary source of data, with tweets collected using the Twitter Streaming API, filtered to only include tweets with location information matching a place in the UK. This location information can take the form of a latitude-longitude coordinate pair, or a rectangular area in which the tweet was located to (see Figure 1).

Figure 1: The two types of ways Twitter attaches location information, and the constituencies these locations overlap with.

The collection was trimmed to a time window relevant to the election, from November 1st (the day after the Early Parliamentary General Election Act 2019 was passed into law, and beginning of the run-up to the General Election), to December 12th (the day of the General Election). Overall in this period 7,649,196 unique tweets were found. The window of collection yielded few server outages or periods of disruption, aside from a major loss of data between the 7th of December and the 10th of December, unfortunately coinciding with the last week of General Election campaigning.

User Timeline Collection

Twitter doesn't apply location tagging to retweets, as there is little relevance of tagging another persons content with the sharing users location. Due to this, no

retweets are observed in the rolling location based collection. As retweets can be seen as a strong measure of endorsement, a decision was made to collect a separate body of tweets using the Twitter User Timeline API, which can collect tweets authored or shared by a particular user for a given time window. As well as retweets, any tweets that didn't make it into the original collection (potentially down to lack of location data, or outages) would also be picked up by this method. One drawback to this method of collection was that the API only returns the most recent 3,200 tweets (including retweets) from a user's timeline, and as this collection was started a number of months after the end of the election this meant users which produced a high volume of tweets since December were excluded. Overall tweets and retweets for 342,568 users were able to be collected via this method.

Political Tweets Dataset

Whilst a lot of political content is shared on Twitter, it's not the main topic communicated on the website, and so in order to eliminate noise from the data it was deemed important to develop a method which could classify tweets as political/apolitical. In order to do this, a dataset of tweets labelled as political or apolitical was first needed. A rudimentary UI was built in Jupyter that would display a tweet, as well as buttons labelled "Political" and "Non-Political", and when buttons were pressed the tweet would be added to a `.csv` file, along with the given label. Tweets were sampled from periods that were assumed to have high levels of political discussion - the TV Debates:

Debate	Channel	Date
Johnson vs. Corbyn	ITV	19th November 2019
Climate Crisis Leaders Debate	Channel 4	28th November 2019
Election Debate	BBC	29th November 2019
Prime Ministerial Debate	BBC	6th December 2019

Whilst the classification of text to be political or non-political is a widely subjective matter, particular care was taken as to treat the question more as "relevant/irrelevant to the 2019 General Election" when labeling tweets. Building the dataset in this way gave a large imbalance between the classes as the vast majority of content on Twitter is non-political, so as to ease this balance somewhat a number of tweets

were added which mentioned politicians from each political party of study, as well as from tweets shared by users which only retweeted a single political party. A balanced split between parties was kept in the addition of these tweets, so as to not add bias towards a specific viewpoint. The resulting dataset comprised of 5003 tweets, 2495 of which were labelled political, and 2508 non-political.

Political Figures Dataset

In order to aid in retweet-based classification of users, a list of political figures from each of the 7 parties of study was required. A list of all candidates standing in the 2019 UK General Election was downloaded from Democracy Club, filtered to only include members of the 7 parties of study, and only relevant fields: name, constituency, political party and Twitter username. Any candidate that didn't have a Twitter account was removed from the dataset. The data was augmented to included current MSPs (Members of Scottish Parliament), current MS's (Members of the Senedd/Welsh Parliament) as well as the now previous UK MEPs (Members of the European Parliament). A web scraper was built in Python using the `requests` and `beautifulsoup` libraries to get ahold of data from the Scottish Parliament[2] and Senedd[3] webpages, and data for each MEP was collected manually.

Preprocessing

Tweets taken from the Twitter API are encoded using JavaScript Object Notation (JSON), and these JSON objects contain a lot of erroneous metadata. Inbound JSON objects from the API were therefore filtered to only include specific fields of interest, namely: the unique **id** of the tweet, the tweet content or **text** that comprises the tweet, the **location** and **location name** of where the tweet was published, the **display name**, **screen name**, and **bio** of the publishing user, the **timestamp** denoting the publishing time of the tweet (converted to UNIX epoch time), as well as the **number of retweets** and **likes** a tweet has, and a list of the **hashtags**, **user mentions**, and **urls** that are present in the tweet. When tweet text is used for analysis, the body of the tweet is stored in the original unprocessed format until further analysis takes place in order to have a preserved version of the original tweet. Where retweets are used the body of the retweet, contained in the JSON field `retweeted_status`, is used as otherwise a prefix is tacked on to the body of the format "`RT @user`". Where extra data have been used - for example in

[2]https://www.parliament.scot/msps/current-msps.aspx
[3]https://senedd.wales/en/memhome/Pages/memhome.aspx

the acquisition of retweets - tweets are pulled from the API using the "extended" format, as without this the tweet body is truncated and key information may be cut off.

For text based classifiers, it was necessary to preprocess the main text of the tweets. This process involved: the replacement of newline characters with spaces, the standardisation of punctuation (there can be multiple characters for things such as apostrophes), the removal of urls, the removal of a set of punctuation, replacing emoji with descriptive tokens, replacing percentages with a descriptive token, and the removal of any remaining unicode artefacts. Most of these processes were performed with regular expressions which would replace any found unclean features with either an empty string or a single space. Multiple whitespace resulting from these replacements are removed and replaced with single spaces. Emoji replacement utilises the Python library `demoji`, which also stores a list of descriptions for each emoji that were used to form descriptive token - for instance the emoji: 🤡 would be given the token "`emoji_clown_face`". The symbols `#` and `@` representing hashtags and user mentions respectively were also stripped out, leaving only the content they represent.

Methods

This section will discuss the method's implemented in building a pipeline to turn support for political parties on Twitter into some semblance of "votes". First the method of inferring a UK's voting constituency is discuss, and then a filtering method to get rid of tweets that contain no political content. Two methods are then proposed to infer a user's voting intentions, a social graph based method in which users are clustered in a network based on who they retweet, and a text analysis method to examine the content of the tweets that a user posted. Finally the process of bias correction is looked at, which aims to infer a demographic transformation, bring the political makeup of Twitter closer to that of the electorate.

Constituency Binning

In order to study the relationship between online political sentiment and UK Parliamentary constituencies, a method to map the embedded tweet locations to constituencies was developed. For most tweets, there is no one to one mapping between location and constituency - it is relatively uncommon to have exact GPS coordinates attached to tweets, and broad location rectangles often overlap multiple constituencies.

This overlapping is particularly observed in more metropolitan areas, in which Twitter location rectangles may cover entire cities (e.g. Newcastle) that may be split into several constituencies (e.g. Newcastle upon Tyne Central, Newcastle upon Tyne East and Wallsend, Newcastle upon Tyne North etc.). To deal with this a fractional weighting system was employed in which a tweet may belong to a number of constituencies, with a weight proportional to the overlap of it's location in that constituency.

For a bounding location rectangle b, the initial weighting w that a tweet receives in a constituency c is:

$$w_c = \frac{\text{Area of } b \cap c}{\text{Area of } b}$$

A tweet whose location is attributed to a single coordinate point, or whose location bounding box falls perfectly in a single constituency is given a weighting of 1 in single constituency it falls in. There are some geographic areas on a map that are not attributed to any constituency, for instance bodies of water, which can fall in the original location bounding box but would not be picked up through this weighting

scheme. Due to this, the constituency weights are normalised, effectively giving the probability that the tweet was authored in each of the observed constituencies. In many circumstances the original bounding boxes can fall over large areas of the UK, some covering large areas such as London or even broader regions like Scotland, in these circumstances the tweet will be fractionally weighted to a large number of constituencies with very small probability.

The size of the location bounding box of the tweet is appended after the constituency binning process so that this can later be used as a parameter to fine tune the mapping between tweets and constituencies. The idea of this being that larger bounding boxes may still be useful to observe regional sentiment to political parties - even aggregating only tweets with broad regional locations England/Scotland/Wales one would hope to observe regional disparities in political attitudes, however by looking at bounding boxes of finer granularities, one would hope to find more localised, constituency level differences in political sentiment.

After each tweet has been assigned constituencies with different weightings, a method is needed to aggregate individual tweets some notion of "votes". Some attempts at electoral prediction using Twitter have used a "1 tweet = 1 vote" rule, to add intensity to the sentiment expressed in a constituency, however this work will consider each Twitter user in the dataset as a single elector. The constituency/weight pairs from each tweet a user makes are summed and normalised to get an assigned user location, under the assumption that a user will author a higher proportion of tweets in their home/registered constituency.

Finally, a decision needs to be made in the aggregation of these votes at the constituency level. Fractional weightings can be kept to represent distributed support for a party around a region, for instance if a Labour voter's location was attributed to three constituencies with weights of $\{0.2, 0.3, 0.5\}$ then we can add support to the Labour party in these constituencies with the respective weights. Alternatively, we can use these weights probabilistically to run a number of simulations in which the full vote (of weight 1) is attributed, constituencies are drawn randomly with respect to their weight. This gives a number of benefits in the eventual forecast, principally we are about to determine that in "k out of n simulations, party x won this constituency", whereas by keeping fractional weights we only have a idea of the vote share of each party.

