Mining Knowledge from News Events

s1034065
Radboud University, Nijmegen, The Netherlands

ABSTRACT

Newspapers are one the most powerful mediums through which information is spread to people across the world. Newspapers can be used as a vital source of knowledge about events, both current and historical. In this paper, we discuss a semi supervised approach to mine a large newspaper corpus pertaining to a specific news event. We first discuss how a large newspaper corpus can be collected from the internet. These unstructured news articles are then converted into a structured format by the means of information extraction. We further perform Topic Modelling to understand what key themes are reported by the these newspapers regarding an event. We demonstrate how the topic model can be evaluated using evaluation metrics that can be correlated with human judgements. Lastly, we perform Sentiment Analysis on the discovered themes to understand what sentiment is conveyed by newspapers to the public regarding these themes.

1 INTRODUCTION

In recent times, due to the rapid growth and advances in information technology, we have access to real time information from every geography of the world at our fingertips. News articles pertaining to any event around the world can be easily accessed by typing in a keyword(s) on the internet to obtain knowledge about the particular event.In this paper, we look at the very timely and influential event 'Brexit'.

On 23rd June 2016, the United Kingdom voted to leave the European Union in a referendum that was dubbed Brexit(Britain+Exit). Though, The Article 50 of the Lisbon Treaty, that formally announces the exit from the European Union was only triggered on the March,29 2017. This allowed Britain a period of 2 years to negotiate a deal with the European union pertaining to trade , immigration , borders etc. This made March,29 2019 the final divorce date. The then Prime Minister Mrs. Theresa May could not get her deal passed in the parliament by the deadline and requested an extension of the negotiation period. The inability to negotiate a deal even after further extensions led to Mrs. May resigning from her post. The new Prime Minister Mr.Boris Johnson requested for a further extension of the period to January 31, 2020 and called for general elections so that the his reformed deal can be passed smoothly with a majority. The Conservative party won the elections and Mr. Johnson was appointed the Prime Minister. The deal was subsequently passed in Parliament and Britain is expected to formally leave the European Union on the January,31 2020.The primary goal of this paper is to understand what are the key topics or themes about 'Brexit' that are being reported by newspaper agencies. Additionally, can these reported themes give us insight into what the future holds?

In this paper, we first demonstrate how we collect several Newspaper articles related to 'Brexit' for a one years period between January 27,2019 to January 27,2020.The articles were collected for this period since this was the crucial period in which the negotiations, extensions and general elections occurred. We used 'The Times Co. UK.' as our source for these news articles as it is one of the most widely read newspapers in the UK. We also collected other relevant information such as date of the article, author and headline apart from the article text. We proceeded by extracting key terms/phrases from the corpus and perform a trend analysis over time. Further, we demonstrate how we used Topic modelling to uncover key topics hidden in the corpus and how these topics can be used to organise and even categorise these articles. Lastly, we perform sentiment analysis to see what sentiment has been conveyed by the authors of these articles about the topics over time.

2 RELATED WORK

Topic modelling approach has been widely used to uncover underlying hidden topics in newspapers. For instance, Newman and Block(2006)[1] uses Probabilistic latent semantic analysis (pLSA) to determine topics in the Pennsylvania Gazette, a major colonial U.S. newspaper from 1728–1800. Also, Yang et al., (2011)[2] uses Latent Dirichlet allocation (LDA) model to find topics in Historical newspapers. The LDA method was introduced by Blei et al. (2003)[3] and is still a popular choice for topic modelling. Though, there is a debate whether the latent space is interpretable. Chang et al., (2009)[4] proposes methods to quantitatively evaluate topic models so as to capture semantic meaning in inferred topics.

The project 'Mining the Dispatch' by Nelson (2010)[5] leverages topic modelling on the Richmond Daily Dispatch historical publication to explore the changes in events in the social and political life of Civil War Richmond. Likewise, in this paper we explore the underlying hidden topics related to 'Brexit' and the changes in these topics over time. We extend this further by also performing sentiment analysis over several topics and understanding the sentiment conveyed by the newspaper about these topics over time.

