Comparing UK newspapers in their perspective on automation and globalization: Methodological appendix *

Bobae Kang†

March 17, 2017

*This is a methodological appendix to my final project for Computational Content Analysis course in Winter Quarter 2017 at the University of Chicago. I express my sincerest gratitude to the course instructor Dr. James A. Evans for his guidance and teachings.

†The University of Chicago, MA Program in the Social Science, bobaekang@uchicago.edu.

1 Introduction

The current project explores the popular discourse in the United Kingdom in 2016 on two different subject categories particularly with respect to their impact on economy and job market. I examine and compare articles published in 2016 by the two major newspapers in the UK, *Daily Mail* and *The Guardian*, regarding the subject categories.

The year of 2016 witnessed a dramatic rise of the "right-wing populism" in the Western world, which led to the referendum vote in the UK for the country's withdrawal from the European Union (better known as Brexit) in June as well as the election of Donald J. Trump as the President of the United States of America in November. At the heart of these great political upsets, it is suggested, lies the frustration of people towards as well as their rejection of the rule of global liberal capitalism and, more specifically, the threats of decreasing standard of living and loss of quality jobs allegedly by foreigners through immigration and outsourcing (e.g., building factories in foreign countries where labor is cheaper). Meanwhile, others stress that increasing automation powered by artificial intelligence is a key threat to job stability, perhaps more significantly so than bad trade deals or immigrant workers. In this project, accordingly, I construct two subject categories based on these two varying perspectives and examine how different newspapers perceive their implications.

The structure of the rest of this report is as follows. Section 2 describes the data as well as how I have collected data. Section 3 details the methods I have chosen to examine the data for the current project as well as justification for such choices. This section also present the results of such examination. Section 4 discusses and offers interpretations for the results as well as their implications. Section 5 then provides a brief conclusion.

2 Data

My corpus consists of articles published in 2016 by two leading news organizations in UK, known to have notably different audiences: *Daily Mail* and *The Guardian*. The choice of these two media organizations is informed by a conventional wisdom that the former is representative of the non-elite conservatives while the latter is favored by the elite liberals in the UK. In addition, these two are among the top newspapers in the country by circulation. Each sub-corpus consists of articles collected programmatically using the following search terms representing the two categories: artificial intelligence, automation and robot for Category 0, and globalization, immigration and international trade for Category 1. The choice of these search terms are admittedly arbitrary but justified by the exploratory nature of the current project.

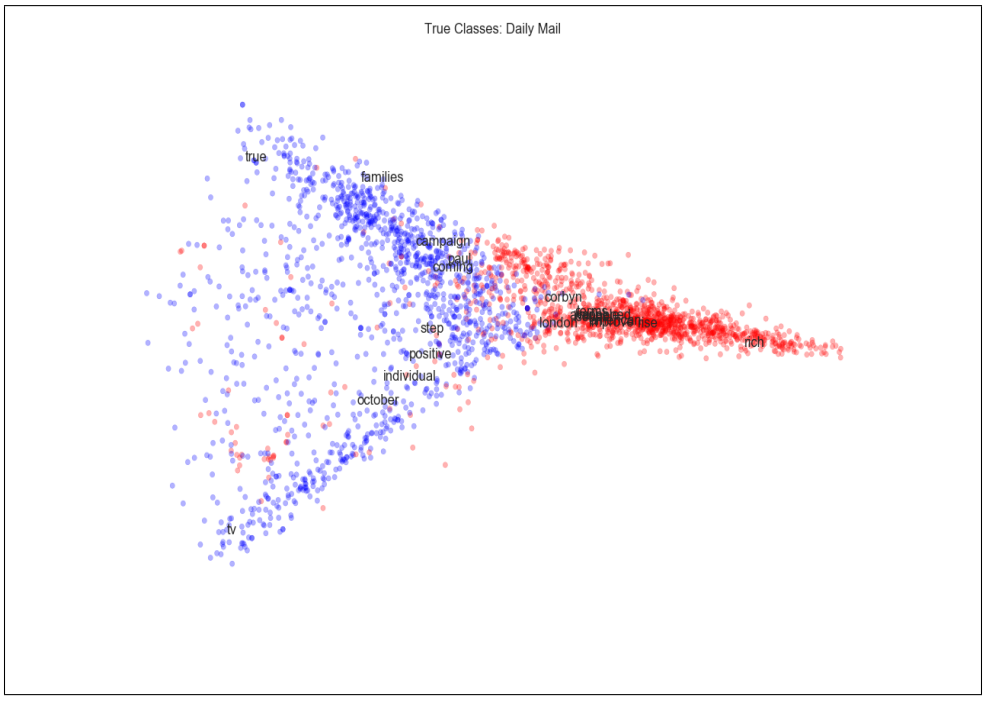
Table 1: Summary of the corpus by source and search term

Newspaper	Search term	Category	Count
Daily Mail	artificial intelligence	0	500
	automation	0	462
	globalization	1	319
	immigration	1	500
	international trade	1	500
	robot	0	500
The Guardian	artificial intelligence	0	500
	automation	0	500
	globalization	1	500
	immigration	1	499
	international trade	1	500
	robot	0	500

