1 Overview

Dataset

Firstly, 477,256 public tweets and 2,337 tweets by Brexit influencers were obtained using GetOldTweets3, a Python package available on pypi.org. The public tweets were obtained from 9 distinct time periods, code named a1, a2, a3, b, …, g (Table A). The influencer tweets were obtained using influencer usernames (Table B) as the main query term, and we obtained 3149 Leave tweets and 1439 Remain.

We initially used the Twitter Search API to obtain data for period f. However, due to the daily limits on live tweets that can be streamed, we decided to go with data obtained from GetOldTweets3 for period f when we were preparing this report.

Research Question

Our main research questions were:

- What was the content of Brexit-related tweets like?
- What proportion of Brexit-related tweets were pro-Leave, pro-Remain, and Neutral? How did this change over time?
- What are the sentiment of Brexit-related tweets? Do they differ based on Leave/Remain camp? Do they differ based on the context of the tweet?
- Can we predict the Leave/Remain alignment of a tweet?

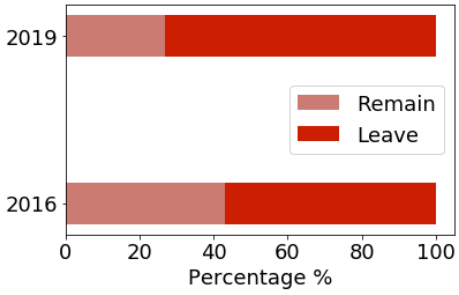
2 Data Analysis

a Camp Analyzer

We wrote the function 'camp_analyzer' to label every public tweet based on its content. Camp_analyzer tracks the number of occurrences of words in the leave-remain dictionary (Table C), and returns camp label:

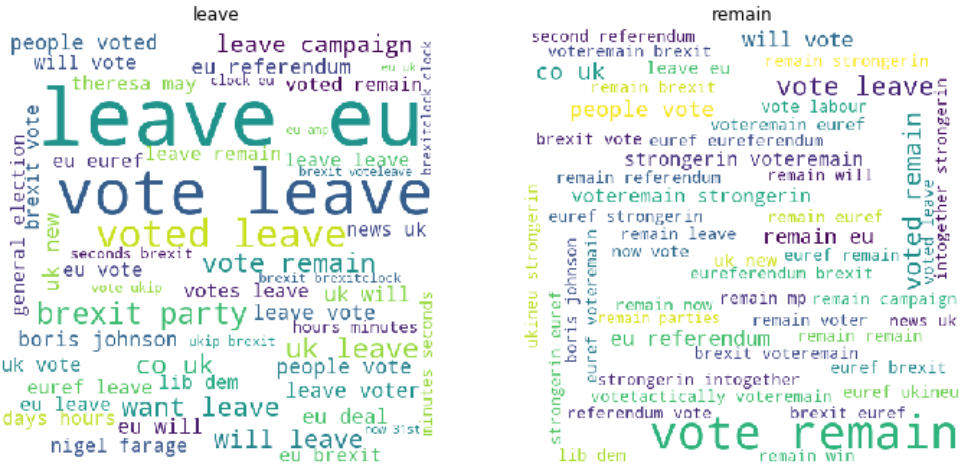
- "Remain", if remain count > leave count
- "Leave", if leave count > remain count
- "Neutral", if remain count = leave count (including cases where both counts are zero)

We compared the Leave & Remain ratio in 2016 to 2019, and found out that our result is comparable to the UK Referendum results of 2016 where "leave" was 52% and "remain" was 48%.

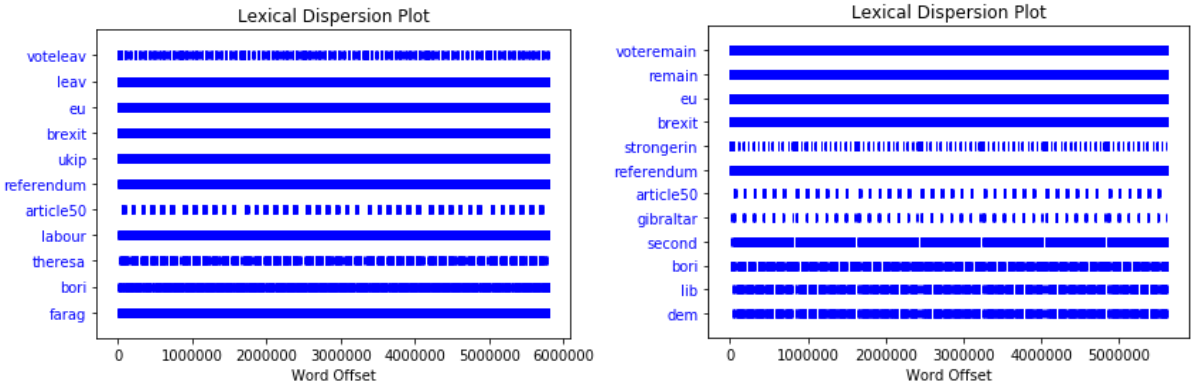


b Wordcloud, Lexical Dispersion

We first conducted preliminary analysis into what the Brexit-related tweets could contain, by creating 3 different word clouds: 1) all public tweets, 2) all Remain tweets, 3) all Leave tweets. Most common words from Leave and Remain were noted for our lexical dispersion plots.



