Topic Modeling to study how terrorist organizations are portrayed in traditional media outlets

CIS 6397 -Text Mining

Mini-Project 3

Authors: Mukesh Reddy Mavurapu, Shiva Sai Ram Prasad Reddy Yelipeddi, Nikhita Peddi

GitHub Link Repo

Abstract: In this research, we delve into the initial processing and analytical examination of a dataset composed of journalistic writings. The endeavor begins by assembling the dataset, delineating the bulk data into separate articles. This dataset is then refined by detaching supplementary information from the main content and identifying key textual elements. Brief overviews and illustrative diagrams are crafted to deepen the analysis. The essence of this investigation is to establish and refine a thematic framework derived from the cleansed dataset, with numerous iterations and adjustments of parameters to compile comprehensive summaries in output files. This discussion assesses these outcomes within the scope of the project's main narrative, leveraging select excerpts from the thematic framework and graphical interpretations to substantiate the conclusions.

I. INTRODUCTION

The surge in online content has flooded the digital landscape with a vast amount of unstructured information, including news pieces. This report tackles the task of text mining within this context, aiming to unearth valuable findings for both journalists and scholars. Our investigation concentrates on the initial stages of data treatment and the preliminary analysis of a collection of news stories, with the objective of establishing a topic model that reveals the themes within these articles.

Commencing with the assembly of our data set, we process the raw text by dividing it into separate articles, while also removing any extraneous metadata. The subsequent phase involves the critical task of preprocessing, where the text is broken down into manageable parts and refined to exclude elements like punctuation and miscellaneous tags. With the refined data, we craft a topic model, fine-tuning the parameters and iterating the process to refine our summaries. These steps are repeated as necessary, with the results and visual evidence supporting the investigative narrative.

II. METHODOLOGY

1. **Corpus Building:** In the data acquisition phase of 2017 research, we accessed a broad collection of journalistic content from the Factiva Global News database. This rich repository included contributions from the Wall Street Journal and The New York Times. We initiated the process by systematically retrieving and splitting the data into .txt files. The segmentation was guided by distinct identifiers such as 'Document NYTF', 'Document INHT', 'Document WSJ', 'Document J000', and 'Document AWSJ', culminating in an initial set of articles.

2. **Data Cleaning of Corpus**: The subsequent phase was dedicated to corpus curation. Our scrutiny was directed at ensuring the articles adhered to a standardized format and that the extraction process was comprehensive. We employed a selection criterion based on the termination of articles with specific copyright declarations, such as "All Rights Reserved", "Japanese Copyright" leading to the retention or exclusion of articles accordingly. Through this process of metadata extraction and content delineation, our corpus was refined to a collection of 1618 articles, each primed for in-depth analytical pursuits.

3. **Data Pre-Processing:** In the pre-processing stage, we undertook several cleaning operations to ensure data uniformity. Punctuation was stripped away, and alphabetic characters were converted to a lower-case state. We then proceeded to tokenize the corpus, breaking down the text into individual word units. Utilizing the stopword list from the NLTK library, we filtered out the linguistic noise, eliminating common but analytically irrelevant words. Subsequently, we applied the Porter Stemming technique, reducing words to their base or root form, thus streamlining the dataset for advanced language processing and examination.

4. **Features Extraction:** In the feature extraction stage of our text analysis procedure, we advanced by creating a document-term matrix to encapsulate the textual data's significant features. We employed a TF-IDF vectorizer, which not only counts word occurrences but also weighs them against their frequency across the entire corpus, thereby highlighting both common and distinctive terms. This quantitative transformation of the corpus facilitated the identification of key terms, which were then graphically represented through a word cloud for immediate visual interpretation. Further analysis was conducted to pinpoint the most frequently occurring terms, distilling the essence of the corpus, and revealing the dominant themes embedded within the data.

