Topic Modeling – Group 4 Report

Sai Phani Ram Popuri
Conceptualization, Formal Analysis, Software
spopuri2@cougarnet.uh.edu

Fahad Mohammed Abdul
Review and Editing, Writing original draft
fmoham20@cougarnet.uh.edu

Purva Dixitkumar Desai
Methodology and Documentation
pdesai3@central.uh.edu

All the code used in this paper can be found at our GitHub repository

Abstract: Topic Modeling has been one of the most used NLP techniques with the increase in text data. The following paper aims at understanding various aspects of it by building a model that segregates how the articles from the Wall-street journal, New York Times portray the terrorist organizations in their writings and the commonly used words these agencies use to name the anti-social elements.

1 Introduction

Text analysis is essential in any large global news database for report writing, as it allows for the extraction of relevant information and insights from vast amounts of data quickly and efficiently. Text analysis involves using various tools and techniques to analyze textual data, such as natural language processing, sentiment analysis, entity recognition, topic modeling, and machine learning algorithms. With text analysis, journalists and researchers can identify patterns, trends, and themes in the news data, helping them to uncover insights that may not be immediately obvious. They can use these insights to produce more in-depth and accurate reports that are based on data-driven analysis. Text analysis can also help to identify bias, misinformation, and fake news in the news data, which is essential for producing accurate and reliable reports. By using text analysis tools to identify the sources of fake news and misinformation, journalists and researchers can help to prevent the spread of false information and maintain the integrity of their reporting. Overall, text analysis is a valuable tool for anyone working with large global news databases, as it can help to extract insights, identify patterns, and ensure the accuracy and reliability of reports.

1.1 Corpus Compilation

After carefully examining every text file in the articles folder, we have concluded that every article therein ends with a phrase that contains the strings "Document NYTF," "Document WSJO," or "Document J." We thoroughly examined each file and found that these strings were always present at the end of each article, which led us to this conclusion. To speed up the pre-processing phase of our research, we combined all the articles into a single corpus file. Now that it has been fully prepared, this file is ready to be used for data processing and analysis.

1.2 Preprocessing

The part of the text mining process first examined in this paper is data preprocessing - a paramount task in the field of text mining. Data is, in its raw form quite dirty. In the context of textual data, this "dirtiness" takes many forms, some of which are due to erroneous data entry and others due to conventions which, though linguistically correct, make data less machine-readable. Punctuation marks, spelling and grammatical errors, capitalization, HTML tags, Metadata and more can contribute to this dirtiness. This paper will look at preprocessing methods which help to cleanse the data of these irregularities.

1.3 Word Frequency

Once data has been cleaned, the task of data analysis can begin. One method of analyzing cleaned data is looking at word frequency. We will be determining how word frequency reflects the content contained in a corpus, and the effects that preprocessing has on the usefulness of word frequency.

1.4 Text Representation

Another critical aspect of text mining is the form of text representation selected. In this paper, we examine the bag of words representation as well as n-grams to determine how text representation affects our analysis.

1.5 Topic Modeling

LDA (Latent Dirichlet Allocation) is a widely used statistical model in natural language processing and machine learning for discovering topics in large collections of text data. It is a generative probabilistic model that assumes each document in a corpus is a mixture of various topics, and each topic is a probability distribution over a set of words. LDA is a powerful tool for automatically extracting underlying themes or topics from unstructured textual data and is commonly used in applications such as topic modeling, sentiment analysis, and information retrieval.

2 Methodology

2.1 Data Preprocessing

Preprocessing of the data involves the following steps:

Converting to lowercase: This allows for comparisons to be made without accounting for case differences.

Removing Metadata: Every article consists of metadata that starts from the first line and ends with 'all rights reserved' Therefore, we are eliminating all the text that forms the metadata.

Removing HTML Tags: When text mining web data, HTML elements are regularly encountered. Because they can make it difficult to examine the text content and we are just removing them.

Punctuation removal: This must be done with consideration as punctuation can sometimes convey meaning (us vs. U.S for example) but very often punctuation does not convey much meaning.

Converting numbers to words: Sometimes text might contain numbers along with the words which might not be helpful for analyzing. The numbers are converted to its word representation for better analysis.

