Social Media Analytics Applied to Tweets About Brexit: Topic Analysis and Sentiments Assessment

Yiran Gao 10462072
Pietro Mazzaglia 10553428
Chris Tansey 10587967
Yulong Wu 10448782

Abstract

Throughout the Brexit process, Twitter has been a popular social media platform for discussing the sociopolitical aspects of the topic. In this paper, we present a pipeline for sentiment analysis of Brexit tweets between different geographic regions of the UK, emotion analysis of the most popular tweets, and topic modelling to gain an understanding of the subtopics of conversation surrounding Brexit. We show that sentiments about the issue still differ between Scotland and England, and that fear is the dominant emotion expressed by the most popular tweets; which is possibly connected with the ongoing COVID-19 pandemic. Upon deeper analysis with topic modelling, we show that discussion of COVID-19 and Brexit are indeed linked.

1 Introduction

The withdrawal of the United Kingdom (UK) from the European Union (EU) following a referendum in June 2016 has proved to be one of the most contentious political decisions in recent history, with a 52-48 split amongst voters (Sampson, 2017). Twitter is a popular social media platform for those wanting to voice their opinions on political and social issues, thus making it a valuable tool for the analysis of topics such as Brexit. In this paper, we present a Social Media Analytics (SMA) pipeline that answers the three following research questions (RQs):

1. How do sentiments in tweets about Brexit differ between Twitter users based in Scotland compared to England?
2. What are the emotions expressed about Brexit by the most popular tweets on Twitter?
3. What subtopics are more frequently discussed when Twitter users tweet about Brexit?

The relevance of RQ1 is based upon the disparity of Brexit referendum votes between England and Scotland, with the countries voting 53% and 38% in favour of leaving the EU, respectively (Sampson, 2017). In the second RQ, our aim is to identify the common emotions expressed in the most popular tweets about Brexit. Finally, in the last RQ, we aim to identify fine-grained subtopics in the Brexit discourse to gain a more specific view of which social issues the public are concerned with.

We note that our research was conducted during the outbreak of COVID-19, meaning that our analysis may be impacted by a change in public discussion during this time. We do believe, however, that analysis of opinions about Brexit in the context of this pandemic is still of high importance.

2 Related Work

Sentiment Analysis (SA) is a subfield of Natural Language Processing that can be used to probe the views of Twitter users using Machine Learning (ML) or lexicon-based approaches, as discussed in the survey conducted by (A. and Sonawane, 2016).

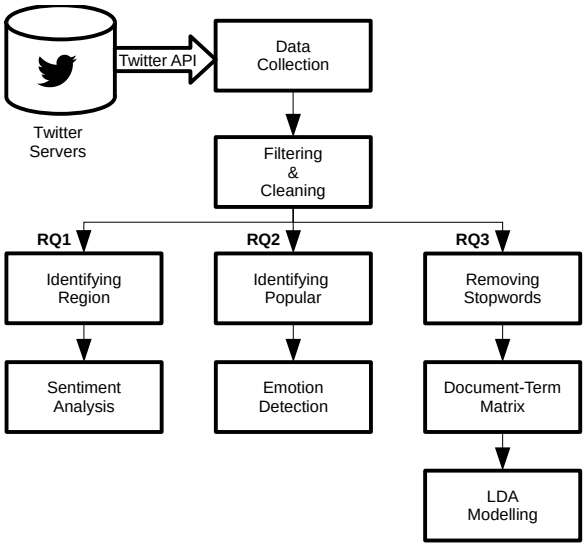
In (Taboada et al., 2011), a Semantic Orientation CALculator (SO-CAL) was developed using dictionaries of words annotated with their semantic orientation which performed well on blog postings and video game reviews without specific training in those domains. In (Saif et al., 2016), another lexicon-based approach called SentiCircles was presented for SA on Twitter. The method takes into account the co-occurrence patterns of words in different contexts in tweets in order to better understand their semantics.