Each set of articles for each keyword for each newspaper was collected separately via web-scrapping (*Daily Mail*) or the application program interface (*The Guardian*). When collecting articles, I have also tokenized each article using `nltk.word_tokenize` function to facilitate the modeling process described in the following Section 3. Each resulting sub-corpus for each source consists of six sets of articles. Table 1 summarizes the number of collected articles for each search term and organization. I sought

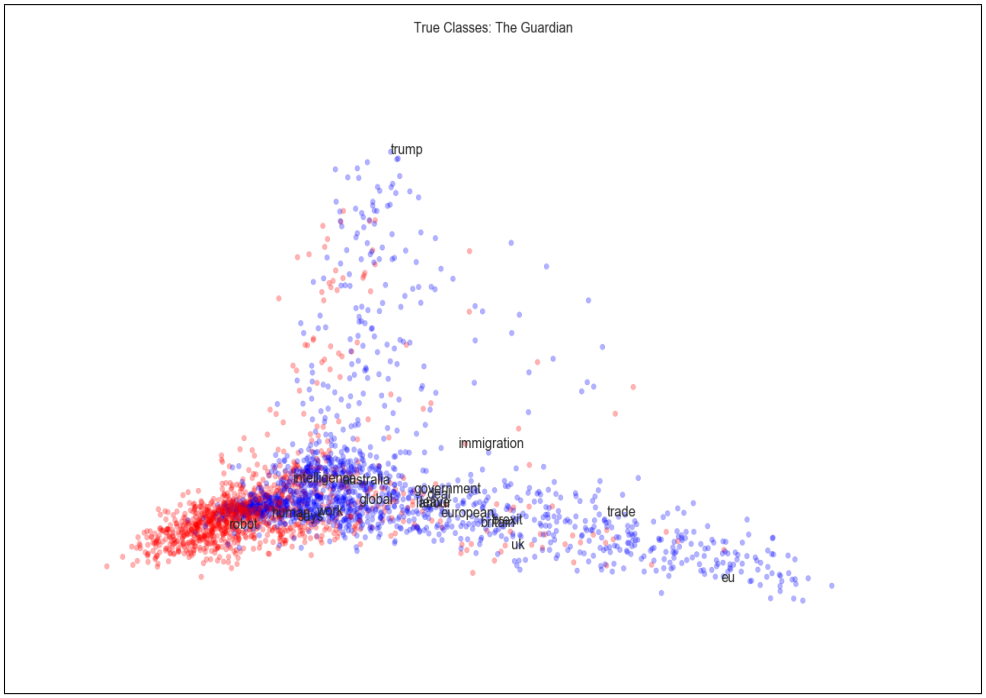
to get maximum five-hundred articles for each search term from each newspaper. Although five hundred for each search term is an arbitrary number chosen mostly for convenience, I believe that this criterion offers a sufficient size of text to conduct a meaningful analysis. For certain search terms, the number of all articles I could collect was less than five hundreds. After tagging each article with its category and removing all duplicates to ensure that each unique article has the same weight in my analysis, the resulting corpus contains total 5,112 articles–that is, 2,585 Daily Mail articles and 2,527 Guardian articles. This is equivalent to over 1,904,703 tokenized words for the Daily Mail corpus and over 2,402,643 tokenized words for the Guardian corpus.

Figure 1: Daily Mail word vectors by category



Before I begin examination of the corpus, I first inspect to ensure that Category 0 and Category 1 are sufficiently distinguishable from one another and therefore provide justification for treating them as separate categories for each set of articles from the same newspaper. For this, I first get word count vectors for each newspaper corpus whereby each word is represented by a vector of its appearance across documents,

Figure 2: Guardian word vectors by category



i.e., news articles. Then I use term frequency-inverse document frequency (tf-idf) method to weigh up words that appear frequently and weigh down words that appear commonly across documents. I also normalize the tokens by removing stop words and stemming the rest. Then I drop all the words not in the vocabulary, which allows to save memory for using PCA to reduce dimension to plot the words on a 2-D plane. After that, I project each words on a 2-D plane of which axes are the first and second principal components to see how these words are distributed. Figure 1 displays the words in the Daily Mail articles and Figure 2 illustrates the words in the Guardian articles.

In each figure, red points represent words from the Category 0 articles and blue points represent words from the Category 1 articles. These figures suggest that in both sets of articles, in general, words from the same category are clustered together while words from different categories are reasonably set apart from one another. The plots, accordingly, proves some justification for me to treat Category 0 and Category 1 as separate categories. I also observe that, for both newspapers, words from Category 0 articles are more tightly clustered than words from Category 1 articles. Such behavior

is expected because the choice of search terms for Category 0 (artificial intelligence, automation, and robot) are likely to have a narrower shared context than search terms for Category 1 (globalization, international trade, and immigration).

3 Method and Results

I use two methods to examine my data as descried in Section 2. The first method is Latent Dirichlet allocation (LDA) topic modeling. LDA is a statistical model that tries to capture the random process by which each document is constructed. Here, a topic is formally defined to be "a distribution over a fixed vocabulary" (Blei, 2012). Each document is assumed to be a mixture of multiple topics, which gives a probability of specific words to appear in that document. In this project, topic modelling offers a means to compare the Daily Mail articles and the Guardian articles in their topics for each category. Understanding these topics may offer some insights into similarity and dissimilarity between the newspapers. The second method is vector semantics. In this method, each word is represented by a vector that contains information concerning the distribution of other words around it. In other words, vector semantics approach assumes that the meaning of each word can be computed from other words around to the word (Jurafksy and Martin, 2016). In this project, this method enables the comparison between two newspapers in their use of specific words.

