Mini Project 3: Topic Model

Mettu Prathima - 2160335
(Formal Analysis, Software, Review Editing)
pmettu@cougarnet.uh.edu

Sai Siddhartha Parvathaneni – 2191129
(Validation, Writing Original Draft, Supervision)
sparvat4@cougarnet.uh.edu

Mahesh Reddy Nalabolu – 2154729
(Validation, Writing Original Draft, Supervision)
mnalabol@cougarnet.uh.edu

Abstract

The purpose of this report is to present the results of a study on how terrorist organizations are portrayed in traditional media outlets using topic modeling. The data used for the study was collected in 2017 from the Wall Street Journal and the New York Times. The report details the pre-processing and exploratory data analysis of the data, the methodology used for topic modeling, the experimental results, and the conclusions drawn from the study.

1. Introduction

Clustering The purpose of this analysis is to comprehend how traditional media outlets like the Wall Street Journal and the New York Times present terrorist organizations. We will employ topic modeling, a method that automatically extracts topics or themes from a corpus of text data without any prior knowledge of the topics, to accomplish this.

The dataset we'll be using, which includes articles from the Wall Street Journal and the New York Times, was compiled in 2017 from the Factiva Global News database. Before constructing a topic model using the Latent Dirichlet Allocation (LDA) algorithm, we will perform exploratory data analysis, create a corpus object, and preprocess the dataset. Finally, we'll talk about the findings and make some inferences from them.

2. Methodology

The first step in analyzing the data was to build a corpus. The corpus was created by loading the data as a large character and then splitting it up into individual articles. The articles were then processed to separate meta-data from the actual articles. The next step was to extract features from the articles. To extract features, the text was tokenized and stop words were removed. The tokenized text was then converted into a document-term matrix.

After the pre-processing and exploratory data analysis, the topic modeling was performed. The topic modeling was done using Latent Dirichlet Allocation (LDA) with the Python library Gensim. For topic modeling, we employed the Latent Dirichlet Allocation (LDA) algorithm. LDA is a statistical generative model that makes the assumption that each document in a corpus is a combination of a small number of topics and a small number of words. LDA employs a Bayesian methodology to infer from the corpus the topics and the word distributions that correspond to them. The LDA model was trained with varying numbers of topics, and the coherence score was used to evaluate the model's performance. The coherence score measures the degree of semantic similarity between the words within a topic. The LDA models with the highest coherence scores were chosen as the final.

3. Experimental Results

To analyze the dataset, we first divided the articles which have sub articles separately and removed the meta data from all the articles which adds noise to the analysis.

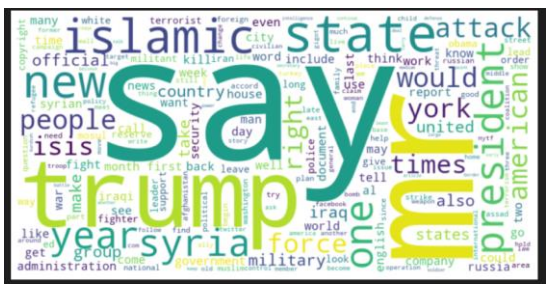
Length of the data after separating all the articles individually is 1691 articles.

After removing the Meta data from all the articles we have merged all the articles into a corpus, and removed stop words , punctuations, and done lemmatization on the data.

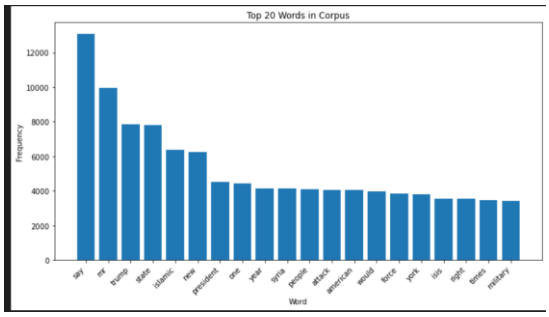
Preprocessed corpus

Word Cloud:

The below word cloud shows the words which have occurred the most in the corpus.

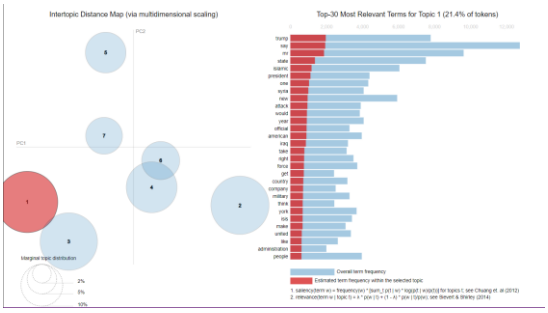


Additionally, we also plotted bar graphs for the most frequent 20 words in the corpus

