Overview:

This research project focuses on gaining comprehensive insights into the emergence, evolution, and diagnosis of the SARS-CoV-2 virus by leveraging advanced text mining and natural language processing techniques. The dataset utilized for this analysis is the CORD-NER dataset, a machine-readable collection of over 29,000 coronavirus-related articles, with more than 13,000 having full text.

Data Preprocessing

Data Collection: Gathered an extensive dataset from the CDC-coordinated effort, providing a rich source of information on COVID-19.
Cleaning and Formatting: Conducted thorough data cleaning to ensure consistency and removed any irrelevant information.

Document Ranking: Implemented a ranking system to prioritize articles based on relevance to the SARS-CoV-2 genome.

Text Mining and NLP

Tokenization and Lemmatization

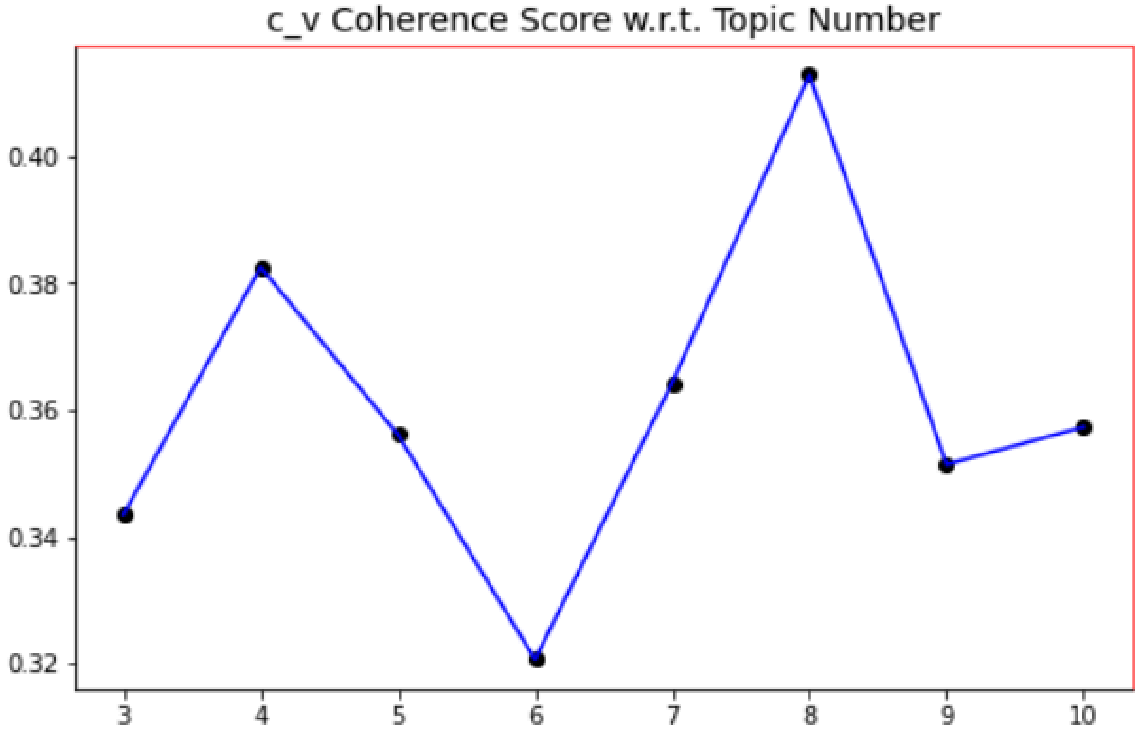
Tokenization: Utilized advanced tokenization techniques to break down documents into meaningful units.

Lemmatization: Applied lemmatization to reduce words to their base form, enhancing the quality of analysis.