First, I describe the use of topic modeling in this project. Topic modeling requires a fixed dictionary, and I use the reduced tokens for the entire corpus (i.e. all articles from both newspapers) to obtain the vocabulary. An alternative is to get a vocabulary for each newspaper. However, I believe that the use of common vocabulary can facilitate the comparison across the newspapers. Then I use the shared vocabulary to create a corpus for each newspaper-category pair. Accordingly, there are four such pairs: *Daily Mail* and Category 0, *Daily Mail* and Category 1, *The Guardian* and Category 0, and *The Guardian* and Category 1. Once the corpus is ready, I train a LDA model for each newspaper-category paring. Table 2 below presents the topic modeling results for two pairs for Category 0, and 3 presents the results for two

pairs for Category 1. These tables show five topics generated using LDA for each newspaper-category pair and ten words with highest probability in each topic.

Table 2: A sample of topic modeling results for Category 0

Daily Mail Word	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Word 0	car	trump	say	robot	use
Word 1	make	app	use	use	car
Word 2	human	use	make	human	robot
Word 3	say	make	ai	research	say
Word 4	use	report	work	work	make
Word 5	group	say	report	make	china
Word 6	work	trade	help	say	test
Word 7	share	china	media	help	work
Word 8	past	work	way	car	game
Word 9	report	group	video	ai	report
The Guardian Word	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Word 0	robot	say	human	say	work
Word 1	say	robot	work	work	car
Word 2	use	use	make	use	use
Word 3	make	make	use	make	job
Word 4	work	war	say	job	say
Word 5	think	trade	report	need	need
Word 6	way	come	power	human	way
Word 7	need	nation	look	way	public
Word 8	job	work	way	robot	robot
Word 9	thing	trump	want	labour	human

One key observation is that, within each category, topics for *Daily Mail* and *The Guardian* are highly similar to each other. This, I believe, is because I use the articles from both newspapers published in the same window of time. That is, both newspapers are likely to cover the same major events although, potentially, with nuanced differences. By design, LDA topic modeling algorithm does not take into account the context in which a word appears across documents, and, consequently, topics for both newspapers covering similar sets of events naturally contain the same words. I must add that for topics for each newspaper-category pair are also similar to

Table 3: A sample of topic modeling results for Category 1

Daily Mail Word	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Word 0	group	trump	trade	trade	trump
Word 1	obama	trade	eu	trump	trade
Word 2	say	obama	global	say	eu
Word 3	state	state	britain	china	state
Word 4	european	say	deal	group	say
Word 5	eu	court	state	eu	campaign
Word 6	report	sunday	market	market	group
Word 7	past	media	group	deal	obama
Word 8	trade	group	trump	sunday	million
Word 9	sunday	deal	past	report	support
The Guardian Word	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4
Word 0	trade	trade	global	say	trade
Word 1	trump	say	uk	australia	eu
Word 2	say	eu	work	uk	uk
Word 3	eu	make	eu	support	min
Word 4	uk	use	say	work	say
Word 5	need	work	report	right	state
Word 6	market	campaign	make	eu	deal
Word 7	make	day	come	nation	right
Word 8	global	global	nation	britain	work
Word 9	nation	trump	brexit	australian	come

one another, which makes it difficult to interpret them as meaningfully distinguishable themes.

Vector semantics, on the other hand, focuses on the context of each word as it determines the meaning of a word based on the distribution of other terms around it. In preparation for performing vector semantics on my data, I first obtain tokenized sentences (lists of tokens corresponding to sentences) and then I normalize them by removing stop words. Then I use `Word2Vec` models from the `gensim` package to get vector representations of the word tokens in both newspaper corpora. Once I have trained `Word2Vec` models, I can compare the meaning of words between two newspapers.

Table 4 presents a set of Category 0 words I choose to examine as well as the set of

four most similar words for each word I examine.[1] Here, similarity between words are measured using pairwise cosine similarity between their vector representations. For terms consisting of more than one word token, I use the semantic equation, where the resulting similar terms are obtained by finding word vectors similar to both first and second word vectors rather than the specific two-word term. This is unavoidable because I use tokenized sentences to train the `Word2Vec` models. Also, notice that I examine more than just the search terms I have used to collect articles. This is to better capture how each newspaper describes subjects of interest that the search terms represent. For Category 0, I examine the word token `ai` in addition to `artificial intelligence` because the former is a popular acronym for the latter. For the search term `automation`, I also examine three closely related words: `automate`, `automated`, and `automatic`. Lastly, for the search term `robot`, I also examine `robotic` and `robots`.

