# Analyzing Public Opinion on Brexit Over Time Using Twitter Sentiment Analysis

Ben Reber, Ibrahim Khan

December 14, 2019

## 1 Abstract

The United Kingdom's decision to leave the European Union, known as Brexit, has been a source of controversy since the June 2016 referendum. Our goal with this project was to determine public sentiment with regards to Brexit between the time of the referendum and now. Twitter has been scraped to gather Brexit related Tweets from July 2016 through November 2019. We investigated certain patterns about popular issues and people for the exploratory data analysis. This data has been pre-processed and classified into the leave and remain side, and then fed into a sentiment analysis tool known as VADER to determine overall public sentiment about Brexit. Results were compared against those of conventional polling methods and their correspondence with realtime events has also been explored.

## 2 Introduction

Traditional polling methods have proven unsatisfactory in recent years to make accurate predictions of public sentiment towards political issues. Twitter provides a large repository of public opinion from millions of users in the form of short Tweets. Recent studies have compared the results from Twitter sentiment analysis with those from traditional polling methods and found it satisfactory (Oliveira et al., Ardehaly and Culotta). Because of this, sentiment analysis using Twitter data is becoming a popular method for determining public opinion (Pandarachalil et al., Ardehaly and Culotta). We have attempted to determine public opinion about Brexit in the time since the referendum using Twitter sentiment analysis.

### 2.1 Background & Related Work

In June 2016, a referendum was held in the United Kingdom to determine whether it should remain in the European Union. The result of the referendum was a 51.9% vote in favor of Brexit. This result defied the predictions of many mainstream media polls (Ray). On March 29th 2017, Prime Minister Theresa May invoked Article 50 of the Treaty on European Union, which allows a country to leave the EU within two years of invocation unless an extension is provided. Since no deal regarding the relationship between the EU and UK following Brexit has yet been agreed upon, the original deadline of March 29th 2019 has been extended twice, first to October 31st 2019, then to January 31st 2020.

As a result of the difficulty of establishing a deal with the EU, current Prime Minister Boris Johnson has shown willingness to push for a "no-deal" Brexit. In this case, the UK would leave the EU without any special agreement in place regarding their post-Brexit relationship. Academics have speculated that this could have severe economic and social consequences (Kierzenkowski et al., Bell). Because of this, there has been volatility in public opinion with regards to Brexit.

Sentiment from Twitter data has been shown to be highly correlated with results of traditional polling methods (O'Connor et al.). Scraping Tweets that are more than a week old is difficult with the Twitter API. Other works have extracted old Twitter data using various methods. Magu and Joshi used the Jefferson-Henrique Python script for gathering old Tweets (Magu et al.).

# 3 Data Gathering & Preprocessing

Using the Jefferson-Henrique script, we gathered data from Tweets containing the keyword "brexit" which were posted between July 2016 and November 2019. We gathered 50,000 Tweets from each month. Tweets were gathered from the end of each month. Each tweet is stored as a row in a csv file with the following columns:

username;date;retweets;favorites;text;geo;mentions;hashtags;id;permalink

We also collected opinion poll data from YouGov for each month to compare against the Twitter analysis result. The sample size for each poll was around 1500. Poll results were adjusted to remove the "Neither/Not Sure" segment and then normalized according to the formula: N(leave) = N(leave)/(N(leave)+N(remain)).

| Month | Remain | Leave | Neither |
|---|---|---|---|
| 07-16 | 43 | 44 | 13 |
| 08-16 | 43 | 45 | 13 |
| 09-16 | 42 | 46 | 11 |
| 10-16 | 44 | 43 | 13 |
| 11-16 | 46 | 42 | 12 |
| 12-16 | 44 | 43 | 13 |
| 01-17 | 43 | 44 | 13 |
| 02-17 | 42 | 44 | 15 |
| 03-17 | 44 | 43 | 14 |
| 04-17 | 45 | 45 | 10 |
| 05-17 | 43 | 43 | 13 |
| 06-17 | 46 | 42 | 13 |
| 07-17 | 46 | 43 | 11 |
| 08-17 | 45 | 43 | 12 |
| 09-17 | 44 | 46 | 9 |
| 10-17 | 44 | 40 | 16 |
| 11-17 | 43 | 43 | 14 |
| 12-17 | 39 | 48 | 13 |
| 01-18 | 46 | 42 | 12 |
| 02-18 | 44 | 41 | 14 |
| 03-18 | 45 | 44 | 11 |
| 04-18 | 45 | 42 | 13 |
| 05-18 | 48 | 47 | 6 |
| 06-18 | 44 | 44 | 12 |
| 07-18 | 46 | 41 | 13 |
| 08-18 | 46 | 42 | 12 |
| 09-18 | 43 | 43 | 13 |
| 10-18 | 46 | 41 | 13 |
| 11-18 | 47 | 39 | 14 |
| 12-18 | 46 | 39 | 15 |
| 01-19 | 45 | 38 | 16 |
| 02-19 | 48 | 38 | 14 |
| 03-19 | 46 | 41 | 14 |
| 04-19 | 44 | 40 | 15 |
| 05-19 | 44 | 42 | 14 |
| 06-19 | 51 | 44 | 5 |
| 07-19 | 46 | 41 | 13 |
| 08-19 | 45 | 40 | 15 |
| 09-19 | 46 | 43 | 12 |