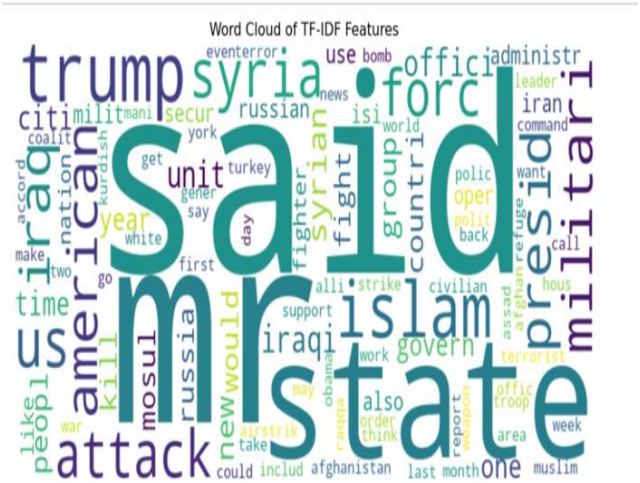
5. **Topic modeling:** Advancing our text analysis framework, we employed Gensim to facilitate topic modeling. Our process began with the formation of bi-gram and tri-gram models, which allowed us to identify and utilize word patterns extending beyond single-term analysis. Next, we conducted lemmatization to refine the words to their canonical forms, ensuring that the variations due to grammatical inflections do not skew the thematic analysis. With these refined models, we then proceeded to construct a Latent Dirichlet Allocation (LDA) model via Gensim, inputting our carefully preprocessed corpus and delineating the number of distinct topics to be extracted for a granular understanding of the textual landscape.

6. **Evaluation of the results:** Perplexity scores gauge the model's capacity to predict new data, with lower scores indicating greater predictive accuracy. In terms of semantic coherence, which is assessed by coherence scores, higher figures suggest more logically consistent topics. Our model's general performance was robust, exhibiting perplexity scores between -9.582998 and -11.253261, denoting a relatively high accuracy in new data prediction. However, coherence scores

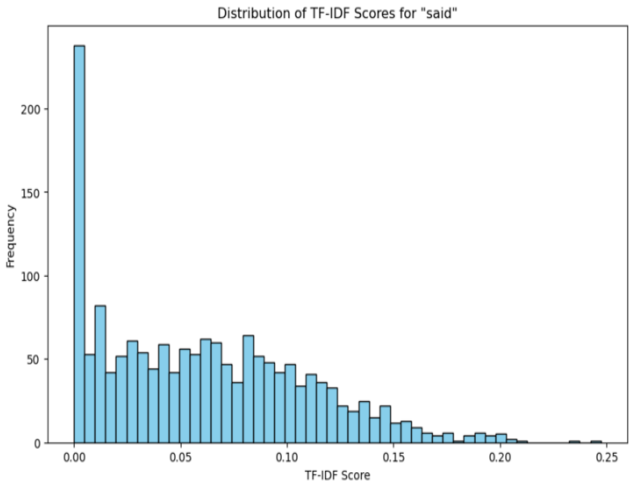
were on the lower side, from 0.407365 to 0.489245, hinting at possible enhancements in topic coherence. Optimal results were obtained with a 16-topic model, which yielded a perplexity score of -9.696508 and a coherence score of 0.489245, illustrating that the model successfully generated topics that were both semantically coherent and predictive. It's noteworthy that increasing the number of topics doesn't necessarily equate to better performance, reflecting the delicate interplay between model intricacy and its functional efficacy.

III. EXPERIMENTAL RESULTS

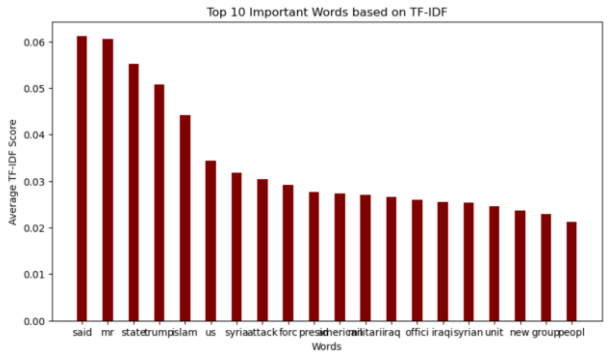
Word Cloud of TF-IDF Features:



Distribution of TF-IDF Scores for the word "said":



Top 20 frequent words based on TF-IDF



Determining Optimal Topic Model Complexity:

Upon evaluating our topic construction, we identified that. 16 topics provided the best balance for our analytical model.

