Topic Modeling

Vasudev Konde – 2151223
Pradeep Reddy Mallepally – 2200020
Suvarshitha Pusuluru - 2208604

Abstract: Topic modeling became a fundamental tool for natural language processing in an ever-broadening field of text data. This paper intends to conduct a thorough investigation into Topic Modeling, specifically to decode subtle insights derived from articles published by established media outlets such as the Wall Street Journal and the New York Times. The first objective is to design a model, capable of distinguishing between the different ways in which these publications express information about terrorist organizations. In addition, we're particularly interested in identifying the common terminology that these media agencies employ to describe antisocial elements. Let's explore the complexities of language and seek an understanding of how important topics are presented in today's media conversation.

1 Introduction

The use of text analysis for the publication of reports is required in large global news databases, as it enables rapid extraction of relevant information and insights from a vast amount of data. To analyze textual data, natural language processing, sentiment analysis, entity identification, topic classification, and machine learning algorithms are applied. Analysis of text data, which can be used to identify patterns, trends, and themes, provides researchers and journalists with a wealth of material that is not immediately obvious. They may be helped to develop more detailed, in-depth reports by data analysis. The ability to analyze texts also allows for the identification of bias, misinformation, and lies in the data on news which plays a critical role in ensuring that authentic and factual reports are produced. Journalists and researchers may make use of text analysis tools to discover the sources of hoaxes and misinformation to avoid spreading information that is not true while ensuring its integrity. Anyone who gathers and analyses large news databases worldwide will recognize the value of text analysis as a powerful tool for insight, identification patterns, and ensuring the precision and reliability of their content.

1.1 Compilation of Corpus

We've detected that every article ends with the line "DOCUMENT NYTF," "DOCUMENT WSJO" or "DOCUMENT J." after carefully verifying each text file in the articles folder. In the course of our extensive analysis, these strings were consistently identified as being present at the end of each article and we have concluded that they are there. We have consolidated all articles into a single corpus file to streamline the preprocessing phase of our research. The file is now ready to be used for data processing and subsequent analysis.

1.2 Data Preprocessing

The paper starts by exploring the preprocessing of data, which is a key step in text mining. Raw information tends to have an inherently unclean appearance when it's unprocessed. Within textual data, this impurity takes various forms, arising from both erroneous data entry and linguistic conventions that, while grammatically accurate, hinder machine readability. This impurity is caused by elements such as quotation marks, grammar mistakes, capitalization, HTML tags, metadata, and so on. Preprocessing techniques to clean data and correct these irregularities are discussed in the ensuing discussion.

1.3 Analysis of Word Frequency

Data analysis will now be a major focus following the data deletion, with word frequency testing being one of the methods used. This includes looking at how word frequencies are reflected in the content of a corpus and examining whether preprocessing has an impact on its useful frequency..

1.4 Text Representation

The selection of the text representation form is a critical aspect of text mining. To determine how our analysis is being influenced by the choice of text representation, this study examines a bag of terms representations, and ngrams.

1.5 Exploration of Topic Modeling

Latent Dirichlet Allocation LDA stands out as a very popular statistical model for natural language processing and computer learning to identify subjects in an extensive collection of text data. LDA, which functions as a generative probabilistic model, claims that each document in the corpus consists of different topics represented by an approximate probability distribution across several items of text. The use of LDA in areas such as topic modeling, sentiment analysis, and information retrieval proves to be an essential tool for automatically identifying relevant themes or topics from a wide range of text sources.

