Topic modelling and sentiment analysis of of 20 years of climate change media coverage

Instructor

Prof. Dr. M. Vojnovic

Teaching Assistant

J. Yi

Candidate number

47602

December 6, 2020

Abstract

Climate change is a pressing issue. How climate change is reported in the media massively shapes public understanding of the issue. It is, therefore, crucial to understand the way in which the press reports climate change. This paper analyses over 20 thousand articles on climate change from 2000-2020 by the U.K. newspaper the Guardian, and attempts to understand its reporting through topic modelling and sentiment analysis. The focus of of these articles has shifted more towards global warming, while the tone of the articles has largely remained negative and unchanged over the course of these 20 years.

1 Introduction

Discourse about global warming and the potential catastrophic effects of climate change are part of the zeitgeist. More and more major international climate change summits appear around the world and many countries are making commitments to net-zero emissions in 2050. The sentiment among the population about global warming and climate change has shifted over the past decades, as more people have accepted the established scientific opinion on this topic and are fearful of what is to come. How this sentiment has changed is hard to capture. However, certain proxies for measuring such sentiment are available.

This paper analyses all 21,001 news articles published in the Guardian from the year 2000 until the year 2020 on the topic of climate change to analyse the frequency of articles published on climate change, visualise and understand topics found in articles over those 20 years and consider how these topics have changed, and lastly analyse the sentiment of these articles, and see if the tone of articles pertaining to climate change has indeed changed over time. The emphasis of this paper is on demonstrating knowledge of fundamental principles and methods of distributed computing for big data, not on qualitative analysis. Latent Dirichlet Allocation (LDA) is used to infer the latent topics of documents, and two different approaches to LDA are compared, one using the Bag of Words approach and the other using word embeddings. Furthermore, two different approaches to unsupervised sentiment analysis are carried out in Spark, one using the vivekn sentiment detector [6] and the other using a pre-trained dictionary sentiment detector. The fundamental structure aiding these analyses is the Machine Learning Pipeline from Spark, allowing multiple operations on dataframes to be chained and carried out in a stepwise fashion.

The results on topic modelling indicate that over the period of twenty years the focus has shifted more toward global warming and emissions, and that science has become more prominent in the climate change coverage. Furthermore, fossil fuels and green energy are increasingly mentioned. In the final period, sea levels and temperature records are mentioned prominently for the first time, marking a new phase in the discourse on climate change. Sentiment analysis indicates that the tone

of articles on climate change has always been negative, and that this tone has not changed significantly over the years. Lastly, using word embeddings instead of a bag of words approach for LDA led to a threefold computational speed-up.

In the next section, the theoretical concepts, that of pipelines, Spark Natural Language Processing (NLP), LDA and climate change coverage, are explained. In the section of method and implementation, the dataset is described and the Jupyter Notebooks are succinctly described. It is recommended to read the Jupyter Notebooks in their respective order after reading this section. In the section thereafter the results are discussed, and in the final section the main findings are highlighted, and possible future research paths are discussed.

2 Theoretical Concepts

In this section the theoretical concepts underlying the analysis in this paper are presented. First, pipelines are described. Then the tools that were implemented from the Spark NLP library [3] are described, namely that of BERT embeddings and sentiment detectors. Afterwards, LDA is explained briefly and lastly climate change coverage in the British newspapers is briefly discussed. As the emphasis of this paper is not on qualitative analysis, this subsection is not separately written as a literature review.

2.1 Pipelines

Pipelines provide a uniform set of high-level Application Programming Interfaces (API) built on top of Spark DataFrames so that multiple operations on these dataframes can be carried out in chain-wise fashion. These pipelines are based on pipelines as first implemented in Python's sci-kit learn library.