Although there are few dramatic contrasts between two newspapers, I observe some minor but notable differences. First, the Guardian corpus appears to be more skeptical of artificial intelligence and automation. Terms such as `surveillance` (0.763) for the vector representation of `artificial intelligence` as well as `unreliable` (0.994) and `inequality` (0.930) for `automatic` in the Guardian corpus do not have their counterparts in the Daily Mail corpus.[2] On the other hand, the Daily Mail articles appear to discuss the automation often in the context of self-driving cars and trucks while the Guardian articles do not have the same focus. In fact, based on the table, the Guardian articles seem to treat the subject more generally. Although terms such as `workplace` (0.966) and `investing` (0.964) for `automated` or `businesses` (0.939) and `consumers` (0.938) for `automation` suggest that *The Guardian* also frequently discusses the subject in the business context, these are not as specific as `driving` and `truck` in the Daily Mail corpus.

[1] Table 4 includes only four most similar terms because of the limited space. With proper input, the `most_similar` method for a `Word2Vec` model can give many more similar words and cosine similarity score for the given term.

[2] Although not included in the table, the term `inequality` has a very high cosine similarity score of 0.9304. The cosine similarity score of the fourth most similar word, `skilled` is 0.940.

Table 4: Words most similar to the terms associated with Category 0

Daily Mail Terms	Similar words
ai	tools, advances, technological, goal
artificial intelligence	ai, researchers, goal, learning
automate	decrease, cope, trucks, occupations
automated	driving, drive, applications, improve
automatic	cope, trucks, code, lane
automation	projects, developing, businesses, productivity
robot	used, pepper, cameras, robotic
robotic	uses, sensors, using, networks
robots	drones, brain, capable, machines
The Guardian Terms	Similar words
ai	tools, understanding, learning, knowledge
artificial intelligence	agencies, robotics, surveillance, intelligence
automate	sophisticated, reflect, behaviours, maximize
automated	creating, improving, flow, workplace
automatic	limbs, exposure, internationally, unreliable
automation	benefit, value, benefits, skilled
robot	story, music, read, stories
robotic	apps, teaching, sensors, abilities
robots	machines, humans, learn, tasks

Likewise, Table 5 presents a set of Category 1 words I choose to examine as well as the set of four most similar words for each word I examine. For the search term `globalization`, I also examine `global`. Furthermore, `immigrant`, `immigrants` are added for the search term `immigration`. Finally, for the search term `international trade`, I also examine the vector representation of `trade`.

In this table, I observe a rather salient contrast between *Daily Mail* and *The Guardian* with respect to the subject of immigration. Most notably, `immigrant` in the Daily Mail corpus appear to be similar to terms such as `crimes` (0.957), `criminals` (0.944), and even `terrorists` (0.927). On the contrary, the same term in *The Guardian* is similar to terms such as `minorities` (0.978) and `abused` (0.976). In addition, *Daily Mail* appears to have significantly more negative attitude toward

Table 5: Words most similar to the terms associated with Category 1

Daily Mail Terms	Similar words
global	growth, investment, industry, markets
globalisation	bureaucracy, uncertainties, shaping, disruption
immigrant	crimes, criminals, protected, teachers
immigrants	deportation, illegally, undocumented, citizenship
immigration	tpp, congress, proposed, tough
international trade	regional, disarmament, pacific, economic
trade	agreement, canada, bilateral, disarmament
The Guardian Terms	Similar words
global	crisis, growth, financial, economic
globalisation	cause, shift, towards, positive
immigrant	phones, autistic, minorities, ethnic
immigrants	undocumented, millions, living, americans
immigration	macron, government, foreign, prime
international trade	nuclear, partnership, organisation, monetary
trade	movement, nuclear, economic, leaders

globalization than *The Guardian*. The term `globalisation` in the Daily Mail corpus has high cosine similarity scores with `uncertainties` (0.966), `disruption` (0.965), `anxiety` (0.957), and `corrosive` (0.957). This sharply contrasts to the Guardian corpus, which generally place neutral and even positive words around the same term: `positive` (0.967), `solution` (0.966), and `achieve` (0.966). However, they both share some sense of change: `uncertainties` (0.966) in the Daily Mail corpus and `shift` (0.970) in the Guardian corpus exemplify this shared sense.

I also compare pairwise cosine similarity between total 13 terms, which include six terms that represent the search terms as well as additional seven selected terms with some economic implications: `ai`, `automat`, `robot`, `globalis`, `immigr`, `trade`, `econom`, `job`, `labour`, `growth`, `decline`, `opportun`, and `risk`. The choice of additional terms are arbitrary and based on intuitions. However, this is justified by the exploratory nature of the current project. Instead of using `Word2Vec` model, I use `Doc2Vec` model. In preparation for training the model, I first obtain a list of tagged documents using the search terms and selected additional terms for each newspaper. Using the tagged

articles, I train a `Doc2Vec` model for each newspaper. Accordingly, the computed cosine similarity is related to the words' co-occurrence on the level of individual articles. That is, two terms, or tags, that have high pairwise cosine similarity are more likely to appear on the same article. This does not guarantee that the two tags are likely to used nearby each other as in `Word2Vec` model.

