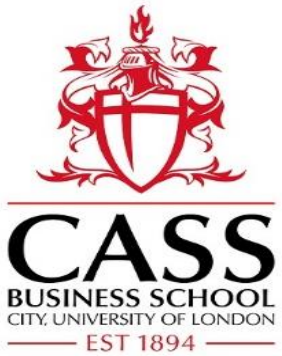


SMM694 FINAL COURSE PROJECT

ANALYSING THE EFFECT OF BREXIT REFERENDUM ON PRESS AND COMPANIES

Group 12

Motiwala, Yousuf

Mwangi, Joseph

Rastogi, Kiran

Wang, Xinye

Warsi, Ahmed

CITY'S BUSINESS SCHOOL

CITY, UNIVERSITY OF LONDON

17th July 2020

Abstract

The United Kingdom European Union membership referendum commonly referred to as the Brexit referendum took place on 23 June 2016 in the United Kingdom. The objective of this project is to analyze a real-world dataset containing press-releases, business reports, and financial analysts reports. And to use the insights emerging after analyzing the performance of British, publicly listed companies in the aftermath of the 2016 Brexit Referendum. Counterfactual data is obtained from a group of publicly listed companies based in France and Germany to estimate how British companies could have performed in case of no-leave. We have applied natural language processing techniques to find the difference in topics from pre to post Brexit referendum in the press documents.

Table of Contents

1. Dataset

We have two datasets for the project:

a. Business Press Corpus

- Articles contain keywords like sustainability, corporate social repo, etc. The choice for the keywords is motivated by the prior literature and empirical evidence showing good companies are better equipped to navigate turbulent competitive landscapes or investments in sustainable initiatives and/or CSR are supposed to create a positive corporate identity, which, in turn, can act as a buffer protecting companies from institutional and market uncertainty
- The period of data is April 23, 2016 - August 23, 2016 (four months around the Brexit Referendum of June 23, 2016)
- The press data contains 603 articles from 4 newspapers namely Financial Times, Telegraph, The Guardian and The Times.

b. Company Level Data

- Economic and financial data for various British and Non-British companies. The financial data includes but not limited to share price, assets held by the company and debt to asset ratio.
- Annual report for British and Non-British companies. It is important to note that some reports are not available in certain years due to mergers and acquisitions that change the legal entity behind the company.
- The period of the data is January 2014 – December 2018.
- The company-level data contains 861 annual reports.

2. Analysis of Business Press Corpus

The Brexit referendum date is 23 June 2016, we have divided the dataset in terms of pre and post Brexit referendum to have a distinction in the topics that emerged from press data. After data cleaning and restructuring we are left with the data source containing time and text of the article.

2.1. Natural Language Processing Pipeline

The NLP pipeline behind our topic model comprises of several steps, the first step is the conversion of all the text to lowercase, converting hyphens and underscores to space. Next, we have used python library spaCy's largest English model for creating NLP pipeline, i.e., en_core_web_lg. Stopwords from python library spaCy along with a couple more of our own are used to make the analysis better.

2.2. Tokenize the text

Tokenization is a way of converting text into smaller units called tokens they can be either word, characters, or subwords. We have used python library NLP to carry out tokenization. After tokenization, every article contains a list of tokens brought out from that piece of text and saved along with the text. We have also removed some commonly used terms like [‘of’, ‘with’, ’without’, ’and’, ‘or’, the’, ‘a’, ‘not’, ‘be’, ‘to’, ‘this’, ‘who’,’ in’].

2.3. Bigrams & Trigrams

N-gram is the sequence of N words, so a 2-gram (or bigram) is a two-word sequence of words like ‘wireless network’, ‘wireless mouse’, or ‘your mouse’, and a 3-gram (or trigram) is a three-word

sequence of words like ‘please turn your’, or ‘turn your mouse’. We have processed press data text and created the trigrams model from the bigrams model.

2.4. Topic Modeling

Topic Modeling is a technique to extract the hidden topics from large volumes of text. Topic models are a type of statistical language models which are used for finding hidden structures in a collection of texts. Several algorithms can be used for topic modelling, we have used Latent Dirichlet Allocation algorithm (LDA).

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

For training the model, we have created text corpus and a data dictionary. After that, we have used Mallet software to estimate a set of competing topic models, each of which retains a unique number of topics ranging 5 to 21 in our case. We have set the maximum number of topics to 21 to maintain efficiency and reduce processing time. Considering the coherence score (a metric that expresses the face validity of the inductively derived topics), we have retained the model with ten topics in pre-Brexit data and eight topics in post-Brexit data.

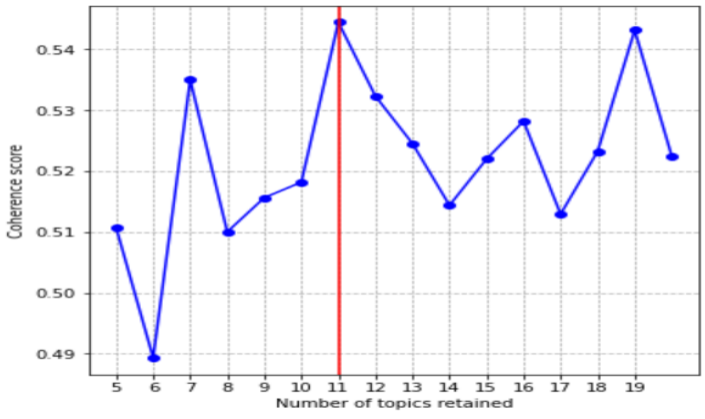


Figure 1: Pre-Brexit text coherence score based on this we have retained ten topics

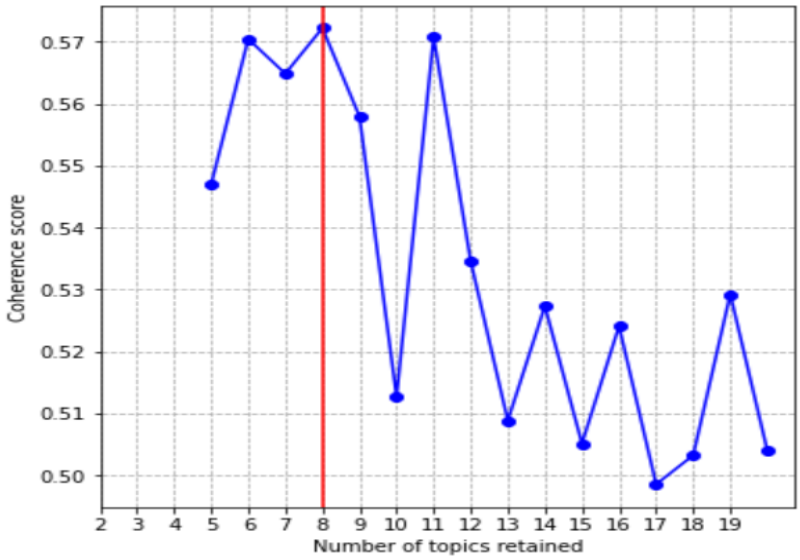


Figure 2: Post-Brexit text coherence score, we have retained eight topics using the above plot

2.5. Pre-Brexit Topic Modeling:

After LDA topic modelling using Mallet software, we have found the following 10 topics:-