Tokenization: Taking the cleaned data and splitting it up into its individual and unique words is a process called tokenization. This set of words or "tokens" is in the form of structured data and can be analyzed as such.

Removing stopwords: Some words such as "the" and "and" which don't convey much information but are used frequently can improve the informativeness of the word frequency distribution. The effect of removing stopwords and stopword selection on results in explored in this paper.

Lemmatization: Since we are trying to find the underlying patterns associated with the corpus, it is essential to consider the dictionary words rather than the stemmed tokens. Lemmatization allows us to reduce the tokens to their dictionary-form and therefore we have implemented it. In specific, we made use of the Word-net Lemmatizer to perform the task. One key aspect while performing the lemmatization is, we only considered the words with the parts-of-speech tagging as either of 'Noun', 'Adjective' or 'Verbs' as these provide information pertaining to the topics.

2.2 Analysis

In short, our process for corpus analysis is documented below in the simplest manner.

1. Read the articles.
2. Preprocessing of the articles
3. Removed stop words from data.
4. Tokenized the text data.
5. Printed top 30 words (k=30)
6. Created top 30 n-grams for (n=2)
7. LDA analysis keeping num_topics = 8.
8. Choosing the optimal number of topics based on coherence score evaluation metric.
9. Visualize the topics using pyLDAvis library.
10. Inference and conclusion.

Analysis of n-grams can provide insights into patterns of co-occurrence and contextual relationship between words in each text. Moreover, it can reveal which words frequently occur together, and which words are more likely to follow or precede certain other words. Additionally, n-gram analysis can help to identify common phrases that might be missed if only single words are considered. Hence, we have analyzed the corpus with n=1 and n=2 for more understanding. We have used ***FreqDist*** function of NLTK library to find the word distributions of our text data. Finally, we have used LDA topic modelling for extracting topics and their related keywords from the corpus.

3 Experimental Results

Once the corpus was read and preprocessed, we began our analysis of the data. The preprocessing was done in python. We have analyzed the corpus with and without stopwords. Later, we employed the LDA for additional analysis, and based on various topic selections and passes, we came to various conclusions.

3.1 Simple Preprocessing

In the first stage, the data was preprocessed, and we removed stopwords from the corpus and after all stop words have been eliminated, the corpus has a processed list of tokens. The sample list displays the first 20 tokens after the stop words have been eliminated. The tokens are shown below.

('istanbul', 'turkish', 'officials', 'accused', 'united', 'states', 'abetting', 'failed', 'coup', 'summer', 'russian', 'ambassador', 'turkey', 'assassinated', 'month', 'turkish', 'press', 'united', 'states', 'attack')

From the above data, we can say that the corpus is related to 'coup' and 'Turkey, and 'united states' is repeated many times. Probably, this corpus is related to a political coup that took place in Turkey. But we are not sure because 'Russia' is mentioned in the corpus, so this might be something else. Also, this corpus mentions 'failed, which means some officials tried to overturn the government and were accused in the process, but they couldn't succeed in doing so. Overall, the data involves countries and is targeted at the governments and officials involved in the coup process.

3.2 Analysis for N-grams

We considered the unigram and bigram analysis for identifying the most commonly occurring words in the articles based on their respective frequencies. It is evident that if a particular word is associated to a topic, then the document containing the topic might also possess the same word in its bucket.