Figures 3 and 4 are heatmaps illustrating the cosine similarity of 13×13 word pairs. Figure 3 is generated by the tagged articles from the Daily Mail corpus, and Figure 4 by the tagged articles from the Guardian corpus. The bar on the right side of each heatmap illustrates the color scheme. Color green indicates that the cosine similarity between two tags are 0.0. The zero cosine similarity indicates that there is no clear relationship between two tags in terms of their occurrence in the tagged documents. Higher cosine similarity score yields yellow and then brown color. When two tags always co-occur, their pairwise cosine similarity score is 1.0 and the corresponding element is colored white. Indeed, the cosine similarity scores one the off-diagonal of both heatmaps are all 1.0 since these are pairs of one word and itself. Negative cosine similarity values suggest that, for the given pair of tags, one tag tends to be present in the documents where the other is absent and vice versa. In the heatmaps, these pairs are colored blue. The more blue the color is, the more dissimilar the paired tags are in this sense.

Figure 3 shows that there are many dissimilar pairs. Examples of highly dissimilar tag pairs are as follows: `ai` and `labor`, `ai` and `decline`, `automat` and `trade`, `robot` and `immigr`, `robot` and `opportun`, and `job` and `growth`. Examples of the few similar word pairs include: `globalis` and `econom`, `globalis` and `growth`, and `job` and `labour`.

Figure 4 shows a somewhat different picture. Particularly interesting pairs are those showing opposite patterns compared to the Daily Mail heatmap. For example, the pair of `ai` and `robot` shows clearly positive cosine similarity in the current heatmap, which makes an intuitive sense, while the same pair was, curiously, colored blue in the previous one. The same observation can be made for the `ai-automat` pair.

Figure 3: Daily Mail word-to-word heatmap

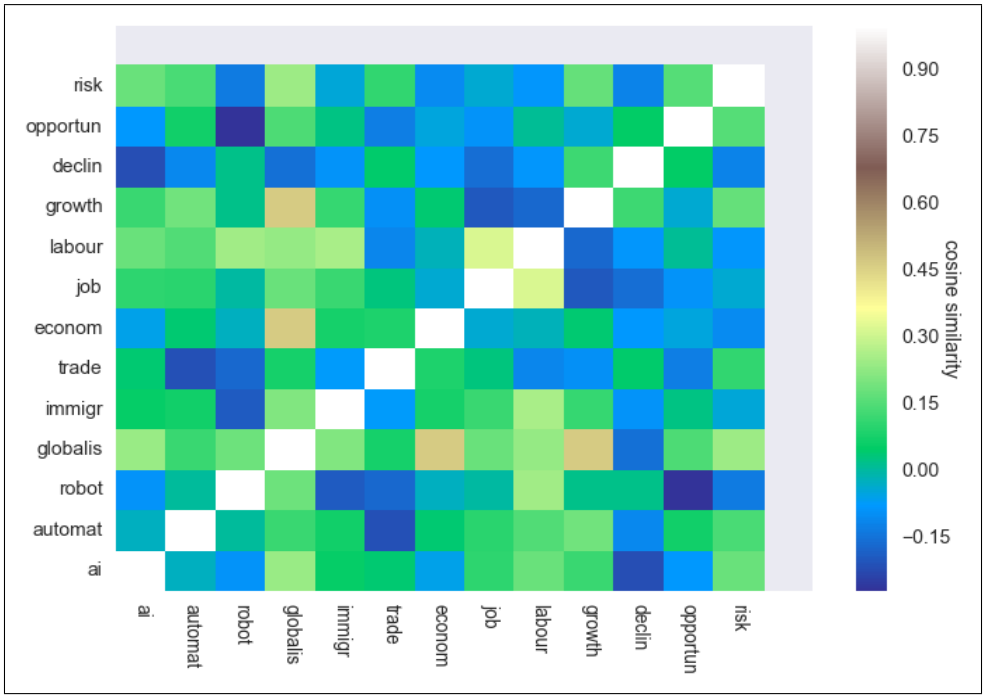
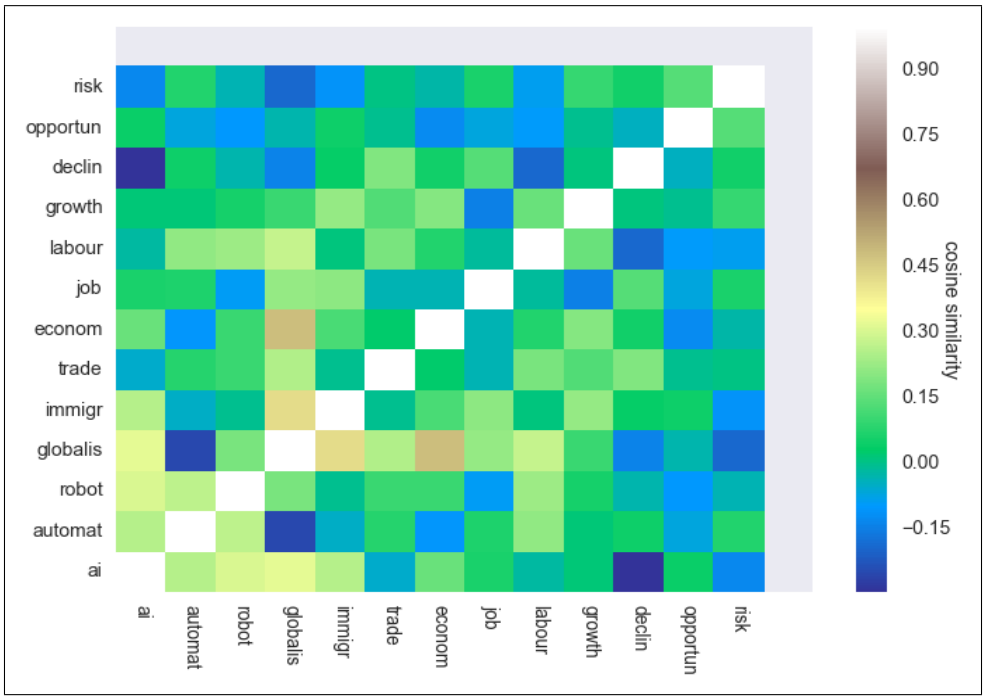