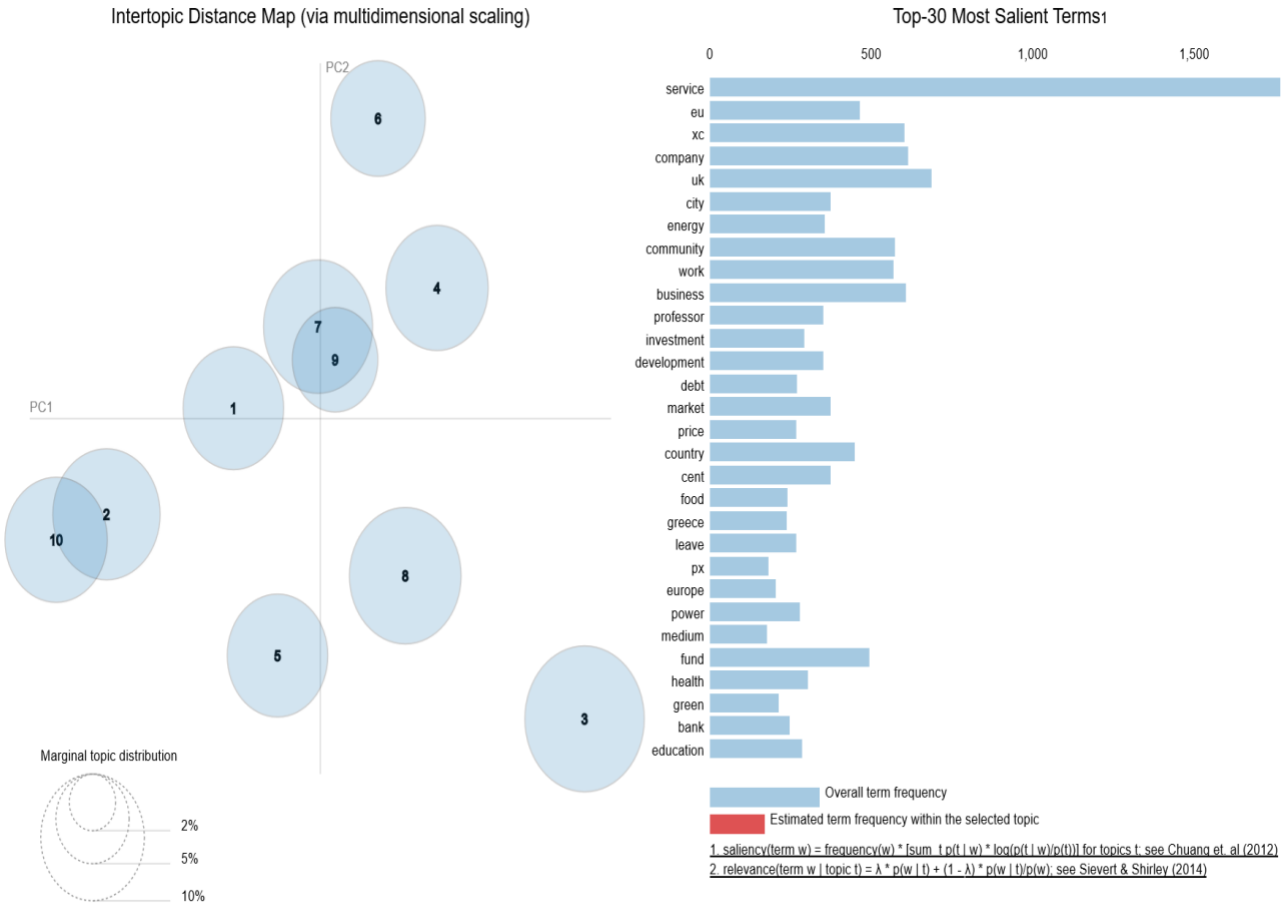
No	Topics
1	0.025*"energy" + 0.017*"food" + 0.016*"power" + 0.014*"green" + 0.012*"company" + 0.012*"industry" + 0.012*"waste" + 0.011*"water" + 0.009*"solar" + 0.009*"carbon"
2	0.025*"company" + 0.018*"market" + 0.018*"investment" + 0.016*"price" + 0.013*"xc" + 0.012*"business" + 0.012*"cent" + 0.011*"pay" + 0.011*"investor" + 0.011*"big"
3	0.088*"service" + 0.021*"community" + 0.018*"professor" + 0.011*"education" + 0.011*"director" + 0.010*"john" + 0.009*"royal" + 0.008*"officer" + 0.007*"founder" + 0.007*"david"
4	0.027*"xc" + 0.025*"city" + 0.011*"home" + 0.010*"child" + 0.010*"build" + 0.010*"house" + 0.009*"day" + 0.006*"people" + 0.005*"urban" + 0.005*"area"
5	0.033*"eu" + 0.029*"uk" + 0.017*"leave" + 0.016*"country" + 0.014*"europe" + 0.012*"britain" + 0.012*"vote" + 0.011*"european" + 0.010*"brexit" + 0.009*"campaign"
6	0.009*"brand" + 0.007*"thing" + 0.007*"sustainability" + 0.007*"year" + 0.006*"good" + 0.006*"people" + 0.006*"design" + 0.006*"island" + 0.005*"start" + 0.005*"meet"
7	0.012*"health" + 0.011*"change" + 0.010*"plan" + 0.009*"people" + 0.008*"cost" + 0.008*"nhs" + 0.008*"report" + 0.007*"social" + 0.007*"work" + 0.006*"future"
8	0.022*"work" + 0.020*"business" + 0.018*"development" + 0.010*"support" + 0.009*"sustainable" + 0.009*"world" + 0.009*"government" + 0.009*"community" + 0.008*"people" + 0.008*"improve"

9	0.018*"px" + 0.017*"medium" + 0.014*"font" + 0.014*"reef" + 0.012*"austin_news_deck_web" + 0.011*"great" + 0.010*"uk_asset_font" + 0.010*"url_http_eip_telegraph" + 0.009*"margin_leave" + 0.009*"format"
10	0.018*"debt" + 0.016*"greece" + 0.013*"bank" + 0.013*"fund" + 0.012*"cent" + 0.012*"imf" + 0.012*"rate" + 0.012*"growth" + 0.011*"finance" + 0.011*"euro"

Topic 5, 10 and 2 revolve around Brexit and the impact of Brexit on industries.

2.6 Visualization

The visualization of topic modelling for pre-Brexit text using python package pyLDAvis:-



From the visualization, we can notice that topics 10, 2 and 5 are closely placed in the intertopic distance map. Diving further into the analysis we have found the term to term probabilities in the text.

