

Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation

Ye Jiang¹, Xingyi Song¹, Jackie Harrison², Shaun Quegan³, and Diana Maynard¹

¹Department of Computer Science

²Department of Journalism Studies

³School of Mathematics and Statistics

University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
{yjiang18,x.song,j.harrison,s.quegan,d.maynard}@sheffield.ac.uk

Abstract

News media typically present biased accounts of news stories, and different publications present different angles on the same event. In this research, we investigate how different publications differ in their approach to stories about climate change, by examining the sentiment and topics presented. To understand these attitudes, we find sentiment targets by combining Latent Dirichlet Allocation (LDA) with SentiWordNet, a general sentiment lexicon. Using LDA, we generate topics containing keywords which represent the sentiment targets, and then annotate the data using SentiWordNet before regrouping the articles based on topic similarity. Preliminary analysis identifies clearly different attitudes on the same issue presented in different news sources. Ongoing work is investigating how systematic these attitudes are between different publications, and how these may change over time.

1 Introduction

Editorial decisions in newspaper articles are influenced by diverse forces and ideologies. News publications do not always present unbiased accounts, but typically present frames reflecting opinions and attitudes which can heavily influence the readers' perspectives (Spence and Pidgeon, 2010). Climate change is a controversial issue in which this kind of framing is very apparent. Although bias among different news sources has been discussed previously (Fortuna et al., 2009; Evgenia and van Der Goot, 2008), sentiment analysis has not been commonly applied to newspaper articles for this purpose.

Sentiment analysis is typically implemented on short documents such as Twitter (Pak and Paroubek, 2010; Agarwal et al., 2011) and customer reviews (Pang et al., 2008; Shelke et al., 2017). However, newspaper articles have diverse context length, so their content is much more complicated than other types of sources, especially as these articles are normally cross-domain. A variety of topics might be discussed in the context of a particular climate change issue. Thus, we need to understand what the target of the opinion is in each case, i.e. which aspect of climate change the opinion is about. For instance, using the methods described in this work, we found in reports about the IPCC 2008 (Intergovernmental Panel on Climate Change) that The Independent talked about carbon dioxide emission, but The Guardian concentrated on issues of rising sea levels.

Furthermore, unlike with short documents where one can just find a single sentiment for that document, in order to understand the overall opinion in articles about climate change, we need to look at each opinion and its target separately, as multiple targets may be addressed in a single article. Additionally, even when reporting on the same event and topic, different newspaper sources will have diverse focuses. However, unlike with tweets or customer reviews, newspaper articles must give at least some semblance of objectivity, and often refrain from using explicit positive or negative vocabulary.

In this paper, we examine a set of articles about climate change in four UK broadsheets during the last decade. It is impractical to manually identify topics and analyse all the opinions about them in this large set. We therefore propose a topic modelling method to generate topics using Latent Dirichlet Allocation (LDA), and then cluster the articles into groups with similar topics. Then we perform sentiment analysis on each cluster, in or-

der to investigate the opinions, how they differ in the 4 sources, and how they may have changed over time.

2 Related Work

Research on sentiment analysis for news articles is not entirely new (Yi et al., 2003; Wilson et al., 2005). Henley et al. (2002) analysed violence-related reports in different newspapers and found that there is a significant difference between the manner of reporting the same violence-related issues. They also found newspaper sentiments reflecting the corresponding ideologies of the editors. However, they applied their content analysis on a limited number of articles, so that the vocabulary for the analysis was also small and strict. Wiebe et al. (2004) applied a classification task for detecting subjectivity and objectivity in newspaper articles. Their work depended on several newspaper datasets which were manually labelled.

Sentiment analysis has been more commonly implemented on newspaper titles. Strapparava and Mihalcea (2007) automatically classified titles with a valence indication, while Burget et al. (2011) proposed a method that classified 6 emotions in Czech newspapers based on their headlines. Burscher et al. (2016) proposed selection and baseline approaches to analyse sentiments in headlines and entire articles respectively, with clustering performed by combining K-means cluster analysis and sentiment analysis. Others have analysed the quotations in newspaper articles. Balahur et al. (2009) extracted annotated quotations from Europe Media Monitor (EMM), and classified them into positive and negative classes using several sentiment lexicons and a Support Vector Machine (SVM) classifier. Both quotations and headlines are short pieces of text, which means that the sentiment analysis is less noisy, and also that the source and target of the sentiment could easily be identified. However, those short pieces of text could not always reveal the insights of news, missing much useful information.