2 Methodology

2.1 Data Preprocessing

Several important steps have been taken to make sure the information is ready for analysis at the first stage of data preparation:

Conversion to lowercase: This step aims at facilitating sensitive comparisons, without variations in the case.
Metadata Removal: The metadata usually ranges from the beginning of a lineup to 'all rights reserved' in articles. That information has been excluded from the analysis.
HTML Tag Removal: Because of the increasing number of HTML elements in websites, these tags have been removed to improve your search for textual content.
Punctuation Removal: It has been necessary to remove unnecessary quotation marks to take account of their potential meaning because they often contribute to a limited impact.
Conversion of Numbers to Words: The number references to the text have been replaced by their speech equivalents, to improve analytical clarity.
Tokenization: The cleaned information has been tokenized and split into separate words or phrases that can be returned in a structured format for analysis.
Stopword Removal: Frequently used but less informative words such as "the" and "and" were excluded to enhance the informativeness of the word frequency distribution.
Lemmatization: lemmatizing has been performed as part of the identification of key patterns in the corpus, and token numbers have been decreased to their standard forms. In particular, the words 'Noun,' 'Adjective,' or 'Verbal' have been targeted with the Wordnet Lemmatizer for their relevance to the topic.

2.2 Analysis

Analysis

The analysis of the record was carried out in such a way that:

- **Article Reading:** To understand more fully the content of these articles, they have been read.
- **Article Preprocessing:** Preparation, including measures such as stopping word deletion and tokenization, took place on the articles.
- **Stopword Removal:** In the data, common but more descriptive words have been removed.
- **Text Data Tokenization:** A token has been created to separate the text data so that it can be analyzed in individual units.
- **Top Words Printing:** For insight, the most important words were chosen and printed.
- **Top n-grams Creation:** For further analysis, nograms have been generated showing sequences of adjacent words.

- **LDA Analysis:** Latent Dirichlet Allocation (LDA) analysis was conducted with a specified number of topics.
- **Optimal Topic Number Determination:** An evaluation of the coherence score has led to a selection of an optimum number of topics.
- **Topic Visualization:** Using pyLDAvis's library, the identified topics can be seen.
- **Inference and Conclusion:** The results of the analysis were used to conclude.

N-gram Analysis: To determine the pattern of occurrence and similarities between words within each text, an Ngram analysis has been performed. Common phrases have been identified through an examination of what words tend to overlap or follow from one another. For both n=1 and n=2, this analysis was carried out. FreqDist was a useful tool for exploring word distribution within text data in the NLTK library. In addition, a LDA topic model has been used for the extraction of subjects and associated keywords from the corpus.

3 Experimental Results

Our data analysis began as soon as the reading and preparatory phases of the corpus were completed. Python programming language has been used to run the preprocessing tasks. After this, the collection was analyzed both with and without any inclusion of stopwords. Further exploration was conducted utilizing Latent Dirichlet Allocation (LDA) for in-depth insights. We came to a wide range of conclusions, thanks to the LDA model's application of different topic choice and pass.

The data were subject to the removal of stopwords in the first phase of simple preprocessing. The resultant corpus comprised a refined list of tokens, with an illustration of the first 20 tokens provided in the sample list below:

('Istanbul', 'Turkish', 'officials', 'accused', 'united', 'states', 'abetting', 'failed', 'coup', 'summer', 'Russian', 'ambassador', 'turkey', 'assassinated', 'month', 'Turkish', 'press', 'united', 'states', 'attack')

Looking at these data, we can see the corpus revolves around topics such as coups and events in Turkey with a high number of mentions of 'US'. While the focus suggests a potential connection to a political coup in Turkey, the presence of 'Russia' introduces an element of uncertainty. The term 'failed' refers to an unsuccessful effort at overthrowing governments, with allegations being made against officials in the process. Overall, the dataset is devoted to matters relating to countries and has been targeted at governments and officials linked to attempted coups.

