Mining Tweets about Brexit Impact on Job Market

Gaurav Gaur (10376766) Muhammad Mohib Khan (10402462) Ma Xiang Xiang (10328541) Nicholas Buttigieg (10354638) Sampriti Mahanty (10379754) Soumya Panda (10356677)

***Abstract*—Brexit has already began to have a significant impact on the job market, especially in the automotive industry. Topic modelling and sentiment analysis provide effective methods of identifying the public opinion towards Brexit, as well as the dominant areas of concern. We then used NER to identify the major geographical areas of interest. Our results showed that the general sentiment towards Brexit is negative and emphasised the significant impact on the automotive industry.**

I. Introduction

The United Kingdom (UK) leaving the European Union (EU) has been an ongoing conversation since 1975, when a referendum was held asking the nation whether the UK should remain in the European Community. Over recent years, this conversation has increased, especially since 2016 when Theresa May gave official notice of the UK's intentions to leave. Countless UK and EU citizens may be affected by this decision, as well as industries relying on trade within the EU. Some of the repercussions are already being observed, with a number of companies within the UK cutting the number of jobs and even relocating.

Twitter has been a vital platform empowering people to voice their opinion and concerns. Given the numerous areas and lives that Brexit (Britain exit) may impact, it is of interest to identify people's attitude towards Brexit and the job market, in addition to any key sectors and geographical areas being discussed.

The following three research questions (RQ) were set out for this paper:

1) Which sectors/industries are being tweeted about in relation to Brexit?
2) What are the general sentiments and emotions towards Brexit and the job situation, as well as the impact of Brexit specifically on the automotive industry job market?
3) What geographical areas are being tweeted about in relation to Brexit?

In the first RQ, the intent is to understand what are the latent themes in the tweets related to Brexit and jobs. We would further like to understand from the detected themes regarding the sectors and industries that are being impacted by Brexit.

In the second RQ, we try to analyse what people believe the impact of Brexit will be on overall employment by analysing their sentiments and emotions. Also, from the results of first RQ, we choose the topic of automotive industry to analyse the impact of Brexit on employment in the industry. Many car manufacturing companies have expressed their concerns over Brexit, while others have hinted to continue investment in the UK regardless of its outcome. Also, automotive industry employs around 856,000 people in the UK, analysing how people think Brexit is going to impact job market can be interesting and important to know.

As for the third and final RQ, understanding which geographical areas are being mentioned alongside Brexit and the job market may give us an indication as to which cities within the UK and countries/cities outside of the UK may be effected the most.

II. Related Work

Topic modelling is an algorithm for finding topics within vast and unstructured collections of literature. This technique infers topics clustering words based on their frequencies and co-occurrences [1]. There have been studies where topic modelling has been used on social media data to identify latent semantic themes. A study on the role of social media in democracy - analysing Brexit and US elections used topic modelling technique to understand the themes discussed with respect to the concerned area of interest [2].

Sentiment analysis is still an evolving area of Natural Language Processing (NLP). There are many different methods used for different levels of granularity, from document level to phrase level. A recent survey on Sentiment Analysis on Twitter data [3] compares two main methods of Machine Learning (ML) and lexicon-based approaches. Some of the early analysis on Twitter data done in [4] used emoticons to find out the sentiment. For example, tweets ending with 'happy' emoticons like ':)' are considered as positive while with 'sad' emoticons like ':(' are considered as negative. Naive Bayes model was then trained to be used as a sentiment classifier. A lexicon-based method for sentiment analysis was used in [5] where dictionaries of words annotated with semantic orientation (polarity and strength) were used.

Similar to sentiment analysis, emotion detection also relies on ML and lexicon-based approaches. A high quality emotion lexicon has been developed in [6] using Mechanical Turk for 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). The annotations obtained were found to be of high quality, motivating their usage for emotion detection for small and large piece of texts alike.

We can use Named Entity Recognition (NER) to find geographical locations from text. There are two main approaches: dictionary-based, and statistical/ML methods. The former requires hand crafted rules, potentially leading to biased results, and are usually incomplete [7]. While ML methods such as hidden Markov model, support vector machines, conditional

random fields [8], and more recently, LSTM nets [9] are the dominating techniques. In this paper, we use the maximum entropy model [10] provided by OpenNLP.

III. Data Collection

This paper focuses on collecting tweet data from Twitter. The provided R pipeline made use of the twitteR package to interface with the Twitter API. The tweets returned via twitteR were truncated, which may alter the results of our analysis since parts of the tweets were missing. Therefore, we opted to use the rtweet [11] package instead, which worked as expected. Data was collected in strict accordance with Twitter's developer terms.

A mixture of popular and recent tweets was collected using various search queries, ignoring retweets. The first and third RQs made use of the same search query, designed to retrieve tweets containing a combination of keywords like 'Brexit', 'industry', 'jobs' or 'unemployment'. The second research question made use of two separate queries to refine the search to the overall job market and the automotive industry job market respectively, including keywords like 'car', 'auto' or 'automotive'. The tweets were stored as CSV files to be used within the pipeline.