Political Relevancy Filtering

There are over 7,000,000 tweets in the original dataset, before appending retweets and tweets without location tags, so in order to perform sensible analysis on such a large amount of data, content filtering is used to skim off only tweets that are detected to be "political". A machine learning classifier was developed to infer the difference between political and non-political tweets. Labelling could have been further subdivided to capture more nuanced categories, for instance parody, opinion or statement, however the process of labelling data is a time intensive task and more complex labels could introduce further bias, a simple binary labelling was deemed enough for filtering.

In the literature, a Naive Bayes classifier has been used for similar tasks[2] and this was first attempted. After cleaning up text, tweets were vectorised by considering the 2500 most common words and counting their occurences in each tweet. TF-IDF vectorisation was considered, although was judged to have a worse performance for this task. A Bernoulli Naive Bayes classifier was used, with smoothing parameter $\alpha = 0.5$. It was found that training on 75% of the data yielded a classification accuracy of around 86% (averaged over 50 trials). 6 fold cross validation was then used to compute a confusion matrix representing the classification of all data, which was found to be:

$$\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} = \begin{pmatrix} 2208 & 300 \\ 489 & 2006 \end{pmatrix}$$

This model looked to be a fairly robust method of filtering, but to improve results a more complex model was adopted. AdaBoost[13] (Adaptive Boosting) is an ensemble learning algorithm, which trains a number of "weak learners" (in this case Decision trees) and takes a linear combination of these to produce a strong classifier. This algorithm has shown great promise for binary classification tasks, and the process of "Boosting" (reweighting misclassified training samples) is able to reduce bias error. After tuning parameters an AdaBoost model with a learning rate of 0.5 utilising 300 weak learners was able to achieve a classification accuracy of 90% (averaged over 50 trials) after training on 75% of the data - 4% better than that of the Naive Bayes. 6 fold validation was again used to yield a confusion matrix for all data, which was found to be:

$$\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} = \begin{pmatrix} 2379 & 129 \\ 474 & 2021 \end{pmatrix}$$

This implies the new classifier has a slight false negative bias, meaning some political tweets will be lost in the filtering process, but the accuracy observed is considerably better than by using Naive Bayes.

Network Analysis

Inspired by the work by Conover et al.[12] a retweet network based method of political orientation classification was developed. After retweet sets were constructed each user A found to have retweets was iterated over and added to a retweet graph connected to a user B if: A had retweeted B, and the message of user B that was shared was classified as political. Edge weights corresponded to the number of times user A had retweeted a political tweet from B. Without political filtering, other communities such as football fans or TV show discussion would have no doubt arisen during node clustering, and the graph would have been too large for salient and efficient analysis.

The resulting network was incredibly large, encapsulating 300,641 nodes and 1,124,511 edges, and a total of 1,951 out of a total 2,546 users from the politician dataset were observed in the network. Community detection was implemented using the Leiden algorithm[26], an improvement on the popular Louvain algorithm which guarantees well connected communities. Leiden can be seeded with an initial partition, so in order to direct the algorithm towards returning party-political communities an initial layout was constructed which allocated politicians of a certain political party to their own cluster, and all non-politician users to a singleton community.

On a single run the Leiden algorithm returned over 1000 communities, the majority of which had a membership of less than five users. However all runs exhibited clusters in which a majority of a party's politicians would be present, with all but two parties having their own distinct cluster; the Conservative Party and Brexit Party sharing a cluster. In order to give a user a party label, and therefore an inferred vote, 50 runs of community detection with the Leiden algorithm were conducted. If a user fell in a cluster which contained a majority of a party's politicians (of those present in the network) a party label was assigned to that user, the users inferred vote being the most common label after 50 runs.

Figure 2 demonstrates the party-political communities found in the network after community detection with the Leiden algorithm, 6 communities were extracted (one for each party with the exception of the shared Conservative/Brexit cluster) and the induced subgraph on those communities were visualised in Gephi using the Force Atlas 2 algorithm[14]. Figure 3 demonstrates the same process, only for the

Conservative/Brexit and Labour communities, showing a striking resemblance to the Democrat/Republican highly partisan structure observed by Conover et al[12].

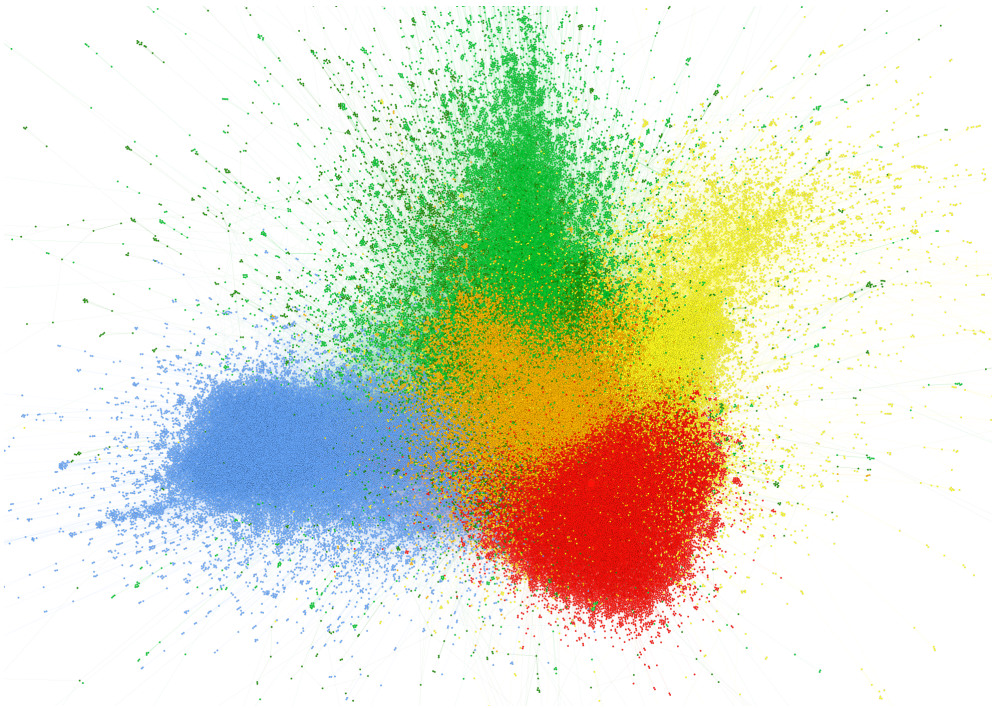


Figure 2: The induced subgraph of party political communities in the retweet network - communities are coloured by political party colour: Conservative/Brexit=Blue, Labour=Red, Liberal Democrat=Orange, SNP=Yellow, Green=Light Green, Plaid Cymru=Dark Green.

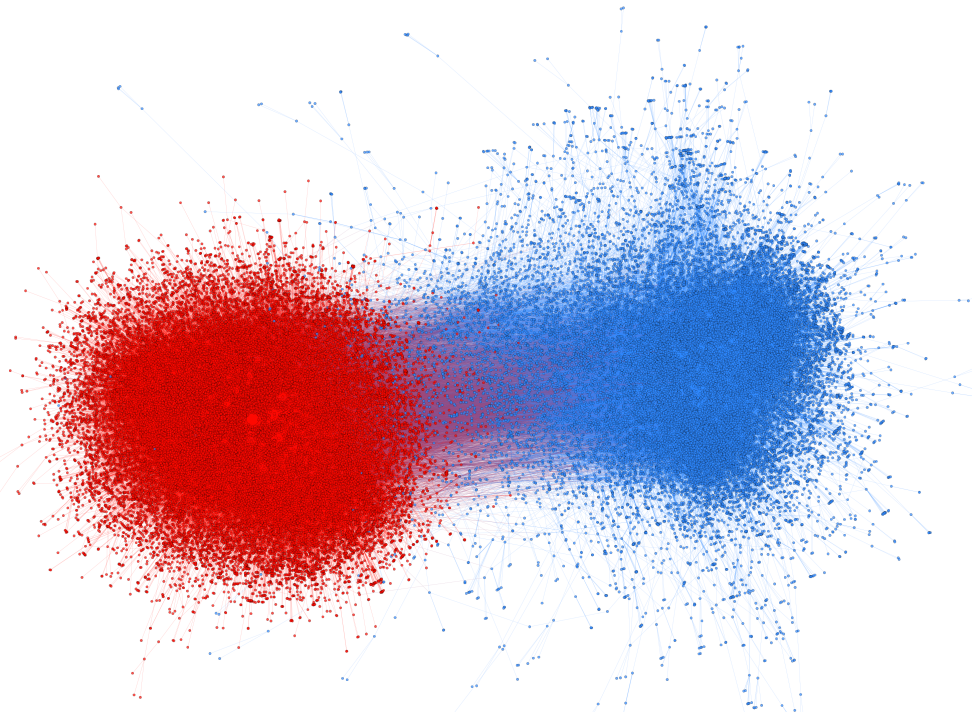


Figure 3: The induced subgraph of Conservative/Brexit and Labour communities, showing a strong partisan structure.