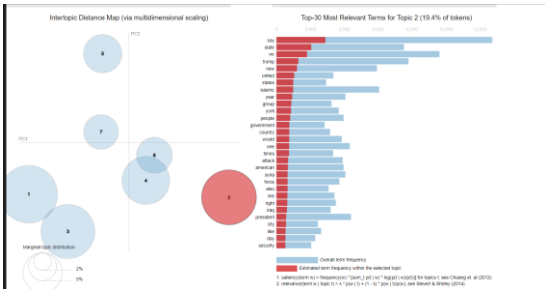


The results of the topic modeling showed that the most prominent topics in the dataset were related to politics and conflict. The top 7 topics identified by the LDA model were:

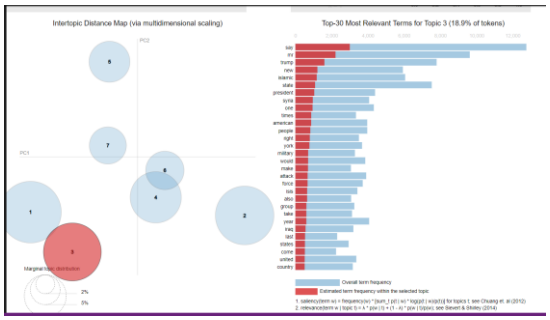
Topic 1 :



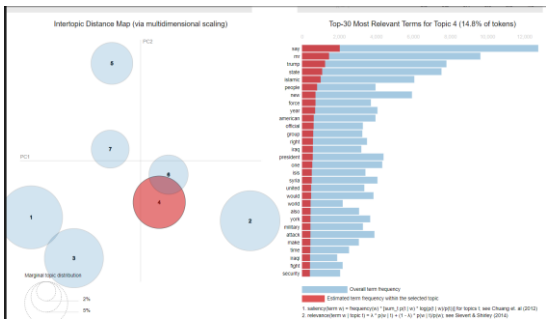
Topic 2:



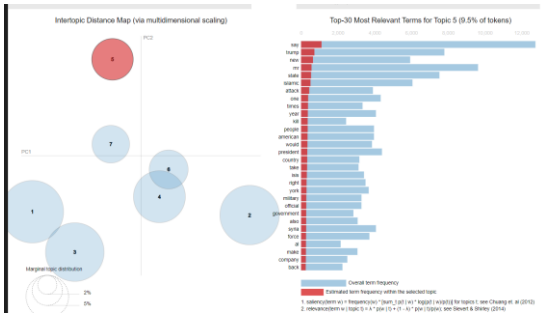
Topic 3:



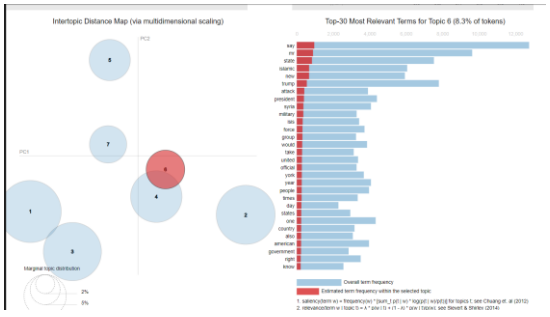
Topic 4:



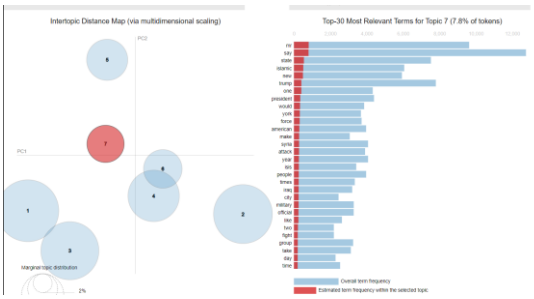
Topic 5:



Topic 6:



Topic 7:



The analysis also revealed that terrorist organizations were mostly portrayed in the media in the context of their violent activities, such as terrorist attacks and military operations against them. The media also covered topics related to the funding and recruitment of terrorist organizations.

Furthermore, it was observed that the portrayal of terrorist organizations in the media was influenced by the political climate of the time. For instance, during the period of analysis, there was increased coverage of political tensions in North Korea, which led to an increase in articles related to North Korean involvement in terrorism.

4. Conclusion

In conclusion, we have successfully used topic modeling to examine how terrorist organizations are portrayed in mainstream media. The dataset was preprocessed, a corpus object was made, and exploratory data analysis was carried out. The LDA algorithm was then used to create a topic model, after which the findings were discussed. We can determine the most prevalent topics in the corpus by looking at the top words in each topic and the distribution of topics across the corpus. Journalists can use this information to better understand how terrorist groups are portrayed in the media and to spot any biases that may exist.

GitHub link for project code

Repository link: Link

5. References

https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/

https://www.tutorialspoint.com/natural_language_toolkit/natural_language_toolkit_stemming_lemmatization.htm