	0	1	2	3	4	5	6	7	8	9
0	energy	food	power	green	company	industry	waste	water	solar	carbon
0	(0.025)	(0.017)	(0.016)	(0.014)	(0.012)	(0.012)	(0.012)	(0.011)	(0.009)	(0.009)
1	company	market	investment	price	xc	business	cent	pay	investor	big
0	(0.025)	(0.018)	(0.018)	(0.016)	(0.013)	(0.012)	(0.012)	(0.011)	(0.011)	(0.011)
2	service	community	professor	education	director	john	royal	officer	founder	david
0	(0.088)	(0.021)	(0.018)	(0.011)	(0.011)	(0.01)	(0.009)	(0.008)	(0.007)	(0.007)
3	xc	city	home	child	build	house	day	people	urban	area
0	(0.027)	(0.025)	(0.011)	(0.01)	(0.01)	(0.01)	(0.009)	(0.006)	(0.005)	(0.005)
4	eu	uk	leave	country	europe	britain	vote	european	brexit	campaign
0	(0.033)	(0.029)	(0.017)	(0.016)	(0.014)	(0.012)	(0.012)	(0.011)	(0.01)	(0.009)
5	brand	thing	sustainability	year	good	people	design	island	start	meet
0	(0.009)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
6	health	change	plan	people	cost	nhs	report	social	work	future
0	(0.012)	(0.011)	(0.01)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)
7	work	business	development	support	sustainable	world	government	community	people	improve
0	(0.022)	(0.02)	(0.018)	(0.01)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.008)
8	px	medium	font	reef	austin_news_deck_web	great	uk_asset_font	url_http_eip_telegraph	margin_leave	format
0	(0.018)	(0.017)	(0.014)	(0.014)	(0.012)	(0.011)	(0.01)	(0.01)	(0.009)	(0.009)
9	debt	greece	bank	fund	cent	imf	rate	growth	finance	euro
0	(0.018)	(0.016)	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)

2.7 Dominant Topic Analysis

We have also tried to find dominant topics in the text along with topic percentage contribution in that article. Some of them are-

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	7.0	0.3024	water, food, environmental, waste, climate_cha...	highlight ed crooks is fascinated by a biograp...
1	1	0.0	0.2042	company, business, price, investment, xc, ener...	highlight 'if china could be persuaded to cons...
2	2	3.0	0.6410	debt, growth, bank, fund, greece, cent, bn, ra...	after months of wrangling eurozone finance min...
3	3	0.0	0.2294	company, business, price, investment, xc, ener...	how can the world best combat global warming m...
4	4	0.0	0.4289	company, business, price, investment, xc, ener...	zaoui co the tiny european advisory firm set u...
5	5	3.0	0.2811	debt, growth, bank, fund, greece, cent, bn, ra...	sir caroline binham 's report fca sets out pla...
6	6	1.0	0.3140	xc, people, day, child, good, brand, world, ag...	it has been a little over a year since i began...
7	7	0.0	0.3441	company, business, price, investment, xc, ener...	the world 's largest government backed investm...
8	8	4.0	0.2532	eu, uk, leave, europe, state, britain, country...	if we take control we can deliver for our mari...
9	9	0.0	0.5131	company, business, price, investment, xc, ener...	i am a year old vicar married but as yet no ch...

2.8 Post-Brexit Topic Modelling

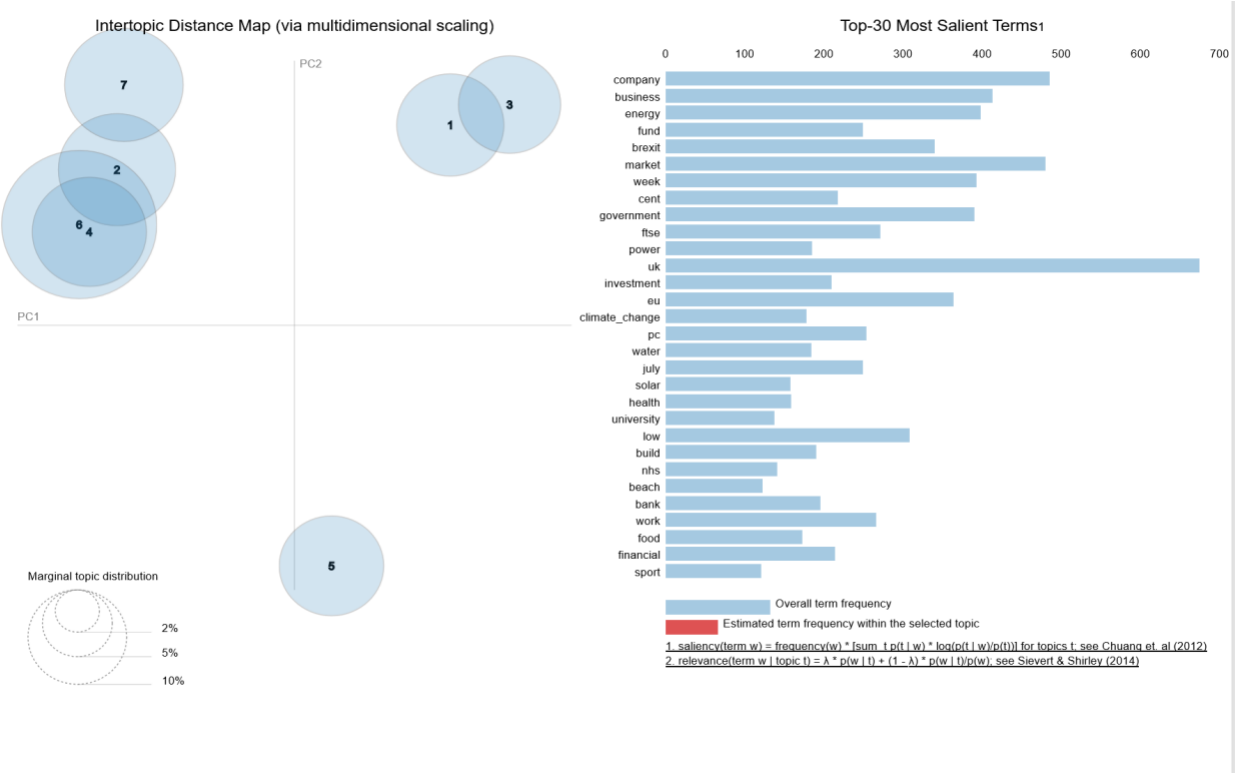
We have selected seven topics for post Brexit text. Further analysing the text using LDA based topic modelling with the help of Mallet software we have topic numbers 6 and 7 related to Brexit.

No	Topic keywords
1	0.015*"water" + 0.013*"build" + 0.012*"work" + 0.010*"home" + 0.010*"city" + 0.010*"live" + 0.009*"people" + 0.008*"px" + 0.008*"waste" + 0.007*"xc"

2	0.018*"fund" + 0.014*"government" + 0.011*"health" + 0.010*"nhs" + 0.010*"service" + 0.010*"financial" + 0.009*"xc" + 0.009*"child" + 0.008*"public" + 0.008*"council"
3	0.013*"include" + 0.011*"beach" + 0.010*"hotel" + 0.009*"contact" + 0.009*"restaurant" + 0.009*"room" + 0.008*"sea" + 0.008*"view" + 0.008*"offer" + 0.007*"uk"
4	0.030*"energy" + 0.014*"power" + 0.013*"climate_change" + 0.013*"uk" + 0.012*"solar" + 0.011*"government" + 0.010*"country" + 0.010*"oil" + 0.009*"world" + 0.009*"policy"
5	0.012*"university" + 0.011*"food" + 0.011*"sport" + 0.009*"programme" + 0.009*"school" + 0.008*"game" + 0.007*"woman" + 0.006*"student" + 0.006*"world" + 0.006*"team"
6	0.015*"uk" + 0.015*"week" + 0.015*"market" + 0.014*"brexit" + 0.011*"ftse" + 0.010*"pc" + 0.010*"eu" + 0.010*"july" + 0.010*"low" + 0.009*"high"
7	0.034*"company" + 0.029*"business" + 0.015*"cent" + 0.014*"investment" + 0.009*"group" + 0.009*"bn" + 0.008*"large" + 0.008*"industry" + 0.008*"deal" + 0.008*"change"

2.9 Visualization

Visualizing the topics using pyLDAvis python library, we can notice that 2,4,6 and 7 lie closely in the intertopic map.