Emotion analysis is similar to SA but represents a more challenging task with subtler distinctions between a large number of possible classifications (Kim and Klinger, 2018). In (Mohammad and Turney, 2010), a lexicon of eight emotions was created and the relationships between them were explored. We used the same set of emotions in this work to answer RQ2.

Popular tweets have been analysed to improve SMA tools in several works, such as in (Suh et al., 2010), where key factors were correlated with higher numbers of retweets. In (Ahmed et al., 2013), a classifier to recognise popular tweets was built exploiting similarity metrics, and in (Kong et al., 2012) the popularity lifespan of tweets was predicted using the first hour of retweet activity.

Topic modelling is a tool to discover the latent topics in document collections. Conventional topic modelling methods such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) have already been used in Twitter-specific contexts (Hong and Davison, 2010; Steinskog et al., 2017), which we also apply in this paper.

Figure 1: Processing pipeline flowchart.



3 Methodology

To answer our RQs, the data collected from Twitter is first filtered, cleaned and finally processed. The overall process is presented in Figure 1.

3.1 Data Collection and Filtering

We collected a dataset of 48,289 tweets from the 6th to the 8th April 2020 using the 'rtweet' package (Kearney, 2019), which provides an API to Twitter's REST.

3.2 Filtering and Cleaning

Our SMA pipeline is devised to work with data in the English language, therefore non-English tweets are filtered out, leaving 42,043 usable tweets.

Afterwards, the text content faces a cleaning process. We perform several text manipulation operations, such as lowering characters, removing punctuation, digits, URLs or special sequences, and expanding abbreviations and contractions.

3.3 SMA Processing

The cleaned data is processed differently for each RQ.

For the first RQ, we first need to identify tweets from England and Scotland. Ideally, this would be performed with a dedicated API, such as GeoNames (Rowlingson, 2019), but limitations on API usage prevent these options from being viable on such a large dataset. Instead, we use the location information from tweet metadata and regular expressions of the most commonly used and most populous areas of England and Scotland to classify most of these locations. We identified 10,643 tweets from England and 1,351 from Scotland, then performed SA using the AFINN lexicon (Årup Nielsen, 2011) on the two sub-datasets to assign a sentiment polarity score to each tweet.

As for RQ2, we analyse the emotions of the popular tweets in our data. In this regard, we provide the definition of 'popular' that we adopted.

Definition (Popular)**.** A tweet is popular if its number of retweets is above the dataset average.

In our dataset, the retweet average is 10.05, meaning that the popular tweets have greater than or equal to 11 retweets, totalling 3,520 tweets (8.37% of the English tweets). We performed an emotion analysis of these popular tweets using the NRC emotion lexicon (Mohammad and Turney, 2013), which provides word associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). In order to account for the COVID-19 issue, we added 'coronavirus' and 'covid' to the lexicon, associating them with the same emotions as the term 'pandemic'.

Finally, for the third RQ, we created a topic model of the subtopics related to Brexit. First, we used the stopword list from the SMART information retrieval system (Bird et al., 2009) to remove irrelevant terms from tweets, also adding the term 'brexit' to the list due to its presence in every tweet. Then, we built a document-term matrix which is fed to the LDA algorithm. Finally, we fine-tuned the parameters of the LDA model on the basis of every round of the LDA visualisation outcome, using the LDAvis package (Sievert and Shirley, 2014).

4 Results

In this section, we analyse our results for each RQ.