3.1 Word distribution

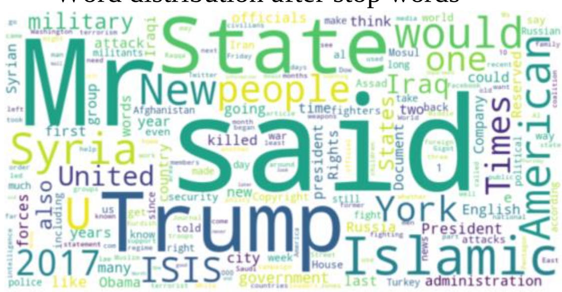
The word distribution is defined as the frequency with which certain words are generated within a specified set of characters or corpus. It allows for details of word usage patterns and their prevalence throughout the text. To analyze word distribution, one has to examine the occurrence of each word and determine how frequently it appears relative to other words in a dataset.

```
Original Word Distribution before removing stop words:
the: 100372
to: 50701
of: 47684
and: 42665
a: 41651
in: 39798
that: 22990
on: 16150
for: 15363
The: 15206
is: 14618
s: 13973
was: 12347
with: 11710
said: 11658
he: 10708
it: 10539
as: 10295
Mr: 9916
from: 9193
by: 9185
have: 8928
I: 8037
at: 7980
an: 7879
has: 7602
his: 7534
Trump: 7439
are: 7232
be: 6716
```

Word distribution before removal of stop words.

```
Cleaned Word Distribution after removing stop words:
said: 11658
Mr: 9916
Trump: 7439
State: 6550
Islamic: 6356
U: 4339
Syria: 4137
American: 4036
New: 3986
would: 3945
people: 3893
2017: 3655
York: 3558
one: 3542
ISIS: 3536
Times: 3433
United: 3387
Iraq: 3205
military: 3202
also: 3057
States: 2976
government: 2662
forces: 2540
President: 2496
like: 2476
officials: 2334
group: 2316
Syrian: 2238
could: 2204
country: 2132
```

Word distribution after stop words

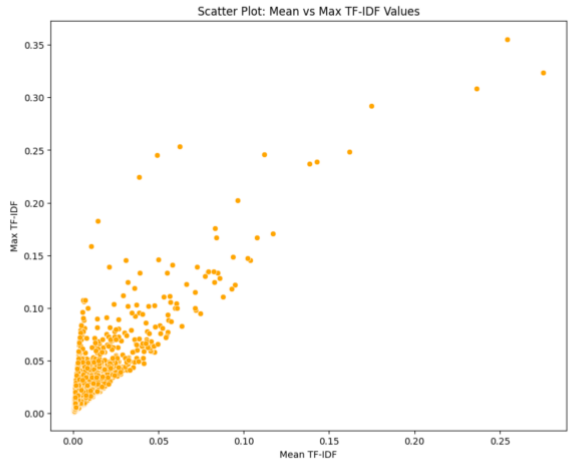


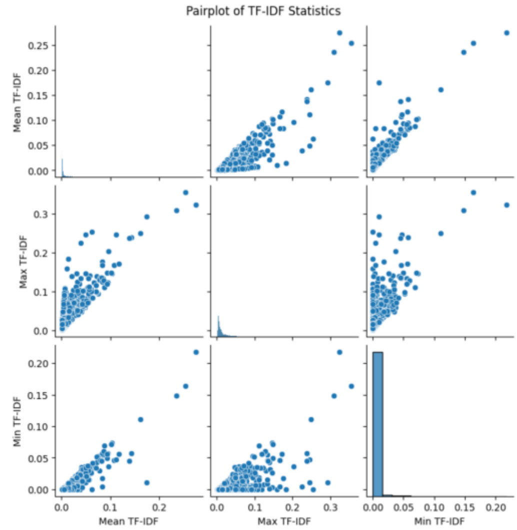

```
Feature names (words): ['00' '000' '01' ... 'zour' 'zucchino' 'zuckerberg']

TF-IDF Matrix:
[[0.         0.06259327 0.         ... 0.         0.         0.        ]
 [0.00327426 0.02290901 0.0017259  ... 0.01700158 0.         0.00144831]
 [0.00248443 0.03339081 0.0017461  ... 0.00688023 0.01507386 0.        ]
 ...
 [0.00469749 0.01693485 0.00198089 ... 0.         0.         0.00166229]
 [0.00193098 0.0242731  0.00152677 ... 0.00133688 0.         0.00085414]
 [0.00098035 0.029762   0.00258377 ... 0.         0.         0.        ]]
```