Looking at the term to term probabilities for post Brexit text-

	0	1	2	3	4
0	water	build	work	city	home
0	(0.015)	(0.013)	(0.012)	(0.01)	(0.01)
1	fund	government	health	nhs	service
0	(0.018)	(0.014)	(0.011)	(0.01)	(0.01)
2	include	beach	hotel	contact	restaurant
0	(0.013)	(0.011)	(0.01)	(0.009)	(0.009)
3	energy	power	climate_change	uk	solar
0	(0.03)	(0.014)	(0.013)	(0.013)	(0.012)
4	university	food	sport	programme	school
0	(0.012)	(0.011)	(0.011)	(0.009)	(0.009)
5	uk	week	market	brexit	ftse
0	(0.015)	(0.015)	(0.015)	(0.014)	(0.011)
6	company	business	cent	investment	group
0	(0.034)	(0.029)	(0.015)	(0.014)	(0.009)

2.10 Dominant Topic Analysis

In this case, we have found some of the below mentioned dominant topics along with the topic percentage-

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	2.0	0.3341	food, sport, game, london, waste, don, people,...	almost trillion in food is thrown away lost or...
1	1	5.0	0.3513	energy, power, solar, oil, company, world, cen...	it 's hard to get a clear picture of the emplo...
2	2	1.0	0.4889	people, work, water, university, community, li...	alive and kicking is an award winning uk chari...
3	3	5.0	0.4382	energy, power, solar, oil, company, world, cen...	elon musk faces a battle with shareholders ove...
4	4	3.0	0.3322	business, company, fund, investment, xc, pay, ...	michael sharp who is stepping down as chief ex...
5	5	3.0	0.4465	business, company, fund, investment, xc, pay, ...	dropbox once one of silicon valley 's fastest ...
6	6	5.0	0.2708	energy, power, solar, oil, company, world, cen...	the mayor of london is pressing ahead with pla...
7	7	5.0	0.4164	energy, power, solar, oil, company, world, cen...	highlight there has been little enthusiasm for...
8	8	7.0	0.4527	eu, uk, economy, growth, economic, market, ban...	greece 's banking system yesterday took a big ...
9	9	4.0	0.5144	government, health, nhs, climate_change, plan,...	there was something akin to desperation in the...

2.11 Whole Press Corpus Topic Modelling

The same methodology, used above for the pre-Brexit and post-Brexit press, is applied to the whole press dataset from April 23, 2016, to August 23, 2016. We use Mallet software to estimate a set of models, each of which retains a unique number of topics ranging 5 to 15. We have set the maximum

number of topics to 15 to maintain efficiency and reduce processing time. Considering the coherence score, we have retained the model with 13 topics in the whole press dataset.

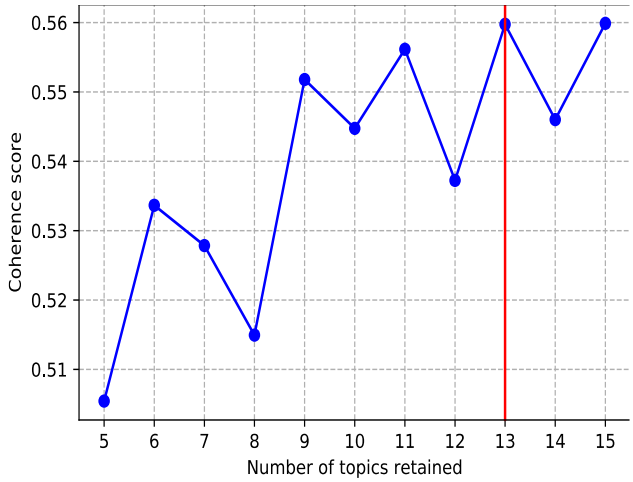


Figure 3: Whole press data text coherence score, we have retained thirteem topics using the above plot

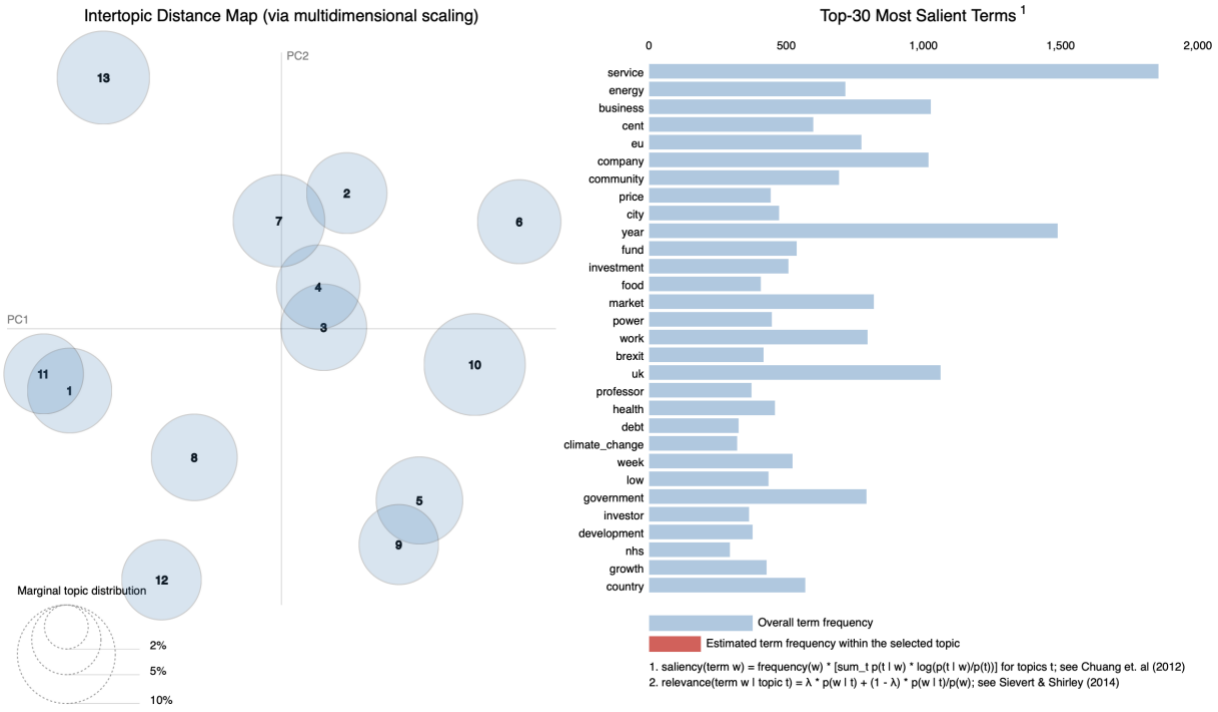
From the model, we have found the following 13 topics, topic numbers 6 & 10 are related to Brexit.