Next, we wrote the function 'remove_stop_stem_contract' to remove stopwords, stem, and decontract (e.g. "won't" to "will not") our tweets, before creating our lexical dispersion plots.

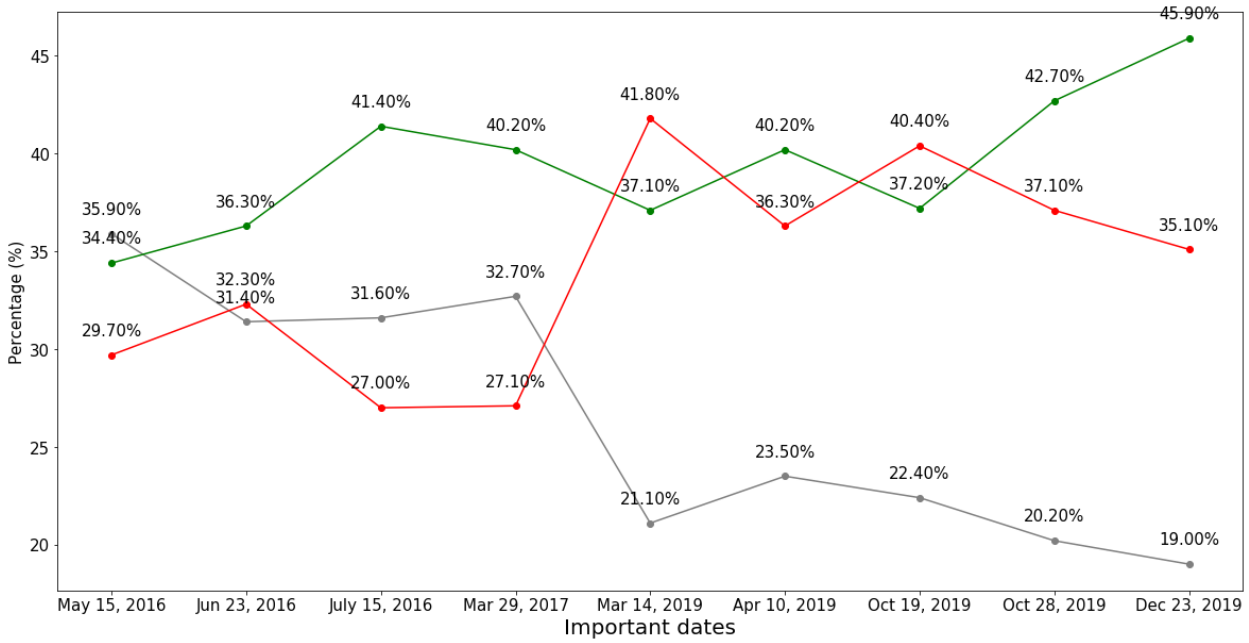


c Vader Sentiment Analysis

Part 1: Sentiments over Time

We were interested in finding out how the proportion of Positive, Negative and Neutral tweets changed over time, from periods a1 to g. We carried out vader sentiment analysis on each tweet and gave it a label based on its compound score (threshold 35). Green refers to Positive, red refers to Negative and grey refers to Neutral in the graph below.

Change in Sentiments over Time



Part 2: Named Entity Detection and Affect Calculator

From this part of the report onwards, we focus on period a2, which is the month of the Brexit Referendum.

After understanding how sentiments changed over time, we also wanted to understand how sentiments change based on the context of the tweet; in particular, if certain people/ organizations/ concepts were mentioned. We hence carried out named entity detection, specifying label types as PERSONs and ORGANIZATIONs. Relevant labels were then used as tags and fed into our affect calculator to find out how tweets which include that tag fare in terms of sentiment.

	tag	affect score
0	voteleave	0.084820
1	leave	-0.001214
2	voteremain	0.099161
3	remain	0.074195
4	brexit	0.024791
5	cameron	-0.077946
6	carney	0.046225
7	farage	0.034675
8	ukip	0.019832
9	labour	0.152621
10	theresa	-0.163750
11	corbyn	0.074268
12	boris	0.017546
13	tories	0.037221
14	gove	-0.076170
15	osborne	0.025633
16	ireland	0.094203
17	trump	0.064246

