Preprocessing and Topic Modelling on News Articles with Political & Social Issues

Kaushal Arvindbhai Dhanani, FNU Syed Sohaib Ali, Sai Jagadeesh Yadavalli
Engineering Data Science
University of Houston
kdhanani@uh.edu, Syedsoha@cougarnet.uh.edu, Syadava2@Cougarnet.uh.edu

Abstract

Topic modeling is a technique used in natural language processing (NLP) to identify topics or themes present in a collection of text documents. It is a statistical method that uses algorithms to uncover hidden patterns and semantic structures within large sets of text data. This report presents an in-depth analysis of how terrorist organizations are portrayed in traditional media outlets by applying topic modeling to a dataset of articles collected from the Wall Street Journal and the New York Times in 2017. We utilize different preprocessing and exploratory data analysis techniques to clean, process and analyze the data, and use Latent Dirichlet Allocation topic modeling technique to identify underlying themes and patterns in the coverage.

1 Introduction

Traditional media outlets, such as newspapers, play a significant role in shaping public opinion and understanding of various issues, including terrorism. In this project we use topic modeling techniques to analyze articles from the Wall Street Journal and the New York Times, two prominent newspapers, to understand how terrorist organizations are portrayed in their coverage. Topic modeling is a powerful unsupervised machine learning technique that discovers hidden thematic structures in large text datasets.

Many applications, such as information retrieval, text mining, and recommendation systems, can benefit from topic modeling. It can help us comprehend the major themes and concepts in a big corpus of text data, as well as making searching and browsing easier, making data organization easy.

2 Methodology

2.1 Data Collection

The data was collected from Factiva which is a global news database that aggregates content from a wide range of sources, including newspapers, magazines, and news wires from around the world. The database includes millions of articles from thousands of sources and covers a broad range of topics such as business, politics, and social issues.

The articles collected in 2017 from the Wall Street Journal and the New York Times cover a wide range of topics related to terrorism, including attacks, threats, and responses to terrorism by governments and other organizations. The articles also cover issues such as the roots of terrorism, the psychology of terrorists, and the impact of terrorism on society.

2.2 Data Pre-processing

It involves preparing and transforming raw text data into a structured and easily understandable format for further analysis, modeling, or processing. The main goal of text pre-processing is to remove noise, inconsistencies, and irrelevant information from the data and convert it into a format that can be effectively used by algorithms and models. It involves various steps such as cleaning the corpus, tokenizing the text, extracting features etc.

2.2.1 Building A Corpus

The input data was loaded as a large string character and then split into individual articles to build a corpus object containing all the articles. To do this we initially created a dictionary which contains the data and then we used a function doc2bow() to convert this dictionary into corpus.

2.2.2 Cleaning the data

The meta-data was separated from the actual articles to ensure that the analysis focused on the content of the articles. We used re.sub('METADATA_PATTERN') function to remove the metadata from the text. Moreover, pre-processed the data by cleaning it, converting the data into lower case and lemmatized the whole dataset by using lemmatize() function.

The Dataset consists of a huge set of articles collected from large Global News database called Factiva. As the articles are collate into multiple files along with the meta data, Cleaning the text files to remove the meta data became crucial for further analysis the data. We started with simple splitting of meta data and articles which are separated by double new line characters. Then each of the observations are observed to find out the patterns of meta data and iteratively filtered the content. We have observed patterns like promotional meta information having words like follow, signup, newsletter, get updates etc. The list includes sentences starting and ending with characters like .,/,,," etc. Further word count of filtered corpus is analyzed to find out the minimum size of sentence that makes sensical in providing useful information for further analysis.

The results were provided after cleansing the articles. Results are also visualized by a word cloud after usual preprocessing like tokenization, lemmatization and stop word filtering.

2.2.3 Feature Extraction

To extract the features from the dataset we utilized the Bag of Words technique (BoW) which converts text into numerical representations.

2.2.4 Data Visualization

The most commonly occurring words in the articles were represented visually by a word cloud. The word cloud can be used to find the terms that appear more frequently, the more prominently it will be shown in the cloud.