LDA is a generative probabilistic model which has been used to extract abstract topics from documents. It investigates the hidden semantic structure from large amounts of text without requiring manual coding, thus reducing time and cost (Blei et al., 2003). Feuerriegel et al. (2016) applied LDA to extract 40 topics from German financial newspaper articles and found that some topics have an

important effect on the stock price market. Xu and Raschid (2016) also developed two probabilistic financial community models to extract topics from financial contracts. However, the implementation of LDA on newspaper articles is less known.

3 Method

3.1 Data

The data for our experiment consists of 11,720 newspaper articles collected from 4 UK broadsheets – *The Guardian*, *The Times*, *The Telegraph* and *The Independent* – between 2007 and 2016. These articles were extracted from LexisNexis by searching all four sources for those containing the keywords “Climate Change” at least 3 times in total.

3.2 Pre-processing

In order to identify the topics that can best represent events and issues with respect to climate change, we use a part of speech tagger to annotate all the words, and only keep the nouns for the LDA model. For the sentiment analysis, all words are included.

3.3 LDA model

Typically, the number of topics in the LDA model is determined by computing the log-likelihood or perplexity. However, Bigelow (2002) has shown that predictive likelihood (or equivalently, perplexity) and human judgment are often not correlated, and even sometimes slightly anti-correlated. In this paper, we therefore treat the topics as clusters, and apply the Silhouette Coefficient instead. This method has been previously used for finding the optimal number of topics (Panichella et al., 2013; Ma et al., 2016), and is suitable for our LDA approach, since LDA is fully unsupervised. Nevertheless, in future work, it may be worth evaluating some probability measures such as log-likelihood and perplexity, and comparing the performance using these methods.

$$Sil = \frac{b - a}{\max(a, b)} \quad (1)$$

where a is the mean distance between a point and other points in the same cluster, and b is the mean distance between a point and other points in the next nearest cluster. In the silhouette analysis (Ma et al., 2016), silhouette coefficients close to +1 indicate that the samples in the cluster are far away

Sources	Topics
The Guardian	copenhagen,world,deal,agreement,summit,president,obama,china,action,treaty
The Times	copenhagen, world, cent, deal, president, summit, agreement, conference, china, year
The Telegraph	world, carbon, copenhagen, summit, deal, cent, agreement, energy, time, president
The Independent	world, carbon, copenhagen, deal, cent, agreement, year, conference, Cancun, government

Table 1: Topics in 2009

Topic_ID	Keywords
Topic 1	0.31*food 0.84*land 0.79*world ...
Topic 2	0.53*year 0.98*science 0.03*time ...
Topic 3	0.29*world 0.21*car 0.18*weather...

Table 2: Example of Topic list in The Guardian 2007

Articles	Topic_ID	Distributions
Article 1	1	0.519842
Article 2	12	0.348175
Article 3	7, 12	0.412394, 0.1492813
Article 4	2	0.249132

Table 3: Example of topic-document matrix

from the neighbouring clusters. In contrast, a negative silhouette coefficient means that the samples might have been assigned to the wrong cluster.

In our case, we repeatedly ran the analysis on the entire dataset with a different number of topics (0-30) and added the silhouette value for each number of topics to the plot in Figure 1. We can see that when the number of topics reaches 20, it has the highest silhouette coefficient score which indicates the best clustering result.

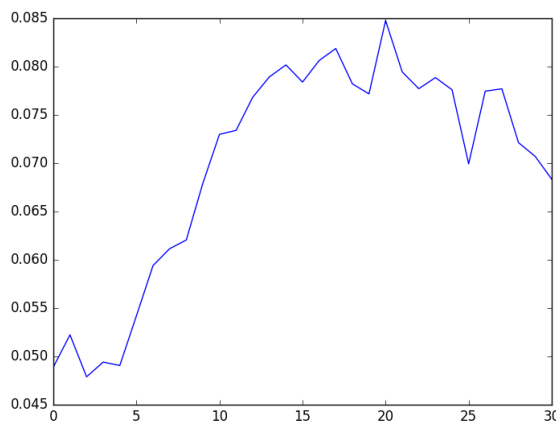


Figure 1: Silhouette analysis for LDA model

Once the number of topics has been determined at 20, the LDA assigns keywords to one of the topics of the news article, based on the probability of the keywords occurring in the topics. This assignment also gives topic representations of all the articles. We repeatedly updated the assignment for 50 iterations to generate both topic distribution in the articles and word distribution in the topics. For each topic in the LDA model, we select the top 10 keywords with their distribution to represent the corresponding topic (see Table 2).