Figure 4: The Guardian word-to-word heatmap



In addition, while the `globalis-immigr` pair has a near-zero cosine similarity score in the Daily Mail heatmap, the same pair has clearly positive cosince similarity score

in the Guardian heatmap. The `globalis-growth` pair sees the opposite change, from clearly positive to near-zero similarity.

The final exploration of the data involves projecting vector representations of the key terms onto different dimensions and search for similarities and dissimilarities between the two newspapers. The projected word embeddings are for eight of the terms previously examined using `most_similar` method for the `Word2Vec` models: `ai`, `automated`, `automation`, `robot`, `globalisation`, `immigrant`, `immigration`, and `trade`. I then construct three dimensions. The first dimension represents the positive-negative spectrum with some economic implications. One side of the first dimension is represented by the following terms: `good`, `better`, `positive`, `opportunity`, and `growth`. The other side consists of the following: `bad`, `worse`, `negative`, `risk`, and `decline`. I choose the same number of terms for both sides for the balance. The second dimension represents the business-labor spectrum. On one side, there are terms such as `business`, `corporation`, and `corporate`. On the other side exist `labour`, `worker`, and `working`. Finally, the third dimension represents the economy-society spectrum. `economic` and `economy` are on one side and `social` and `society` are on the other.

Figure 5: Projection of key terms on selected dimensions for the Daily Mail corpus

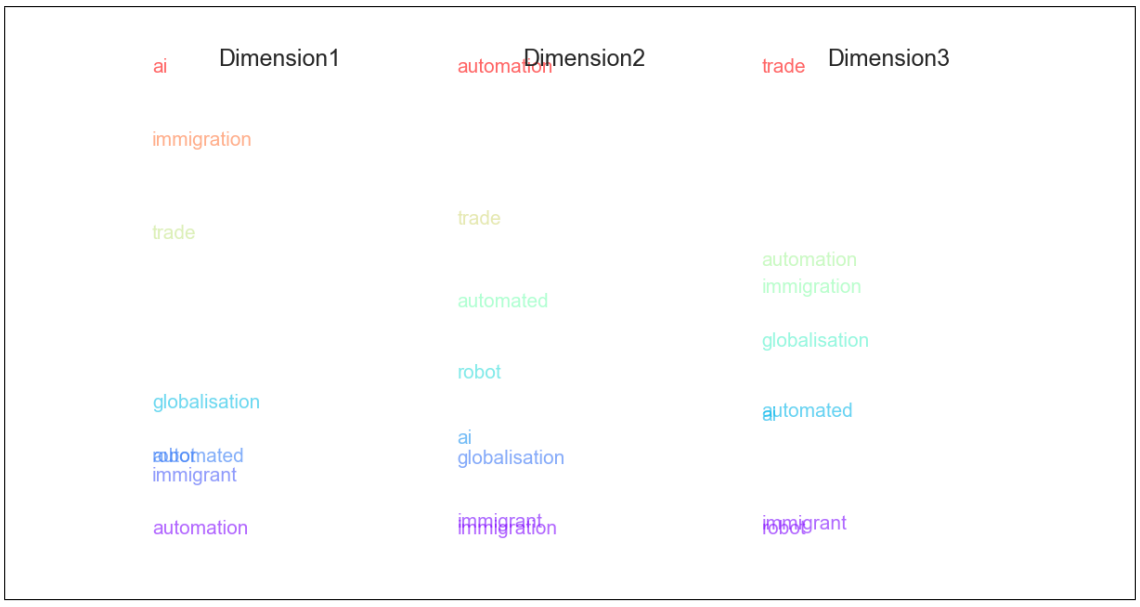


Figure 5 shows the projection of eight key terms on the three dimensions for the Daily Mail corpus. The plot for the first dimension presents a curious pattern in which `ai` is the most 'positive' and `immigration` is the second 'positive' of all. `automation` is the most 'negative,' followed by `immigrant`. This is not so intuitive to me since I would expect `ai` and `automation` to be close to each other and I would expect the same for `immigration` and `immigrant`. `robot` and `automated` also appear to have 'negative' implications. The curious separation between `ai` and `automation` word vectors continues to hold in the second dimension. In this case, `automation` seems to be highly 'corporate' term while `ai` is more 'labour' term. However, the locations of `immigrant` and `immigration` makes intuitive sense in this dimension. More specifically, first, they are close to each other and second, they are more close to 'labour' terms. `trade` is more 'corporate' term as expected. In the last, economy-society dimension, the placement of `trade` again makes intuitive sense. `ai` and `automation` are still set apart, the former being more 'social' and the latter more 'economic'. `immigration` and `immigrant` also shows a similar pattern wherein the former term appears more 'economic' and the latter more 'social.'