The pipeline API adopts the DataFrame from Spark SQL in order to support a variety of data types. The operations that can be carried out on the dataframe are either of a transformer or estimator type. A transformer is an abstraction that includes feature transformers and learned models. Technically, this transformer converts one DataFrame into another, generally by appending one or more columns. This appended column is usually obtained after performing some operations on one or more existing columns. An estimator implements a method fit(), which accepts a DataFrame and produces a model, which is in turn a transformer. For instance, a learning algorithm such as CountVectorizer, often used in a Bag of Words approach, is an Estimator, and calling fit() trains a CountVectorizerModel, which is a model and therefore a transformer. This transformer can subsequently convert one dataframe into another. In essence, a pipeline consists of a sequence of PipelineStages (transformers and estimators) to be run in a specific order.

2.2 Spark NLP

Spark NLP is a library from John Snow Labs and is the state-of-the-art for Natural Language Processing in Spark. It is open source, and built natively on Apache Spark

[3]. From this library I used multiple transformers and estimators, and most notable are that of word embeddings and sentiment detection.

Word Embeddings In the-state-of-art of the NLP field, embedding is the most successful way to resolve text related problem and this approach outperforms the Bag of Words approach significantly. Bag of Words has strong limitations such as that of a large feature dimension, and sparse representation, which are resolved by using word embeddings. To obtain these word embeddings, I used BERT, a deep neural network, one of the best unsupervised feature-based approaches to converting text. BERT stands for Bidirectional Encoder Representations from Transformers, and is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [4]. Using BERT Embeddings leads to much faster training of Machine Learning algorithms that are fed processed text.

Sentiment detection To detect the sentiment in different news articles, I used the ViveknSentimentDetector [6]. This detector uses the Naive Bayes classifier for sentiment analysis, and achieves an accuracy of 88.8% on the popular IMDB movie reviews dataset.

2.3 LDA

Latent Dirichlet Allocations trains a generative statistical model that allows sets of observations to be explained by latent groups that explain why part of the data to be explained are similar to each other. The model is in a sense hierarchical: as input it is fed a collection (corpus) of documents, and these documents contain words. Topics are to be inferred from these documents. Each document is in this sense associated with a distribution over (latent) topics, and each topic is associated with a multinomial distribution over words, and each word is drawn from a topic distribution. All of this is inferred by training an LDA model. The model then produces topics with words that are most associated with these topics, and these topics have to be labelled manually. To find the posterior distribution for LDA, marginalising the relevant variables is computationally extremely expensive. In fact, it is not likely that commodity hardware could do this for any reasonably sized dataset. Therefore, variational Bayes inference is used, which is such that the posterior is in this way approximated by a simpler distribution [1].

2.4 Climate change coverage

How climate change is covered in the media strongly shapes public understanding of the issue. It is therefore crucial to understand how the press reports on climate change. Apart from the complex recipe that decides what ends up being published in newspapers, many climate change stories are related to political, ecological or meteorological events. Scholars have indeed identified that peaks in climate change coverage coincide with political and meteorological events [2].

The Guardian is owned by a charitable trust and is renowned for strong coverage

of environmental issues. Of all the U.K. broadsheets, it has the most left-wing readership. Coverage of climate change in the Guardian takes noticeable leaps in the years 2005, 2007, 2009 and 2015 [5]. There were peaks in climate change coverage in 2009, with many stories about the Copenhagen UNFCCC conference (COP15) and Climategate as well as in 2015 around the time of the Paris Agreement.

We look at four different time periods in climate change reporting, the first from 2000-2004, the second from 2005-2009, the third from 2010-2014 and the final from 2015-2019. A similar subdivision is made by [7].

2000-2004 coincides with the time period used in many other studies of climate change coverage [7]. Second, the period from the adoption of the Kyoto Protocol, to the negotiation of the post-Kyoto regime in Copenhagen in 2009 is considered one period (2005-2009). After Copenhagen up to 2015 is considered one period, with the Paris COP in 2015, which represent significant peaks and troughs in coverage. Finally, 2015 until the most recent complete year is considered one period.