Looks like there were very negative sentiments towards 'leave', David Cameron, Theresa May and Michael Gove in the month of the referendum!

d Topic Modelling

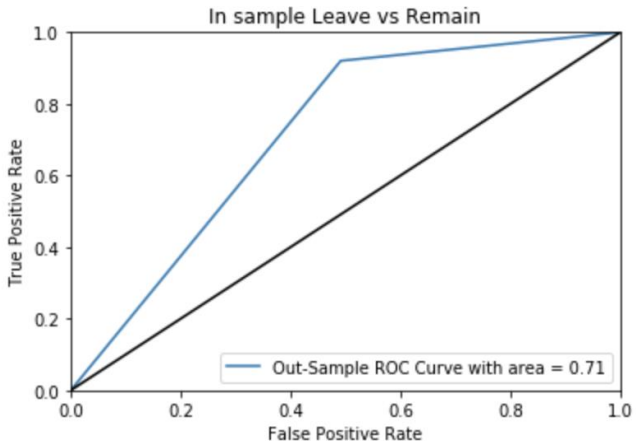
Next, we incorporated some unsupervised learning (Latent Dirichlet Allocation) to discover if there are distinct topics across the tweets. After running our model with 10 topics, we labelled each tweet's lda_topic based on its topic of highest probability. Using gensim, we then generated a summary of the tweets from each topic, in order to understand what each topic might be about. This is a good complement to checking the top probability words for each topic, as it provides more information about the context in which the most common words are found.

In general, we found that Brexit-related tweets are very politically-charged, with many references to specific politicians and their parties.

e Machine Learning

In order to apply the Machine Learning techniques we learnt in class, we utilized sklearn's CountVectorizer to generate bag-of-words vocabulary from the entire corpus of tweets, and vectorised each tweet in the corpus. Each tweet vector has label '1' corresponding to Leave, and '0' corresponding to Remain. Where necessary (e.g. for neural networks), the label on each tweet vector is one-hot encoded. We used our public tweets as training data (labels determined by camp_analyzer), and our influencer tweets as testing data (labels determined by the political leaning of said influencer).

Linear Regression ROC



Classification Tree, Random Forest

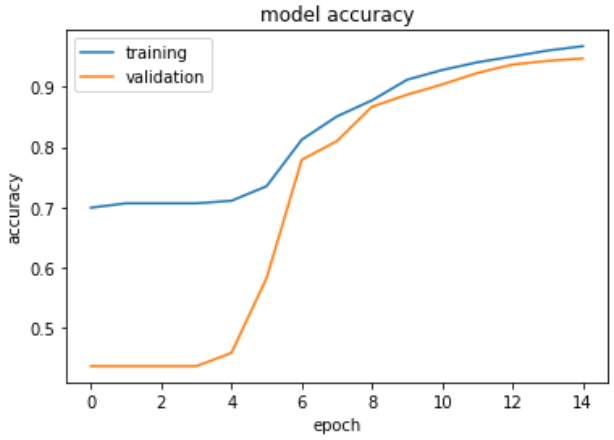
For both methods, GridSearch was used to identify the best parameters, bringing our accuracy up from an initial 0.653 to 0.708.

The sklearn python package was used to fit the classification tree by entropy minimization and with a maximum depth of 10.

Neural Networks

The keras python package was utilized to build the neural networks used in this paper. We used the Sequential model to sequentially build Dense hidden layers for a fully-connected neural network, and a softmax output layer to give the probability of some input being equal to either Leave or Remain.

We found the best outcomes with 4 hidden layers and 32 nodes per layer, using the tanh activation function. This gives us an accuracy of 0.788.

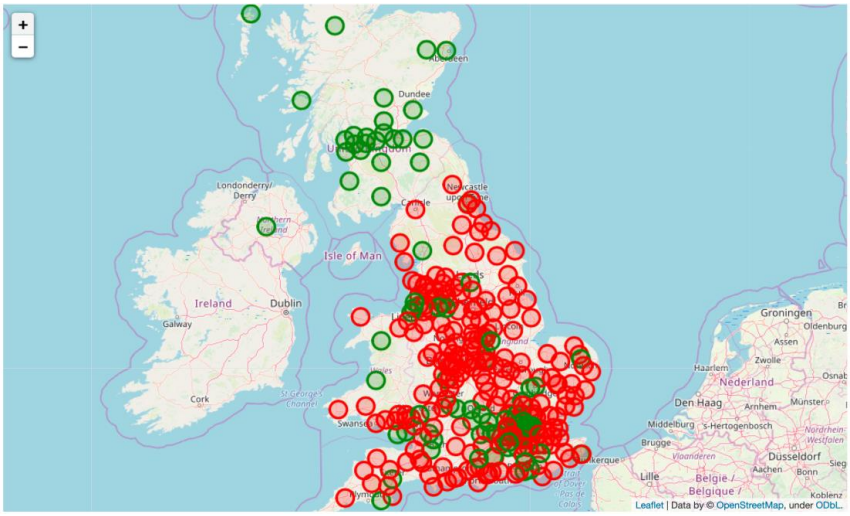


Support Vector Machines

Finally, we experimented with using linear SVM, and obtained accuracy of around 0.69.

f Further Visualizations

Using Folium mapping technique, we analyse that the EU referendum 2016 shows majority of the areas in the UK voted “leave”. Below, red represents Leave and green represents Remain.