The goal of a word cloud is to quickly convey the most common words in a text document or set of documents, allowing the viewer to quickly identify the most relevant or important themes. Word clouds are often used in marketing, social media, and other areas where it is important to quickly identify key topics or themes from large amounts of text data.

We use matplotlib and pyLDAvis to visualize the topic models from the analysis by creating a word cloud.

We have also visualized the length of different articles, meaning summary of the corpus after cleaning in fig 2.

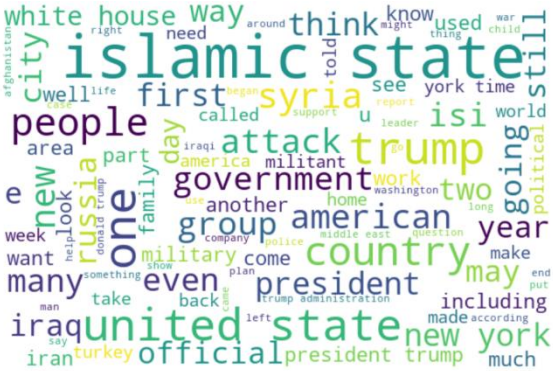


Fig.1: Word cloud of Corpus

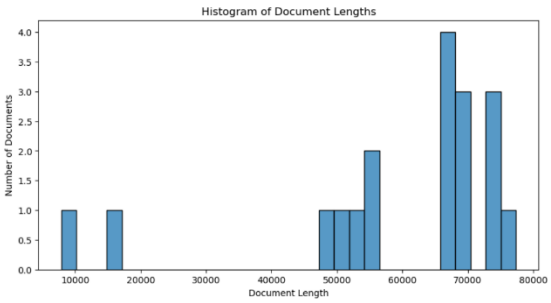


Fig 2: Summary of corpus after cleaning

2.2.5 Topic Modelling

To create a topic model using the cleaned corpus, we utilized Latent Dirichlet Allocation (LDA)

algorithm provided by the Gensim library. The goal of topic modeling is to identify the underlying topics present in a collection of documents.

Its algorithm assigns each word in each document to one of the latent topics, then iteratively refines these assignments by looking at the topic distribution of each document and the word distribution of each topic. It adjusts the topic assignments of each word based on the probability of the word occurring in each topic and the probability of the document containing each topic.

3 Experimental results

We have preprocessed the data first and then visualized the top words of whole corpus. We have also applied topic modelling algorithm to generate the topics from our corpus.

3.1 Preprocessing

The first stage in our project is to import the dataset and transform it into a corpus of individual articles. On the corpus, we perform data preprocessing and exploratory data analysis (EDA) to derive observations and insights.

The corpus was loaded as a list of articles and combined into a single string object.

- We build the corpus of text data by extracting metadata (such as the document's title, author, and date) and cleaning the text.
- The cleaned text is then tokenized (split into individual words), and the punctuation and stop words (common words like "the" and "and") are removed.
- The remaining words are lemmatized (reduced to their base form) and stored in a new list.

The purpose of this process is to prepare the text data for use in natural language processing (NLP) tasks, such as sentiment analysis or topic modeling. By cleaning and preprocessing the text, the data is easier to analyze and more accurate results can be obtained.

Table 1 shows the top ten words in the corpus and their frequencies. These words represent some of the most popular themes and issues discussed in the articles.

Word	**Frequency**
Said	11693
State	10457
Trump	7730
Islamic	6289
New	6230
Time	5932
President	4372
One	4283
American	4088
syria	4082

Table 1: Word frequency of corpus

We have also generated wordlcoud for most common words in whole courpus. Fig 1 shows the representation of the wordcloud of our corpus.

3.2 Topic Modelling

Topic modeling is a technique used in natural language processing and machine learning to identify topics or themes within a collection of text data. It is an unsupervised learning method that can be used to discover latent patterns and relationships within a large corpus of text.

The output of a topic modeling algorithm is a set of topics and their corresponding word distributions, which can be used to summarize the main themes or concepts within the text data. Topic modeling can be applied to a wide range of applications, including text classification, document clustering, and content recommendation.

Some popular tools and libraries for topic modeling include Gensim, Mallet, and Scikit-learn in Python.

3.2.1 LDA Algorithm

Latent Dirichlet Allocation (LDA) is a widely used generative model for topic modeling.