No	Topics
1	0.025*"city" + 0.014*"university" + 0.012*"home" + 0.011*"building" + 0.010*"water" + 0.009*"build" + 0.009*"world" + 0.008*"year" + 0.008*"student" + 0.008*"design"
2	0.019*"health" + 0.017*"nhs" + 0.013*"plan" + 0.012*"scheme" + 0.012*"pension" + 0.012*"government" + 0.012*"council" + 0.012*"year" + 0.010*"care" + 0.009*"trust"
3	0.010*"campaign" + 0.010*"people" + 0.008*"party" + 0.006*"vote" + 0.006*"it" + 0.006*"change" + 0.006*"thing" + 0.005*"state" + 0.005*"day" + 0.005*"claim"
4	0.042*"energy" + 0.022*"power" + 0.019*"climate_change" + 0.016*"green" + 0.012*"solar" + 0.012*"fuel" + 0.010*"emission" + 0.010*"climate" + 0.010*"uk" + 0.010*"wind"
5	0.033*"cent" + 0.032*"year" + 0.024*"price" + 0.016*"growth" + 0.014*"rise" + 0.014*"increase" + 0.012*"low" + 0.011*"demand" + 0.011*"big" + 0.010*"company"
6	0.030*"eu" + 0.019*"debt" + 0.014*"european" + 0.013*"country" + 0.013*"uk" + 0.013*"government" + 0.011*"greece" + 0.010*"deal" + 0.010*"europe" + 0.010*"euro zone"
7	0.017*"work" + 0.016*"development" + 0.015*"people" + 0.013*"community" + 0.012*"social" + 0.011*"country" + 0.009*"international" + 0.009*"local" + 0.009*"relate" + 0.009*"global"

8	0.039*"business" + 0.017*"work" + 0.011*"team" + 0.011*"company" + 0.010*"small" + 0.008*"good" + 0.007*"it" + 0.007*"model" + 0.007*"award" + 0.007*"general"
9	0.030*"company" + 0.025*"investment" + 0.024*"fund" + 0.015*"investor" + 0.014*"group" + 0.013*"invest" + 0.011*"oil" + 0.011*"year" + 0.010*"money" + 0.010*"return"
10	0.018*"uk" + 0.018*"market" + 0.017*"brexit" + 0.016*"week" + 0.011*"ftse" + 0.011*"eu" + 0.011*"july" + 0.010*"referendum" + 0.009*"low" + 0.008*"high"
11	0.012*"include" + 0.011*"beach" + 0.010*"hotel" + 0.009*"offer" + 0.008*"good" + 0.008*"restaurant" + 0.007*"room" + 0.007*"child" + 0.007*"contact" + 0.007*"read"
12	0.024*"food" + 0.016*"waste" + 0.013*"product" + 0.011*"consumer" + 0.011*"sustainability" + 0.010*"good" + 0.010*"produce" + 0.009*"company" + 0.009*"industry" + 0.009*"brand"
13	0.088*"service" + 0.020*"community" + 0.018*"professor" + 0.012*"john" + 0.011*"education" + 0.011*"director" + 0.010*"royal" + 0.008*"david" + 0.008*"officer" + 0.006*"chief_executive"

2.12 Visualization using pyLDAvis

Visualizing the topics using pyLDAvis python library, we can notice that 2,3,4 and 7 lie closely, while topic 6 and 10 lie on the same side in the intertopic map.



Looking at the term to term probabilities for the whole press data around Brexit:

	0	1	2	3	4
0	city	university	home	building	water
0	(0.025)	(0.014)	(0.012)	(0.011)	(0.01)
1	health	nhs	plan	scheme	pension
0	(0.019)	(0.017)	(0.013)	(0.012)	(0.012)
2	campaign	people	party	vote	it
0	(0.01)	(0.01)	(0.008)	(0.006)	(0.006)
3	energy	power	climate_change	green	solar
0	(0.042)	(0.022)	(0.019)	(0.016)	(0.012)
4	cent	year	price	growth	rise
0	(0.033)	(0.032)	(0.024)	(0.016)	(0.014)
5	eu	debt	european	country	uk
0	(0.03)	(0.019)	(0.014)	(0.013)	(0.013)
6	work	development	people	community	social
0	(0.017)	(0.016)	(0.015)	(0.013)	(0.012)
7	business	work	team	company	small
0	(0.039)	(0.017)	(0.011)	(0.011)	(0.01)
8	company	investment	fund	investor	group
0	(0.03)	(0.025)	(0.024)	(0.015)	(0.014)
9	uk	market	brexit	week	ftse
0	(0.018)	(0.018)	(0.017)	(0.016)	(0.011)
10	include	beach	hotel	offer	good
0	(0.012)	(0.011)	(0.01)	(0.009)	(0.008)
11	food	waste	product	consumer	sustainability
0	(0.024)	(0.016)	(0.013)	(0.011)	(0.011)
12	service	community	professor	john	education
0	(0.088)	(0.02)	(0.018)	(0.012)	(0.011)

3. Analysis of Company Data

The analysis will involve determining whether the inclusion of topic modelling information will give insight emerging after analyzing the performance of British, publicly listed companies in the aftermath of the 2016 Brexit Referendum.

This will involve having analyzed the share performance using only financial information and then analyzing with both financial and topic modelling. It is important to note that financial data was made annual to be compared with the annual report. In this case, we used 2014, 2015 and 2016 as pre-Brexit and 2017 onwards as post-Brexit.

3.1. Exploratory Analysis

Some exploratory analysis is conducted in the financial data to get some insight on the general performance of the companies. The simple returns, $\frac{P_t}{P_{t-1}}$, where P_t is the price at time t, is used as a measure of company performance.

A comparison of the average returns by British and Non-British companies over the four years is conducted to determine how companies have been impacted by Brexit.

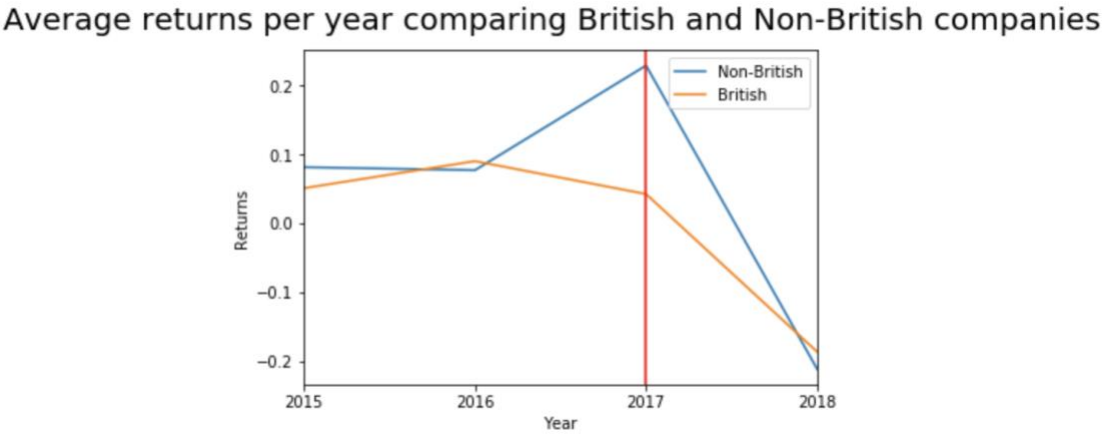


Figure 4: Average returns per year comparing British and Non-British companies.

Figure 4 shows that the performance of both British and Non-British companies declined post-Brexit. This shows that Brexit harmed both types of companies. Wielechowski (2016) suggests that poor performance by companies could be due to uncertainty in the economic environment in Europe. It is not yet known if the UK will have access to the European free market.

