BREXIT: PREDICTING THE BREXIT UK ELECTION RESULTS BY CONSTITUENCY USING TWITTER LOCATION BASED SENTIMENT AND MACHINE LEARNING

James Usher and Pierpaolo Dondio

School of Computing Technological University Dublin, Dublin, Ireland

ABSTRACT

After parliament failed to approve his revised version of the 'Withdrawal Agreement', UK Prime Minister Boris Johnson called a snap general election in October 2019 to capitalise on his growing support to 'Get Brexit Done'. Johnson's belief was that he had enough support countrywide to gain a majority to push his Brexit mandate through parliament based on a parliamentary seat majority strategy. The increased availability of large-scale Twitter data provides rich information for the study of constituency dynamics. In Twitter, the location of tweets can be identified by the GPS and the location field. This provides a mechanism for location-based sentiment analysis which is the use of natural language processing or machine learning algorithms to extract, identify, or distinguish the sentiment content of a tweet (in our case), according to the location of origin of said tweet. This paper examines location-based Twitter sentiment for UK constituencies per country and aims to understand if location-based Twitter sentiment majorities per UK constituencies could determine the outcome of the UK Brexit election. Tweets are gathered from the whisperings of the UK Brexit election on September 4th 2019 until polling day, 12th December 2019. A Naive Bayes classification algorithm is applied to assess political public Twitter sentiment. We identify the sentiment of Twitter users per constituency per country towards the political parties' mandate on Brexit and plot our findings for visualisation. We compare the grouping of location-based sentiment per constituency for each of the four UK countries to the final Brexit election first party results per constituency to determine the accuracy of location-based sentiment in determining the Brexit election result. Our results indicate that location-based sentiment had the single biggest effect on constituency result predictions in Northern Ireland and Scotland and a marginal effect on Wales base constituencies whilst there was no significant prediction accuracy to England's constituencies. Decision tree, neural network, and Naïve Bayes machine learning algorithms are then created to forecast the election results per constituency using location-based sentiment and constituency-based data from the UK electorate at national level. The predictive accuracy of the machine learning models was compared comprehensively to a computed-baseline model. The comparison results show that the machine learning models outperformed the baseline model predicting Brexit Election constituency results at national level showing an accuracy rate of 97.87%, 95.74 and 93.62% respectively. The results indicate that location-based sentiment is a useful variable in predicting elections.

KEYWORDS

BREXIT Election, Twitter, Sentiment, UK Election.

David C. Wyld et al. (Eds): CMLA, CSEIT, NETCoM, NLPD, GRAPH-HOC, WiMoNE, CIoT, NCS - 2022
pp. 01-16, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121101

1. INTRODUCTION

One area that has experienced an increase in use of Twitter is that of electoral campaigning and political strategy formulation. With the increasing prominence of Twitter as a political communication tool, politicians and political parties now maintain an active presence on same. Twitter provides an optional static data field in the user profile which allows the user to provide their location. Twitter users have an option to fill in this field thus providing their location. Mobile devices now pick up the user location from GPS coordinates and provide a location coordinate for the user to choose from a dropdown menu also. In Twitter, tweets can be posted with the location address field which identifies the user's current position. These geographical tweets, so to speak, with text content have been utilised to detect real-time events, such as estimating Typhoon trajectory or Earthquake location [1]. Sentiment analysis of Twitter messages, is the act of retrieving opinions from tweets. Twitter users express sentiments about specific topics or entities with different strengths and intensities, where these sentiments are strongly related to their personal feelings and emotions. Computational sentiment analysis methods attempt to measure different opinion dimensions. By classifying polarity estimation using Natural Language Processing (NLP) into three polarity classes namely, positive, negative, and neutral or supervised and unsupervised machine learning algorithms fulfil the objective of classification. Sentiment analysis has been accomplished in a variety of genres of communication, including professional, media-like news articles [2], web forums [3,4] and Facebook [5,6]. The growth in sentiment analysis has projected itself to politics, as political strategist and research firms pursue the valuable opinions of large populations to help formulate political strategies. Twitter sentiment principally has become a widely explored foundation for election forecasting [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21], due to instance data availability in conjunction with the extensive use of Twitter globally.

This paper examines location-based Twitter sentiment per UK constituencies and aims to understand if location-based Twitter sentiment majorities per UK constituencies could determine the outcome of the UK Brexit election from the period of September 4th 2019 until polling day December 12th 2019. We build a model for classifying "tweets" into binary features for classifying positive or negative sentiment based on location. We build three machine learning models a decision tree, neural network and Naïve Bayes model using location-based sentiment and constituency-based data from the UK electorate to predict the Brexit election results by first party. In section II, we present the findings of our literature review. In section III, we examine the political landscape, the datasets, python machine learning libraries, data filtering, visualisation, location modelling, bassline modelling, prediction and evaluation. Section IV looks at the results of each model and section V concludes and looks at future research.