Assumptions:

- Documents are generated from a mixture of topics.
- Each topic is characterized by a distribution over words.
- The distribution of topics in a document follows a Dirichlet distribution.

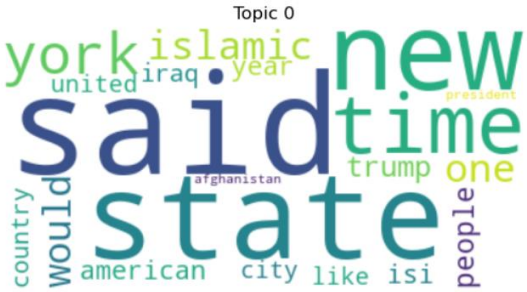
Process:

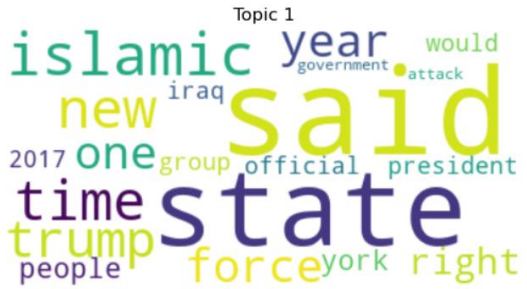
- Choose the number of topics K.
- Initialize the parameters for the model, including the topic-word distribution and the document-topic distribution.
- Iterate through each document and for each word: Calculate the posterior distribution of topics for the word. Sample a topic from the posterior distribution.
- Update the counts for the word, topic, and document accordingly. Repeat the iteration process until convergence.

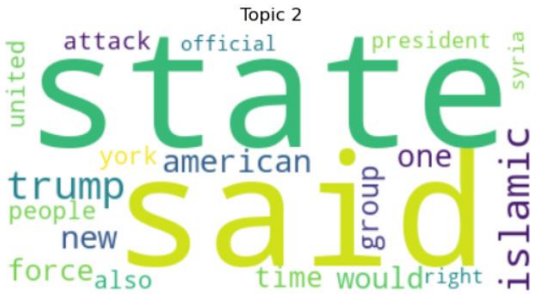
We have employed the following steps for our topic modelling:

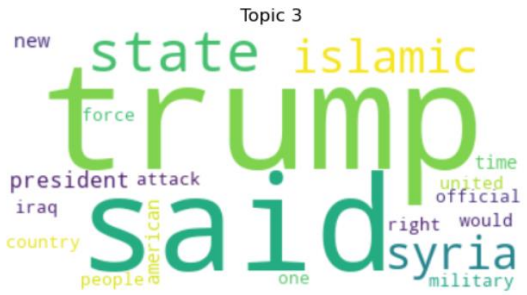
- The documents were preprocessed by tokenizing them into words and removing stop words.
- The preprocessed documents are stored and we have used necessary libraries like gensim for topic modeling and matplotlib and pyLDAvis for visualization.
- Then, we have employed topic modelling using Latent Dirichlet Allocation (LDA) for 10 topics and visualized the results with word clouds and pyLDAvis library.

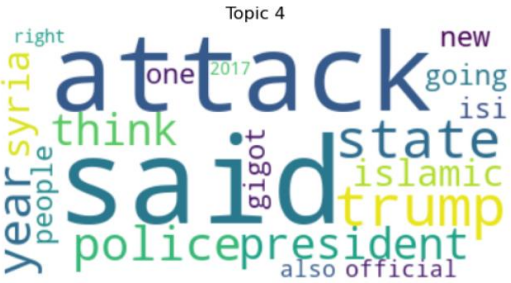
Word-clouds for each topics:











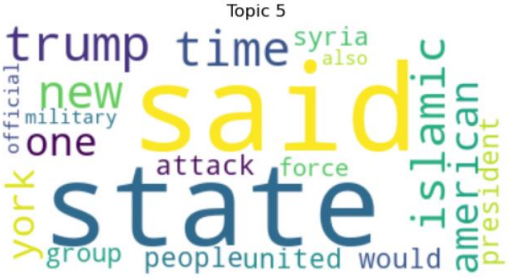


Fig 3: Wordclouds for each of 10 topics

Topic modelling with different parameters

We have selected the ranges for number (3,5,7,10) of topics and number of passes (5,10,20) and run the LDA again to see how it performs. We have saved the summary of different number of topics too. We have stored the summary of these different model in txt files. We have shown the summary for 10 topics with 10 passes below.

```
Topic 0: 0.003*"said" + 0.002*"trump" + 0.002*"state" + 0.002*"new" + 0.002*"time" + 0.
Topic 1: 0.002*"state" + 0.002*"said" + 0.001*"trump" + 0.001*"islamic" + 0.001*"new" +
Topic 2: 0.011*"said" + 0.009*"trump" + 0.009*"state" + 0.006*"islamic" + 0.005*"presid
Topic 3: 0.011*"said" + 0.009*"trump" + 0.009*"state" + 0.007*"islamic" + 0.006*"syria"
Topic 4: 0.011*"state" + 0.011*"said" + 0.009*"islamic" + 0.005*"group" + 0.005*"iraq"
Topic 5: 0.002*"said" + 0.002*"state" + 0.001*"islamic" + 0.001*"one" + 0.001*"new" + 0
Topic 6: 0.010*"said" + 0.010*"state" + 0.007*"new" + 0.007*"time" + 0.006*"trump" + 0.
Topic 7: 0.012*"state" + 0.011*"trump" + 0.010*"said" + 0.007*"new" + 0.007*"united" +
Topic 8: 0.010*"said" + 0.008*"state" + 0.008*"new" + 0.008*"time" + 0.005*"york" + 0.0
Topic 9: 0.011*"said" + 0.010*"state" + 0.007*"time" + 0.007*"new" + 0.006*"islamic" +
```

Fig 4: Output Summary of Topic model 10 with 10 passes

Like this, we have saved the summary of topic models with different inputs.

- The model is finally utilized to further visualize with an interactive visualization package called pyLDAvis, which displays the topics produced by the LDA model (refer to Figure 5), together with the words that best describe each subject and the connections among them.

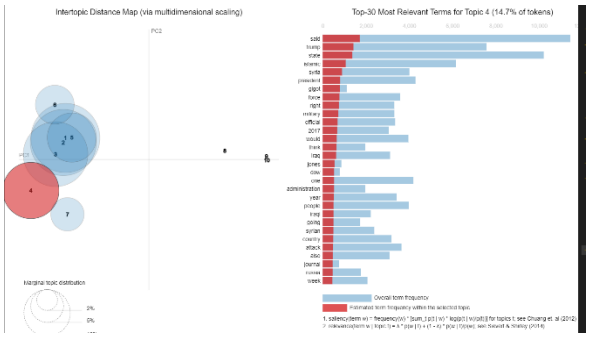


Fig 5: Topic modelling visualization

4 GitHub link for project

Repository link:

https://github.com/CIS-6397-Textmining-Spring-2023/miniproject3-miniproject3_group-12

5 Conclusion

The data for this study was gathered in 2017 from Factiva, a huge Global News database, and includes items from the Wall Street Journal and the New York Times. Tokenizing and eliminating

stop words are used to preprocess the articles. To extract topics from the corpus, the study used topic modeling with the LDA algorithm.

We were successfully able to remove the Meta data from the articles in the corpus and by using the Latent Dirichlet Allocation (LDA) algorithm we were able to determine the meaning topics from all of the articles from the given corpus.

Our in-depth analysis using topic modeling provides valuable insights into how traditional media outlets portray terrorist organizations. By identifying common themes and patterns in the each of the 10 topics , we can help journalists and researchers better understand the narrative surrounding terrorism in traditional media and potentially improve the quality of reporting on this complex and sensitive topic.

Our findings could expand the analysis to other media outlets or time periods to provide a more comprehensive understanding of the evolution of terrorism coverage over time.

6 Author Credit Statement

Kaushal Arvindbhai Dhanani – Methodology, Software, Project administration, Writing: review & editing.

FNU Syed Sohaib Ali– Investigation, Formal analysis, Theoretical conceptualization

Sai Jagadeesh Yadavalli – Writing: editing & review, Data curation, Validation, Visualization