Text Analysis

A method of content based analysis analysis was also developed which took inspiration from the remark made by Rao and Spasojevic[22] that in many text classification scenarios knowing the sentiment of a piece of text isn't always useful enough to draw salient insights. They instead suggest that contextual models should be used, and where Rao and Spasojevic[22] used an LSTM model with word embeddings, here a fastText classifier was used. fastText[15] is an open source text classification model provided by Facebook AI, which is able to return comparable accuracy to deep learning models at a fraction of the training time due to the use of hierarchical softmax in the computation of class probabilities.

To train the fastText classifier a set of tweets labelled with party leaning needed to be constructed. In order to do this a set of users was collected who only retweeted members of a single political party, with an assumption being made that these users political beliefs would align with that party. Then, any tweets or retweets that were classified as political that were shared by these users were assigned the assumed party of the user. This resulted in a set of around 71,240 tweets, 14,993 which were labelled as Liberal Democrat, 12,869 as Green, 12,735 as SNP, 12,439 as Labour, 10,445 as Conservative, 5,232 as Brexit Party and 2,527 labelled as Plaid Cymru. Two of the minority parties, the Brexit Party and Plaid Cymru didn't yield enough data to give a more balanced split between the parties.

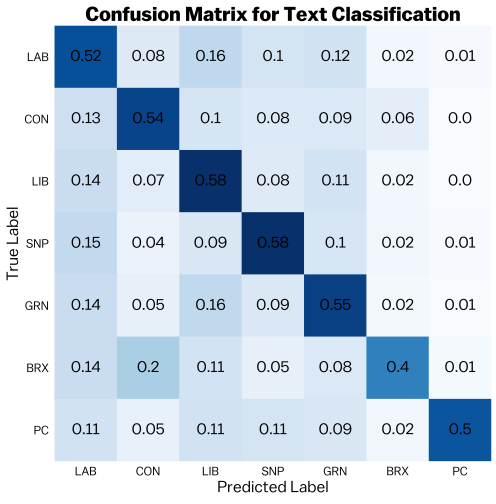


Figure 4: Confusion matrix for results of classifying tweets in the test set with political labels.

Data was preprocessed for input to the fastText model, by appending class labels in the format `__label__PARTY` at the front of the cleaned tweet body. A supervised

fastText model was trained on 70% of the data, and hyperparameters were autotuned by fastText on a validation set containing 20% of the data, with class labels balanced between each split of the data. The model was tested on the remaining 10%, and a confusion matrix was computed demonstrating misclassifications as can be seen in Figure 4. Some examples of correctly classified tweets as well as misclassifications can be found in Table 1.

On it's own the results of text classification seem poor, the Liberal Democrat and SNP classes achieve the best precision with 58% of tweets being correctly classified, and the worst precision was achieved by the Brexit Party class which wasn't even able to classify a majority of the tweets correctly. This is most likely due to a noisy dataset, as the construction of the training set involved assumptions about the partisan orientations of the users, relied on the political filtering to only let through tweets relevant to UK politics, and assumed that all content shared by a partisan user aligned with their assumed party vote. Accuracy is improved by classifying all political tweets that a user makes, producing probability vectors for each tweet, and then taking the highest party score of the geometric median of these vectors. Aggregation using the geometric median rather than a simple average made the method more robust to outliers. To cull any non-political users that may get incorrectly classified, a decision was made to label users with a "`NONE`" label if they shared three or less tweets classified as being political.

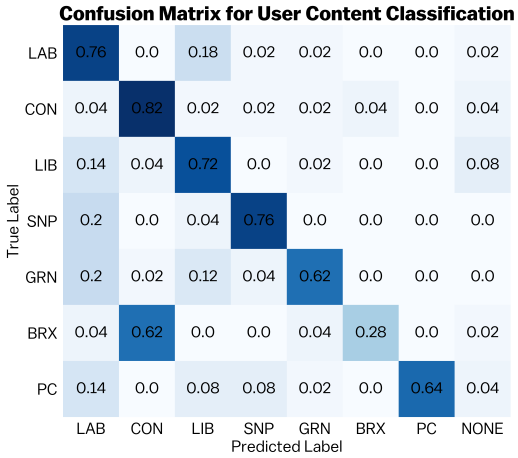


Figure 5: Confusion matrix for results of classifying tweets in the test set with political labels.

Tweet	True Label	Predicted Label
The only way we can move this country forward, and get Parliament working for you again, is with a majority Conservative parliament.	CON	CON
#BBCDebate I wish @BorisJohnson would remind @jeremycorbyn of the origin of austerity: three terms of labour overspend.	CON	BRX
I will always stand for the many not the few no matter which category I fall into ♥ We need look after each other & the planet ♥	LAB	LAB
THE 👏 NEXT 👏 PRIME 👏 MINISTER 👏	LAB	CON
This isn't hard. If you're frustrated by two deeply flawed parties, there is a proper alternative. A party led by a young, internationalist, progressive woman who will stop Brexit and take action on the climate crisis. I'll be voting for @joswinson and the @LibDems	LIB	LIB
It's official: Labour are throwing Remainers under the big red bus	LIB	LAB
There's not a single Tory/Labour marginal in Scotland. Only @theSNP can beat the Tories in Scotland. Voting Labour will help the Tories.	SNP	SNP
First Minister @NicolaSturgeon has marked #WorldAIDSDay with a message to commemorate those who have lost their lives to AIDS and to explain the steps Scotland is taking to stop the spread of HIV.	SNP	GRN
.@TheGreenParty is standing down in many seats to help Labour win them. Not a word of thanks from @UKLabour. Instead Labour is trying to unseat the UK's sole Green MP @CarolineLucas in Brighton. So selfish & sectarian!	GRN	GRN
"You want to get out of the European Union but you're still happy to take your salary as an MEP" @jon_bartley calls our Nigel Farage's hypocrisy #bbcqt	GRN	LIB
We can't trust the Tories to deliver Brexit on their own. Vote for The Brexit Party today!	BRX	BRX
FINALLY hidden on the BBC Website factual pushback against Bojo's Brexit Lies. He must be challenged over this. This is the Brexmas Election and yet every time Boris talks about his 'great new deal' and 'quick FTA' everyone just swallows it.	BRX	CON
Ireland isn't perfect, but there's not even a sad-bloke-in-a-bedroom Facebook group that wants to go back under Westminster rule. Wonder why? Maybe this "Mother of Parliaments" isn't such a good thing after all. Wales - independence is better	PC	PC
#IndyWales is now in the mainstream of Welsh politics. And it's because of the thousands who've marched this year, the hundreds who have leafleted, everyone who's done banners on bridges. We are the ones growing this movement - keep it going! 👍	PC	SNP

Table 1: Examples of tweets correctly and incorrectly classified with political affiliation.

A set of users were found who retweeted multiple parties, retweeting more one than other - 50 users fitting this description were found for each party. Each users tweets were classified and aggregated as described to produce a single party label, or their inferred vote. The results of this can be found in the confusion matrix shown in Figure 5, which shown an improvement in precision across all classes except for the Brexit Party, in which a majority of tweets are misclassified as relating to the Conservative Party.

Bias Correction

Many works in the literature discuss the inherent bias present in Twitter data, with a number of researchers[18] claiming that this is the primary reason that Twitter cannot be used for electoral prediction. It was assumed that significant bias would be present in the data used for this study, and this was later confirmed in the results (discussed in the next section). With such bias present, even an 100% accurate measure of assessing a Twitter user's voting intention and location would not translate into a vote share or seat count reflective of the real results. This is due to the fact that: a number of Twitter users will not vote, a number of voters will not tweet/exist on the platform, and the demography of Twitter users does not reflect that of the electorate. In a study on the political representativeness of Twitter users in the 2012 US Presidential Election, Barbera and Rivero[5] found that Twitter users that tend to tweet about politics tend to be males who live in urban areas, and their political ideologies tend to be more extreme than that of the electorate, whilst in a survey of Twitter users conducted during the 2013 Italian General Election Vaccari et al.[28] found that users were younger, better educated, male and left wing. Mellon and Prosser[17] studied the demographics and political representativeness of British Twitter users, and concluded that there may be hope for electoral prediction methods utilising Twitter, so long as they control for demographic differences through appropriate weighting methods.