The results for SA in RQ1, in Figure 2, show that the distributions appear to be roughly similar, with users in both countries mainly tweeting with neutral sentiments (about 25%). This neutral band of sentiments is, however, slightly more common amongst users in Scotland. Other key differences can be seen in the tails of the distributions, where extremely negative sentiments are mainly expressed by users in Scotland, and extremely positive sentiments are mainly expressed

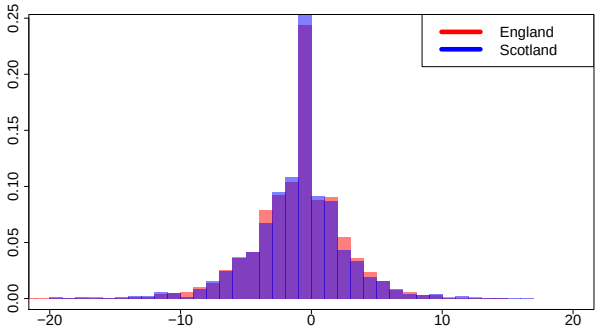
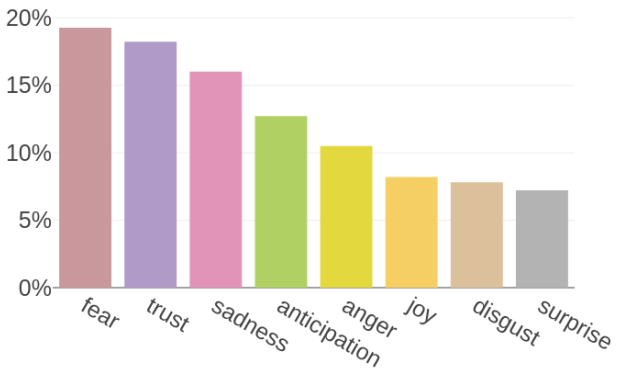


Figure 2: England - Scotland sentiments comparison.

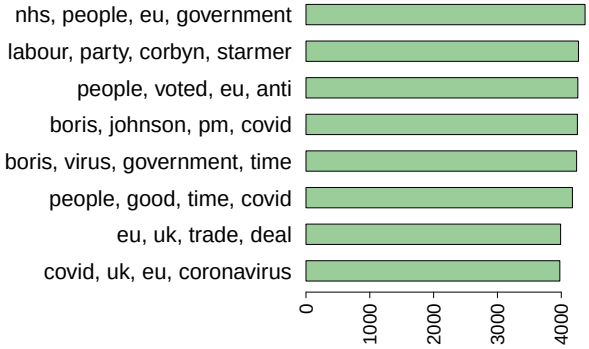
Figure 3: Emotion analysis of popular tweets.



by users in England. These results may correlate with the 2016 referendum result where England voted in favour and Scotland voted against Brexit, but it must be noted that these sentiments are not necessarily directed towards any particular political beliefs since we are not using a domain-specific lexicon; owing to the difficulty of building this (Hamilton et al., 2016).

For the second RQ, the result of emotion detection on the popular tweets is presented in Figure 3. The plot shows that the dominant emotion is fear, with 19.28%, followed by trust (18.23%) and sadness (16.02%). All the other detected-emotions spread is lower than 15% in the tweets. Negative emotions are likely to dominate because of the influence of COVID-19. We then identified words linked to the individual emotions in order to gain further insight into the words that were contributing to the feelings expressed in popular tweets. We present this information in Figure 4. Among the dominating words for 'fear' and 'sadness', there are 'coronavirus' and 'covid', which are proxies for 'pandemic'; corroborating the hypothesis that current attitudes are driven by the pandemic. Furthermore, the significant presence of the term 'deal' under the 'trust' emotion could mean that many people demand a profitable agreement with the EU.

Figure 5: Eight most frequently discussed subtopics.



As for RQ3, Figure 5 illustrates eight subtopics, represented by the terms that are highly related to them, and their frequency of occurrence in the dataset. By analysing the frequencies, we observe that there are two main subtopics. The first relates to the high-level themes of government and the National Health Service, whilst the second relates specifically to the Labour party and its recent change of leadership. Slightly less common subtopics relate to the COVID-19 pandemic, Boris Johnson, and trade deals.