3 Limitations & Future Research

Limitations: A significant obstacle we faced was the lack of reliable tweet geo-location data, which prevented us from filtering our tweets based on location, say if we only wanted Brexit-related tweets from the UK. We also had to tackle the complexity of tweets being not just text, but it including links, pictures, retweets etc. With basic text mining tools, we were unable to extract deeper meaning from these other information that tweets contain, which might provide good insight to our research questions.

Further Research: With further insight into Natural Language Processing, we would like to gain deeper insight into what the content of these Brexit tweets may be, which we felt was not adequately discovered by LDA. Given access to premium Twitter APIs, we would also love to analyze the networks between tweet contributors, specifically how Brexit influencers are connected to public users (whether pro-Leave influencers are largely only connected to pro-Leave public users etc).

4 Appendix

Table A: Tweet Data

Corpus No.	Source of Data	Date Range	Query	Number of Tweets
Famous	GetOldTweets3	06/01/2016 - 06/30/2016	Influencer usernames	2,337

a1		05/01/2016 - 05/31/2016	Brexit-related terms	79,653
a2		06/01/2016 - 06/30/2016	Brexit-related terms	170,997
a3		07/01/2016 - 07/31/2016	Brexit-related terms	91,249
b		03/29/2017 - 04/05/2017	Brexit-related terms	16,912
c		03/14/2019 - 03/21/2019	Brexit-related terms	23,367
d		04/10/2019 - 04/17/2019	Brexit-related terms	21,710
e		10/19/2019 - 10/26/2019	Brexit-related terms	27,074
f		10/28/2019 - 11/04/2019	Brexit-related terms	24,308
g		12/10/2019 - 12/16/2019	Brexit-related terms	21,986
-	Twitter Search API	11/28/2019 - 12/04/2019	Brexit-related terms	2,076

Table B: Query Terms

Query	Terms
Brexit-related terms	brexit, LeaveEUOfficial, LeaveEU, LabourLeave, ukip, no2eu, notoeu, britainout, leaveeu, voteleave, ukineu, VoteRemain, LabourInForBritain, euref', eureferendum, betteroffout, eureform, betteroffin, yes2eu, yestoeu, intogether, vote_remain
Influencer usernames	Leave: @Borisjohnson, @leaveeuofficial, @jacob_rees_mogg, @michaelgove, @matthew_elliott','jwhittingdale', @giselastuart , @douglascarswell, @bernardjenkin, @annietrev, @catharinehoey, @bylukejohnson, @stevebakerhw, @v2DominicRaab, @john4brexit,@v2andrealeadsom,@patel4witham Remain: @donaldtusk, @theresa_may, @david_cameron, @jeremycorbyn, @george_osborne, @philiphammonduk, @yvettecoopermp, @dannyalexander, @damiangreen, @carolinelucas, @wdjstraw, @sajidjavid, @aluncairns, @JustineGreening, @jeremy_hunt, @GregClarkTW, @markfielduk, @MayorofLondon Filtered by:

	brexit, LeaveEUOfficial, LeaveEU, LabourLeave, ukip, no2eu, notoeu, britainout, leaveeu, voteleave, ukineu, VoteRemain, LabourInForBritain, euref', eureferendum, betteroffout, eureform, betteroffin, yes2eu, yestoeu, intogether, vote_remain, 'vote, remain, EUDebate, EU, Referendum, brexitclock, leave, strongerin, independenceday, union, immigration, independence

Table C: Leave-Remain Dictionary
Leave hashtags: #leaveeuofficial, #labourleave, #ukip, #no2eu, #notoeu, #betteroffout, #voteout, #eureform, #britainout, #leaveeu, #voteleave, #beleave, #loveeuropeleaveeu, #leaveeu, #vote_leave, #leave Remain hashtags: #intogether, #labourinforbritain, #catsagainstbrexit, #yes2eu, #yestoeu, #betteroffin, #votein, #ukineu, #bremain, #strongerin, #leadnotleave, #voteremain, #vote_remain, #remain

5 References

https://pypi.org/project/GetOldTweets3/

https://www.bbc.com/news/uk-politics-eu-referendum-35616946