2. RELATED WORK

Election forecasting using a 'Twitter Tracker' for the Irish General Election 2011 was explored by [8] allowed users, and journalists, to tap into the content on Twitter pertaining to the election through an accessible dashboard-style interface. The approach assumed that the percentage of votes that a party receives is related to the volume of related content in social media. Larger parties will have more members, more candidates and will attract more attention during the election campaign. Smaller parties, likewise, will have a much smaller presence. Volume was based on the measure as the proportional share of party mentions in a set of tweets for a given time period. They found that Twitter does appear to display a predictive quality which is marginally augmented by the inclusion of sentiment analysis. [15] election predictions use a similar method whereby the prediction is based on the number of times the name of a candidate is mentioned in tweets prior to elections. The approach was successful in predicting the winner of

the Venezuelan, Paraguayan and Ecuadorian Presidential elections held in Latin America during the months of February through April 2013. These findings contrast severely with [14] who found that simple methods for predicting election results based on sentiment analysis of tweets text are no better than random classifiers. They recommend that, in order to improve the accuracy of sentiment analysis, a method is needed to go beyond the reliance on word polarity alone. Pre-processing techniques such as POS tagging and word sense disambiguation might be necessary, as well as the inclusion of non-lexical features. Similarly, [16,20] found that party mentions had no relevance to the predicting the outcome of the German elections 2009. [9] computed the number of Twitter messages referring to a particular political party as an indication of the eventual winner. The analysis achieved an 86% classification accuracy. [10] analysed the on-line popularity of Italian political leaders throughout 2011, the voting intention of French internet-users in both the 2012 Presidential ballot and subsequent Legislative election, and found a remarkable ability of social-media to forecast on average electoral results. Findings also uncovered sentiment analysis of social media seems to provide more accurate predictions when focusing on the most popular leaders or on mainstream parties.

Twitter sentiment can be used also for other electoral purposes such that [17] examined the political preference of voters using Twitter and found that Twitter-generated content and user behavior during the election campaigns contain useful knowledge that can be used for predicting the political preference of those users. In addition, they showed the predicted preference changes over time and that these changes co-occur with campaign-related events. This type of analysis is quite useful, too, when taking into account the controversial 350m Brexit Bus claim by PM Boris Johnson [25], perceived as a deliberate attempt to swing voters to the support the Conservative mandate of 'Getting Brexit Done'. The benefits of monitoring allows party strategists to measure the reaction of campaign-related events via Twitter sentiment and the polls and tailor responses accordingly. This is further demonstrated by [19]. However, [18] finds that not all Twitter sentiment corresponds to the poll's predictions. Alternative election prediction methods and concepts do exist such that [26] examined the use of forecasting a Conservative Party victory through the pound using ARIMA and Facebook's Prophet. [27] successfully used location-based Twitter sentiment to predict the US presidential elections of 2016 and UK general elections of 2017. The study extracted location data provided by users from tweet meta-data and this was used to plot state-wise subjectivity and polarity on a map of the US. For the UK elections, tweets using two different filters (keywords and geo-location) were plotted for visualisation. Findings showed that sentiment based on location did reflect on-ground public opinion. In UK elections, the Labour party performed better than expected and also had a more positive sentiment on Twitter. The study observed that tweets mentioning Donald Trump had higher subjectivity than ones discussing Hillary Clinton. The study concluded that the ability to map user sentiment provided tremendous benefit in accurately predicting public opinion as this allows distinction of user opinions based on geographical location. [28] performed location-based sentiment analysis on 650,000 tweets in order to understand trends and patterns regarding the Indian elections. Discoveries saw both positive and negative sentiment change from one location to the other and 'social events' can trigger a sharp rise in both negative and positive sentiments regarding a political party.

3. DATA COLLECTION

3.1. Political Climate

In the countdown to the Brexit election on 12th December 2019, UK Prime Minister Boris Johnson and his Conservative Party vouched to leave the EU with the 'Withdrawal Agreement' settled. Labour pledged to renegotiate the 'Withdrawal Agreement' although accepting Brexit

and held a referendum, letting voters choose between the renegotiated Withdrawal deal and remaining in the EU. While Labour's election strategy early on was to emphasise that the vote was about more than Brexit, the party changed its focus. The message was that Labour's leadership was not opposing Brexit. By opposing Mr. Johnson's deal, it wanted to find what it believed to be a better one [22]. The Liberal Democrats guaranteed to revoke Article 50, while the SNP proposed to hold a second Brexit referendum, however, revoking Article 50 if the alternative was a no-deal exit. The DUP and the Brexit party supported the Conservative party stance to 'Get Brexit Done'. Plaid Cymru and the Green Party supported a second Brexit referendum, supporting the belief that the UK should stay in the EU.

The incumbent Conservative government Brexit strategist, Domnic Cummings, was known to have spent a considerable amount of time using artificial intelligence to tackle Brexit activities [23]. In his infamous 'weirdos and misfits' blog, he made reference to statistical and ML forecasting [24]. In what appeared to be a daring move to call a snap election, it appears feasible that Prime Minister Boris Johnson's strategists may have conducted their own location-based sentiment analysis to determine if the country backed his 'Get Brexit Done' mandate. Conventional wisdom would suggest that through user location information, a more accurate understanding of the on-ground location public opinion would be advantageous. By consuming location data, a political strategist could candidly gauge the support levels for candidates and policies for each region or constituency, thus, giving a more detailed picture of public opinion ultimately defining political strategy.

Hypothesis 1: Can Twitter location-based sentiment per constituency predict the outcome of the Brexit Election?

3.2. Brexit Twitter Election Dataset

The dataset is filtered to contain tweets from September 4th until 12th December 2019. The Brexit Election Twitter dataset contains over 7.3 million individual tweets collected daily within said period. Each tweet is identified by a tweet identifier, the date-time-seconds of the submission (GMT), location, verified indictor, the text content, #Hastag, number of followers, twitterhandle, and a sentiment score derived from the Native Bayes machine learning algorithm which ranges from -1 = negative. 0 = neutral and 1= positive. The data is stored on a remote server which houses an SQL LITE database to store the retrieved data. The server environment consists of an 8 Core Intel R Xeon (R) CPU E5, a 2630 v4 2.20 GHz Intel processor with 16 GB RAM, 400GB memory and a 64-bit Windows 10 Operating System.