3 APPROACH

In this section, we discuss the flow of our semi supervised approach. We begin by collecting a large corpus of newspaper articles regarding 'Brexit' from the internet. These downloaded articles are highly unstructured and we use information extraction techniques to retrieve the relevant information. The articles are then pre-processed and cleaned for our analysis. We further perform topic modelling to find the key themes being reported in our corpus and perform sentiment analysis on them.

TxMM'19-'20, Radboud University, Nijmegen, Netherlands

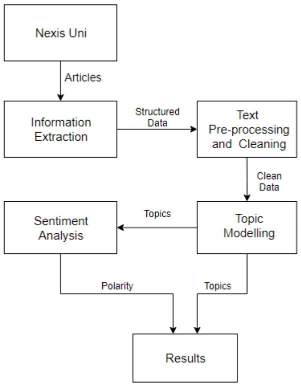
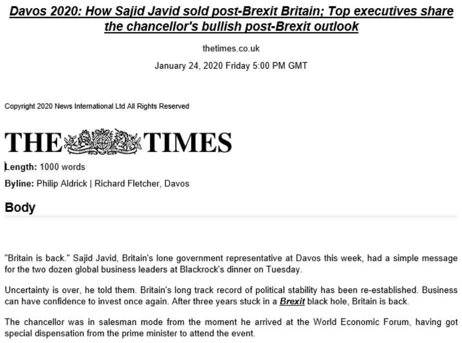


Figure 1: The Flow Chart depicting the high level approach.

3.1 Nexis Uni

To begin with our analysis, we needed to extract relevant news articles from the web. Nexis Uni[6] is one such tool that allows us to search for news articles pertaining to any topic by merely typing in the keyword(s). Nexis uni also allows us to search for articles between two dates and from various online publications. We used the keyword 'Brexit', chose the dates between January 27, 2019 and January 27, 2020 and filtered articles from 'The Times Co. UK' for our analysis. The tool allows the download of these unstructured articles using either Rich Text Format(RTF), Word or PDF and has a limit of only 100 articles at a time.



Davos 2020: How Sajid Javid sold post-Brexit Britain; Top executives share the chancellor's bullish post-Brexit outlook

thetimes.co.uk

January 24, 2020 Friday 5:00 PM GMT

Copyright 2020 News International Ltd All Rights Reserved

THE TIMES

Length: 1000 words

Byline: Philip Aldrick | Richard Fletcher, Davos

Body

"Britain is back." Sajid Javid, Britain's lone government representative at Davos this week, had a simple message for the two dozen global business leaders at Blackrock's dinner on Tuesday.

Uncertainty is over, he told them. Britain's long track record of political stability has been re-established. Business can have confidence to invest once again. After three years stuck in a ***Brexit*** black hole, Britain is back.

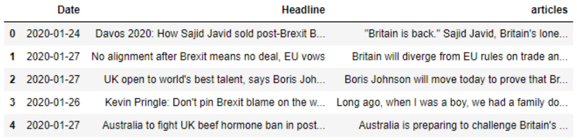
The chancellor was in salesman mode from the moment he arrived at the World Economic Forum, having got special dispensation from the prime minister to attend the event.

Figure 2: An example of an unstructured article downloaded from Nexis Uni.

We downloaded the files in the word format. As seen in figure 2, the articles were unstructured, contained a logo and other irrelevant information.

3.2 Information Extraction

To start with any kind of analysis we need to first convert the unstructured articles into a structured format.To do this, we built a rule based extractor to extract information such as headline, date, authors and the article body and create a structured data frame from it.We evaluated the extractor by checking for the precision on the headline, date and article text columns as these three columns are crucial to our analysis. The precision of our extractor was around 97%.