3 Method and implementation

This section describes the data and notebooks very briefly. All relevant descriptions and further discussion are in the Jupyter Notebooks.

3.1 Data

Each observation in the Guardian dataset consists of 3 variables, namely that of

- webPublicationDate: Indicates the datetime the article was published in the Guardian.
- bodyText: Contains the full text of the article.
- wordcount: Indicates the wordcount of an article

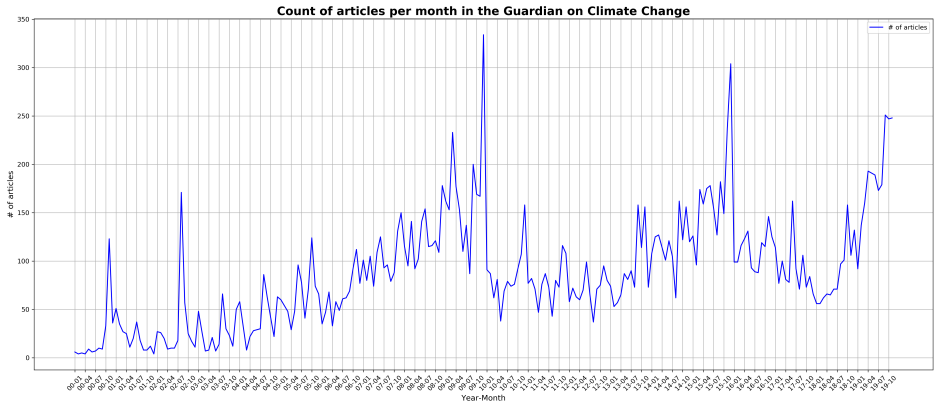
The bodyText is the raw input for all analyses, and the publication date is mostly relevant for plotting purposes. The data was gathered using the Guardian API, and the code for this is to be found in `0.Gathering data.ipynb`.

3.2 Jupyter Notebooks

There are two remaining Jupyter Notebooks. In `1.BERT Embeddings and Sentiment Analysis.ipynb` three pipelines are created, one to convert the raw text into word embeddings, and two to measure sentiment in the articles. Afterwards the plots to demonstrate sentiment over time are created. In `2.LDA in PySpark.ipynb`, a pipeline is created to process the raw text into a format for LDA. First, a tutorial-like exposition of launching PySpark on Google Cloud Platform is given. After performing LDA for all four periods, in order to infer different topics for the four different periods, efficiency gains of performing LDA on word embeddings instead of Bag of Words are analysed.

4 Results

In this section the results of the LDA and sentiment analysis are succinctly analysed, after which the efficiency gains of performing LDA on word embeddings instead of bag of words are given. The plot below demonstrates the count of articles per month covering climate change in the Guardian.



Topic modelling The output of all four LDA models contains ten topics each, each of which in turn contain a list of words with assigned weights. For each topic, I manually labelled them based on the seven most prominent words, i.e. words with the highest weights for each topic.

In the period from *2000-2004* two topics relate to global warming and emissions, with words such as 'climate', 'change', 'world', 'emission', 'global' and notably 'Kyoto'. This is understandable as the Kyoto protocol was heavily discussed on a global scale in these years. Two topics pertain to weather, with words such as 'weather', 'rain', 'flood', 'river', and 'risk'. Another topic pertains to natural disasters ('earthquake', 'rescue', 'kill'), another to developing countries ('africa', 'aid'), and another to energy and power.