To investigate this claim, and test whether localised political sentiment can be observed on Twitter, as well as probing if there is any efficacy in Twitters use as an instrument for electoral prediction, an attempt was made at learning a vector of bias correcting weights to correct the party share found at the constituency level. The hope is that this weighting vector will act as a demographic transformation from Twitter to the electorate. A 7-dimensional weighting vector is aimed to be learned through an optimisation process, each element of the vector corresponding to a weight which will be multiplied with the party vote in each constituency when results are calculated. In order to do this a random sample of 50% of the 631 (ex-

cluding the constituency of the Speaker seeking re-election) British constituencies is taken, and party share vectors are constructed from the Twitter results found in each constituency as well as the ground truth results, with elements of the party share vectors organised in the following way:

$$\begin{bmatrix} x_{\text{CON}} & x_{\text{LAB}} & x_{\text{LIB}} & x_{\text{SNP}} & x_{\text{GRN}} & x_{\text{BRX}} & x_{\text{PC}} \end{bmatrix}$$

A decision was made to not normalise the ground truth results, as this could lead to large spikes which could throw off the optimisation process. For instance in East Devon, the second place candidate was an Independent receiving 40.4% of the vote, so a vector of party share without normalisation would look like this:

$$\begin{bmatrix} 0.508 & 0.045 & 0.028 & 0 & 0.011 & 0 & 0 \end{bmatrix}$$

and with normalisation like this:

$$\begin{bmatrix} 0.858 & 0.076 & 0.047 & 0 & 0.019 & 0 & 0 \end{bmatrix}$$

By normalising vectors for anomalous constituencies such as these, party support can be vastly over or under inflated. Simulated Dual Annealing was used to solve an optimisation problem, in order to minimise the total squared error across the constituencies sampled between the corrected vote share and real vote share, and the bias weighting vector was constricted to only values between 0 and 2, with the annealing process initialised with a vector of all ones. There is an obvious drawback to this line of study, in that this could clearly not be conducted during an election. The ground truth constituency vote shares were used to fit the vector, and so without knowing this beforehand it's not possible to apply this process mid-election period. Coletto et al.[11] make an acknowledgement of this in their work, but state that once learned, the weighting vector should assumedly stay fairly stable given no major changes in the demographic makeup of Twitter, the electorate and without any major political shake-up.

Results

Data Preparation

Of the x tweets found in the election period, y of them were classified as political after the filtering step. Figure 6 shows an hour by hour breakdown of the tweet volume over the election period, as well as the number of tweets that were retained after political filtering. Some spikes can be seen in the overall tweet count which relate to major events throughout the election period, with four particularly noticeable spikes in the political tweet volume being labelled on the plot. Through a visual inspection of tweets cross referenced with news articles, these were found to be: (1) the spike on November 19th: the first head-to-head debate between Boris Johnson and Jeremy Corbyn[4], (2) the spike on November 22nd: "BBC Question Time leaders special" in which members of the public were able to question the party leaders on TV[5], (3) the spike on December 6th: "The Prime Ministerial Debate" another BBC election broadcast in which Boris Johnson and Jeremy Corbyn went head-to-head for the second time[6], and (4) the spike on December 12th which corresponded with the date of the election[7].

[4]BBC News: General Election 2019: First head-to-head debate on 19 November https://www.bbc.co.uk/news/election-2019-50268753

[5]The Guardian: BBC Question Time leaders special: who came out on top? https://www.theguardian.com/politics/2019/nov/22/question-time-leaders-special-who-won-johnson-corbyn-swinson-sturgeon

[6]BBC News: General election 2019: Boris Johnson and Jeremy Corbyn clash over Brexit in BBC debate https://www.bbc.co.uk/news/election-2019-50681321

[7]BBC News: UK set for 12 December general election after MPs' vote https://www.bbc.co.uk/news/uk-politics-50229318

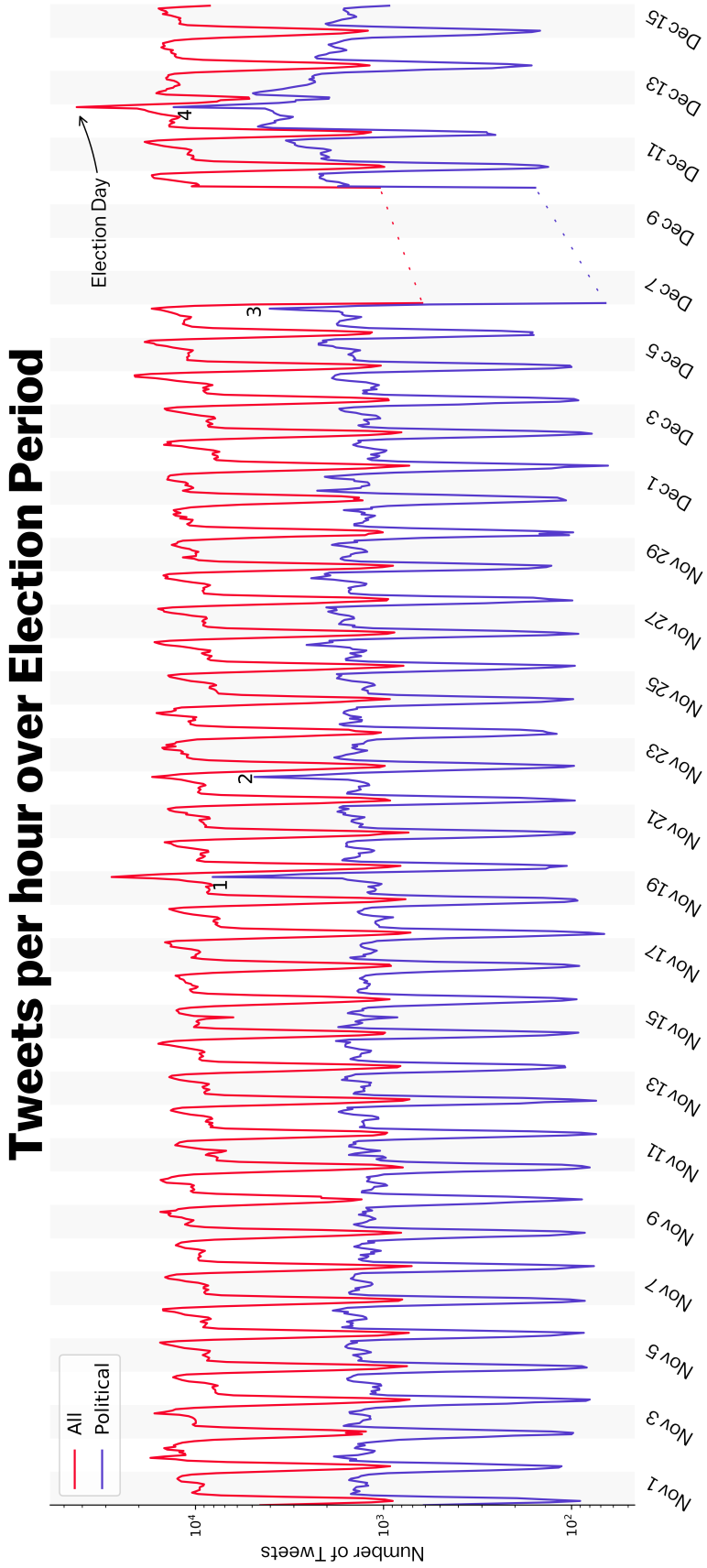


Figure 6: Number of tweets collected per hour during the election period, as well as the number of tweets classified as political per hour.

Political Classification

Figure 7 shows the share of votes found using each method, with the actual vote share each party achieved in the General Election added for reference. An additional method not discussed in the results features here, which will be termed the "retweet method". This simply gave a user a label of the political party that they retweeted the most, with ties broken randomly. Out of the three methods used to infer votes, the network clustering method assigned the most votes to users, with 58,096 users being assigned a vote, followed by the text analysis method with 45,127 users assigned a vote, and lastly the retweet based method in which 40,387 users were assigned a vote.

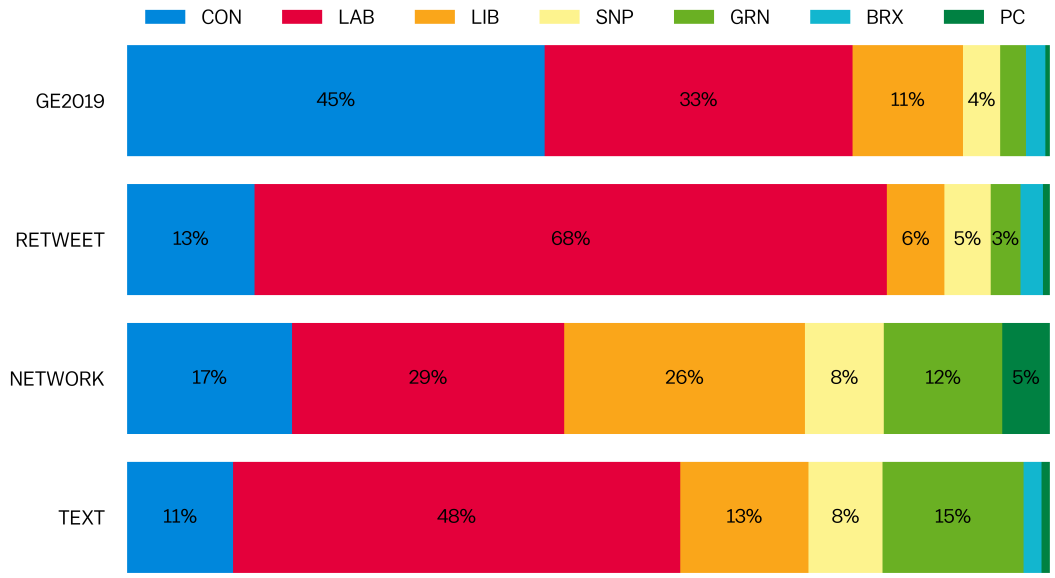


Figure 7: The results of political filtering on all tweets over the election period.