	Date	Headline	articles
0	2020-01-24	Davos 2020: How Sajid Javid sold post-Brexit B...	"Britain is back." Sajid Javid, Britain's lone...
1	2020-01-27	No alignment after Brexit means no deal, EU vows	Britain will diverge from EU rules on trade an...
2	2020-01-27	UK open to world's best talent, says Boris Joh...	Boris Johnson will move today to prove that Br...
3	2020-01-26	Kevin Pringle: Don't pin Brexit blame on the w...	Long ago, when I was a boy, we had a family do...
4	2020-01-27	Australia to fight UK beef hormone ban in post...	Australia is preparing to challenge Britain's ...

Figure 3: A sample of the structured data after extraction

3.3 Text Pre-processing and Cleaning

After extracting the relevant information into a structured dataframe, we were left with 12,946 articles. We now need to prepare the article texts for topic modelling and sentiment analysis. We initially inspected the article text and found it to be noisy. There were unnecessary punctuation's , white spaces and special characters that added no value to the analysis. Hence, we clean the data by removing the noise. We further tokenise the paragraphs into sentences and then sentences into words to get a bag of words representation.

We further explored our text corpus by looking for the most frequent terms used in our corpus. We saw that words like 'the','an','and' etc. are the most popular words as expected since they are widely used in any sentence structure. These words are called stopwords and removed since they add no additional information to our analysis. We also removed other words that are not stopwords but are widely used in articles. Also, the corpus contained several words that meant the same but were used in different grammatical forms/tenses. We performed lemmatization on these words to transform them into their base form which helps with the analysis.

3.4 Topic Modelling

Topic modelling refers to the task of identifying hidden topics within a set of documents.In this paper, we use Latent Dirichlet allocation for topic modelling. We chose this method because it has shown to be highly productive and has performed better than other methods on several newspaper datasets for instance Yang et al., (2011)[2] and Nelson (2010)[5].

LDA is a generative probabilistic model and can be thought of as three-layer hierarchical Bayesian model. Firstly, each word of a document collection is modeled as a mixture over an underlying hidden set of topics. Each topic is, thereby, modeled as an infinite mixture over a set of topic probabilities.

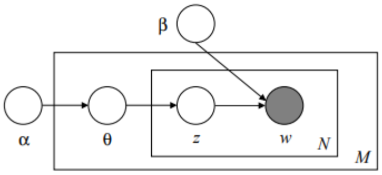


Figure 4: Graphical model representation of LDA. Blei et al. (2003)[3]

Further, to evaluate the topic model we use the Coherence measure (Cv) proposed by M. Röder et al.(2015)[7]. This measure combines several semantic coherence measures such as the indirect cosine measure, Normalised Pointwise Mutual Information(NPMI) and the boolean sliding window. This measure has shown to exceed other benchmarks with respect to correlation with human judgements. Additionally, LDA also has two hyperparameters namely alpha and beta that can be tuned to obtain better results. Alpha represents document-topic density. The higher the value of alpha, documents are made up of more topics and vice versa. Beta represents topic-word density. The higher the value of beta, topics are made up of most of the words in the corpus and vice versa.

3.5 Sentiment Analysis

We further organise the articles based on the topics found in the topic modelling phase. We can then perform sentiment analysis on these topics to see whether the newspaper reports positively or negatively on them.The key point to note here is that we cannot build a sentiment classifier here since we do not have a labelled dataset. To get around this, we use the TextBlob[8] library in python. TextBlob uses PatternAnalyzer from the pattern package. The PatternAnalyzer leverages SentiWordNet which is a publicly available lexical resource for opinion mining and finds words and phrases it can assign polarity and subjectivity to. It then averages them all together for a longer text.

The polarity score outputted by TextBlob is within the range of -1 to 1. Hence we define positive sentiment between the 0.1 to 1 range, negative sentiment between the -0.1 to -1 range and neutral between the -0.1 to 0.1 range. To evaluate the performance of the sentiment model we manually tag random 100 articles and evaluate the performance using the accuracy measure.