Each article is assigned to a set of topics, and each topic generates a set of keywords based on the vocabulary of the articles. After acquiring the topics from the LDA model, we convert the bag-of-words model into a topic-document matrix, which can be seen as a lower dimensionality matrix (Table 3).

We then select the highest distribution topic among 20 topics from each news article in different news sources.

3.4 Applying SentiWordNet

To automatically annotate the articles with sentiment labels, we use SentiWordNet¹, which contains roughly 155,000 words associated with positive and negative sentiment scores. The keywords in each topic indicate the sentiment targets to be annotated with the corresponding score from SentiWordNet. For each article, the scores for all targets are combined and normalised (to a score between -1 and +1) to deal with the fact that some clusters have more articles than others. The different attitudes of each news source on the same climate change issue can then be analysed once we have a score for each article. For this, we manually check the keywords in the topic lists in each news source in each year, and group those topics containing at least two of the same keywords. Specifically, we analysed every keyword in each topic ID from 2007 to 2016 in each news source, and extract the keywords which occur in each topic. Then we also extract the topic IDs based on those keywords, and group the IDs based on the topics that contain at least two identical keywords. We assume that those news articles have similar or the same topics, as well as sentiment targets, though this also requires verification. We note that

¹<http://sentiwordnet.isti.cnr.it/>

Detected Sentences	
Positive	China itself defended its crucial role in saving the Copenhagen conference from failure. (The Guardian, 28 Dec, 2009) Don't panic. Copenhagen really wasn't such a disaster. (The Independent, 15 Dec, 2009)
Negative	The move emerged from the chaotic Copenhagen conference on climate change. (The Telegraph, 21 Dec, 2009) Copenhagen puts nuclear options at risk. (The Times, 23 Dec, 2009)

Table 4: Example sentences with sentiment polarity detected in the four news source in 2009.

the current method of grouping similar topics between news sources manually could introduce human bias. Future work will look at ways to avoid this.

4 Results and Discussion

We compared the 4 news sources by analysing the clusters we identified. For some years, there was no single topic that appeared in the clusters (probably because different newspapers attached different levels of importance to most topics). One example that stands out, however, is the reporting by all 4 broadsheets of the Copenhagen Summit in 2009 (see Table 1). The clusters all contain the keywords “copenhagen” and “agreement”, which refer to the Copenhagen Summit explicitly. This feature identified the main topics that also can be seen as the sentiment targets. We utilised this feature to compare the different attitudes toward the same issue (Copenhagen Summit) between four news sources. However, the keywords are mostly different between the sources in other years. For instance, some topics in *The Guardian* and *The Times* have large numbers of keywords such as “gas” and “energy” in 2012, but topics in the *The Telegraph* in that year are associated with the keyword “wind”, while *The Independent* has keywords like “government” and “investment”.

In Figure 2, we show how sentiment differs between the reports about the Copenhagen Summit in 2009 in the 4 newspapers. Table 4 gives also some examples of positive and negative sentences found. A manual check of a random selection of the relevant articles confirms the general tendency. Most of the articles used some negative words, such as “failures”, “collapse”, “drastic”. However, Figure 2 indicates that the overall sentiment is relatively impartial to positive (the average sentiment score across all sources is +0.15). *The Guardian* is the most positive, while *The Times* is the most negative. We suspect that some of the keywords may be a bit misleading (e.g agreement is typically positive), which might influence the sentiment analy-

sis.

However, there are some clear indications that match the automatic analysis results. While *The Guardian* does have some quite negative reports about the summit, mentioning things like “catastrophic warming”, it also tries to focus on the hope aspect (“The talks live. There is climate hope. A bit. Just.”). *The Independent* tends also towards the positive, talking about leaders achieving “greater and warmer agreement”. The Telegraph, on the other hand, plays more on the fear and alarmist aspect, talking about “drastic action” and “imminent dangerous climate change”, although also about positive steps towards the future. The Times, on the other hand, emphasises the role of honesty; although its overall tone is not overwhelmingly negative, it does mention repeatedly the fear and alarmist aspect of climate change and some of the negative points about the summit (for example that Obama will not be there).