3.3. Westminster Parliamentary Constituency Dataset

This dataset contains all of the results of the Westminster Parliamentary Constituencies for the Brexit Election 2019 [29]. The dataset contains the ons_id, ons_region_id (which act as an identifier for the constituency within the British Isles), the result by First party (i.e. the party that won the seat with the most votes) using the first past the post voting system, country, region name, constituency name, majority votes, electorate valid_votes, invalid_votes, majority and gender of candidate.

3.4. Twitter API

The most common way to access Twitter data is through the Twitter REST API. Using the secure tokens obtained via the Oauth process, this provides authentication and thus allows the user to receive the requested Twitter data. We utilise the Twitter API, the python Tweepy library, the

python Naïve Bayes textblob library, and we use a SQL LITE database as a repository for the Twitter data.

3.5. Python Tweepy Library

We utilise the "Tweepy" python library to accept the Twitter data by creating a tweepy.py file [30] Twitter offers several streaming endpoints, each customised to certain use cases. These streams are categorised as follows:

- Public Streams: Streams of the public data flowing through Twitter. Suitable for following specific users or topic or data mining.
- User Streams: Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
- Site Streams: The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

For this study, we are concerned with Brexit public streams. We customise our Stream Listener API from the Tweepy library to capture the incoming tweets from the Brexit #Hashtags contained in Table I.

Table 1. Hashtags

#Hashtags	***Sample Tweets***
#Brexit	*RT @IsolatedBrit: To avoid a fascist revolution. That's why we're leaving the EU? #Brexit*
#BrexitChaos	*No-deal Brexit could put public at risk, warns Met chief #Brexit #BrexitChaos #BrexitCrisis*
#BrexitShambles	*New Labour Leader desperately needed. Preference would be Yvette Cooper, David Lammy or Chuka #BrexitShambles*

3.6. Python Naives Bayes using TextBlob

Naive Bayes is a straightforward model for classification. It has been proven to be effective in text categorisation. *Text blob* employs a multinomial Naive Bayes classifier, where the assumption is that each feature is conditional independent to other features given the class. Bayes theorem is illustrated in equation (1)

$$P\left(\frac{C}{T}\right) = \frac{P\,(C)\;\; P\,(T/C)}{P\,(T)} \tag{1}$$

where *c* is a specific class, in our context either *positive* sentiment or *negative* sentiment, and *t* is a tweet text we want to classify. *P(c)* and *P(t)* is the prior probabilities of a sentiment class *c* and a text *t*. *P(t |c)* is the probability the text appears given this class. The goal is choosing value of *c* to maximise *P(c|t)*. Using a *bag of words* approach, each text *t* can be represented as a vector features $\{w_1, w_2, .., w_n\}$m, where $w_i$represent the occurrency of each word w_i in the text *t*, usually weighted. Therefore $P(w_i|c)$ is the probability of the i^{th} feature in text *t* appears given class *c*. In the Naive Bayes approach, each feature w_i is independent from each other. Therefore $P(w_i|c)$ is the probability of the i^{th} feature in text *t* appears given class *c*. In order to classify a text *t*, we

need to compute the maximum likelihood estimation of each one. When making prediction for a new text *t*, we calculate the log likelihood $\log P(c) + \sum_i \log P(w_i|c)$ of different classes, and take the class with highest log likelihood as prediction.

3.7. Data Filtering

The objective of data filtering is reducing the noise from the Twitter dataset concerning neutral sentiment and non-UK constituency locations. We have collected 7,332,842 tweets between September 4th 2019 and December 12th 2019. We are only interested in the 650 Westminster Parliamentary Constituencies in the UK. To reduce the convolution between the domestic tweets (users who live in the election constituencies) and non-domestic tweets (tweets outside elections constituencies), we can apply geo-location filtering: users whose tweets originate from the election constituency are included in the prediction model inclusive of positive and negative location-based sentiment, the remainder are ignored. Figure 1 illustrates the unique values from the Twitter dataset and shows that there are 120,530 locations.

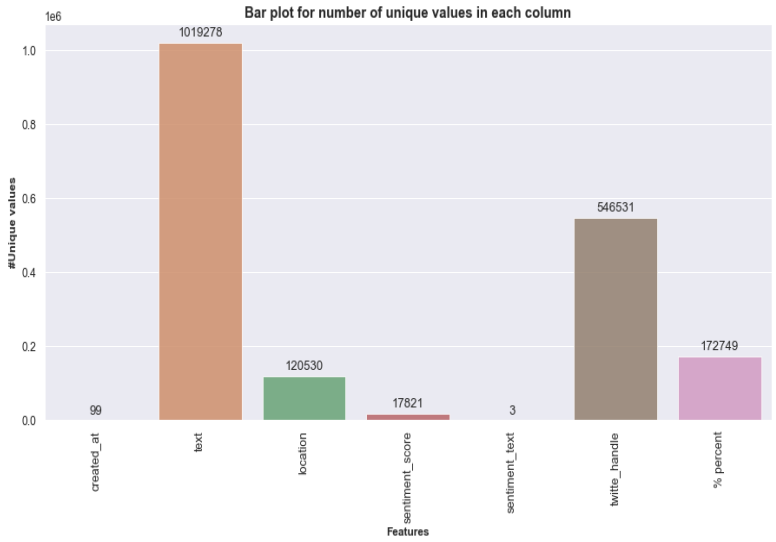

Figure 1. Bar Plot for Unique Values

Looking closer at the ratio on location we can deduce from the Twitter dataset that 61.2% have no location inserted. Furthermore, users have input generic locations such as London, England, Scotland, United Kingdom. Figure 2 illustrates same.