This leads us to investigate which sectors have been most impacted by Brexit.

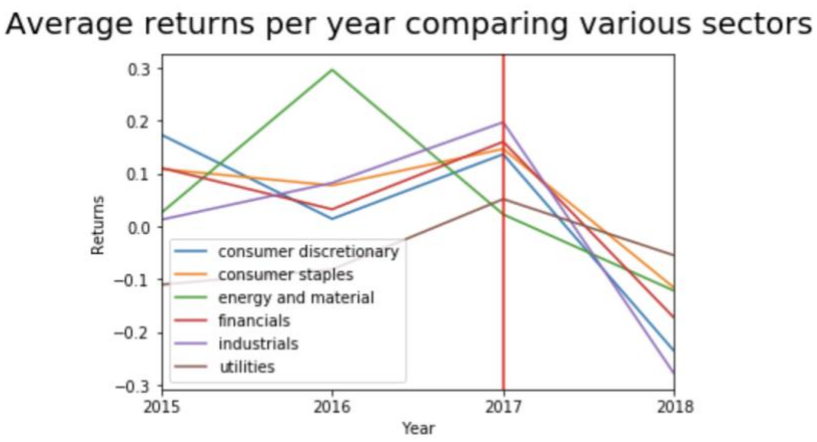


Figure 5: Average returns per year comparing various sectors

Figure 5 shows that all industries suffered worse performance post Brexit. The industry which performed the worst is industrials. Stewart (2019) believes that the industrial sector like clothing and textile, chemicals, pharmaceuticals and manufacturing will be the highest exposed because of this sector exports high shares of output to the EU, or purchase high levels of inputs from the EU (or both). Therefore, the uncertainty about the free movement of people and goodwill impact this sector the most.

3.2. Data preparation for topic-probabilities features

For generating the topic-probability features, we passed the documents through the NLP pipeline discussed in the lectures. We noticed additional words that were impacting the topic modelling and therefore added a custom dictionary of stop words that included the name of the companies, positions in the companies, months, etc. We then removed 5 documents that had the length of >320,000 words for the speed efficiency and also sampled 33% of the documents. Data has been sampled by pre and post Brexit so that there is an equal ratio of representation of both the events.

In the original data, ~37% of the companies are British and the same has been taken care in the pre and post Brexit sampling (40% and 43% respectively).

Additionally, we have taken 6 topics even though the coherence score returned a better value at 10 topics. By plotting 10 topics we realised that there is a high overlap, then we moved to 7 and then to 6 because it gives us the lowest overlap while covering the major topics.

3.3. Topic Analysis

The six topics clustered through Mallet shows different sentiments. The topic first actually revolved around the negative sentiment of loss, risk, liability and loan. While the second topic discussed the employee's benefits, plan, execution and compensation. Topic three had mixed sentiments of asset and cost, risk and performance. The fourth topic was a bit related to topic 1 i.e. negative sentiments but discussed interest rates. Topic fifth seems about French companies and discussed service, strategy, employees and performance. For the topic sixth, it was a bit hard to distinguish because it had the mixed features of topic third and fourth.

3.3.1. Regression 1 (Logistic): Can the impact of Brexit referendum on British companies be identified through topic-probabilities features?

In this analysis, we have tried to fit a logistic model where the dependent variable is the interaction of two variables, i.e. British/Non-British & Pre/Post-referendum. Therefore, the dependent variable will be '1' if and only if it is a British company and has been observed post-referendum. The independent variables are the topic probabilities of each document.

Result: The p values for all variables are not significant at 5%, therefore we could not suggest that the topic patterns have been changed between British ad Non-British firms during pre and post-referendum.

However, we should keep in mind that these values are dependent upon only 33% of the sample and taking the whole data may change the opinion.

Generalized Linear Model Regression Results

Dep. Variable:	british_and_brexit	No. Observations:	276
Model:	GLM	Df Residuals:	269
Model Family:	Binomial	Df Model:	6
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-86.079
Date:	Thu, 16 Jul 2020	Deviance:	172.16
Time:	09:09:55	Pearson chi2:	159.
No. Iterations:	9		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-42.5107	27.236	-1.561	0.119	-95.893	10.872
feat_0	40.6165	27.492	1.477	0.140	-13.266	94.499
feat_1	38.2975	27.425	1.396	0.163	-15.454	92.049
feat_2	43.2801	27.627	1.567	0.117	-10.868	97.428
feat_3	39.8144	27.469	1.449	0.147	-14.024	93.653
feat_4	39.3130	27.433	1.433	0.152	-14.455	93.081
feat_5	36.0829	26.876	1.343	0.179	-16.593	88.759

Regression 1: Logistic regression where the dependent variable is if the company is British and it is post-Brexit against topic modelling probabilities.

3.3.2. Regression 2: Can topic probability features demonstrates an impact on share prices?

The idea of which is taken from the journal by Peng, Y. and Jiang, H., (2015) Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In this analysis, we will try to fit a simple linear regression where the y variable will be share price and x variables will be documented to topic probabilities.

The share price data has been taken from the long_term_financial dataframe. One point to note is, the union of topic probabilities and financial dataframe is only 75 observations and it is due to inconsistency in the company names (E.g. Adidas and Adidas ag; Greggs & Greggs Plc etc.). Going forward we can try to make these names consistent to make our analysis more insightful.

Results: Considering Peng, Y. and Jiang, H. (2015) we had assumed that topic probabilities may affect the movement of the share price. But in the above linear regression, we can observe none of the topic probability features is statistically significant to prove our hypothesis.

Additionally, the co-efficient of those features are large and in negative therefore this regression might not be the optimum way of analyzing this data.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.285
Model:	OLS	Adj. R-squared:	0.221
Method:	Least Squares	F-statistic:	4.508
Date:	Thu, 16 Jul 2020	Prob (F-statistic):	0.000669
Time:	09:30:03	Log-Likelihood:	-578.04
No. Observations:	75	AIC:	1170.
Df Residuals:	68	BIC:	1186.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7636.3008	1.19e+04	0.642	0.523	-1.61e+04	3.14e+04
feat_0	-7136.0640	1.2e+04	-0.595	0.554	-3.11e+04	1.68e+04
feat_1	-7711.3985	1.21e+04	-0.640	0.524	-3.18e+04	1.63e+04
feat_2	-6937.7990	1.21e+04	-0.573	0.568	-3.11e+04	1.72e+04
feat_3	-7622.9228	1.13e+04	-0.674	0.503	-3.02e+04	1.49e+04
feat_4	-7555.9728	1.2e+04	-0.628	0.532	-3.16e+04	1.65e+04
feat_5	-7473.5269	1.2e+04	-0.623	0.535	-3.14e+04	1.65e+04

Omnibus:	47.562	Durbin-Watson:	0.809
Prob(Omnibus):	0.000	Jarque-Bera (JB):	156.596
Skew:	2.038	Prob(JB):	9.90e-35
Kurtosis:	8.788	Cond. No.	554.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Regression 2: Logistic regression where the dependent variable is if the company is British and it is post-Brexit against topic modelling probabilities.

3.3.3. Regression 3: Can topic features improve the linear model of share price prediction along with other financial variables?

