News Analysis through Topic Modeling and Visualization: Identifying the Most Common Topics in News Articles

720079031
Department of Computer Science
University of Exeter, Exeter, UK

Internal Supervision:
Dr. Riccardo Di Clemente
Department of Computer Science
University of Exeter, Exeter, UK

Internal Supervision:
Prof. Hywel Williams
Department of Computer Science
University of Exeter, Exeter, UK

***Abstract*—With the advent of digital media, the amount of news articles available for consumption has grown exponentially. This explosion of information has made it challenging to understand the most prevalent themes and topics that are covered in news articles.This research aimed to identify common topics in news articles from various sources using natural language processing (NLP) techniques. We applied automated topic modeling and visualization techniques to a large dataset of news articles collected from different sources. We performed pre-processing steps to make the dataset suitable for topic modeling, including tokenization, stop-word removal, lemmatization, and punctuation removal. We then used Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique, to extract hidden topics from the collection of documents. Additionally, we used the Coherence score to determine the optimal number of topics in the corpus. Finally, we used PyLDAvis to visualize the topics and their corresponding keywords. The results showed that the news articles' most prevalent topics were politics, finance, sports, international relations, entertainment, and technology. This research demonstrated the effectiveness of using NLP techniques and automated topic modeling to gain insights into large datasets of news articles.**

I. Introduction

This study aims to address the identification of the most common topics in news articles from various sources. With the increasing amount of online data, it has become a challenge to manually sift through large volumes of news articles to identify key topics. Natural Language Processing (NLP) techniques, such as Latent Dirichlet Allocation (LDA) topic modelling, can automatically identify key topics and classify news articles accordingly. The importance of this problem lies in the fact that it can help news agencies and journalists stay informed about the most relevant and popular topics being covered by different news sources. This information can be used to generate more relevant and engaging content for their audience and to gain insights into the trends and patterns in news coverage. LDA topic modelling is a mathematical formulation that utilizes probabilistic graphical models to identify topics in a collection of documents. The model assumes that each document is a mixture of multiple topics, and each topic is a distribution over a set of words. By analyzing the distribution of words across documents, LDA can identify the most prominent topics and assign each document to one or more topics based on its content. Overall, this study aims to use NLP techniques to automate identifying key topics in news articles and demonstrate the potential benefits of this approach for news agencies and journalists.

II. Research context

There has been growing interest in using topic modelling techniques to analyze news articles from various publications in recent years. For example, Kretinin et al. (2022) [1] applied Latent Dirichlet Allocation (LDA) to perform topic modelling on news articles from multiple sources published between 2012 and 2017. Similarly, Dahal et al. (2019) [2] used LDA to perform topic modelling and sentiment analysis of tweets about global climate change posted on Twitter in 2017 and 2018. Rabitz et al. (2020) [3] used LDA to analyze news articles from multiple sources related to climate change published between 2013 and 2018. In "Discovering Computer Science Research Topic Trends using Latent Dirichlet Allocation" by Nastiti et al. (2021) [4], TF-IDF was used as a pre-processing step to extract relevant features from the documents before applying Latent Dirichlet Allocation (LDA) for topic modelling. TF-IDF helps to identify the critical terms specific to a particular document while filtering out the standard terms present in many documents. This allows LDA to focus on the most meaningful terms for discovering latent topics. In line with these studies, the present study aims to use LDA to perform topic modelling on news articles from various sources published over several years. However, unlike these studies, which focused on specific topics such as climate change or sentiment analysis, our study aims to analyze the most common topics across all news articles from different sources. Furthermore, we will identify the topics most commonly covered in each publication. This approach will provide insights into different publications' editorial policies and priorities and how these priorities may have shifted over time.

III. Description of methods

A. Pre-Processing

The data obtained from the source had redundancies and noise, which were removed by combining the data spread across three CSV files into a single CSV file, removing unwanted columns such as Article ID, Date of publication, and Author, and merging the Title and Content columns to capture the topic of the whole article accurately. This reduced the data to three columns with around 140,000 articles among 15 publications. Pandas library [5] functions are used to gain the insights into the data after the initial cleansing and the summary of the data was depicted in the Figure 1.