Figure 6: Projection of key terms on selected dimensions for the Guardian corpus

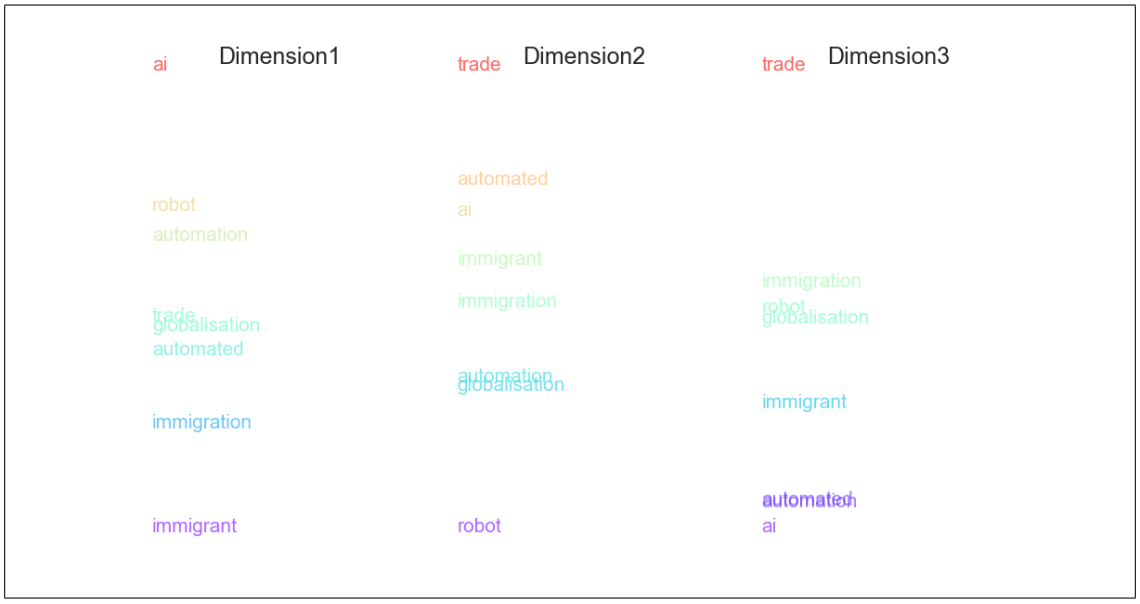


Figure 6 shows the projection of key terms on the same three dimensions–this time, for the Guardian corpus. The placement of key words on the dimension makes more intuitive sense when compared to the previous plot. `ai`, `robot`, and `automation` are all generally 'positive' terms. `automated`, however, is relatively more 'negative.' `trade` and `globalization` are in the middle. Both `immigration` and `immigrant` appear as 'negative' terms. In the second dimension, `trade` is the most 'corporate' term, followed by `automation` and `ai`. The change in the position of `immigrant` and `immigration`, now in the middle of the spectrum, is worth mentioning. That is, compared to the Daily Mail articles, the Guardian articles may be treating these two terms less as a working-class subject. Interestingly, `robot` now appears to be the most 'labour' term. Finally, when projected on the third dimension, `ai`, `automation`, and `automated` all appear as highly 'social' terms. I observe `robot`, `immigration`, and `globalisation` in the middle of the spectrum. `trade` appears to be the most 'economic' of all terms by far.

4 Discussion

In Section 3, I have tried a variety of methods to explore the similarity and differences between two newspapers, namely *Daily Mail* and *The Guardian*, concerning their perspective on two categories. Based on the characteristics of both newspapers and their audience, it might be assumed that *Daily Mail* would describe words associated with immigration, globalization, and international trade with negative economic implications while *The Guardian* would emphasize the economic challenge posed by artificial intelligence, automation, and robot.

First, the LDA topic modeling results indicate that the two newspapers, on either category, may be more similar than different from each other. That is, for both categories, many of the top 10 words in each topics for all topics in the Daily News corpus can be found in the top words in the Guardian corpus. As I have suggested above, this similarity may have resulted from the fact that both newspapers covered more or less the same set of major events in 2016. I must add that the topic modeling

approach is largely limited in terms of interpretability. For example, in 2, how can one summarize Topic 0 for the Daily Mail corpus, in which words with high probability include `car`, `make`, `human`, `say`, `use`, `group`, `work`, `share`, `past`, and `report`? Most other topics generated by LDA share the same problem with interpretability. On a related note, it is difficult to make any intuitive distinction among different topics. Tables 2 and 3 show that, for the given category, all topics appear to result from permutation of the same few words.

I also try other methods based on vector semantics to examine the same data. First, I train a `Word2Vec` model for each newspaper corpus and use `most_similar` method to compare vector representations of selected terms across newspapers. Table 4 for Category 0 shows that the same terms are used somewhat differently in different newspapers. However, the difference between two corpora is not oppositional. I have identified two noticeable differences. First is that *Daily Mail* often relates automation-related terms with driving and truck while *The Guardian* mentions the same set of words in a more general, business-related context. Second is that *The Guardian* appears somewhat skeptical of artificial intelligence and automation as suggested by some of their related terms: `surveillance` and `unreliable` for `artificial intelligence` and `inequality` for `automation`. These findings provide some, though not conclusive, evidence for the assumed characteristics of *The Guardian* that the newspaper may stress the challenges of the latest technological innovations.