4 RESULTS

In this section, we discuss our results and findings after performing the analysis. As a part of exploring the data we first searched and extracted the mentions of a few named entities such as 'Boris Johnson', 'Theresa May', 'Conservative Party' and 'Labour Party'. We further performed simple trend analysis over time on these entities to analyse if there were any patterns in their mentions by the newspaper. The result of this analysis can be seen in figure 5.

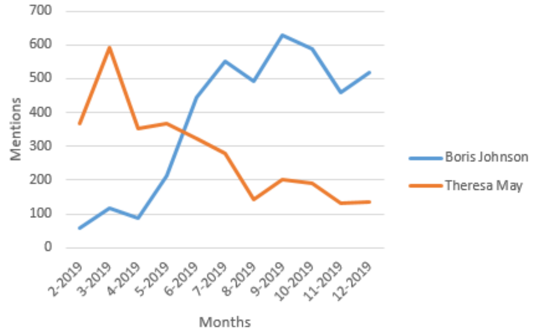


Figure 5: Frequency of articles mentioning Boris Johnson and Theresa May atleast once over time.

In figure 5, we see that Mrs.Theresa May was mentioned in about 600 articles in March, 2019. This could be attributed to the fact that 29 March, 2019 was the original date for Britain's divorce with the Europe and the country was in a state of panic. After this period, Mrs. May's mentions have decreased while Mr. Boris Johnson's shot up as he became the leader of the Conservative Party after Mrs. May quit.

We further used topic modelling on the article corpus. We performed a grid search on the hyperparameters - the number of topics, alpha and beta and checked the coherence score(Cv) for every combination of the hyperparameter. The results can be seen in figure 6.

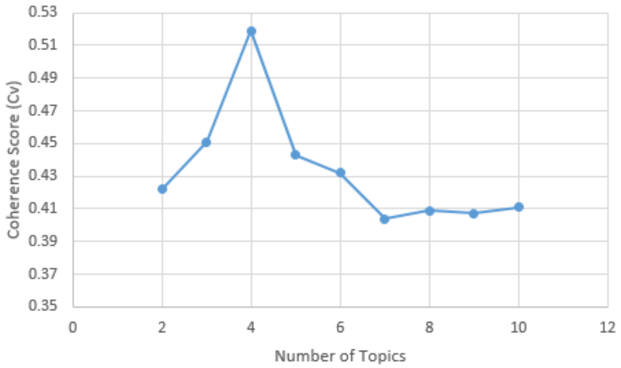


Figure 6: Coherence Score for different number of topics.

After performing the grid search for the hyperparameters and analysing figure 6, we found the optimal number of topics to be 4 with a alpha of 0.91 and beta of 0.31. We further also tried to understand the semantic meaning captured by the topics and their underlying distribution. After carefully analysing the results, we found that the topics closely represented 'Business', 'General News', 'Politics' and 'Constituents of UK'. The topics and their word distribution is shown in table 1.

Table 1: Topics discovered by LDA

Business	General News	Politics	Constituents of UK
cent	say	deal	country
company	time	party	ireland
business	go	vote	scotland
market	make	government	border
rise	work	election	northern
fall	come	labour	trade
price	get	leave	kingdom
share	good	leader	government

From table 1 it becomes clear that 'The Times Co. UK' essentially reports about these 4 topics when they write any article regarding 'Brexit'. The key point to remember here is that these topics are not present in articles independently of each other. In other words, it is possible for an article to contain 2 or more topics. Though, these underlying topics not only can be used to organise articles according to their themes, but also to recommend articles to users. For example, for a investor who is interested in reading an article about the topic 'Business', can be suggested other articles in this category as future reads rather than articles from other categories. Also, this method was able to vaguely classify articles into respective broad categories in an unsupervised manner which saves a lot of time compared to manually annotating each article and training a classifier. Alternatively, the topic modelling output can also be used as an input feature to a classifier which might aid better classification results.