Bias Correction

From the results of the political classification vote shares, demonstrated in Figure 7, a clear bias can be observed. In the retweet and text methods this bias inflated Labour Party support, and in the network based method deflated Conservative and Labour Party support and seemed to give the minority parties a bigger voice. Bias can also be observed at the constituency level, and Figure 8 shows scatter plots for the vote share observed across constituencies using an example execution of the retweet based method vs. the real vote share for the Labour and Conservative

parties. Table 2 shows the calculated Pearson coefficient between computed vote share and true vote share for the Conservative and Labour parties, for each method and it's bias corrected counterpart. Positive correlations were observed across all methods for both parties, with a weaker positive correlation observed for Labour vote share using the text analysis method. The retweet and text based methods saw an increase in correlative strength when bias corrective was applied, although the network based method saw a decrease. The strongest correlations were observed for the retweet based method with bias correction.

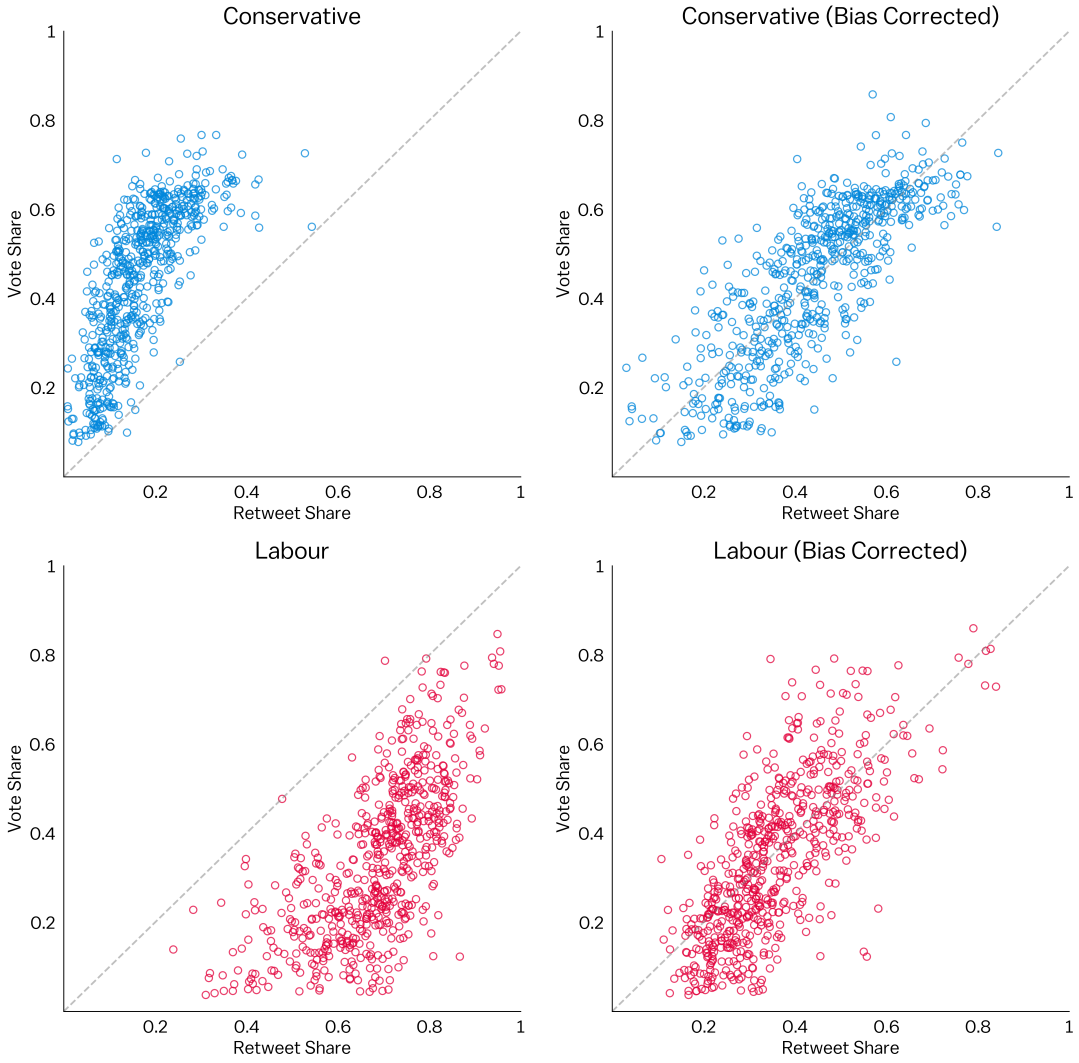


Figure 8: Scatter plots showing the results computed at the constituency level for the Labour and Conservative parties, along with the results after bias correction. The grey dotted line on each plot corresponds to the perfect calibration that a model could have - in which each computed vote share is perfectly reflective of the real vote share.

Method	Conservative Vote	Labour Vote
Retweet	0.756	0.670
Retweet (Bias Corrected)	**0.797**	**0.731**
Network	0.706	0.639
Network (Bias Corrected)	0.658	0.659
Text	0.702	0.549
Text (Bias Corrected)	0.760	0.665

Table 2: Pearson coefficients calculated for Conservative and Labour computer vote shares against the true vote shares they received at the constituency level.

Bias correcting vectors found for in various runs of the methods are as follows; the retweet based method:

$$\begin{bmatrix}1.464 & 0.245 & 0.922 & 0.539 & 0.438 & 0.802 & 0.394\end{bmatrix}$$

the network based method:

$$\begin{bmatrix}1.806 & 0.951 & 0.403 & 0.637 & 0.208 & 0.799 & 0.222\end{bmatrix}$$

and the text based method:

$$\begin{bmatrix}1.126 & 0.222 & 0.280 & 0.290 & 0.074 & 0.561 & 0.182\end{bmatrix}$$

It can be observed that Conservative support was inflated and Labour support deflated across all methods, and all methods also damped the vote share of minority parties.

Constituency Level Results

Figure 10 shows maps demonstrating the result of combining users inferred party vote with inferred location, with a user's vote being applied fractionally to each location they were observed in. A map of the seats won by each party in the General Election is provided in the Appendix for comparison. Each method is demonstrated in the figure, with and without bias correction, and in each map, the constituency in which the Speaker was seeking re-election was automatically marked as a victory for the Speaker, and was hence left out of any further analysis. Assigned user locations were computed by filtering out the upper quartile of distinct bbox sizes, as this was found to provide the best results over a limited search range. All three maps without

Method	MAE	Seat Accuracy	Top-2 Seat Accuracy	Correct Party Rankings (%)
Retweet	0.161	0.352	0.923	0.155
Retweet (Bias Corrected)	**0.061**	**0.802**	**0.984**	**0.498**
Text	0.152	0.396	0.641	0.024
Text (Bias Corrected)	0.066	0.786	0.978	0.464
Network	0.150	0.439	0.688	0.068
Network (Bias Corrected)	0.066	0.781	0.978	0.369

Table 3: A number of metrics calculated for each method with and without bias correction.

bias correction give a political landscape unrepresentative of the United Kingdom, drastically overstating Labour's support and understanding Conservative support. Regional differences in political support are still observed across the Scottish border for methods with and without bias correction, yet retweet and text methods without bias correction overstate Labour support in Scotland. The network based method without bias support dramatically overstates support for minority parties, which was also observed in the total vote share in Figure 7. Localised party support for methods with bias correction seem to reflect that of the real constituency wins than their non-corrected counterparts. A number of metrics were also computed for each example execution, for each method, which can be seen in Table 3. The metrics computed were the MAE or Mean Absolute Error between the computed vote share and real vote share for each party which stood in each constituency, the seat accuracy; or percentage of seats in which the correct outcome was computed, the top-2 seat accuracy; or the percentage of seats in which the correct outcome and the percentage of how many constituencies the rank of parties by computed vote share matched the real rank.

It can be observed that the bias corrected retweet method outperformed the other methods in every category, with the other bias corrected methods only slightly behind in some. The best method was able to infer over 80% of seat results correctly, as well as almost 99% of seats in which either of the top 2 computed parties won. This corresponded to only 10 seats in which the top 2 predicted parties didn't win in the constituency, out of those 10 seats in 7 the real result was Liberal Democrat, 1 was Green, 1 was Plaid Cymru and 1 was Conservative.

In order to determine that bias correction wasn't overfitting, and that there was some representativeness in the underlying data that was exploited with the bias correcting

Metric	No Change	N_0	N_1
MAE	**0.061**	0.076	0.091
Seat Accuracy	**0.802**	0.737	0.624
Top-2 Seat Accuracy	**0.984**	0.975	0.906
Correct Party Rankings	**0.498**	0.437	0.392

Table 4: Computed metrics for Null Models vs. data with no changes made.

method, two null models were constructed with the bias corrected retweet based method applied. Null results were constructed as follows: in the first null model (dubbed N_0) half of the user locations were swapped before votes were assigned to constituencies, and in the second null model (dubbed N_1) in which half of the seat results were swapped. Both of these took place before bias correction, to verify that bias correction was not the only thing responsible for overall vote share/seat counts comparable to the real election and that some element of localised political sentiment was at play in the data. Table 4 shows the results of computing metrics on the null models, and comparisons to bias correction on retweet data with no changes, and Figure 9 shows maps of resulting seats for the two null models.