Due to Twitter's API limitations, we could only retrieve tweets up to 9 days old. The tweets collected for the first and third RQs contained a total of 11,400 tweets posted between 24/03/19 and 06/03/19. The tweets collected for job market analysis were around 14,350 while 5,170 tweets were used for auto industry analysis.

IV. Methodology

A number of common data cleaning steps were carried out for all the RQs, as well as some question-specific steps. Most of the columns in the tweet datasets were redundant, so we kept a subset of the datasets containing the 'created_at' and 'text' columns. The data cleaning steps were carried out on the tweet text stored in the 'text' column. We removed problematic characters, twitter mentions, digits, special characters, URLs and unnecessary whitespace. The cleaned tweet corpora were then saved as CSV files.

A. RQ 1 - Topic modelling for sectors/industries

We have used an unsupervised text mining technique, Topic Modelling, to answer the first RQ and identify the key themes that are discussed via tweets on Brexit. Latent semantic themes/topics are modelled by analysing how frequently groups of words co-occur in a given body of text corpus. Latent Dirichlet Allocation (LDA), a topic modelling technique has been used in this case. LDA is a generalisation of probabilistic latent semantic indexing that allows a particular text corpora to contain a mixture of multiple topics. For parameterized models such as LDA, the number of topics K is the most important parameter and this information has to be supplied in advance. In this case, we fix the number of topics to 5 by trial and error method. If K is too small, the collection is divided into a few, very general semantic contexts and if K is too large, the collection is divided into too many topics, which may overlap or are sometimes hardly interpretable.

B. RQ 2 - Sentiment and emotion analysis for jobs and the automotive industry

For Sentiment Analysis, we split a tweet sentence into words. These words are then compared to the dictionaries of positive and negative words to get a score for each sentence. The score of positive words is subtracted from score of negative words to get a final score. A sentiment of 'Positive' is then assigned if the final score is greater than 0, 'Negative' for scores less than 0 or 'Neutral' if scores are equal to 0. The process is repeated for all the tweets. Visualization of sentiments is then created from scores, as discussed in Section IV.

For emotion detection, we use the NRC dictionary described in [6]. The dictionary is used to calculate the presence of eight different emotions (discussed in section II) and their corresponding valence.

C. RQ 3 - NER for geographical locations

NER tools are needed so that we can search through the tweet corpus and find locations (cities, countries, etc) mentioned in the tweets. The NLP [12] and Apache openNLP [13] packages were used to carry out the NER. Firstly, we used the Apache OpenNLP Maxent tokenizer to generate three annotators to compute word token annotations, sentence token annotations and entity annotations, specifically location annotations. The individual tweets were tokenised into words and sentences, after which the location tokeniser produced a list of locations from the tweet text data.

It was observed that the word 'Brexit' had a high frequency count and was being labelled as a location, so it was removed from the list of locations. When attempting to remove it during the data cleaning steps, the NER step produced worse results. The list of locations was appended to the tweet corpus, meaning that we now had all the tagged locations of the individual tweets.

V. Results

The objective to conduct topic modelling was to understand the key themes in Brexit tweets and identify particular sectors of interest. The top 10 words for each of the topics have been grouped into word clouds in Fig. 1 and each word cloud is representative of a particular topic. The topics can be grouped into the following themes by analysing the most frequently occurring words within each topic: (1) Brexit voting (Fig. 1(a)) (2) Service sector and trade (Fig. 1(b)) (3) Generic discussions (Fig. 1(c)) (4) Industrial Impact - construction, factories and unemployment (Fig. 1(d)) and (5) Industrial Impact on German Industries/Automobile manufacturers (Fig. 1(e)).

We also tried to rank the topics based on their proportion in the overall tweets. This is indicative that some topics are way more likely to occur in the corpus compared to the other topics. In this case, generic topics regarding Brexit such as losing jobs are the most commonly discussed which is 27.7%

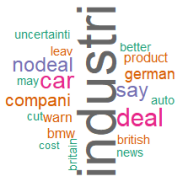
of the overall corpus followed by discussion on Brexit voting (21.2%), Service sector and trade (20%), Industrial impact on German industries (15.9%) and impact on construction, factories and unemployment (15.1%).

(a) Brexit voting

(b) Service sector and trade

(c) Generic discussions

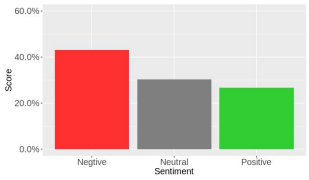
(d) Industrial Impact - Construction, Factories and Unemployment



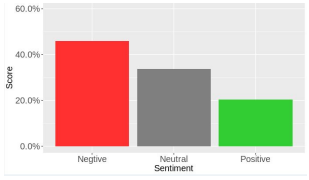
(e) Industrial Impact - German companies / Automobile manufacturers

Fig. 1. Topic Modelling results.