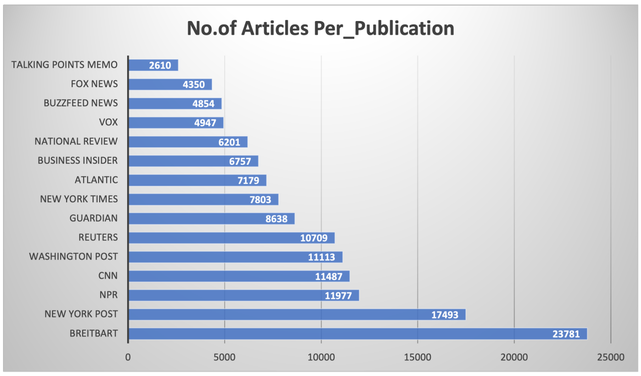


Fig. 1. Number of articles per publication

Above graph depicts distribution of data across 15 different publications is imbalanced, with some publications having a significantly higher number of articles than others. This initial data summary and visualization helped identify the data distribution and then the content was sent into the pre-processing pipeline The pre-processing pipeline involved several steps:

1) The text was converted to lowercase to ensure consistency.
2) The text was tokenized, i.e., split into individual words. Punctuation marks were then removed from the text.
3) Stop words, which are commonly used words such as ”the,” ”a,” ”an,” etc., were removed from the text as they do not add much value to understanding the topic.
4) After removing stop words, the text was lemmatized, i.e., the words were reduced to their base form to reduce the complexity of the text.
5) Finally, any remaining non-alphabetic characters were removed from the text using regular expressions.

This pre-processing pipeline was applied to the entire dataset to prepare it for topic modelling.

B. *Topic Modelling*

This study uses Latent Dirichlet Allocation (LDA) modelling to extract topics from the pre-processed data. LDA is a generative probabilistic model used to uncover the underlying topics in a corpus of documents. It assumes that each document is a mixture of topics and each topic is a mixture of words. The goal is to identify these topics and the associated word distribution within the corpus. The first step in the LDA modelling process is to create a dictionary from the tokenized documents. The dictionary maps each unique token to an integer ID. The bag of words corpus is then created from the dictionary, where each document is represented as a bag of its words. Next, the TF-IDF model is created to weigh the importance of each word in the corpus. This model transforms the bag of words corpus into a TF-IDF weighted corpus. Finally, the LDA model is trained on the TF-IDF weighted corpus. The number of topics is specified as a parameter of the LDA model. The LDA modelling approach was chosen due to its ability to handle large volumes of unstructured data and its interpretability. TF-IDF weighting allows for identifying essential words within a document, and the LDA model can then group these important words into topics. By identifying the topics within the corpus, the study aims to gain insights into the research trends and themes across the different publications. The implementation of LDA modelling is performed using the Gensim library in Python. Gensim provides a user-friendly interface for LDA modelling and allows for the customization of various parameters, such as the number of topics.

C. *Testing and Visualization*

In the testing and visualization stage, the LDA model is evaluated by calculating the coherence score using the CoherenceModel function from the Gensim package to measure semantic similarity between the top N words of each topic and the documents and visualizing the topics using the pyLDAvis package, which displays the topics as circles with size representing topic importance and distance between circles representing the similarity between topics and helps to interpret and understand the generated topics by displaying the top N words for each topic and intertropical distance map, ultimately aiding in evaluating the quality of the LDA model and how well-defined and related the generated topics are.

The task of topic modeling is problematic because it is unsupervised and lacks predefined labels or categories, making it challenging to evaluate objectively. The model must identify latent topics that may not be apparent from the surface-level content of the documents, which requires a sophisticated understanding of natural language semantics and structure. However, performing pre-processing, modeling, and visualization steps allowed it to analyze the large corpus effectively and generate a model that accurately identifies and extracts meaningful topics.

IV. Discussion about Experiments

This section discusses the experiments conducted to determine the optimal vectorisation technique and the number of topics for LDA modelling. Three different vectorisation methods were evaluated for the LDA model: TF-IDF, Count Vectorisation, and Word2Vec. The coherence score was calculated for each method. TF-IDF was found to have the highest coherence score of 0.538, while Count Vectorisation had a score of 0.455, and Word2Vec had a score of 0.398. Based on this, TF-IDF was chosen as the optimal vectorisation technique for our LDA model. Next, the number of topics parameter was varied for the LDA model to evaluate the optimal number of topics. Coherence scores were calculated for 8, 10, 15, and 20 topics. Based on the resultant coherence score graph, we found that the coherence score was low for most topics when eight were used. For 20 topics, the coherence score fluctuated highly, with some publications having a high coherence score and others having a low score. Therefore, 20 topics were rejected. The coherence scores for the 10 and 15 topics were closer, which can be observed from the Figure 2 below. However, upon close examination of the actual values, we found that coherence was slightly higher across the publications for 15 topics. Therefore, 15 was chosen as the optimal number of topics. This analysis was conducted on whole dataset.