The results show a worse performance in all metrics for the two null models, with better results for N_0.

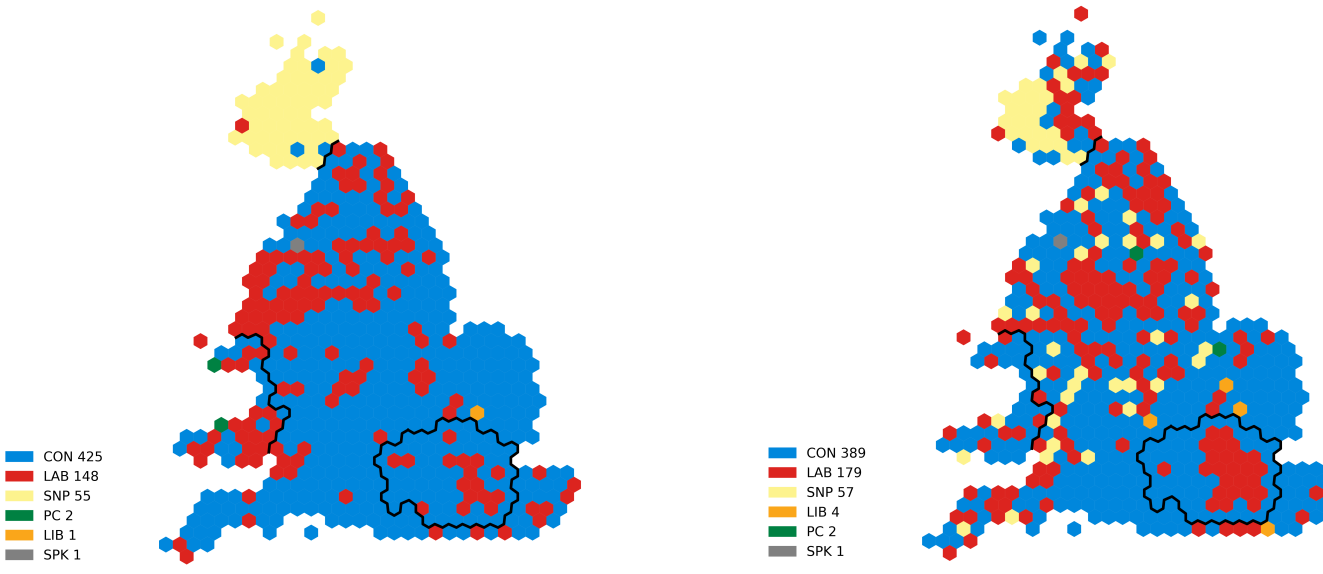


Figure 9: Computed seat maps for the two null models, Left: N_0, Right: N_1.

Finally results were obtained by applying an alternative vote counting process, in

which a user's vote is assigned in totality to a single constituency with the probability determined by the probability in which the users location was found in that constituency. This gave the added benefit that probabilities could by inferred for each constituency, rather than a simple share. Figure 11 shows a histogram of the total seats computed by the bias corrected retweet method for the Conservative and Labour parties across 1000 runs, in which a new bias correcting weight vector was computed from the fractional results every 10 runs. The figure shows that an outright majority (a total number of seats greater than 326) was obtained by the Conservative party in every single simulation, and that the peaks of each distribution fall closely to the (marked) real seat count that each party received.

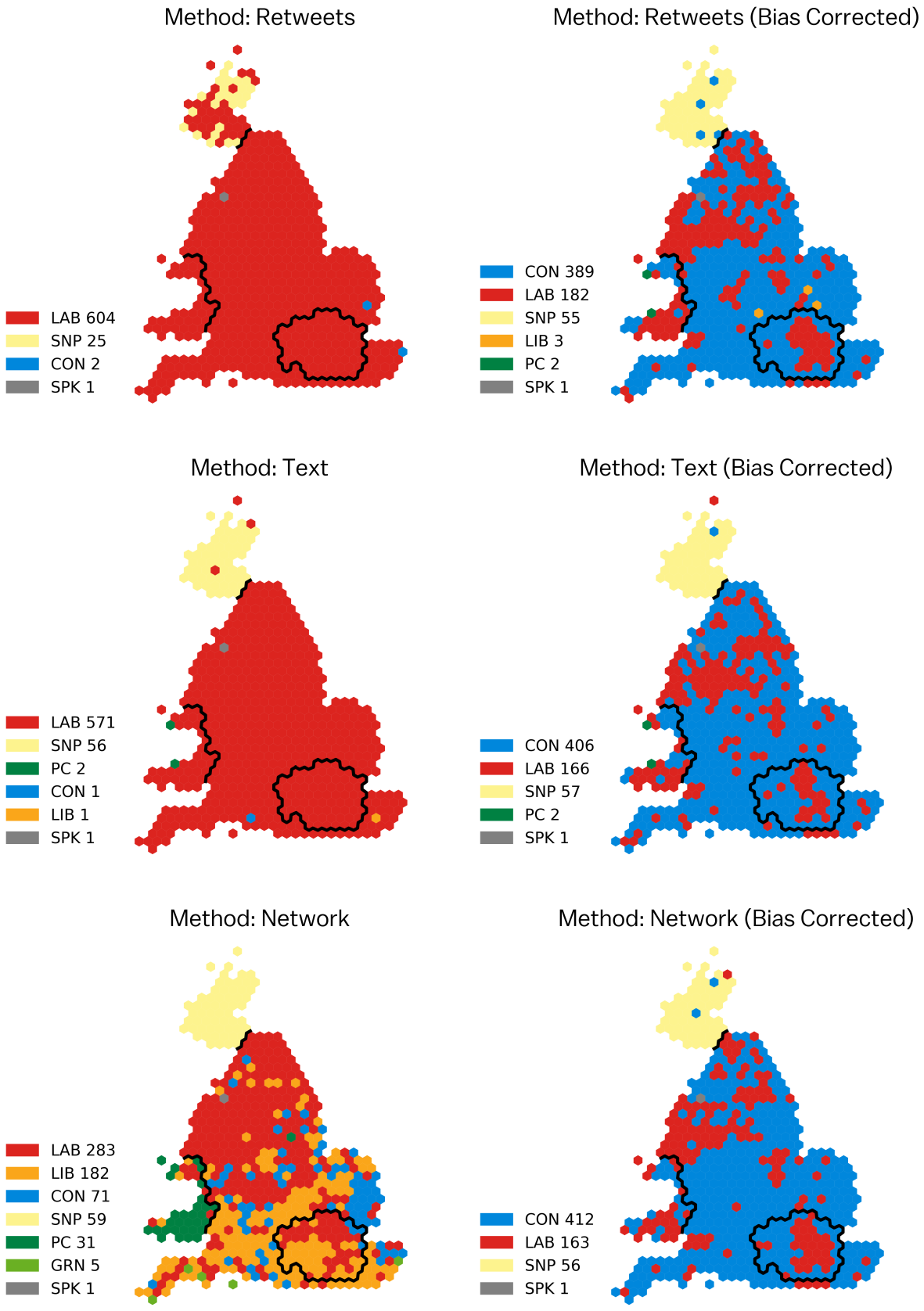


Figure 10: Election maps for the various methods.

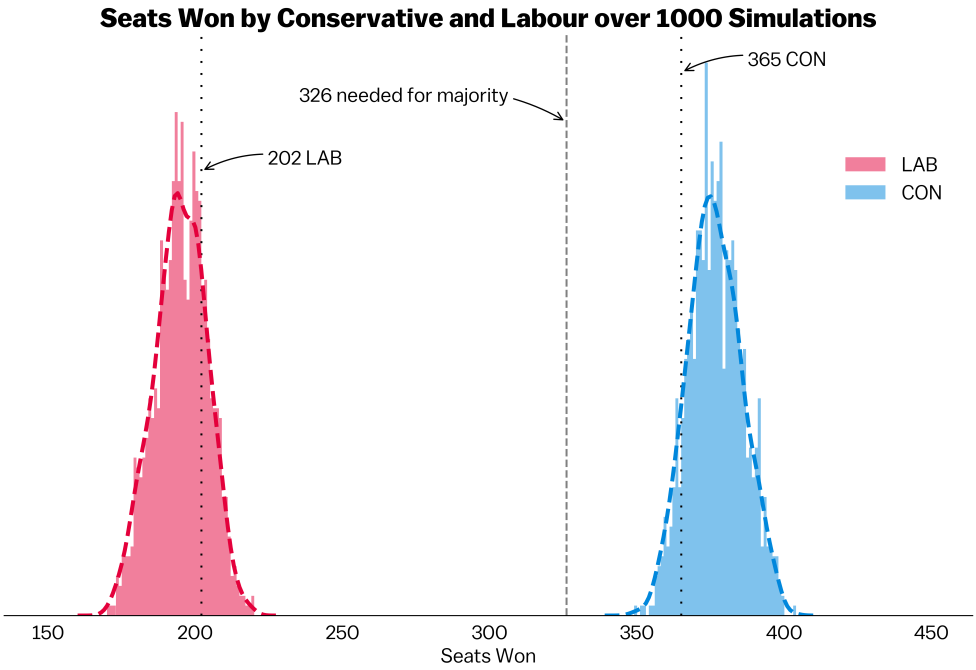


Figure 11: Histogram of seats obtained by the Conservative and Labour party after running 1000 simulations sorting users into different constituencies stochastically.