Table 1: YouGov poll data.

We performed some pre-processing operations to the data. First, we sliced each record into a list of values of attributes and extracted the text of the tweet. Then we cleaned the data by removing hyperlinks and special symbols. We also removed the hashtag symbols attached to certain important terms to make it easier to match those terms. We also did regular expression matching on the tweets to make sure that each word contains only letters.

For the task of computing frequent itemsets, we maintained a separate repository of tokenized data by creating a list of 30 important Brexit related terms. Then, in this repository, we converted each tweet into a list of words and filtered out all the words that don't belong in the list of important words.

# 4 Exploratory Data Analysis

We performed frequent itemset mining on the list of tokenized data to determine the most frequent 1 and 2 sized itemsets. This was done using the Python library pyfpgrowth.

| Itemset | Support |
|---|---:|
| ireland | 11960 |
| conservative | 12758 |
| business | 13774 |
| economy | 15006 |
| bbc | 24062 |
| parliament | 25138 |
| britain | 28623 |
| corbyn | 34513 |
| corbyn, labour | 10391 |
| government | 37136 |
| remain | 37450 |
| johnson | 38535 |
| boris, johnson | 21790 |
| labour, tory | 11148 |
| tory, vote | 12032 |
| boris | 51718 |
| leave, vote | 12293 |
| eu, leave | 13738 |
| may | 58360 |
| labour, labour | 12951 |
| labour, vote | 22766 |
| vote | 100077 |
| vote, vote | 18707 |
| eu | 134110 |
| eu,eu | 17971 |

Table 2: Frequent itemsets of size one and two.

We fed all scraped tweets into VADER and plotted the baseline overall sentiment in our dataset, ignoring all Tweets which were classified as neutral (Figure 1).

We examined the frequency of mentions of Boris Johnson, Theresa May and Jeremy Corbyn (figure 2), and the overall sentiment of Tweets containing mentions of these political figures (figure 3). This was done by collecting the subset of our data which contained the first or last names of each politician.

# 5 Methods and Experiments

In order to determine overall sentiment toward Brexit, we first needed some way to classify Tweets as referring to sentiment towards the leave side versus sentiment towards the remain side. This is because positive or

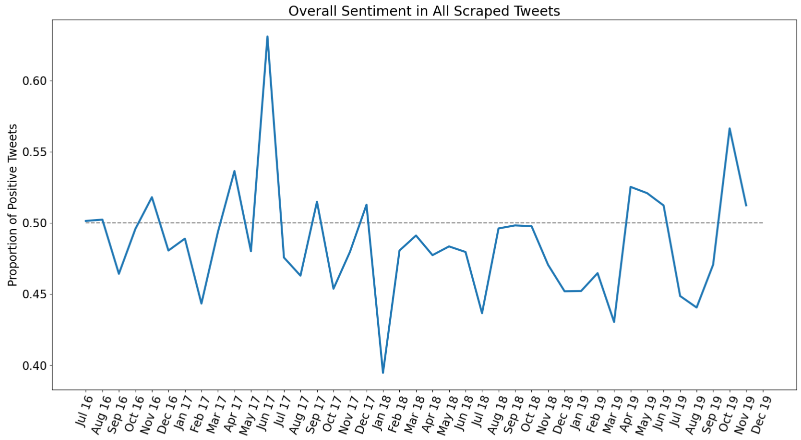

Figure 1: Raw sentiment in all scraped Tweets regarding Brexit.