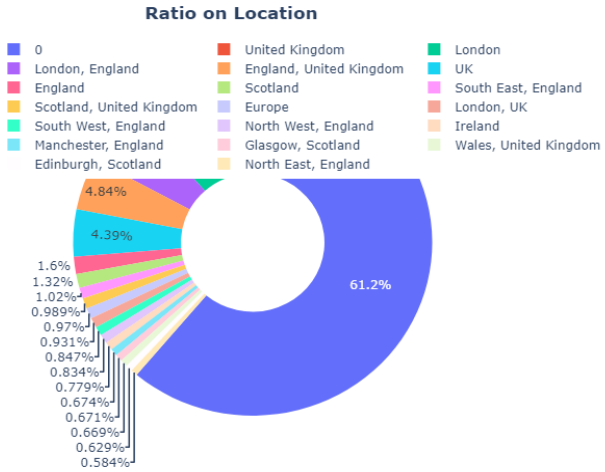

Figure 2. Ratio on Location for No Location

To eliminate non-domestic locations, we merge our Twitter locations with the 'Westminster Parliamentary Constituencies' to filter out the noise, essentially, removing all non-Westminster Parliamentary Constituencies'. Figure 3 shows that we matched 469 constituencies out of 650 for the four UK countries from the Twitter dataset.

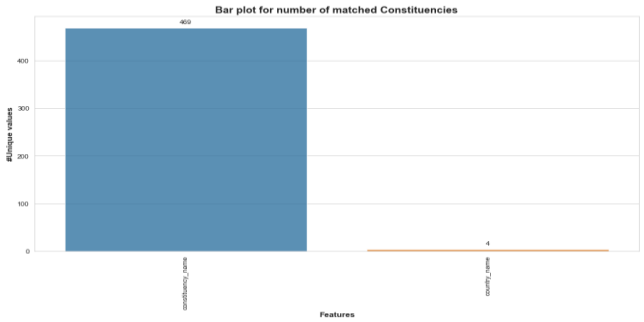

Figure 3. Bar Plot for number of matched constituencies

We show the top 10 ratios by location for Westminster Parliamentary Constituencies in Figure 4 by percentage of tweets after completing the filtering process.

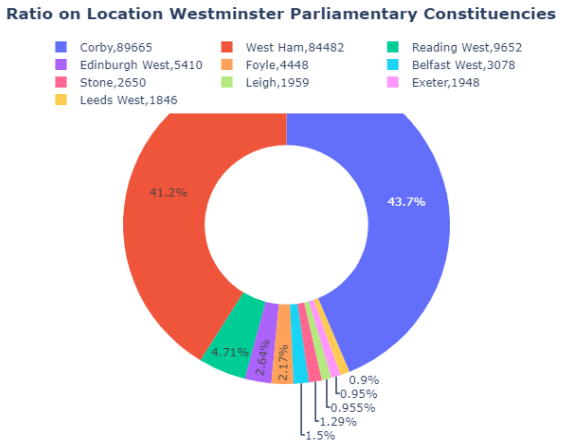

Figure 4. Ratio on Location for Westminster Parliamentary Constituencies.

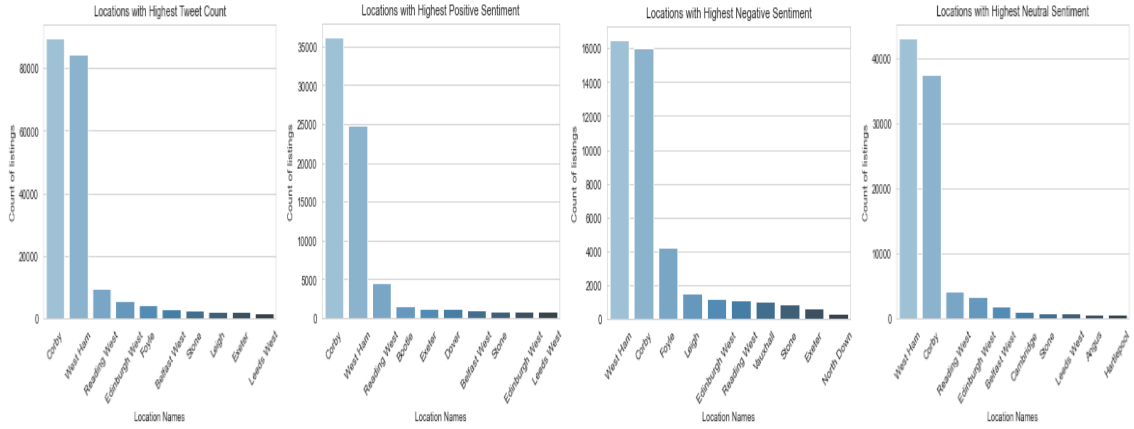

Figure 5. Sentiment Breakdown for Westminster Parliamentary Constituencies

The breakdown of location-based sentiment is illustrated in figure 5. Interestingly, from an English perspective Corby and West Ham feature predominately as locations with high tweet activity; they represent almost 84% of total tweet activity and score highly in positive, negative and neutral tweet sentiment. Corby is a Conversative constituency and West Ham is a Labour constituency. The Scottish constituency of Edinburgh West, which is Liberal Democrats party territory opposed to Brexit, captures almost 0.95% of tweet activity. From a Northern Ireland perspective, Belfast West and Foyle are Sinn Fein and SDLP party respective constituencies both parties opposed to Brexit. Both Northern Ireland constituencies represent 3.67% of the total tweet activity.