Coherence Scores

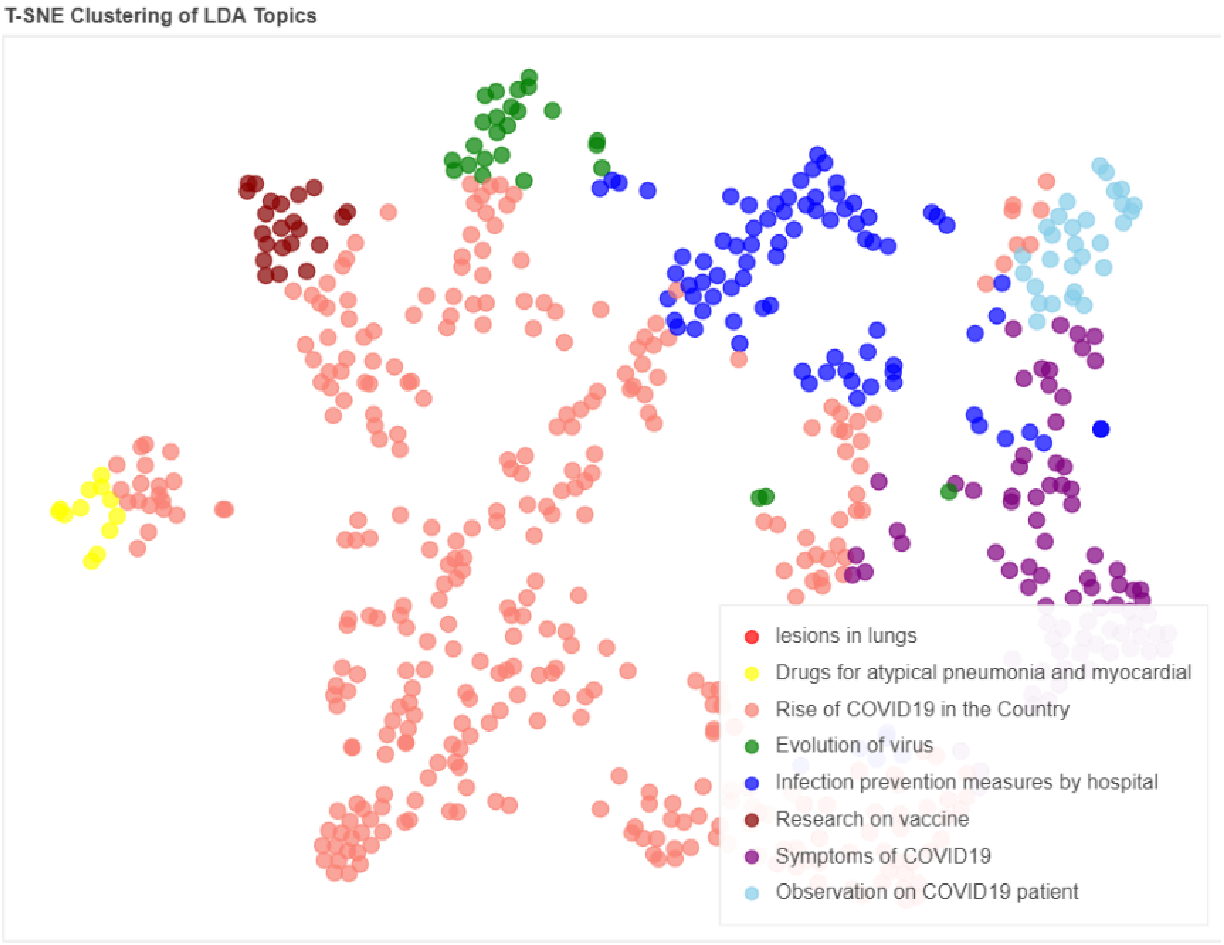
Topic Modeling: Employed topic modeling to extract meaningful topics from the dataset.

Coherence Scores: Calculated coherence scores to evaluate the interpretability and relevance of identified topics



Dimensionality Reduction

t-SNE: Implemented t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize high-dimensional data, facilitating a better understanding of document relationships.

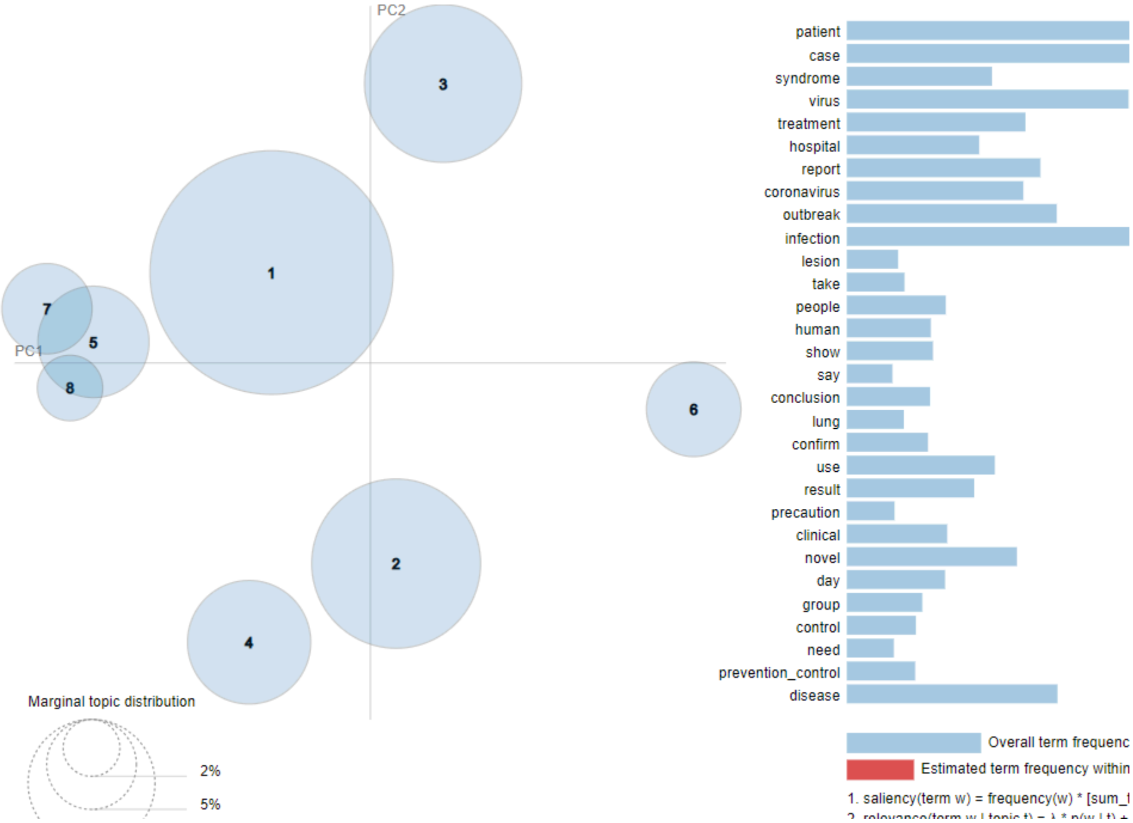


Word Embeddings

Word2Vec Embedding: Utilized Word2Vec embeddings to represent words in vector space, capturing semantic relationships.

Intertopic Graph

Intertopic Graph Analysis: Constructed a graph to visualize relationships and interactions between identified topics, providing a holistic view of the dataset.



Spacy Parser

Spacy Integration: Incorporated Spacy for advanced natural language processing tasks, enhancing document understanding.