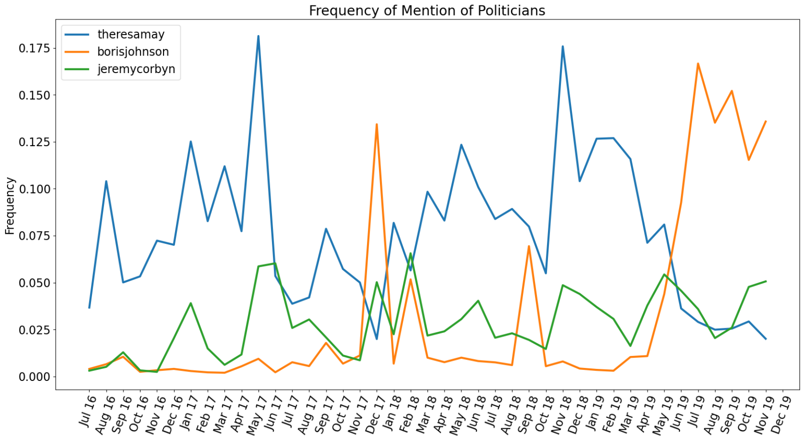

Figure 2: Frequency of mentions of political figures over time.

negative sentiment in a Tweet does not necessarily imply the same sentiment towards Brexit. It is possible to reflect negative feelings toward Brexit with a positive Tweet, (e.g. "Remaining in the E.U. is the best thing we could do for this country!"), and it is also possible to show positive feelings toward Brexit with a negative Tweet (e.g. "People who voted remain are so stupid!" ). A Tweet containing the word "Remain" is considered descriptive of sentiment toward the remain side, and a Tweet containing the word "Leave" is considered decriptive of sentiment toward the leave side. In order to increase the amount of data considered, we assume that the same is true for mentions of the the Conservative/Tory party and Labour party, since Brexit is on Conservative's platform and Labour stands on the opposite side. Tweets which contain keywords from exactly one side (leave, conservative and tory versus remain and labour) were considered, and all else were removed. We found that between 4000 and 6000 Tweets for each month fit this criteria.

A Tweet reflecting positive sentiment toward the leave side or negative sentiment toward the remain side was classified as pro-Brexit, and one with negative sentiment toward the leave side or positive sentiment toward the remain side was classified as anti-Brexit. We recorded the number of positive and negative Tweets which were classified as reflecting pro-Brexit and anti-Brexit sentiments.

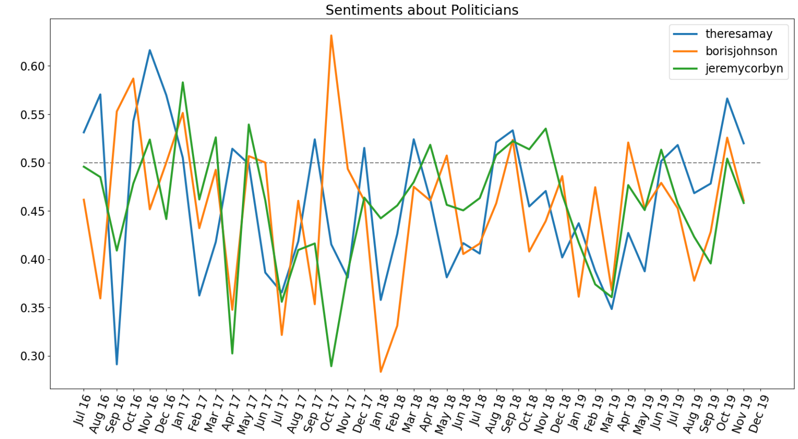

Figure 3: Sentiment about political figures over time.

# 6 Results

Figure 4 shows the output of our model for determining pro versus anti-brexit sentiment. Table 3 shows how the overall sentiment of Tweets matched their sentiments towards Brexit. Tweets in the main diagonal were classified as pro-Brexit and Tweets in the other quadrants were classified as anti-Brexit.