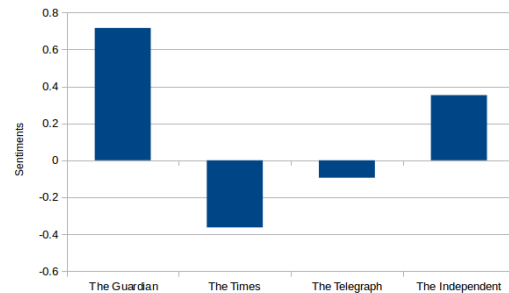


Figure 2: Attitudes of four news sources to the Copenhagen Summit in 2009

In future work, we plan a number of improvements. SentiWordNet is not ideal because it does not cover all the terminology in the specific domain of climate change, nor does it deal with context (see (Maynard and Bontcheva, 2016) for a discussion on these points). We will therefore develop a semi-supervised learning approach, based on a small corpus of manually annotated news articles that we will create, combining lexicon-based and corpus-based methods with co-training,

in order to take the best of each. The lexicon-based method will combine LDA with word-embeddings to build a domain-specific lexicon, while the corpus-based method will use a stacked denoising auto-encoder to extract features from news articles. The preliminary results demonstrate the comparison of attitudes between different publications in a single year. However, the attitude towards such climate change topic may change over time. Ongoing work is investigating how the attitudes may change over time between different publications.

5 Conclusion

In this paper, we have described a methodology and a first experiment aimed at understanding the attitudes expressed by different newspapers when reporting about climate change. Traditionally, these kind of analyses have only been carried out manually, and are therefore limited to small case studies. Our aim, however, is to apply such techniques on a large scale, looking at thousands of documents and studying the differences over time, geographic area and newspaper type. While this is only one example about different attitudes to an event, it nevertheless shows a nice case study about how we might use the approach to analyse the different attitudes expressed in the news about the same topic.

Due to the difficulty of annotating news articles manually, and the fact that existing labelled data is rare, an unsupervised approach is more suitable in this case. In contrast to most of the existing sentiment classification approaches, our method is fully unsupervised, which provides more flexibility than other supervised approaches. The preliminary results demonstrate that our method is able to extract similar topics from different publications and to explicitly compare the attitudes expressed by different publications while reporting similar topics.

The methodology is domain-independent and could also be applied to different languages given appropriate lexical resources. Besides the co-training approach mentioned above, there are a number of other ways to extend this work: in particular, we aim to extend the sentiment analysis to consider not just positive and negative attitudes, but also the emotions expressed, and to analyse the effect this might have on readers. The current method also ignored word ordering, so that issues

like negation are not considered. We therefore will extend our method to include higher order information in our future experiments.

References

- Apoorv Agarwal, Boyi Xie, Iliia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, pages 30–38.
- Alexandra Balahur, Ralf Steinberger, Erik Van Der Goot, Bruno Pouliquen, and Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*. IEEE, volume 3, pages 523–526.
- Cindi Bigelow. 2002. Reading the tea leaves. *New England Journal of Entrepreneurship* 5(1):1.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.
- Radim Burget, Jan Karasek, and Zdeněk Smekal. 2011. Recognition of emotions in czech newspaper headlines. *Radioengineering* 20(1):39–47.
- Bjorn Burscher, Rens Vliegthart, and Claes H de Vreese. 2016. Frames beyond words: applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review* 34(5):530–545.
- Belyaeva Evgenia and Erik van Der Goot. 2008. News bias of online headlines across languages. *The study of conflict between Russia and Georgia* 73:74.
- Stefan Feuerriegel, Antal Ratku, and Dirk Neumann. 2016. Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE, pages 1072–1081.
- Blaz Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detection of bias in media outlets with statistical learning methods. *Text Mining* page 27.
- Nancy M Henley, Michelle D Miller, Jo Anne Beazley, Diane N Nguyen, Dana Kaminsky, and Robert Sanders. 2002. Frequency and specificity of referents to violence in news reports of anti-gay attacks. *Discourse & Society* 13(1):75–104.
- Shutian Ma, Chengzhi Zhang, and Daqing He. 2016. Document representation methods for clustering bilingual documents. *Proceedings of the Association for Information Science and Technology* 53(1):1–10.

- Diana Maynard and K. Bontcheva. 2016. Challenges of Evaluating Sentiment Analysis Tools on Social Media. In *Proceedings of LREC 2016*. Portoroz, Slovenia.
- Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*. volume 10.
- Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.
- Annibale Panichella, Bogdan Dit, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. 2013. How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, pages 522–531.
- Nilesh Shelke, Shriniwas Deshpande, and Vilas Thakare. 2017. Domain independent approach for aspect oriented sentiment analysis for product reviews. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*. Springer, pages 651–659.
- Alexa Spence and Nick Pidgeon. 2010. Framing and communicating climate change: The effects of distance and outcome frame manipulations. *Global Environmental Change* 20(4):656–667.
- Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pages 70–74.
- Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30(3):277–308.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.
- Zheng Xu and Louiqa Raschid. 2016. Probabilistic financial community models with latent dirichlet allocation for financial supply chains. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*. ACM, page 8.
- Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, pages 427–434.