From Fig. 2, it is notable that the dominant sentiment towards Brexit impact on jobs is negative. This shows more people believe Brexit will result in reduction in employment opportunities. Similar trend is observed for auto industry job market in Fig. 2(b) with negative sentiment being the most dominant and even lesser positive sentiments than the overall results in Fig. 2(a).

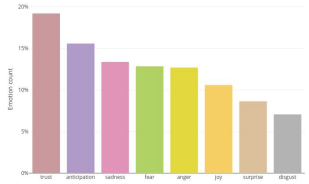


(a) Overall Jobs

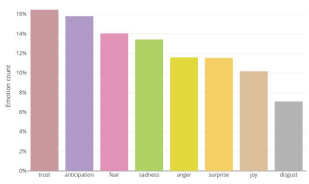


(b) Auto-industry Jobs

Fig. 2. Sentiment Analysis for Brexit impact.



(a) Overall Jobs



(b) Auto-industry Jobs

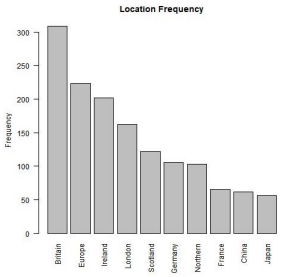
Fig. 3. Emotion Detection for Brexit impact.

The two most common emotions detected in Fig. 3 are 'trust' and 'anticipation', as people are very anxious of how the outcome of Brexit will be on job market. The next three emotions are 'fear', 'sadness' and 'anger', indicating people are generally not happy over Brexit impact on job market.

The frequency axis in Fig. 4(a) shows how many times a location has been mentioned in the tweets, where a single tweet can contain multiple location tags. Fig. 4(b) shows a wordcloud based on the frequency. The most popular locations are mostly inside of the EU. Britain is, by a significant margin, the most discussed location, followed by Europe and other locations inside of the UK. Ireland, Germany and France are the most frequent European countries, while China and Japan the only non European locations in the top ten. Both Germany and Japan have strong ties to the UK in the car manufacturing industry, emphasising how it is one of the industries with the largest impacts.

However, by analysing the data, we noticed two limitations of the current NER procedure: locations with compound names are split, for example, United Kingdom into "United" and "Kingdom", or Northern Ireland into "Northern" and "Ireland", which explains the "Northern" tag in Fig. 4(a), suggesting an overlap of Ireland with Northern Ireland. The second limitation is the inability of recognising smaller regions such as Oxfordshire.

Furthermore, we found that the five most spoken about cities within the UK were London, Manchester, Cambridge, Ellesmere and Sunderland. The inclusion of Ellesmere and Sunderland reinforces the impact on the car industry, as Vauxhall Motors and Nissan have released statements detailing the uncertainty of job safety and the decision to abandon the manufacturing of a new model respectively.



(a) Location frequency plot

(b) Location wordcloud

Fig. 4. Named Entity Recognition results.

VI. Conclusion

We present an analysis on twitter data about Brexit's impact on the job market. Our topic modelling results show that people are mainly tweeting about Brexit in relation to voting, trade and the service sector, German car factories, construction and factories. The general sentiment and emotions towards Brexit are negative. Numerous countries within and outside of the EU were mentioned, like Ireland, Germany, France, China and Japan. It was observed that the automotive industry is one of the areas seeing the greatest negative impact. Future research may be carried out to explore Brexit's impact on other areas, such as health, education and finance.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] W. Hall, R. Tinati, and W. Jennings, "From brexit to trump: social medias role in democracy," *Computer*, vol. 51, no. 1, pp. 18–27, 2018.

[3] V. A. Kharde and S. Sonawane, "Sentiment analysis of twitter data : A survey of techniques," *CoRR*, vol. abs/1601.06971, 2016.

[4] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," vol. 10, 01 2010.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[6] S. M. Mohammad and P. D. Turney, "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*, CAAGET '10, (Stroudsburg, PA, USA), pp. 26–34, Association for Computational Linguistics, 2010.

[7] M. Allahyari, E. D. Trippe, and J. B. Gutierrez, "A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques," 2017.

[8] S. Sekine, "Named Entity: History and Future," *Project notes, New York University*, p. 4, 2004.

[9] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," *Aclweb.Org*, 2018.

[10] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition," *Sixth Workshop on Very Large Corpora*, pp. 152–160, 1998.

[11] M. W. Kearney, *rtweet: Collecting Twitter Data*, 2018. R package version 0.6.7.

[12] K. Hornik, *NLP: Natural Language Processing Infrastructure*, 2018. NLP package version 0.2-0.

[13] I. F. Kurt Hornik, *openNLP: Apache OpenNLP Tools Interface*, 2010. openNLP package version 0.0-8.