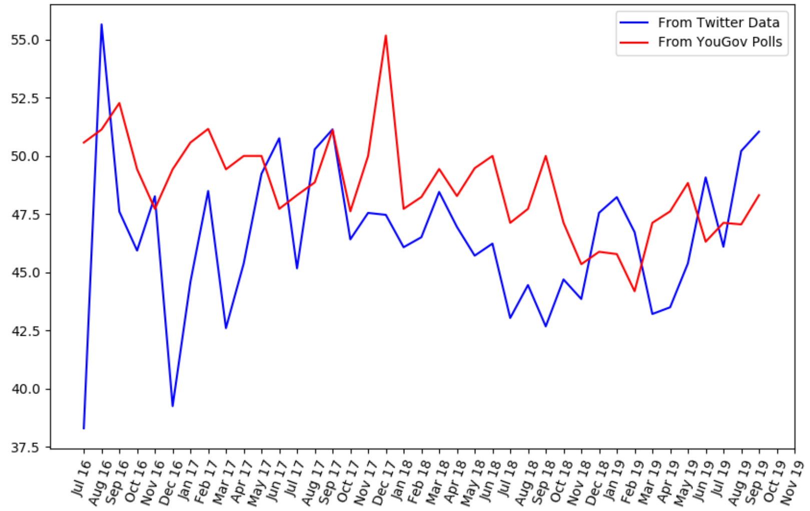

Figure 4: Predicted approval of Brexit from Twitter data compared to YouGov polls.

# 7 Discussion

Figure 2 shows the frequency of mentions of various politicians over time. The spike in popularity of Theresa May in May 2017 was likely caused by the general election that occurred around that time. May had another spike around November 2018, likely because of the publishing of the Brexit Withdrawal Agreement and its endorsement among many E.U. member states. Around the time of May's resignation, mentions of May started decreasing and mentions of Johnson started increasing as he came into office.

| | | Classification | |
|---|---|---|---|
| Sentiment In Tweet | | Leave | Remain |
| | P | 34,750 | 67,242 |
| | N | 49,311 | 66,207 |

Table 3: Distribution of Tweets according to positive/negative sentiment and leave/remain classification.

Figure 4 shows our overall model result. Our model tends to predict more negative sentiment towards Brexit than the YouGov polls. This could be because Twitter is skewed toward a younger demographic (pewresearch.org), and younger demographics tend to be more liberal. Right after the referendum, there is a strong anti-brexit sentiment on Twitter. It may be because of the anger towards the result of the referendum which motivated many anti-brexit people to express their disapproval online. There is a spike in popularity of brexit for April - May 2017, perhaps due to the general election in which Conservatives managed to win the most seats. In 2018, Brexit popularity started to decline, which is evident in both our results and the polling. It might be due to various setbacks May's Brexit strategy suffered. Just ahead of the first meaningful vote on the Withdrawal Agreement in January 2019, Brexit popularity increased a little. But that increase was short lived as both the first and second votes were defeated in the parliament. Brexit popularity saw a sudden increase after the resignation of May and inauguration of Johnson as the new Prime Minister.

We can see from this that there is more negativity among Tweets supporting Brexit than there is among Tweets supporting continued EU membership.

## 8 Bibliography

Pandarachalil, R., Sendhilkumar, S. & Mahalakshmi, G.S. Cogn Comput (2015) 7: 254. https://doi.org/10.1007/s12559-014-9310-z

Daniel José Silva Oliveira, Paulo Henrique de Souza Bermejo & Pâmela Aparecida dos Santos (2017) "Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls", Journal of Information Technology & Politics, 14:1, 34-45, DOI: 10.1080/19331681.2016.1214094

E. M. Ardehaly and A. Culotta, "Mining the Demographics of Political Sentiment from Twitter Using Learning from Label Proportions," 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, 2017, pp. 733-738.

Kierzenkowski, R., et al. (2016), "The Economic Consequences of Brexit: A Taxing Decision", OECD Economic Policy Papers, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/5jm0lsvdkf6k-en.

Bell, Christine. 2016. 'Brexit, Northern Ireland and British-Irish Relations.' European Futures, 26 March.

Jefferson-Henrique Script: https://github.com/Jefferson-Henrique/GetOldTweets-python

Ray, Sarah C. "The Brexit Upset: Why Was the U.K. Vote a Surprise?". Middlebury University, October 2016, http://www.middlebury.edu/newsroom/archive/2016-news/node/542305. Accessed December 2019.

F. Franch, "Wisdom of the crowds: 2010 UK election prediction with social media," J. Inf. Technol. Politics, vol. 9, no. 4, pp. 57–71, 2012.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in Proc. ICWSM, 2010, pp. 178–185.

O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in Proc. ICWSM, 2010, pp. 122–129.

https://www.pewresearch.org/internet/fact-sheet/social-media/

https://pypi.org/project/pyfpgrowth/