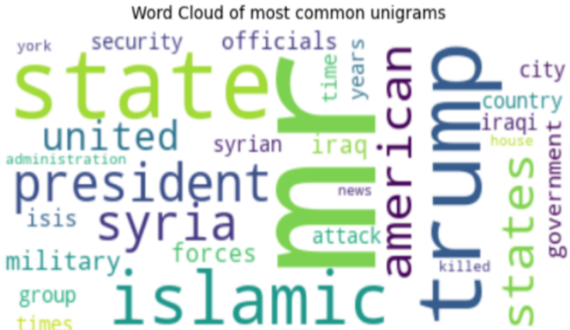

Fig 1: Word Cloud for Unigrams

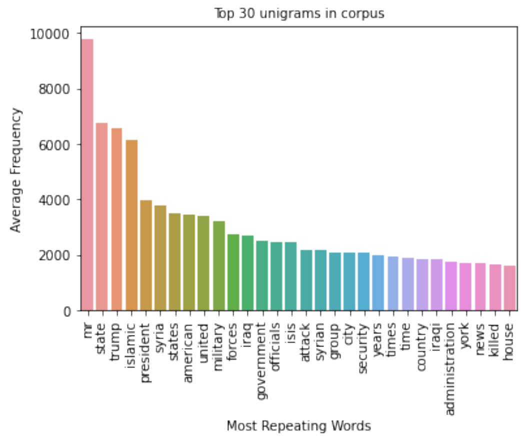

Fig-2: Distribution for unigrams

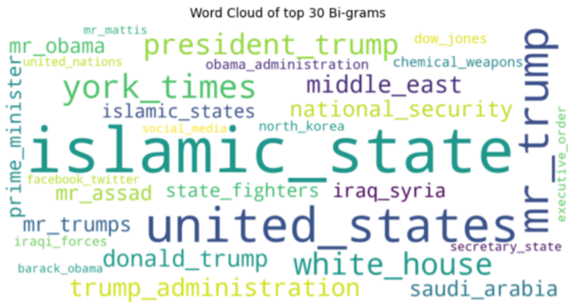

Fig-3: Word cloud for bi-grams in the corpus.

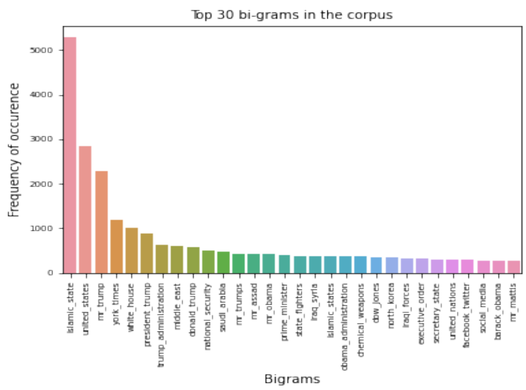

Fig-4: Distribution for corpus(n=2)

From Fig 1 and Fig 2, the top 30 one-word phrases in each text, based on their frequency. The resulting list displays the most common words, including "mr", "state", "trump", "islamic", and "president". These words suggest that the text may be related to politics, international affairs, and security, with a particular focus on the United States and the Middle East. This information could be useful in analyzing the text and understanding its content and context.

From Figs 3 and 4, the top 30 bigrams, which are pairs of two words that commonly appear together in each text corpus. The most frequent bigrams in the corpus include "islamic state", "united states", "mr trump", "white house", and "President Trump", suggesting that the text may be related to politics and international affairs, especially the Trump administration. Other frequently occurring bigrams include "middle east", "saudi arabia", "chemical weapons", and "north korea", indicating a focus on security and conflict in the Middle East and Asia. The list provides insights into the important topics and themes discussed in the text corpus and can be used for further analysis.

3.3 LDA (Latent Dirichlet Allocation)

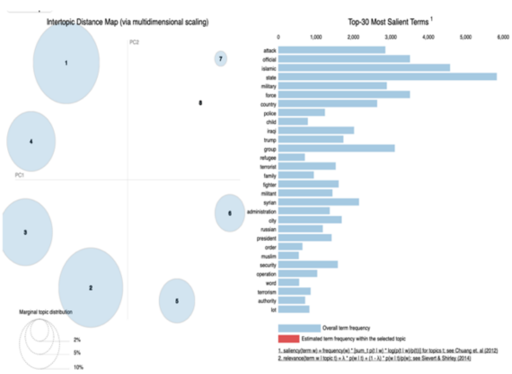

Fig 3: LDA visualization for 5 topics

Here we have selected 5 topics randomly and from Fig. 3, we can see the marginal topic distribution, and when you choose the topic, it gives the total token percentage of each topic, with the highest having 32.6%, and we have analyzed this for the top 30 terms in the corpus. The lowest is Topic 8, with 0.7%. The top ten most important words and their respective weights have been taken for analysis. The weights represent the importance of each word in its corresponding topic. Topic 0 is related to national security and immigration, with words such as "country," "refugee," "executive order," "immigration," and "security." Topic 1 relates to social issues, with words such as "family," "child," "police," and "woman." Topic 2 is more general and discusses broader themes such as time, story, politics, and world events. Topic 3 revolves around President Trump, his administration, and his policies. Topic 4 focuses on terrorism and violent attacks, with words such as "attack," "police," "terrorist," and "terrorism." Topic 5 is related to the military and conflicts, with words such as "state," "force," "military," "syrian" and "iraqi" Topic 6 is related to art and culture, with words such as "art," museum," "artist," and "painting." Finally, topic 7 is related to legal matters and constitutional issues, with words such as "affidavit," "legal," "detainee," "court," and "constitutional." Overall, the topic model provides insight into the variety of topics that could potentially be found in a large corpus of text and the relative importance of certain words within each topic.