Red Wall Seats

There is no standarised definition of the red wall, a popular term in British political analysis first coined in a tweet[8] by James Kanagasooriam, but it seems to have come to be known as the collection of Labour seats, lost in the 2019 Election predominantly falling in the North of England or Midlands, which had a high pro-Brexit vote in the 2016 EU Referendum. To standardise the definition for this work, we take the red wall to be any seat Labour lost in the 2019 Election that fell in: the North East/West of England, the East/West Midlands as well as Yorkshire and the Humber - this results in a collection of 44 Parliamentary constituencies, the full list of which can be found in the Appendix.

[8]https://twitter.com/JamesKanag/status/1161639307536457730

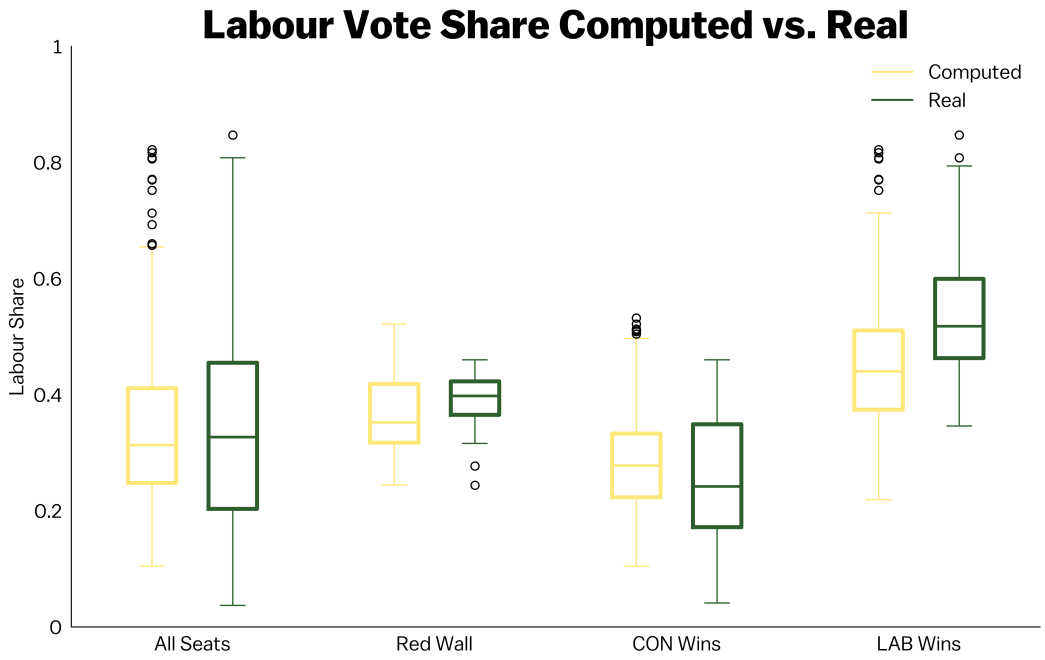


Figure 12: The share of Labour share found in various groupings of constituencies..

Over 100 runs, the retweet method with bias correction was used to compute the result in red wall seats, with a different bias correcting vector computed each time and the resultant average accuracy being 73%.

Figure 12 shows Labour vote share across various groupings of seats: all constituencies, our definition of the red wall, all real Conservative wins and all real Labour wins. Seats in the red wall were found on average to have less Labour support than Labour wins, more than Conservative wins, and slightly more than all constituencies (the majority of which were Conservative wins), this aligns with the ground truth results. Median computed vote share in red wall results after bias correction was also found to be lower than the median of the real vote share.

Discussion

The results of using AdaBoost[13] for political filtering seems to have offered promising results, although it's hard to exactly quantify. There are a number of minor spikes in the graph of political tweet volume which don't correspond to a spike in overall tweet volume, suggesting at the very least the filtering process isn't just a reflection on the number of tweets authored that hour. Furthermore the spikes do seem to line up with relevant political events (as demonstrated in the results), with the most noticeable being the election in which the peak hourly political tweet count was comparable to hourly tweet counts for all tweets across various times during the election period. A better way to evaluate political filtering would be to demonstrate the loss in accuracy presented when methods didn't filter for only political content, for instance in the construction of the network. However, with the amount of data processed in this work the constructed network would be too large, and these results would likely be unattainable.

The vote share obtained from political classification, whilst wildly different to the real results proved explainable. The network based method classifies a user as supporting a party if they appear in the same community as a majority of politicians, and here an assumption is made that clustering in this way gave fully partisan clusters, although other variables are likely to be at play. The cluster labelled as Plaid Cymru featured more Labour politicians than Plaid Cymru politicians, yet was classified as a Plaid Cymru cluster as it contained a higher proportion of all Plaid Cymru political figures than Labour figures. It is suspected that regional retweeting patterns have influenced the formation of clusters, and that the Plaid Cymru community would be more accurately represented as a "Welsh Politics" cluster. Further evidence can be seen for this, as the leader of the Welsh Labour Party is assigned to the Plaid cluster rather than Labour's. The same seems to happen with the SNP cluster, in which a number of politicians from other parties feature (including the Twitter account for the Scottish Green Party), yet the majority of SNP members fell into the community and so it was labelled as SNP. This seems to serve as a baseline explanation for an over-inflation of minority party support, especially felt by the regional parties. Perhaps in a future work to counter this regional networks should be instead considered, which would potentially be more reflective of party support in Scotland and Wales, although a drawback of this is that it would lead to smaller snapshots of the true retweet network, and this could affect accuracy. Moreover, the networked method was able to classify more users than any other method, something which was at first assumed would be a strength but may have actually resulted in the misclassification of non-political users. For

instance, a user may have only retweeted a certain newsreader, or radio host that was retweeted more by members of a certain party, and if the shared tweets were able to loosely pass through the political filter, then such a user would be given a disingenuous inferred political label.

Inspiration for the network classifier was drawn from the paper by Conover et al.[12] in which users were marked as "left" or "right", groupings far more polarising than those found when classifying users in a multi-party system. Due to this, the methodology of the classifier relies on the principle of homophily, in the hope that well connected groups will form clusters of like minded political support. In multi party systems in which some parties hold an affinity for each other, often like minded political beliefs are found across parties, with information passed across the divide between partisan clusters. For this reason the Brexit Party wasn't picked up by this classifier at all; the patterns of retweets by supporters of the Brexit Party were judged to be so similar to that of Conservative Party supporters that they were assigned the same cluster. Also, without the method which seeded the classifier with an initial layout in which party politicians belonged to separate partition, the largest or second largest cluster was often found to be a mix of broadly, pro-EU centre left parties that can be described as the "progressive alliance", and included the then leader of the Liberal Democrats, as well as several well known figures from the Labour party. It could be argued that British politics has become more partisan and fragmented, yet in many cases less explicitly party political, and it seems this is why the network based classifier proved to have the worst performance of the three.

The text and retweet based methods, whilst still appearing unrepresentative of the real spectrum of political support, seem to possess less non-demographic differences, which Mellon and Prosser[17] suggest lead to errors when reweighting. This is observed in the results for bias correction (see Table 2), in which bias correcting weights make the correlations between online share and real share stronger for retweets and text, yet weaker for the network method. Without significant reconfigurations it seems like a networked approach as laid out in this work is unsuitable for political inference in the scope of a task such as this. In regards to the text classification method, it seems there are still inherent sub-biases that would need to be ironed out before suitability for a task such as this, the group with most noticeable over-inflation being the Green Party. This could be due to pronounced misclassification in tweets relating to Climate Change, which was an especially pronounced topic throughout the General Election[9]. It's assumed that many issues attributed to the text analysis method could be fixed using cleaner datasets, but it was found that good labelled datasets are difficult and time consuming to build.

[9]https://www.bbc.co.uk/news/election-2019-50594593

Overall the retweet method looks to have provided the most robust methodology as a predictor for constituency vote share, total seat counts, predicting a users political allegiances, and thus inferred vote. This isn't particularly surprising as it's arguably the most simple and effective indicator of endorsement on Twitter. Retweeting a political account/politician without tacking on other text through a quote tweet, means that the tweet will be shared with a users followers without any further comment, or argument, meaning that the user must have some level of affinity with the message/source in order to share it. Using a comprehensive dataset of political figures from each party seems to have proved robust, although without comparison to a more constricted version of the politician user set it's difficult to comment. Restricting the politician set to only party leaders and official party accounts could prove to lessen the effects of the "long tail"[20] that may be present using this retweet method, in which the average Twitter user will probably not recognise the vast majority of political figures present in the seed set.