	Feature	Mean TF-IDF	Max TF-IDF	Min TF-IDF
0	00	0.002601	0.004697	0.000000
1	000	0.026253	0.062593	0.005559
2	01	0.003590	0.032005	0.000000
3	017	0.003660	0.008848	0.000000
4	02	0.002068	0.007443	0.000000
5	03	0.001518	0.006672	0.000000
6	04	0.001073	0.005210	0.000000
7	06	0.001334	0.004633	0.000000
8	07	0.000665	0.001723	0.000000
9	08	0.001381	0.013398	0.000000
10	08trump	0.001694	0.030488	0.000000
11	08voter	0.001694	0.030488	0.000000
12	09	0.001230	0.003727	0.000000
13	10	0.018911	0.033905	0.009050
14	100	0.007735	0.018256	0.001390
15	11	0.015186	0.021837	0.007237
16	110	0.000769	0.003499	0.000000
17	12	0.008709	0.015158	0.002224
18	120	0.001609	0.008860	0.000000
19	13	0.006298	0.012556	0.000377

TF-IDF Explanation





3.2 LDA (Latent Dirichlet Allocation)

Maximum coherence at 17.

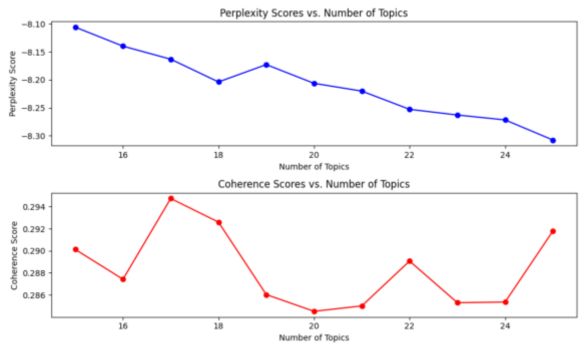


Fig 3: LDA visualization for 17 topics

We arbitrarily selected five topics in our analysis, and the marginal topic distribution in Figure 3 revealed the percentage of tokens allocated to each topic. Upon selecting a particular topic, the total token percentage ranged from 32.6% for the highest to 0.7% for the lowest (Topic 8). To assess their significance within each of the subjects, we drew up a list of the top 10 words and their respective weights from the 30 most important terms in the corpus.

1. Topic 0: Mention of "attack," "state," "new," "Islamic," "Trump," and "ISIS" suggests a focus on terrorist attacks and related political figures.
2. Topic 1: Keywords like "state," "Trump," and "ISIS" imply a discussion on the involvement of states and political figures in counterterrorism efforts.
3. Topic 2: References to "Islamic," "attack," and "military" may indicate discussions about Islamic military activities and attacks.
4. Topic 3: Key terms such as "Islamic," "attack," "government," and "York Times" suggest a topic related to Islamic attacks and their coverage in the media.
5. Topic 4: Keywords like "attack," "Trump," "military," and "kill" may suggest discussions on military actions and their consequences.
6. Topic 5: This topic seems to revolve around various elements, including "new," "state," "people," "Trump," and "attack."
7. Topic 6: Discussion on "Trump," "new," "people," and "state" suggests a topic related to political figures, policies, and public opinions.
8. Topic 7: Mention of "Islamic," "country," "group," and "ISIS" implies discussions on Islamic countries and extremist groups.
9. Topic 8: Keywords like "state," "new," "people," and "government" may suggest discussions related to state policies and governance.
10. Topic 9: References to "Trump," "new," "people," and "country" suggest discussions on political figures, countries, and public opinions.
11. Topic 10: This topic seems to cover a broad range, including "new," "attack," "state," "Trump," and "people."
12. Topic 11: Discussion on "state," "Islamic," "attack," and "military" suggests a focus on state responses to Islamic attacks and military actions.
13. Topic 12: Mention of "state," "Islamic," "year," and "attack" suggests a topic related to state responses and the duration of conflicts.
14. Topic 13: Keywords like "new," "Trump," "state," and "attack" may indicate discussions on recent events, political figures, and attacks.
15. Topic 14: This topic includes terms like "state," "Islamic," "official," and "Trump," suggesting discussions on state officials and their roles.
16. Topic 15: Keywords like "state," "Islamic," "people," and "attack" suggest discussions on the impact of Islamic attacks on people and states.