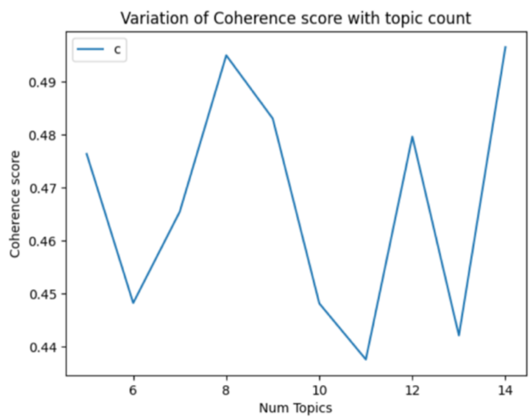

Fig 4: Plot for finding optimal number of topics.

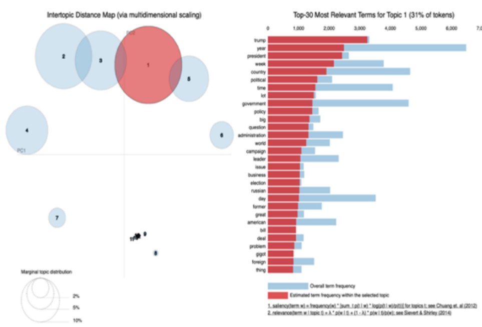

Fig 5: Frequency of terms

Here we are finding the optimal number of topics, which is 13, as we can see from Figs. 4 and 5. The topics include Iranian politics, terrorism, the military, art, the news media, and others. One of the most prominent topics in this dataset is related to Iranian politics, with words such as "Iranian," "Tehran," and "Iran" being the most important keywords. This suggests that the dataset contains articles that discuss Iranian politics and its various aspects, including the Iranian government and its policies. Another important topic in this dataset is terrorism, with keywords such as "attack," "police," "Islamic," and "terrorist" being the most important. This suggests that the dataset contains articles that discuss acts of terrorism, their perpetrators, and their impact on society. The dataset also includes topics related to the military and security, with words such as "force," "military," and "security" being prominent. This suggests that the dataset contains articles that discuss military operations, security policies, and other related topics. Other topics in the dataset include art, news media, and various miscellaneous topics such as sports, exhibitions, and other events. Overall, the dataset appears to contain a wide variety of topics, with a focus on politics, terrorism, and security-related topics. The given keywords and their weights provide insights into the most important aspects of each topic, allowing for a better understanding of the content within the dataset.

4 Conclusion

In this project, we have preprocessed the data and analyzed the corpus after removing stopwords. Specifically, we have performed LDA topic modelling after carefully analyzing the data. In conclusion, the 8-topic model is superior to the 13-topic model because it has non-overlapping and distinct topics, with each topic having a clear focus and set of important keywords. The 8 topics cover a range of important issues such as national security, social issues, art and culture, and legal matters. This model provides valuable insights into the variety of topics that can be found in a large corpus of text and the relative importance of certain words within each topic.

The below table highlights the topic names for 8-topic model as it provided with the non-overlapping solution.

Topic 1	National Security
Topic 2	Social and Law Enforcement
Topic 3	Prominent global events
Topic 4	President Trump administration
Topic 5	Terrorist organizations and vices
Topic 6	Military face-offs
Topic 7	Art and Culture
Topic 8	Legal issues and Constitutional Affairs

5 References

1. https://github.com/fastflair/Tutorials/blob/master/NLP_ML/NLP_Tutorial.ipynb
2. https://github.com/adashofdata/nlp-in-pythontutorial/blob/master/4-Topic-Modeling.ipynb
3. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
4. https://www.youtube.com/watch?v=nNvPvvuPnGs&t=619s
5. https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13