5 Conclusion

In this paper, we implemented a pipeline to answer three RQs about Brexit using data from Twitter. Our methodology uses lexicon-based approaches and LDA to explore geographic differences in sentiments about Brexit, emotions expressed in the most popular tweets about the topic, and identify the most relevant subtopics alongside it. We infer that our results might be affected by the COVID-19 pandemic.

As for future work, we might consider developing a domain-specific lexicon or training an ML-based system on a labelled dataset for sentiment and emotion analysis.

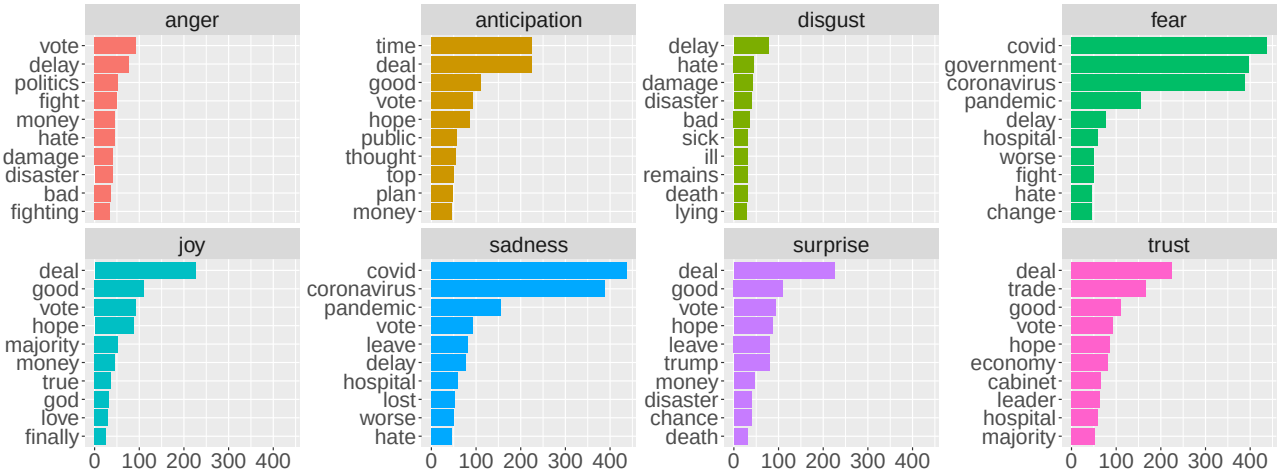


Figure 4: Word-based emotion analysis of popular tweets.

References

Vishal A. and S.S. Sonawane. 2016. Sentiment analysis of twitter data: A survey of techniques. *International Journal of Computer Applications*, 139(11):5–15.

H. Ahmed, M. A. Razzaq, and A. M. Qamar. 2013. Prediction of popular tweets using similarity learning. In *2013 IEEE 9th International Conference on Emerging Technologies (ICET)*, pages 1–6.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 80–88, New York, NY, USA. Association for Computing Machinery.

Michael W. Kearney. 2019. rtweet: Collecting and analyzing twitter data. *Journal of Open Source Software*, 4(42):1829. R package version 0.7.0.

Evgeny Kim and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *CoRR*, abs/1808.03137.

Shoubin Kong, Ling Feng, Guozheng Sun, and Kan Luo. 2012. Predicting lifespans of popular tweets in microblog. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 1129–1130, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Barry Rowlingson. 2019. *geonames: Interface to the "Geonames" Spatial Query Web Service*. R package version 0.999.

Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. *Inf. Process. Manage.*, 52(1):5–19.

Thomas Sampson. 2017. Brexit: The economics of international disintegration. *Journal of Economic Perspectives*, 31(4):163–84.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Asbjørn Steinskog, Jonas Therkelsen, and Björn Gambäck. 2017. Twitter topic modeling by tweet aggregation. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 77–86, Gothenburg, Sweden. Association for Computational Linguistics.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, page 177–184, USA. IEEE Computer Society.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.