3.8. Data Visualisation and Location Modelling

UK choropleths were obtained from the Open Geography portal from the Office for National Statistics (ONS) to create each of the UK maps [31] The matplotlib python library was used to create each visualisation in addition to the 'Westminster Parliamentary Constituencies' data provided by the House of Commons library. The UK choropleths are imported into a static web page which takes the data from the 'Westminster Parliamentary Constituencies' dataset and plots the constituencies. Sentiment is established by computing a value count majority for each constituency from the Twitter dataset. The value count majority is then identified as the majority

sentiment for the particular constituency and applied accordingly to the constituencies for each of the visualisations.

3.9. Baseline Model

Exploratory Data Analysis (EDA) and Linear Discriminant Analysis (LDA), which is a supervised machine learning technique used to find a linear combination of features that separates two or more classes of objects, is undertaken to understand the vote distribution between the UK Brexiteer and Bremain constituencies. The results of which indicate that there is an imbalance, that is, where the class distribution is not equal or close to equal and, is instead, biased or skewed. We see from the target distribution and the linear discriminate analysis that there is an overwhelming majority of UK constituencies in favor of 'Getting Brexit Done'.

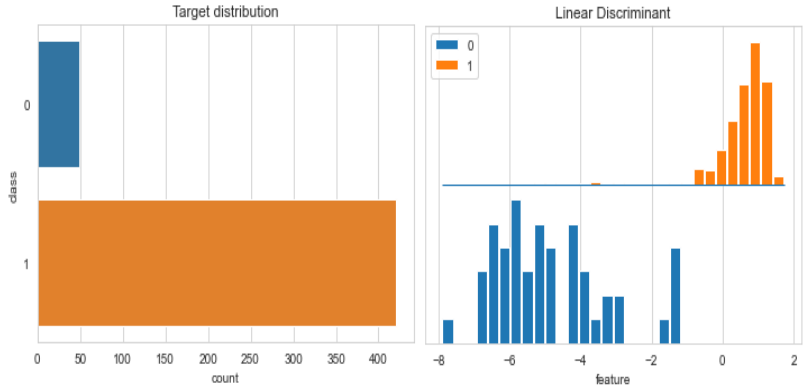

Figure 6. Target and LDA Analysis

Knowing that the EDA and LDA have identified an imbalanced classification, building a baseline metric to evaluate our models would prove meaningful. Particularly in imbalanced classification models, it can appear that the final model isn't really doing much better than guessing. Hence, we need to establish what accuracy is adequate to call our model significant. We import Sklearn dummy classifier library and use the "uniform" strategy which generates predictions uniformly at random. The objective of balancing classes is to either increase the frequency of the minority class or decrease the frequency of the majority class. This is done in order to acquire approximately the same number of instances for both the classes. Baseline accuracy is computed without location-based sentiment and is established for each of the UK countries as illustrated in table 2. Baseline accuracy is also computed for the combined UK constituencies at national level for comparable assessment to all machine learning models.

Table 2. Baseline Metrics

Country	Method	Baseline Accuracy
NI	Uniform	.40
Scotland	Uniform	.44
Wales	Uniform	.32
England	Uniform	.45
UK	Uniform	.54

3.10. Predicting the Brexit Election

To determine the sentiment of a constituency *k*, each tweet that contains the constituency's location, as derived from the location field, is considered a vote for that constituency *k*. If a tweet contains positive or negative sentiment, it counts as a vote towards constituency *k* or, otherwise, it is ignored. The defining vote per constituency *k* is defined by the majority sentiment indicator be that positive or negative sentiment such that positive sentiment is a vote to 'Get Brexit Done' as defined by the 'Brexiteer' parties. This represents the Brexit political mandates of the Conversative, Labour, DUP and the Brexit Party (also known as UKIP). Negative sentiment is a vote to reject Brexit as defined by the 'Bremainer' parties. This is indicative of the Liberal Democrats, Green Party, Plaid Cymru, Sinn Fein, SNP and the smaller independent parties. We pre-process the tweets by removing emoticons. The tweets are turned into word vectors and a standard Naive Bayes classifier setup is employed for classification as outlined in section 3.6. Each visualisation illustrates the Brexit constituency sentiment per 'First Party' result (i.e., the first party past the post or the first party to get the required constituency vote majority, coloured by Brexit stance) and the associated location-based sentiment predicted result split out per country such that the Brexiteer parties are denoted in blue and Bremainer parties are denoted in red.