To answer the questions that were laid out at the beginning of this work, firstly: the political environment found on Twitter does seem to exhibit a left-skew, this was found in each method aimed to infer a user's voting intentions, even with the exception that the network based method deflated Labour support, which seemed to be made up for in the inflated support of the centre/centre-left Liberal Democrat party, centre-left SNP, left wing Green Party and left/centre-left Plaid Cymru. Without bias correction the localised support in terms of the winning party was incredibly far from that of reality, but the null model analysis (see Table 4) seemed to suggest underlying localised political support present in the data, which was able to be exploited with bias correction, rather than bias correction being entirely responsible for end results being comparable to reality.

Secondly, it seems to have been possible to observe decreased support for Labour in red wall constituencies. This would be a much more easier claim to prove (or disprove) if enough data could be recovered for tweets and retweets from the 2017 Election for comparison, but without this an attempt was made at comparative analysis of red wall seats against other groups of constituencies, which found decreased levels of support of corrected Labour vote share in red wall constituencies than real vote share, with the median computed red wall vote share found to be marginally closer to the share in Conservative won seats, as opposed to Labour won seats. Repeated runs of various methods also were able to predict red wall seats as won by the Conservative Party with a high degree of accuracy.

Finally, the efficacy of a bias correcting method not only seems viable, but integral to a process of electoral prediction using Twitter, and the results of bias correction seem to have been positive in when applied to datasets that didn't feature large

amounts of non-demographic differences. A higher pearson coefficient was observed post bias correction in this cases as well as lower MAE and better seat accuracy, seemingly implying that the corrective process brings results closer to reality.

Conclusion and Future Work

In this work, an investigation into the efficacy of Twitter as an instrument of electoral prediction was conducted, in which novel methods for political inference in a multi-party framework were developed. Firstly a political filtering method was developed using AdaBoost[13] and was seen to provide robust results on validation sets. Then three predictors of inferring a Twitter users vote were established, one using a content based analysis of what a user tweeted, one using a metric of which party the user most retweeted utilising a large set of British political figures, and one using a network based analysis, clustering users on a retweet graph. To the author's best knowledge, this is the first work which has combined traditional methods of inferring political orientation, with methods taken from social sensing to infer a user's location in order to combine this in the production of a seat by seat vote count. With the wealth of tweets marked with location data available to researchers, it's surely vital that any further work tackling the possibility of electoral prediction using Twitter take advantage of location data, especially in cases in which a simple global vote share is often not indicative of the result - notably the United Kingdom's First Past The Post electoral method, and the United States' Electoral College. This is also the first work that the author is aware of that extended methods developed for bias correcting factors to study a multi-party General Election.

Overall positive results were gathered, but it must be stressed that a prediction was never produced, merely that hope was given to the cause of building a predictive system using Twitter. Seat by seat election counts produced in this work had no information baked in of the previous election history in each constituency, incumbency, constituency demographics, tactical voting methods/party affinities or even of the parties that stood in each constituency, and were still able to produce a seat count with an 80% accuracy to the ground truth. It is hoped that this work may show that the problem of electoral prediction using Twitter may be relaxed to the rather inexpensive collection of a located user's retweets, and the inference of a vector of bias correcting weights, which could potentially be seeded with other traditional polling, or be produced using demographic samples.

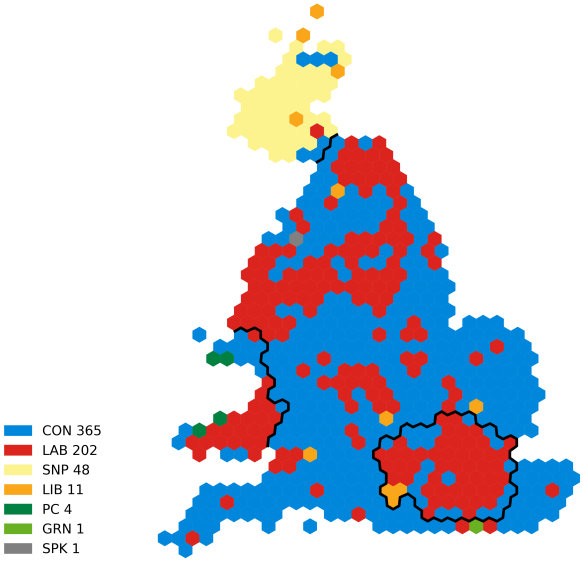
It is of the belief of the author that major improvements in the field of electoral prediction can be undertaken, firstly by leveraging more state of the art methods in areas such as location inference and text analysis than this work was able to, but secondly by going further with the process of bias correction. It could potentially be possible to infer more demographic attributes about a user, which could lead to demographic sampling methods being utilised that are already used in election forecasting, such as Multilevel Regression and Post-stratification. Text analysis

methods could also be improved using the current state of the art developments in fields such as Aspect Based Sentiment Analysis, as the attempt in this work was only able to capture the party affiliation a piece of text may have had, when in many cases targeted sentiment for/against a party/politician would be more useful.

Finally, by following the doctrine of Gayo-Avello[4] who makes a number of incredibly pertinent recommendations for future work into electoral prediction using Twitter, it's hoped that new opportunities and discoveries in the field could arise. The integration of a number of these approaches could hopefully open up a number of interesting new research challenges.

Appendix

Election Map of Real Results



Red Wall Seats

Ashfield
Barrow and Furness
Bassetlaw
Birmingham, Northfield
Bishop Auckland
Blackpool South
Blyth Valley
Bolsover
Bolton North East
Burnley
Bury North
Bury South
Colne Valley
Crewe and Nantwich
Darlington
Derby North
Dewsbury
Don Valley
Dudley North
North West Durham
Gedling
Great Grimsby
Heywood and Middleton
High Peak
Hyndburn
Keighley
Leigh
Lincoln
Newcastle-under-Lyme
Penistone and Stocksbridge
Redcar
Rother Valley
Scunthorpe
Sedgefield
Stockton South
Stoke-on-Trent Central
Stoke-on-Trent North
Wakefield
Warrington South
West Bromwich East
West Bromwich West
Wolverhampton North East
Wolverhampton South West
Workington

2

References

[1] Nasser Alsaedi, Pete Burnap, and Omer Rana. Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology*, 17:1–26, 03 2017.

[2] Rudy Arthur, Chris A. Boulton, Humphrey Shotton, and Hywel T. P. Williams. Social sensing of floods in the uk. *PLOS ONE*, 13(1):1–18, 01 2018.

[3] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *CoRR*, abs / 1003.5699, 2010.

[4] Daniel Gayo - Avello. "i wanted to predict elections with twitter and all i got was this lousy paper" - a balanced survey on election prediction using twitter data. *CoRR*, abs / 1204.6441, 2012.

[5] Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6):712–729, 2015.

[6] Johan Bollen, Huina Mao, and Xiao - Jun Zeng. Twitter mood predicts the stock market. *CoRR*, abs / 1010.3003, 2010.

[7] Chris Boulton, Humphfrey Shotton, and Hywel Williams. Using social media to detect and locate wildfires. 04 2016.

[8] Pete Burnap, Rachel K. Gibson, Luke Sloan, Rosalynd Southern, and Matthew L. Williams. 140 characters to victory?: Using twitter to predict the uk 2015 general election. 05 2015.

[9] Michael Cogley. The rise of parler, team trump's new favourite social network, 2020.

[10] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 91–99, 01 2013.

[11] Mauro Coletto, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Electoral predictions with twitter: A machine-learning approach. *CEUR Workshop Proceedings*, 1404, 01 2015.

[12] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. 01 2011.

[13] Yoav Freund and Robert Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9:1–12, 06 2014.

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[16] Vasileios Lampos, Tijl Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. volume 6323, pages 599–602, 09 2010.

[17] Jonathan Mellon and Christopher Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[18] Panagiotis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. pages 165–171, 10 2011.

[19] Katy Minshall. Uk election conversation attracts over 15m tweets, 2019.

[20] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int'l Conference on Social Computing (SocialCom)*, pages 103–110. IEEE Computer Society, 10 2011.

[21] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: User classification in twitter. pages 430–438, 08 2011.

[22] Adithya Rao and Nemanja Spasojevic. Actionable and political text classification using word embeddings and LSTM. *CoRR*, abs / 1607.02501, 2016.

[23] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. pages 851–860, 01 2010.

[24] A. Schulz, Aristotelis Hadjakos, Heiko Paulheim, J. Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. 01 2013.

[25] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*, 18(3):74–84, 2013.

[26] Vincent Traag, L. Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 03 2019.

[27] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. volume 10, 01 2010.

[28] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T. Jost, Jonathan Nagler, and Joshua Tucker. Social media and political communication: A survey of twitter users during the 2013 italian general election. *Rivista Italiana di Scienza Politica*, 43(3):381–410, 12 2013.

[29] Dong Wang, T. Abdelzaher, and Lance Kaplan. Social sensing: Building reliable systems on unreliable data. *Social Sensing: Building Reliable Systems on Unreliable Data*, pages 1–213, 01 2015.

[30] Faiyaz Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. 01 2012.