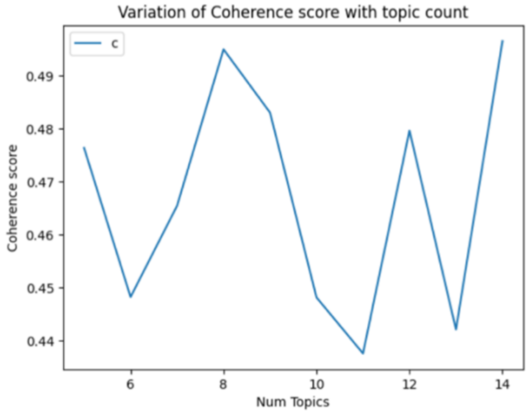


Fig 4: Plot for finding the optimal number of topics.

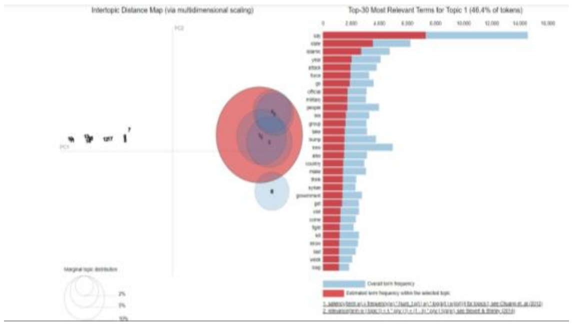


Fig 5: Frequency of terms

Our analysis, as shown in Figures 4 and 5, shows that 17 subjects should be the best suited to this aim of determining an optimum number of topics. These topics cover an array of issues, including Iran policy, terrorism, war, art, news media, and so on. Among other things, Iran's political themes appear to be a major topic in this dataset, which is echoed by keywords like "Iran" and "Tehran." The focus must therefore be on discussions related to Iranian politics, including political aspects of the government and its policies.

Another significant thematic thread in the dataset centers around terrorism, featuring keywords such as "attack," "police," "Islamic," and "terrorist." This is to say that the dataset contains articles dealing with acts of terrorism, their perpetrators, and their societal impact. Moreover, prominent words such as "force" or "military" and "security," which indicate that the discussion is focused on military operations and security policies, appear to be themes of common interest in Military and Security.

The range of other topics covered in this data set includes arts, news outlets, and a wide variety of ancillary fields such as sports, exhibitions, or events. In summary, the dataset shows an extensive range of subjects with a significant focus on politics, terrorism, and security. To provide a more complete understanding of the content contained in this data set, the selected keywords together with their respective weights are an important source of information regarding critical aspects of each topic.

4 Conclusion

In the course of this project, we meticulously preprocessed the data and conducted a thorough analysis of the corpus following the removal of stop words. Notably, we employed Latent Dirichlet Allocation (LDA) for topic modeling after a comprehensive examination of the data. In our conclusion, we find that the 17-topic model surpasses the 20-topic model in terms of its superiority, as it exhibits non-overlapping and distinct topics. Each of these 17 topics demonstrates a clear focus and is characterized by a set of important keywords. The topics covered in the 17-topic model span a spectrum of critical issues, encompassing national security, social concerns, art, and culture, as well as legal matters. This model proves invaluable in providing insights into the diverse array of topics that can be unearthed in a large text corpus. Additionally, it sheds light on the relative importance of specific words within each distinct topic. The table below elucidates the names of the topics within the 17-topic model, emphasizing its 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://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
2. https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/
3. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
4. https://towardsdatascience.com/topic-model-visualization-using-pyldavis-fecd7c18fbf6
5. https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13