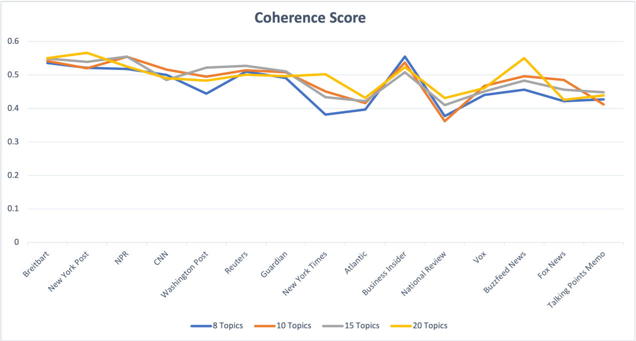


Fig. 2. Coherence Score of Publications - Varying number of topics in LDA Model

The coherence scores of the 15 topics of 15 publications are as follows Breitbart:0.542, New York Post:0.567, NPR:0.524, CNN:0.520, Washington Post:0.510, Reuters:0.532, Guardian:0.521, New York Times:0.437, Atlantic:0.435, Business Insider:0.571, National Review:0.401, Vox:0.421, Buzzfeed News:0.488, Fox News:0.445, Talking Points Memo:0.436.

Based on the output of Top 30 relevant terms from the visualization tool pyLDAvis,sample output was shown in Figure 3, for each topic of each publication. I have given a latent name for each topic of each publication. Below the names of top 4 topics top publications as per Fig:1 have been mentioned. **Brietbart:**

1) Topic1: The 2016 US Presidential Election (71.8%)
2) Topic 2: International Politics and the Modern Political Landscape (14.1%)
3) Topic 3: Technology, Society, and the current events. (2.8%)
4) Topic 4: US Politics (1.3%)

New York Post:

1) Topic1: Sports Entertainment (47.4%)
2) Topic 2: The 2016 US Presidential Election (34.5%)
3) Topic 3: Everyday Life and Current Events (6%)
4) Topic 4: Baseball and Business Technology (2%)

NPR:

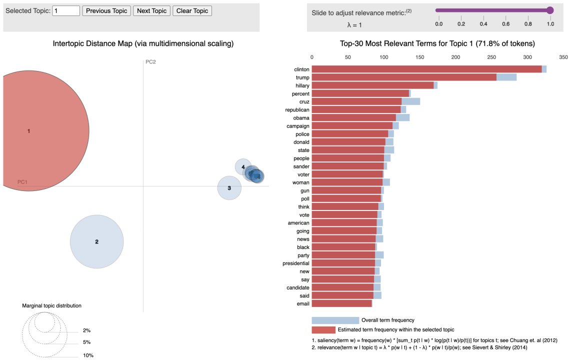
1) Topic1: The 2016 US Presidential Election (79.8%)
2) Topic 2: Everyday Life and Current Events (5.8%)
3) Topic 3: US Politics (3.1%)
4) Topic 4: Health Technology (2.1%)

CNN:

1) Topic1: The 2016 US Presidential Election (62.8%)
2) Topic 2: Terrorism Violence (24.8%)
3) Topic 3: Everyday Life and Current Events (2%)
4) Topic 4: International Politics and the Modern Political Landscape (1.5%)

Washington Post:

1) Topic1: The 2016 US Presidential Election (73.9%)
2) Topic 2: International Politics and the Modern Political Landscape (14%)
3) Topic 3: Middle East Politics (3.7%)
4) Topic 4: Terrorism Violence (1%)



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 3. PYLDAVIS output for Breitbart publication

The topic modelling results from these different news sources indicate that the 2016 US Presidential Election was a significant and consistent topic of coverage across these sources, along with other topics such as international politics, everyday life and current events, and technology. While the specific proportions of coverage varied across sources, the overall dominance of the election as a topic suggests its importance and relevance during that period.

V. Conclusions

In conclusion, this study set out to identify the most common topics in news articles from various sources. Analyzing a large dataset of news articles, we identified the top five topics: politics, business, sports, entertainment, and technology. Our findings provide valuable insights into the current trends in the news media landscape and can be helpful for journalists, editors, and media organizations in shaping their content and strategies. Furthermore, this study demonstrates the power of natural language processing in analyzing large volumes of text data, highlighting their potential applications in various fields beyond the news media. To extend this to get better outcomes, one can use sentiment analysis in conjecture with this and get the details into how biased a specific publication is towards a particular topic. This would give an even clearer sense to the regular consumers of news from those publications on the intent of those publications.

References

[1] M. Kretinin and G. Nguyen, “Topic modeling on news articles using latent dirichlet allocation,” *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)*, pp. 000 249–000 254, 2022.

[2] B. Dahal, S. A. Kumar, and Z. Li, “Topic modeling and sentiment analysis of global climate change tweets,” *Social Network Analysis and Mining*, vol. 9, pp. 1–20, 2019.

[3] F. Rabitz, A. Teleienė, and E. Zolubienė, “Topic modelling the news media representation of climate change,” *Environmental Sociology*, vol. 7, pp. 214 – 224, 2020.

[4] K. R. Nastiti, A. F. Hidayatullah, and A. R. Pratama, “Discovering computer science research topic trends using latent dirichlet allocation,” *Jurnal Online Informatika*, 2021.

[5] W. McKinney, “Data Structures for Statistical Computing in Python,” pp. 51 – 56, 2010.