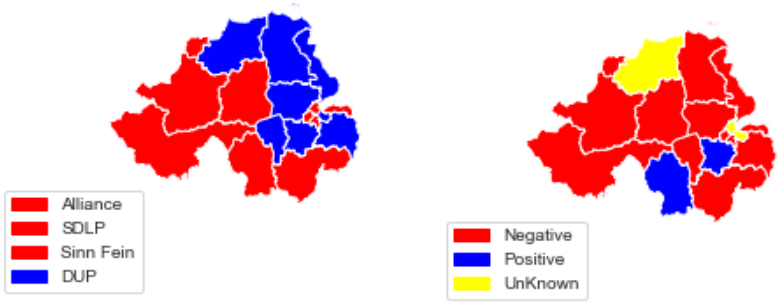

Figure 7. Northern Ireland Location based sentiment

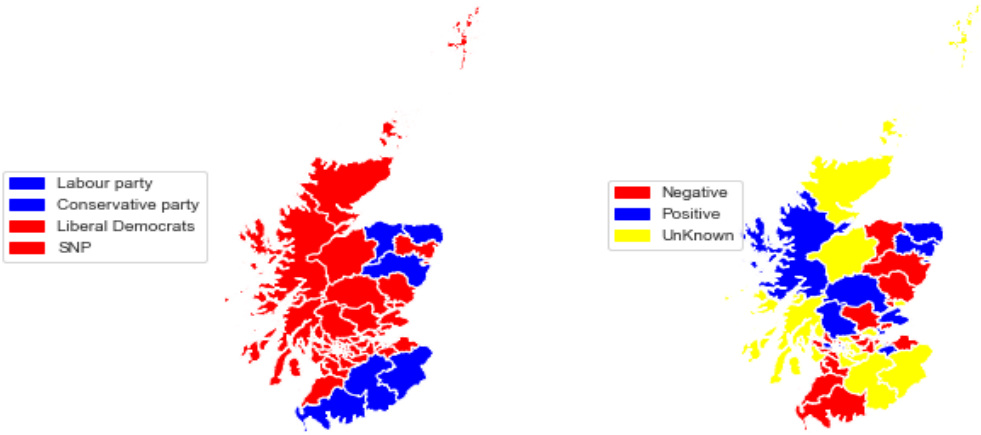

Figure 8. Scotland Location based sentiment

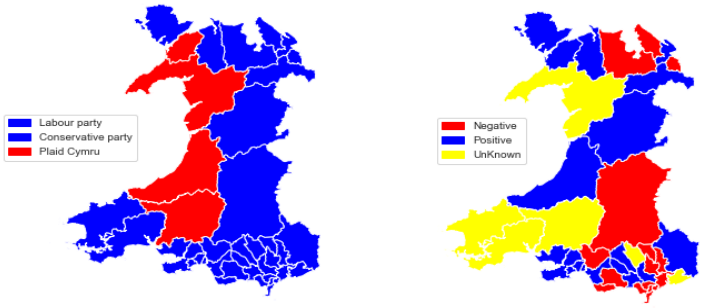

Figure 9. Wales Location-based Sentiment

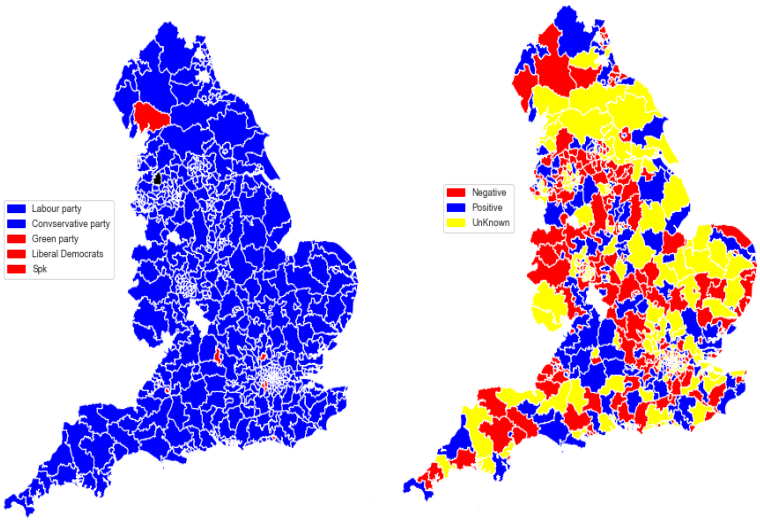

Figure 10. England Location-based Sentiment

3.11. Evaluating the Forecast

We evaluate the location-based sentiment forecast for all constituencies per UK country in Table 3. To measure this, we allocate binary values to the First Party where 1 refers to the Brexiteer parties and 0 refers to the Bremainer parties. We also allocate binary values to the location-based sentiment prediction for the constituency where 1 refers to a predicted vote for the Brexiteer parties and 0 is a vote for the Bremainer parties. The Pearson correlation coefficient is used to detect the degree of linear correlation between two continuous variables, in this case the political party stance and the location-based sentiment. The Pearson correlation coefficient values range from -1 to 1. Positive values mean the selected variables have a positive correlation with the target. Negative value means the selected variable has a negative correlation with the target. The Pearson correlation coefficient between two variables is defined as the quotient of the covariance and standard deviation between two variables. The equation defines the population correlation coefficient and is denoted by:

$$p = \frac{cov(x_1, X_2)}{(\sigma x_1, \sigma X_2)} \quad (2)$$

The Northern Ireland and English constituencies have a positive correlation between location-based sentiment and the First Party. This means that an increase or decrease in the value of the location-based sentiment variable is generally followed by an increase or decrease in the value of the First Party variable. Scotland and Wales show a negative correlation meaning that an increase (decrease) in the value of the location-based sentiment variable is generally followed by a decrease or (increase) in the value of the First Party variable. We use a confusion matrix to compute the performance measurement for our location-based sentiment machine learning classification. Where P = Pearson, B= Baseline= Accuracy, P= Precision, R= Recall and F1 =F1. Accuracy indicates that our highest number of constituencies that were predicted correctly was Northern Ireland 60%, followed by Scotland 58%, Wales 44% and then England at almost 38%.

Table 3. Accuracy of Predictions per UK Country

Country	P	B	A	P	R	F1
NI	0.204	.40	**.60**	.50	.167	.250
Scotland	-0.121	**.**44	**.58**	.143	.40	.21
Wales	-0.271	.32	**.441**	.993	.438	.596
England	0.068	.45	.378	.973	.377	.543

3.12. Machine Learning Models

Hypothesis 2: Can UK Twitter location-based sentiment per constituency combined with Westminster Parliamentary Constituencies' data increase the accuracy of the Brexit Election Prediction baseline? Decision tree, neural network and Naive Bayes models are created.