Table 5 also presents that the terms in Category 1 also have different uses in different newspapers. Perhaps the most notable contrast between the two corpus involves immigration-related terms as well as `globalisation`. In the Daily Mail corpus, immigrant frequently shows up with unmistakably negative words such as `crimes` and `criminal`. The same can be said for globalisation as suggested by the terms such as `anxiety`, `fears`, and `corrosive`. These findings are generally compatible with the expectation that *Daily Mail* may hold a critical view towards Category 1 issues. It is however, unclear whether the newspapers' negative view on immigrant and globalization is based on their economic implication. In fact, crime-related terms

suggest that *Daily Mail* is likely to characterize immigrants as a social problem than an economic challenge. On the other hand, *The Guardian* appears more generous towards immigrants. Terms such as `minorities` and `abused` even suggest that *The Guardian* may perceive them, in general, as victims rather than aggressors. In addition, *The Guardian* appears to see globalization in significantly more positive terms than *Daily Mail*. On trade-related terms, the two newspapers do not show little noticeable difference.

The second vector semantics-based method uses `Doc2Vec` model to explore the co-occurrences of certain words across articles. The examined words include search-term related ones in addition to a few others with economic implications. Overall, two newspapers show different patterns. Understanding what such difference suggests is a more difficult work. One notable finding is that Category 0 terms (`ai`, `automat`, and `robot`) in the Daily Mail corpus, surprisingly, do not have high cosine similarity scores among one another. In other words, it appears that the newspaper do not treat these words as closely related terms, suggesting that *Daily Mail* may not recognize the broad trend that encompasses all three subjects. The opposite is true for the Guardian corpus as illustrated by a small block of positive cosine similarity scores on the bottom left corner in Figure 4.

It must be added that the interpretation of the cosine similarity scores between `labour` and other terms must take into consideration the fact that the same term is used as the name for a major political party in UK. For example, the high dissimilarity between `labour` and `growth` in Figures 3 and 4 may suggest that either the plight of the working class in UK or the Labour Party has little to do with growth.

Finally, I use the trained `Word2Vec` models for both newspapers and project key terms onto three dimensions I construct with other selected terms. The first dimension represents, roughly, the positive-negative spectrum. The second dimension represents the business-labour spectrum. Finally, the third dimension represents the economic-social spectrum. The plot for the first dimension in Figure 5 supports the previous finding that *Daily Mail* may not see Category 0 terms as mutually related terms. The same plot, when combined with the plot for the third dimension, also supports

the finding that *Daily Mail* describes immigrants as a social problem. The second dimension adds to this observation that the 'immigrant problem' are more relevant to the working class than to businesses. Another term that appeared with negative words in the previous analysis using `most_similar` method was `globalisation`. The second and third plot adds to our interpretation of such negative view. That is, `globalisation` is negative to workers but not particularly in economic terms.

Figure 6 illustrates how the key terms are located in the same three dimensions in the Guardian corpus. First, I observe that, as suggested by the `Doc2Vec` pairwise heatmap, *The Guardian* sees Category 0 terms to be related to one another in general. However, the second and third plots show that `robot` has implications distinct from those of the other Category 0 terms: more 'labour' and less 'social' than others, to be specific. In addition, based on the previous analysis, I can interpret that the location of `immigrant` in the first plot in Figure 6 from that in Figure 5. That is, in the Guardian articles, the 'negative'-ness of the term has more to do with the plight of immigrants themselves as opposed to their social impact.

It must be noted that the current project has at least the following two sets of non-trivial limitations. The first set of limitation involves selection of its search terms or, more generally, choice of its corpus. My choices are motivated by a series of intuitions. For example, I made an a priori assumption that the two newspapers are distinct from each other in terms of their audience. I further assumed, without much theoretical or empirical basis, that such difference would lead to diverging attitudes toward the selected subjects between the newspapers. In addition, my choice of search terms to represent the chosen subjects was largely arbitrary. The second set of limitations concerns the choice of methods. In brief, all the methods used in this project to explore the corpus lack statistical rigor. However, since the purpose of this project was largely exploratory, I believe the findings in this project are still valuable and may provide some guidance for future research.

5 Conclusion

In this project, I conduct an exploratory analysis of selected articles from two major UK newspapers, *Daily Mail* and *The Guardian*. In selecting the articles, I use six search terms divided into two categories. I begin with an intuition that the two newspapers would have different attitude toward the two categories. That is, I expected that *Daily Mail* would stress the negative economic implications of globalization, immigration, and international trade while *The Guardian* would emphasize the economic challenge posed by artificial intelligence, automation, and robot. The findings of this projects provide some evidence for such expectations although not conclusively. More specifically, *Daily News* appears to be critical of immigrants and globalization and *The Guardian* appears to be aware of potentially negative implications of artificial intelligence and automation. However, whether such critical views are concerned with the economic implications of the subject matters is not clear.

References

Blei, David M., "Probablistic Topic Models," *Communications of the ACM*, 2012, *55* (4), 77–84.

Jurafksy, David and James H. Martin, *Speech and Language Processing* 2016.