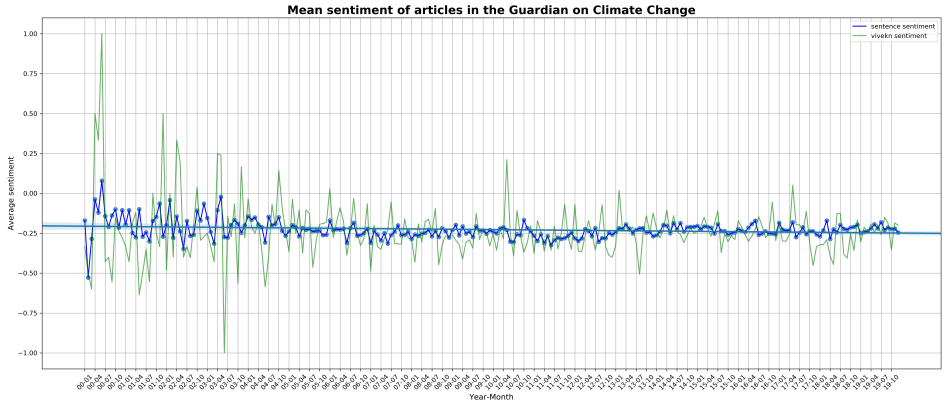
In the period from *2005-2009*, four topics relate to global warming and emissions. The word 'reduce', 'oil' and 'carbon' now feature in these topic descriptions, words that did not appear in the first period. Again, one topic pertains to the weather. A new topic is introduced here, pertaining to green energy instead of energy generally ('power', 'car', 'government', 'plan', 'green') and another new topic is introduced, pertaining to politics specifically ('need', 'political', 'take', 'issue'). The urgency of global collaboration has seemingly become more accepted.

In *2010-2014*, there are three topics relating to global warming and emissions. The word 'china' is now mentioned for the first time in a topic description, and the words 'science' and 'scientist' and 'change' and 'public' appear frequently, showing that

the scientific opinion of global warming is becoming more established. The words 'change' and 'public' likely relate to the Guardian's interest in public opinion being changed on climate change, as in this period denying climate change was still a rather accepted position. Two topics pertain to energy, with the words 'fossil' and 'fuel' now appearing, and the last topic pertains to politics.

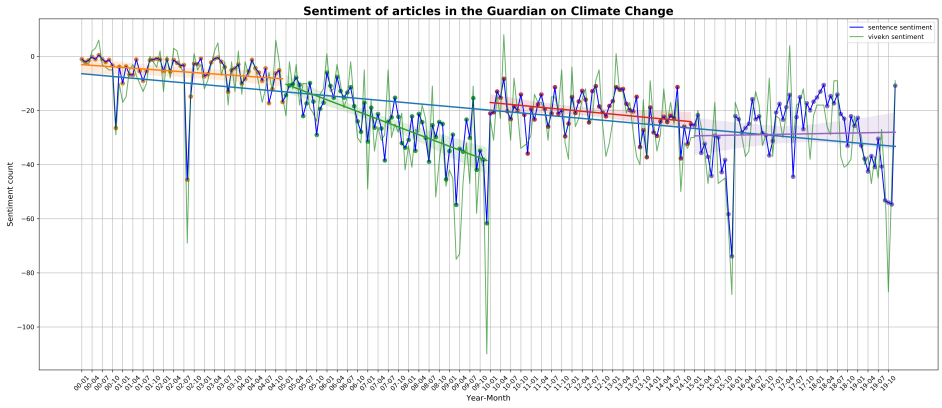
In *2015-2019* four topics relate to global warming and emission, with the word Paris now appearing (from the Paris COP), and the word Trump also appearing for the first time. In this period, many temperature records were broken, and now one topic pertaining to weather includes the words 'fire', 'temperature', 'record' and 'city'. Another topic again pertains to scientific opinion, and a new topic is introduced: that of sea levels and damage to nature, with the words 'ice', 'warm', 'sea' and 'reef' appearing now.

Sentiment Analysis The plot below demonstrates the results from the sentiment analysis. Three lines are drawn in the plot. The plot was created by taking the mean of the sentiment scores for all articles published in one month. The green line is based on the sentiment detector that assigns +1 and -1 to an article, while the dark blue line assigns a score between +1 and -1, as described in the Jupyter Notebooks. Therefore, the mean is more volatile for the green line. The mean sentiment score is around -0.25 consequently, and almost never reaches a score higher than 0 for a given month. This confirms that climate change is in itself a topic with a negative connotation. A trendline is drawn for the dark blue line, and it shows that over the course of twenty years the average sentiment of climate change articles has decreased slightly, but not significantly.