3.12.1. Sequential Neural Network Model

Sequence classification is predictive modelling where you have some sequence of inputs over space or time and the task is to predict a category for the sequence. In this instance we are predicting the election result per constituency such that the vote can be to 'Get Brexit Done' or to reject Brexit and remain in the European Union. We use both the Twitter and Westminster Parliamentary Constituencies dataset. The combined dataset consists of 32 features and we need to predict the results by constituency. The following categorical values constituency_name, country_name, region_name, country_name, constituency_type, mp_gender and first party are converted into numerical values as neural networks algorithms expect numerical values to achieve cutting-edge results. We use one-hot encoding with Pandas and pass the data into the get_dummies Panda's function; this converts the text or categorical data into numerical data with which the model expects and perform better. The new columns are column binded into the preexisting dataset. The dataset is now split into training and testing. The training data will have 90% samples and test data will have 10% samples. The neural network is build using Keras and Tensorflow. Keras is a deep learning API written in Python, running on top of the machine learning platform TensorFlow. It was developed with a focus on enabling fast experimentation [32] Tensorflow is an end-to-end, open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets

researchers push the state-of-the-art in ML and developers easily build and deploy ML-powered applications [33]. The dataset has an input layer of 32 and output layer of 1. The Dense layer is used to specify the fully connected layer to the neural network. The arguments of the Dense layer are the output dimension which is 32 in the first case, and an input dimension of 479. The activation function used is relu. The loss function and the optimiser are binary_crossentropy which specifies that we have binary classes. The optimiser is Adam. It is an adaptive learning rate optimisation algorithm that's been designed specifically for training deep neural networks [34] The neural network is trained using 100 epochs. The model's accuracy score is computed by the sklearn metrics accuracy score library.

3.12.2. Decision Tree Model

A decision tree is a supervised machine learning algorithm. Decision tree builds a classification model in the form of a tree structure with a root node (the top node) and underlying branches. It breaks the dataset into smaller and smaller subsets whilst simultaneously creating and developing a tree structure. Once the tree is finalised it will have a number of branches also known as decision nodes and each branch will have an underlying node also commonly known as Leaf nodes which represents the classification or decision. Sklearn's decision tree library is imported to compute the decision tree using the same training and test data ratios used to compute the neural network

3.12.3. Naive Bayes

We use a Naive Bayes supervised machine learning algorithm as referenced in section 3.6.

3.13. Evaluating the Models

The results indicate that the decision tree model performed better in terms of accuracy than the neural network and the Naive Bayes models. All three models were significant in terms of the calculated baseline of 54%.

Table 4. Machine Learning Modelling Results

Model	Country	Accuracy
Decision Tree	UK	. 9787
Neural Network	UK	.9574
Naive Bayes	UK	.9362
Baseline	UK	.545

4. RESULTS

Table 3 illustrates the results at the individual country constituency level using location-based sentiment versus the calculated baseline models. Where X_1, X_2, X_3, and X_4, represent the location-based sentiment from the respective UK constituencies at country level "*Northern Ireland", "Scotland", "Wales"* and *"England"*. We can reject the null hypothesis that Twitter location-based sentiment does not predict the Brexit election with a reasonable degree of accuracy such that $\beta_1, \beta_2, \beta_3, . \eta \neq 0$. Evidence presented herein confirms a relationship relates to location-based sentiment for predicting the Brexit election in the case of the constituencies for X_1, X_2, X_3, where X_1 *("Northern Ireland")* exhibits the highest accuracy, followed by

X_2,(‘*Scotland*) and X_3, (‘*Wales’*) respectively. X_4 (‘*England’*) is not indicative of any significant predictive relationship. Table 4 shows the decision tree, neural network and Naive Bayes machine learning models’ accuracy using location-based sentiment at national level exceeds the baseline accuracy result significantly with a 43%, 41% and 39% improvement from the calculated baseline accuracy.

5. CONCLUSIONS

In this paper, we aimed to establish if location-based Twitter sentiment could predict the Brexit Election results at constituency level for each of the four UK countries and, similarly, at a national level. Our results indicate that Twitter location-based sentiment had the single biggest effect on constituency result predictions in Northern Ireland and Scotland and a marginal effect on Wales-based constituencies whilst there was no significant prediction accuracy to England’s constituencies. We further established that Twitter location-based sentiment improves machine learning prediction accuracy for our decision tree, neural network and Naive Bayes models of up to 43%, 41% and 39% respectively from the calculated baseline accuracy. Owing to the constraints of Twitter location-based sentiment, there were 181 constituencies not represented within the dataset. That may have been as a result of non-user location input or the possibility that there was actually no representation from said constituencies on Twitter. In some cases, users opted to shorten their constituency name rather than use the full constituency name for example ‘Cities of London and Westminster’ is inserted as ‘Westminster’. The issue here is that the constituencies would not match to the constituency naming convention contained in the Office for National Statistics choropleth mapping thus providing reconciliation differences. To combat this, our future research will look to counteract this shortcoming by identifying same and produce a model to compensate for said constraints. In the case of non-user location input, further research is warranted on local language detected within the Twitter text to infer said missing locations as a result of non-user location input. Essentially, Twitter text at the word level can be extracted and word selections (based on identified word distributions per known constituencies) can be aligned to the missing constituencies to yield the location taken from our original Twitter dataset. This would help overcome the constituency location sparsity problem and allow for a probabilistic framework for estimating a Twitter user’s constituency-level location based purely on the content of the user’s tweets in the absence of geospatial cues. This would be heavily reliant on a classifier which identifies words in tweets with a local geographic scope such that the observed geographical distribution of the words in tweets correlates to the geo-locations. [35] created a similar type of model using city-level locations. With continuing chunter intensifying of the breakup of the Union and Scottish Independence, location-based sentiment would prove a very useful tool in strategising against the breakup of what once was a sense of British identity that bound the United Kingdom together and now appears to be disintegrating.