For this analysis, we calculated three regression. We have first calculated the impact of variables *(British/Non-British, Pre/Post Referendum, British/Non-British X Pre/Post Referendum, age, operating and sector)* on the share price for all 10860 observation. We calculated the same on our new union data frame of 75 variables to see if the result on larger data still holds on the smaller set.

We then added topic features to this model to see if these new variables improve the share price, prediction model?

Results: The analysis of total financial data shows that the interaction of post-referendum & British companies is statistically significant to the share price. But prior the analysis, we have assumed that this relation will have a negative co-efficient, but to our surprise the relation is positive. the co-relation (R-squared) for this analysis is 0.3 which may not be a good fit.

With the limited data of 75 observation, the results observed were not significant. In fact, after adding the variables from topic probabilities, it reduced the R square indicating that topic probability features may not be important to the share price prediction.

We would like to emphasize that these observations have been made from **33%** of the *ar__docs* data and we have selected **6 topics** basis our judgement (as discussed earlier, coherence score for 10 topics were higher than 6 topics but the topics were highly overlapping).

Also, the union of financial and document data only yielded **75** observations due to inconsistency in company names. Going forward, provided with better computation limits we might be able to analyze the whole data and then the insights may vary significantly.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.305
Model:	OLS	Adj. R-squared:	0.304
Method:	Least Squares	F-statistic:	475.4
Date:	Thu, 16 Jul 2020	Prob (F-statistic):	0.00
Time:	10:21:21	Log-Likelihood:	-88152.
No. Observations:	10860	AIC:	1.763e+05
Df Residuals:	10849	BIC:	1.764e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-143.2720	21.295	-6.728	0.000	-185.014	-101.530
if_british	877.6046	22.498	39.008	0.000	833.504	921.705
if_post_brexit	5.9721	21.735	0.275	0.783	-36.632	48.576
british_plus_brexit	66.9327	31.155	2.148	0.032	5.864	128.002
operating	2.8220	0.237	11.910	0.000	2.358	3.287
age	0.5237	0.110	4.764	0.000	0.308	0.739
sector_consumer_discretionary	309.5736	22.704	13.635	0.000	265.069	354.078
sector_consumer_staples	652.1774	28.414	22.952	0.000	596.480	707.875
sector_energy_and_materials	118.5540	24.784	4.784	0.000	69.973	167.135
sector_financials	-79.5690	26.041	-3.056	0.002	-130.614	-28.524
sector_utilities	173.8279	35.150	4.945	0.000	104.927	242.728