We further went ahead and performed sentiment analysis on the articles in our 'Business' topic category. We used the TextBlob[8] package from python to get sentiment polarity scores for each article text since we did not have a labelled dataset. To evaluate the results we manually annotated 100 random articles within the 'Business' category and calculated the accuracy. The accuracy was fairly low at 61%. After carefully analysing the results, we found that article texts contained various sentiments and the aggregation performed by TextBlob could not capture the overall sentiment correctly. To get around this, we instead performed the sentiment analysis on the article headlines. The accuracy on the article headlines was found to be 78% which is a major improvement.

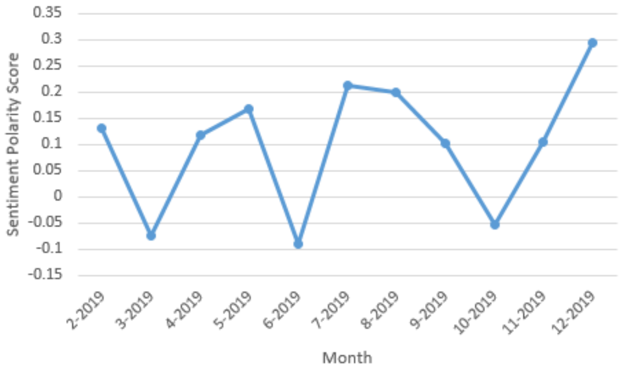


Figure 7: Average sentiment polarity score over time.

Lastly, we averaged the polarity scores for every month and performed a trend analysis. The results can be seen in figure 7. In the figure, we can see that the average polarity scores are marginally positive usually and sharply fall to slightly negative in March, 2019. This maybe attributed to the first Brexit extension requested and the country being in a state of panic.The other two sharp dips take place in June, 2019 and October, 2019 which coincide with Mrs.May quitting as the Prime Minister and Britain requesting a second extension respectively. Additionally, the trend tends to be highly positive in December and can be attributed to the fact that Mr. Johnson's deal was finally passed in parliament.

5 DISCUSSION AND OUTLOOK

In this paper, we discussed how we can mine knowledge about prominent news events using news articles in a semi supervised way. We used topic modelling to uncover underlying hidden topics within the articles that not only can be used to organise and recommend news articles, but also may be used as an input to classifiers to increase their classification power. We also discussed ways to evaluate the topic models by using a combination of coherence measures and performing a grid search on the model hyperparameters. The output of the LDA model was fairly good but in future we would like to perform a extensive comparison using different topic models and different combinations of coherence measures.

Also, we discussed how we can use sentiment analysis on the articles within a specific discovered topic category. The model performed badly on the article text. This was probably because a news article contains comments from multiple people, the writers assessment and facts which is difficult to model. Hence, we went ahead and performed the analysis on the article headlines which was a smaller text corpus but was effective. The accuracy of our sentiment model could be further improved by using a list of positive and negative financial lexicons, since we were working with articles that included the topic 'Business'. This could have potentially increased the accuracy.Additionally, sentiment analysis on articles regarding 'Business' can be highly useful for investors and businessmen and this can also be an input into algorithmic trading systems or other financial systems. In future, we could look also at including articles from other news publications and compare and the contrast the results.

It is also important to note that the claims made about the ups and downs in the trend analysis purely resembles correlation and the paper in no way implies causation of any kind. Also, the dataset used in this paper was from one particular news source and hence the analysis does not represent the complete knowledge about 'Brexit'.

REFERENCES

[1] David J Newman and Sharon Block. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767, 2006.

[2] Tze-I Yang, Andrew Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, 2011.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[5] Robert K Nelson. Mining the dispatch. *Mining the dispatch*, 2010.

[6] Nexis Uni. (2018). nexis uni™: Empowering academic research for digital natives, 2014.

[7] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015.

[8] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 3, 2014.