ACKNOWLEDGEMENTS

My thanks to the Technological University Dublin School of Computing for their support on this paper.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, WWW ’10, pages 851–860, New York, NY, USA, 2010. ACM.

[2] P. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. J. Finance 62 (2007), 1139–1168.

[3] S. Das and M. Chen. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Manage. Sci. 53, 9 (2007), 1375–1388
[4] Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Info. Syst. 26, 3 (2008)
[5] C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro. 2013. Sentiment analysis of Facebook statuses using naïve bayes classifier for language learning. In Proceedings International Conference in Information, Intelligence, Systems and Applications (IISA'13). 1–6.
[6] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs", Proceedings of International Conference on Weblogs and Social Media, 2007.
[7] N. Beauchamp. Predicting and interpolating state-level polling using Twitter textual data. In New Directions in Analysing Text as Data Workshop, 2013.
[8] A. Bermingham and A. F. Smeaton. On using Twitter to monitor political sentiment and predict election results. In SAAI '11, 2011
[9] A. Boutet, H. Kim, E. Yoneki, et al. What's in Your Tweets? I Know Who You Supported in the UK 2010, General Election. In ICWSM '12, pages 411{414, 2012.
[10] A. Bruns, J. Burgess, et al. # ausvotes: How Twitter covered the 2010 Australian federal election. Communication, Politics & Culture, 44(2):37{56, 2011.
11] A. Ceron, L. Curini, S. M. Iacus, and G. Porro. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, 16(2):340{358, 2014
[12] M. Choy, M. Cheong, M. N. Laik, and K. P. Shung. US presidential election 2012 prediction using a census corrected Twitter model. arXiv preprint arXiv:1211.0938, 2012
[13] M. Choy, M. L. Cheong, M. N. Laik, and K. P. Shung. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. arXiv preprint arXiv:1108.5520, 2011
[14] J. E. Chung and E. Mustafaraj. Can collective sentiment expressed on Twitter predict political elections? In AAAI '11, pages 1770{1771, 2011.
[15] M. Gaurav, A. Srivastava, A. Kumar, and S. Miller. Leveraging candidate popularity on Twitter to predict election outcome. In SNA-KDD Workshop, 2013.
[16] A. Jungherr, P. Jurgens, and H. Schoen. Why the pirate party won the German election of 2009 or the trouble with predictions: A response to ... Social Science Computer Review, 30(2):229{234, 2012.
[17] A. Makazhanov, D. Ra_ei, and M. Waqar. Predicting political preference of Twitter users. Social Network Analysis and Mining, 4(1):1{15, 2014.
[18] Y. Mejova, P. Srinivasan, and B. Boynton. GOP primary season on Twitter: popular political sentiment in social media. In WSDM '13, pages 517{526, 2013
[19] F. Nooralahzadeh, V. Arunachalam, and C. Chiru. 2012 Presidential Elections on Twitter - An Analysis of How the US and French Election were Reflected in Tweets. In CSCS '13, pages 240{246, 2013.
[20] E. T. K. Sang and J. Bos. Predicting the 2011 Dutch Senate Election Results with Twitter. In Workshop on Semantic Analysis in social media, pages 53{60, 2012.
[21] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In ICWSM '10, pages 178 -185, 2010.
[22] General election 2019: Brexit - where do the parties stand? https://www.bbc.com/news/uk-politics-48027580
[23] https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000
[24] https://dominiccummings.com/tag/machine-learning/
[25] https://www.bbc.com/news/uk-42698981
[26] J. Usher and P. Dondio. 2020. BREXIT Election: Forecasting a Conservative Party Victory through the Pound using ARIMA and Facebook's Prophet. In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020). Association for Computing Machinery New York, NY, USA, 2020. ACM
[27] Ussama Yaqub, Nitesh Sharma, Rachit Pabreja, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. 2020. Location-based Sentiment Analyses and Visualization of Twitter Election Data. Digit. Gov.: Res. Pract. 1, 2, Article 14 (April 2020), 19 pages
[28] Maima Almatrafi, Suhem Parack, and Bravim Chavan. 2015. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In Proceedings of

the 9th International Conference on Ubiquitous Information Management and Communication (IMCOM '15).
[29] https://commonslibrary.parliament.uk/research-briefings/cbp-8749/
[30] Tweepy Documentation, http://docs.tweepy.orglen/v3.S.0/
[31] https://geoportal.statistics.gov.uk/datasets/ons::westminster-parliamentary-constituencies-december-2019-boundaries-uk-bfc-v2/explore?location=55.450000%2C-2.000000%2C5.78
[32] https://keras.io/about/
[33] https://www.tensorflow.org/
[34] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. 2014. arXiv:1412.6980v9
[35] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10). Association for Computing Machinery, New York, NY, USA, 759–768.

AUTHORS

James Ushe is an Irish native and dwells in Dublin when not spending his time in the countryside of Co Meath. He spends a lot of his free time uncloaking patterns in geopolitical events such as Brexit. Currently he has produced four published papers on Brexit. When not conducting Brexit experiments you will find him enjoying music.