The second plot contains the count of all sentiment scores in one month. As most articles have a negative sentiment score, this plot is heavily influenced by the number of articles published on climate change in a given month. Five trendlines are drawn, each for one period and one for the entire period from 2000-2019. Here a serious

connection can be made between a decrease in positive sentiment and year. Strangely, the sentiment seems to be increasing in the final period. This can, however, be explained by the fact that the number of articles on climate change from the year 2015-2019 is slightly decreasing for the first two and a half years.



Efficiency gains Using BERT embeddings instead of the CountVectorizer Bag of Words leads to a decrease in computation time. It is standard practice in state-of-the-art applications to use embeddings. However, BERT embeddings come with one strong limitation. The vectors cannot be translated back into words (at least not with my approach) and therefore inferring topics is not possible anymore. However, other uses of LDA are not excluded with this approach, such as organising/clustering documents, or featurization and dimensionality reduction.

5 Discussion

In this section the main findings are highlighted, its limitations are discussed and paths for future research are laid out.

Twenty years of coverage of climate change in the U.K. newspaper the Guardian is analysed in this paper through topic modelling and sentiment analysis. The 20 year coverage is subdivided into four five-year periods. Emphasis is on the fundamental principles and methods of distributed computing for big data. The results on topic modelling indicate that over the years the focus has shifted more to global warming and greenhouse gas emissions, and that scientific opinion has become more prominent in the climate change coverage. Furthermore, green energy becomes more prominent over the years, and fossil fuels are increasingly mentioned. In the final period, sea levels and temperature records are mentioned prominently for the first time, marking a new phase in the climate change debate. The sentiment analysis indicates that the sentiment (or tone) of articles on climate change has always been negative, and that this tone has not changed significantly over the

years, although there appears to be a slight decrease in sentiment. Lastly, using word embeddings instead of a bag of words approach for LDA led to a threefold computational speed-up.

This paper has three important limitations. The first is that of qualitative analysis, the second is in its sentiment analysis and the third in its topic modelling analysis. A proper qualitative analysis would require a thorough literature review on newspaper coverage of climate change, but this is beyond the scope of this paper. The qualitative results could be justified better after such a literature review. Furthermore, the sentiment analysis was performed using pretrained models. The output of the pretrained models is sometimes limited, as demonstrated in Jupyter Notebook 1. The general consensus is that a sentiment analyser should be trained for a specific task. This is possible but requires manual labelling of sentiment of a large number of articles, a task beyond the scope of this paper. Lastly, using LDA for topic modelling is not necessarily state-of-the-art anymore. Using recurrent neural networks or deep LDA for this task usually outperforms LDA.

These three limitations also point at possible directions for future research, with a more thorough literature review, a more sophisticated sentiment analysis after manual labelling, and different approaches to topic modelling.

References

[1] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3(Jan)*, 993-1022.

[2] Boykoff, M. T., Mansfield, M. (2008). 'Ye Olde Hot Aire'*: reporting on human contributions to climate change in the UK tabloid press. *Environmental research letters, 3(2)*, 024002.

[3] Butch, Q. *Next-Generation Machine Learning with Spark.* 1st ed., Apress, 2020.

[4] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[5] Doulton, H., Brown, K. (2009). Ten years to prevent catastrophe?: Discourses of climate change and international development in the UK press. *Global Environmental Change, 19(2)*, 191-202.

[6] Narayanan, V., Arora, I., Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 194-201). Springer, Berlin, Heidelberg.

[7] Saunders, C., Grasso, M. T., Hedges, C. (2018). Attention to climate change in British newspapers in three attention cycles (1997–2017). *Geoforum, 94*, 94-102.