Omnibus:	7598.411	Durbin-Watson:	0.040
Prob(Omnibus):	0.000	Jarque-Bera (JB):	142403.348
Skew:	3.163	Prob(JB):	0.00
Kurtosis:	19.573	Cond. No.	625.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Regression 3: Regression of financial variables on share price (10,860 observations)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.441
Model:                            OLS   Adj. R-squared:                  0.364
Method:                 Least Squares   F-statistic:                     5.697
Date:                Thu, 16 Jul 2020   Prob (F-statistic):           8.58e-06
Time:                        09:33:01   Log-Likelihood:                -568.79
No. Observations:                  75   AIC:                             1158.
Df Residuals:                      65   BIC:                             1181.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      -61.5757    146.910     -0.419      0.676    -354.976     231.825
if_british                     646.7349    169.739      3.810      0.000     307.743     985.727
if_post_brexit                  93.5083    182.787      0.512      0.611    -271.543     458.559
british_and_brexit            -113.6372    247.960     -0.458      0.648    -608.848     381.574
operating                        3.1666      5.096      0.621      0.536      -7.010      13.344
sector_consumer_discretionary  632.6847    222.492      2.844      0.006     188.338    1077.031
sector_consumer_staples       -133.6336    189.455     -0.705      0.483    -512.001     244.734
sector_energy_and_materials    -21.4133    192.684     -0.111      0.912    -406.229     363.402
sector_financials              113.6016    223.174      0.509      0.612    -332.109     559.312
sector_utilities               168.0754    228.776      0.735      0.465    -288.821     624.972
==============================================================================
Omnibus:                       26.039   Durbin-Watson:                   0.852
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.623
Skew:                           1.101   Prob(JB):                     1.53e-14
Kurtosis:                       6.938   Cond. No.                         95.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Regression 4: Regression of financial variables on share price (75 observations)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.486
Model:                            OLS   Adj. R-squared:                  0.356
Method:                 Least Squares   F-statistic:                     3.724
Date:                Thu, 16 Jul 2020   Prob (F-statistic):           0.000141
Time:                        09:33:18   Log-Likelihood:                -565.62
No. Observations:                  75   AIC:                             1163.
Df Residuals:                      59   BIC:                             1200.
Df Model:                          15
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     981.6807   1.17e+04      0.084      0.934   -2.25e+04    2.45e+04
if_british                   1188.6403    813.426      1.461      0.149    -439.022    2816.303
if_post_brexit                147.6709    192.375      0.768      0.446    -237.271     532.613
british_and_brexit           -188.1816    258.613     -0.728      0.470    -705.664     329.301
operating                       0.7984      5.330      0.150      0.881      -9.867      11.464
sector_consumer_discretionary 1021.0767    318.400      3.207      0.002     383.959    1658.195
sector_consumer_staples         33.2056    219.274      0.151      0.880    -405.561     471.972
sector_energy_and_materials    -43.5180    274.095     -0.159      0.874    -591.982     504.946
sector_financials              313.2555    355.042      0.882      0.381    -397.183    1023.694
sector_utilities                74.2460    374.080      0.198      0.843    -674.287     822.779
feat_0                       -1788.5899    1.2e+04     -0.149      0.882   -2.58e+04    2.22e+04
feat_1                       -1042.2253   1.19e+04     -0.088      0.930   -2.48e+04    2.27e+04
feat_2                       -1821.7199    1.2e+04     -0.151      0.880   -2.59e+04    2.22e+04
feat_3                        3718.9345   1.19e+04      0.314      0.755      -2e+04    2.74e+04
feat_4                       -2085.0088   1.18e+04     -0.177      0.860   -2.56e+04    2.14e+04
feat_5                       -1536.2020   1.18e+04     -0.130      0.897   -2.52e+04    2.21e+04
==============================================================================
Omnibus:                       19.213   Durbin-Watson:                   0.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.082
Skew:                           0.670   Prob(JB):                     4.03e-13
Kurtosis:                       7.058   Cond. No.                     8.36e+03
==============================================================================
```

Regression 5: Regression of financial variables and topic features on share price (75 observations)

3.3.4. Regression 4: Estimating returns (share price change) through financial data

Linear regression is conducted to determine which variables have an impact on the performance of companies. Three new variables are included in the model: whether the performance is for a British company, whether the performance is post-Brexit and if the performance is both from a British company and post Brexit.

Results: The regression result above shows that British companies, post-Brexit and both from a British company and post-Brexit hurts the performance of a company. The variable with the highest coefficient is whether the company is British and post-Brexit, on average this led to a 7.53% worse performance in the share price of British companies. However, this variable is not statistically significant.

If the performance was post-Brexit is statistically significant which means that Brexit led to a 7.10% worse performance in the share price of all companies investigated.

OLS Regression Results

Dep. Variable:	returns	R-squared:	0.054
Model:	OLS	Adj. R-squared:	0.041
Method:	Least Squares	F-statistic:	4.074
Date:	Thu, 16 Jul 2020	Prob (F-statistic):	1.81e-05
Time:	23:37:18	Log-Likelihood:	-199.39
No. Observations:	724	AIC:	420.8
Df Residuals:	713	BIC:	471.2
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1009	0.074	1.365	0.173	-0.044	0.246
if_british	-0.0149	0.035	-0.424	0.672	-0.084	0.054
if_post_brexit	-0.0710	0.033	-2.133	0.033	-0.136	-0.006
british_plus_brexit	-0.0753	0.048	-1.575	0.116	-0.169	0.019
assets	-0.0132	0.017	-0.785	0.433	-0.046	0.020
operating	0.0010	0.000	2.778	0.006	0.000	0.002
sector_consumer_discretionary	0.0242	0.035	0.696	0.487	-0.044	0.092
sector_consumer_staples	0.0590	0.044	1.337	0.182	-0.028	0.145
sector_energy_and_materials	0.0662	0.037	1.772	0.077	-0.007	0.140
sector_financials	0.0070	0.046	0.153	0.878	-0.083	0.097
sector_utilities	-0.0365	0.055	-0.666	0.505	-0.144	0.071

Omnibus:	399.799	Durbin-Watson:	2.349
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6823.580
Skew:	2.092	Prob(JB):	0.00
Kurtosis:	17.446	Cond. No.	282.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Regression 5: Regression of financial variables on share returns (724 observations)

3.4 Analysis using features generated through Df-Idf and PCA analysis

Since the above results were not significant, we tried to do the the feature extraction through TF-IDF. We then reduced the 1M+ features to 10 Principal Component and then used these features for the Binomial and Linear regression models.

To our surprise, the results gave a **better fit and significant 'p' values** compared to the previous models.

As per our understanding, tf-idf only focuses on the words frequency without considering its position and semantic meaning. Tf-Idf does not just rely on the term frequency but it also consider the 'importance' of the word by penalizing (give low score) the words that occur in most of the document and rewarding the words (give high score) that occur in few document

3.4.1 Impact of document features on "British and Post Brexit" variable using Logistic Regression

Here we have performed logistics regression with "British and Post Brexit" variable as x and all the PCA features as y to see if these features have an impact on the dependent variable.

In the result, we can see that P Value of first principal component is **statistically significant**. As we know, first principal component contains the most information about the data, thus we can evidently observe that topic features of **British companies post refrendum has indeed changed**.

Generalized Linear Model Regression Results

Dep. Variable:	british_and_brexit	No. Observations:	276
Model:	GLM	Df Residuals:	265
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-84.531
Date:	Fri, 17 Jul 2020	Deviance:	169.06
Time:	18:00:13	Pearson chi2:	138.
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.3053	0.601	-5.502	0.000	-4.483	-2.128
pc0	-7.9825	2.171	-3.676	0.000	-12.238	-3.727
pc1	-8.3826	3.485	-2.405	0.016	-15.214	-1.551
pc2	8.9882	3.901	2.304	0.021	1.343	16.633
pc3	5.0890	2.348	2.168	0.030	0.487	9.691
pc4	-5.7616	4.597	-1.253	0.210	-14.772	3.249
pc5	-3.2087	4.321	-0.743	0.458	-11.678	5.261
pc6	3.4562	4.084	0.846	0.397	-4.549	11.462
pc7	-1.0854	2.564	-0.423	0.672	-6.111	3.940
pc8	5.9817	3.619	1.653	0.098	-1.112	13.075
pc9	1.2456	2.770	0.450	0.653	-4.183	6.675

Regression 6: Logistic regression of variables returned by PCA on “British and Post Brexit” variable

3.4.2 Analysing impact of newly generated features on share price

In terms of share prices, the model generated by Tf-idf features shows a better fit compared to the previous model. Also, it is worth to note that PC4, PC5 and PC6 are statistically significant with respect to the share price. Negative co-efficient of PC6 may indicates that it may be about loss/risk/liability etc.

In the notebook, we have performed the same analysis including other financial variables such as ‘asset’, ‘sector’, ‘operating’. But we couldn’t find the significant improvement by adding these features thus its results have not been included in the report.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.519
Model:	OLS	Adj. R-squared:	0.416
Method:	Least Squares	F-statistic:	5.055
Date:	Fri, 17 Jul 2020	Prob (F-statistic):	5.82e-06
Time:	18:01:21	Log-Likelihood:	-563.18
No. Observations:	75	AIC:	1154.
Df Residuals:	61	BIC:	1187.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	470.9699	343.107	1.373	0.175	-215.116	1157.056
pc0	-227.3038	878.103	-0.259	0.797	-1983.180	1528.572
pc1	664.4770	1564.510	0.425	0.673	-2463.955	3792.909
pc2	4933.6142	1402.797	3.517	0.001	2128.549	7738.680
pc3	460.1283	601.597	0.765	0.447	-742.839	1663.095
pc4	7025.3894	2139.454	3.284	0.002	2747.287	1.13e+04
pc5	6965.2909	1923.902	3.620	0.001	3118.212	1.08e+04
pc6	-4903.7210	1698.129	-2.888	0.005	-8299.339	-1508.103
pc7	684.7157	1980.843	0.346	0.731	-3276.224	4645.656
pc8	-1526.6700	1799.432	-0.848	0.400	-5124.857	2071.517
pc9	1212.9400	1474.034	0.823	0.414	-1734.573	4160.453
if_british	-8.5121	456.379	-0.019	0.985	-921.098	904.074
if_post_brexit	167.3894	194.734	0.860	0.393	-222.005	556.783
british_and_brexit	50.6603	249.546	0.203	0.840	-448.337	549.658

Omnibus:	29.235	Durbin-Watson:	1.030
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57.255
Skew:	1.392	Prob(JB):	3.69e-13
Kurtosis:	6.251	Cond. No.	80.3

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Regression 7: Linear regression of PCA generated features on share price

4. Conclusion

Analysing the press data before and after Brexit referendum shows that the topics are somewhat similar but the ranking in topics have changed. Energy, service, eu, uk, business are some similar topics and post-Brexit referendum topics are more inclined towards sustainability and business.

We can observe that the features extracted through Mallet could not fit neither provide significant values over all three variables, i.e. Share price, Return of share price and British + Post-Referendum. This may have happened due to less sample of 33% or since we took the topics as 6

rather than 10 as suggested by the coherence score or the Mallet LDA analysis was not a fit to this problem.

On the other hand, features generated through Tf-Idf and PCA, showed a better fit and indeed showed an impact on the share price of the companies as well as it fit to a considerable extent to understand wether the observation belongs to a British post referendum company.

5. References

Kapadia, S., 2019. *Topic Modeling in Python: Latent Dirichlet Allocation (LDA).* [Online] Available at: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

Peng, Y. and Jiang, H., 2015. Leverage financial news to predict stock price movements using word embeddings and deep neural networks. *arXiv preprint arXiv:1506.07220*.

Prabhakaran, S., 2018. *Topic Modeling with Gensim (Python).* [Online] Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#14computemodelperplexityandcoherencescore

Stewart, K., 2020. Who will suffer most from Brexit? Effects by region, sector, skill level and income group. [Blog] *London School of Economics and Political Science*,.

Wielechowski, M. and Czech, K., 2016. Brexit related uncertainty for United Kingdom economy. *Acta Scientiarum Polonorum. Oeconomia*, *15*(4).