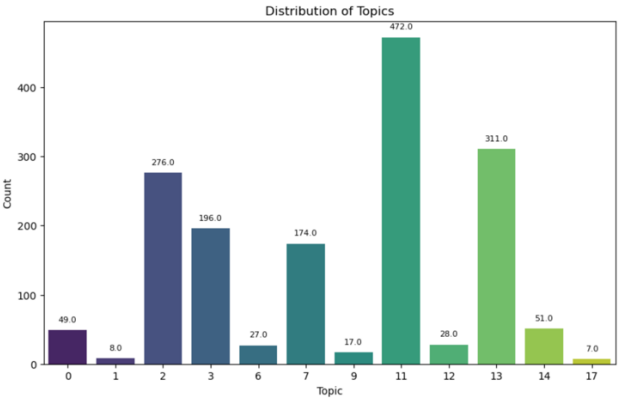
	Topics	coherence_values	perplexity_value
0	15	0.457009	-9.582998
1	16	0.489245	-9.696508
2	17	0.459985	-9.834827
3	18	0.435964	-9.949514
4	19	0.421342	-10.074754
5	20	0.448453	-10.185260
6	21	0.471243	-10.299457
7	22	0.463671	-10.420741
8	23	0.474031	-10.545910
9	24	0.448122	-10.667059
10	25	0.460018	-10.772767
11	26	0.433633	-10.910795
12	27	0.443289	-11.008675
13	28	0.465885	-11.144233
14	29	0.407365	-11.253261

Modeling Strategy and Thought Process:

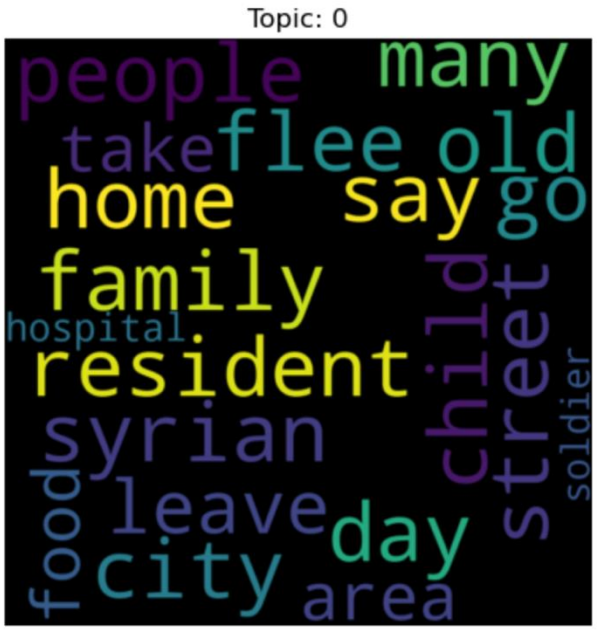
In crafting our model, we prioritized blending analytical insights with our understanding of the content, which predominantly features discussions on terror incidents. Recognizing that themes such as governance, terrorism, and national security often intersect, we extended our topic range from 15 to 30 during evaluation. Given the suboptimal coherence scores, we ultimately chose to construct a model with 16 topics.

Visualization and Inference

The bar chart illustrates the frequency distribution of topics. Notably, Topic 13 predominates with a count of 472, suggesting a high prevalence or interest. In contrast, Topic 17 is least represented with a count of 7, indicating its rarity or lesser focus. The chart shows a significant variation in counts across different topics, reflecting diverse levels of engagement or occurrence.

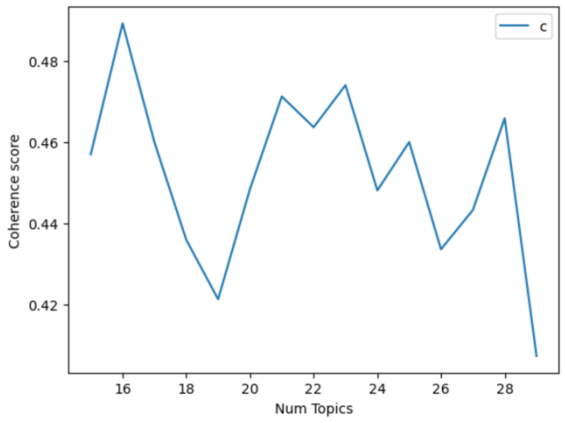


The word cloud is for Topic 0 from an LDA model with 20 topics. The word cloud visually emphasizes the most common words associated with the topic, with "people," "family," "home," and "hospital" being most prominent, suggesting a focus on social and healthcare aspects. Terms like "Syrian" and "soldier" hint at a possible context related to Syrian affairs or conflicts. This visualization aids in interpreting the key themes of the textual data analyzed.

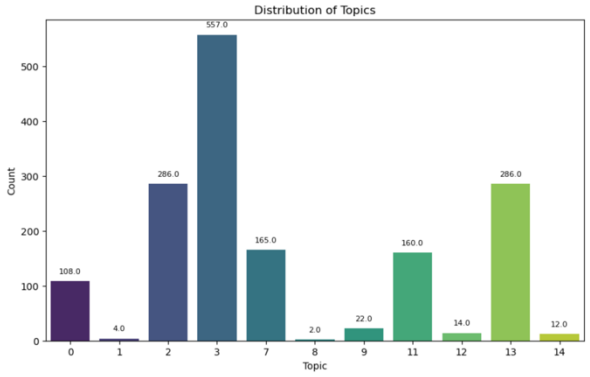


The word cloud is visualization for 'Topic: 0', from a text analysis using Latent Dirichlet Allocation (LDA). This graphical representation highlights the most prominent words associated with the topic, where the size of each word corresponds to its frequency or importance. Words like "war," "show," "expansionism," and "sarin" feature prominently, suggesting they are key terms within the topic.

The line graph depicting coherence scores as a function of the number of topics in a topic modeling analysis. The coherence scores range from about 0.42 to 0.48, and the graph shows that the highest coherence score is achieved when the number of topics is set to 16. This peak suggests that 16 is the optimal number of topics for the model to achieve the best interpretability or separation between the topics.



The bar chart titled "Distribution of Topics" created using Python's seaborn library. It showcases the frequency distribution of different topics, with Topic 3 having the highest count at 557, indicating a significant emphasis on this theme in the analyzed dataset. The chart underscores the disparities in topic occurrence, highlighting variations in relevance.



IV. CONCLUSION

This project highlights the critical role that initial data processing and exploratory analysis play in evaluating newspaper content, culminating in topic modeling. By developing a corpus, refining it, and pinpointing key features, we created concise summaries and insightful visual aids for a deeper understanding of the material. Topic modeling proved invaluable in isolating core themes within the corpus, offering journalists tools for deeper data insight to refine their storytelling. The findings reinforce the necessity of meticulous data preparation in journalism. Despite the intensive effort required in data processing and analysis, the resulting depth of understanding can significantly empower journalists to craft stories with profound impact. Future research could benefit from excluding frequently used words and incorporating novel analytic methods or applying this established methodology to varied data sets for comparative analysis.

AUTHOR CONTRIBUTIONS

Mavurapu Mukesh Reddy:

Concept Formulation, Approach, Development, Verification, Examination and Refinement, Oversight

Yelipeddi Shiva Sai Ram Prasad Reddy:

Detailed Examination and Research, Software, Crafting the Initial draft, Critical Review and Refinement,

Peddi Nikhita: Draft Reviewing, Creation of Visual Representations, Direction, Refining and Enhancing the Text.