# Exploration of topic modelling using Natural language processing techniques

# Introduction

Topic modeling is an NLP technique used to uncover hidden themes within a large set of documents. It operates under the assumption that every document is a blend of a finite number of topics and that each word in the document is associated with one of these topics.

This helps in organizing, understanding, and summarizing the information within the text data, making it easier to analyze and interpret large text corpora.

Topic modeling is used in the context of demonetization in India. The stated objectives were to combat black money, reduce counterfeit currency, and promote a shift towards a cashless economy. This decision had far-reaching impacts on the economy and the daily lives of millions of people.

Topic modeling in the context mentioned, carries the following relevance:

* **Capturing Reactions:** The demonetization event generated a vast amount of public discourse across various platforms, including social media, news articles, blogs, and forums. Topic modeling can help analyze this discourse to understand the public's reaction, sentiment, and opinions about the policy.
* **Identifying Key Concerns:** By applying topic modeling to the text data related to demonetization, the main concerns and issues raised by the public, such as cash shortages, impacts on businesses, and the effectiveness of the policy in curbing black money can be identified.

# Methodology

The methodology for topic modeling can be described through the figure below:

A diagram of a process

Description automatically generated

# Data Collection

The source of the dataset is Kaggle [1]. The dataset Contains 14940 records. Following are the primary reasons for the selection of the dataset:

* With a large volume of tweets collected over the period of demonetization, the dataset offers a substantial amount of text data that is ideal for performing topic modeling. The richness of the data allows for the extraction of diverse topics and trends.
* Being publicly available on Kaggle, the dataset is easily accessible for research purposes. This ensures transparency and reproducibility of the analysis.
* The data is preprocessed to a certain extent and anonymized, which is essential for maintaining the privacy of individuals while allowing researchers to focus on the content of the tweets.
* The dataset contains a wide range of reactions, opinions, and discussions, making it suitable for topic modeling to uncover hidden thematic structures and sentiments. This can provide deep insights into the societal impact of demonetization.

# Data Preprocessing

Preprocessing in machine learning refers to the process of transforming raw data into a clean and understandable format. This step is crucial because the quality and relevance of the data directly impact the performance of the machine learning models. Following pre-processing steps are carried out for topic modeling:

* Removal of Mentions, Hashtags, URLs, and Special Characters: These elements are not useful for topic modeling as they do not contribute meaningful content and could introduce noise.
* Tokenization and Lowercasing: This step splits the text into individual words (tokens) and converts all characters to lowercase.
* Removal of Stop Words: This step removes common words that are generally not useful for topic modeling (e.g., "and", "the", "is"). Stop words do not carry significant meaning and can dilute the importance of more meaningful words.
* Lemmatization: This step reduces words to their base or root form (e.g., "running" to "run"). Lemmatization helps in grouping together different forms of a word so they can be analyzed as a single item.

Also, the 5000 data are sampled from total data from topic modeling and evaluation. The sampling is done through random selection with the goal of reproducibility.

# Topic Modeling with LDA

Although many algorithms are available for topic modeling, LDA (Latently Dirichlet Allocation) is preferred in this analysis. LDA's reign in topic modeling can be attributed to its well-rounded capabilities. Unlike some competitors, it offers a strong mathematical foundation based on probability, ensuring reliable topic discovery. This probabilistic approach also translates to clear and understandable topics, each identified by the words most likely associated with it. Additionally, LDA scales well, making it adept at handling the large datasets that are commonplace in areas like social media analysis. This popularity is further bolstered by the wealth of research surrounding LDA and the ease of use provided by numerous tools available for various programming languages. Finally, LDA serves as a standard for evaluating newer topic modeling algorithms. While it does have limitations, such as assuming words are independent and requiring careful selection of the number of topics, LDA's strengths often make it the preferred choice in topic modeling.

LDA requires the few parameters to be set during its implementation of modeling. The parameters set were:

|  |  |
| --- | --- |
| **Parameters** | **Value** |
| num\_topics | 10 |
| random\_state | 42 |
| update\_every | 1 |
| chunksize | 100 |
| Passes | 50 |
| alpha | ‘auto’ |
| per\_word\_topics | True |

**Gensim** (<https://github.com/piskvorky/gensim>) library is used for the purpose as it has large community support and is actively maintained. The algorithms in Gensim are memory-independent w.r.t. the corpus size (can process input larger than RAM, streamed, out-of-core).

# Model Evaluation

The problem of topic modeling is unsupervised. Unlike supervised learning, which uses labeled data to predict outcomes or classify data points, unsupervised learning aims to find hidden patterns or intrinsic structures within the data.

Perplexity and Coherence are the metrics used to evaluate the model.

The results of the model evaluation metrics are

|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Perplexity | -6.609823647541917 |
| Coherence | 12.866791517489753 |

Perplexity is a measure of how well the model predicts a sample. Lower perplexity indicates a better model. However, perplexity alone may not fully capture topic quality.

The perplexity value shown above means the model is relatively good at predicting the words in the corpus. Lower (more negative) values indicate better performance.

Coherence score measures the semantic similarity of words in each topic. A coherence score measures how interpretable and meaningful the topics are, based on how frequently words co-occur in the documents. Higher coherence scores indicate more interpretable and coherent topics.

A coherence score shown above suggests a certain level of interpretability and meaningfulness in the topics generated by the model.

# Results, Visualization, and Interpretation

The results of the analysis are visualized through heat maps, word clouds and interactive html visualization.

Seaborn is used for heatmap visualization. Seaborn is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. The result of the visualization is shown below

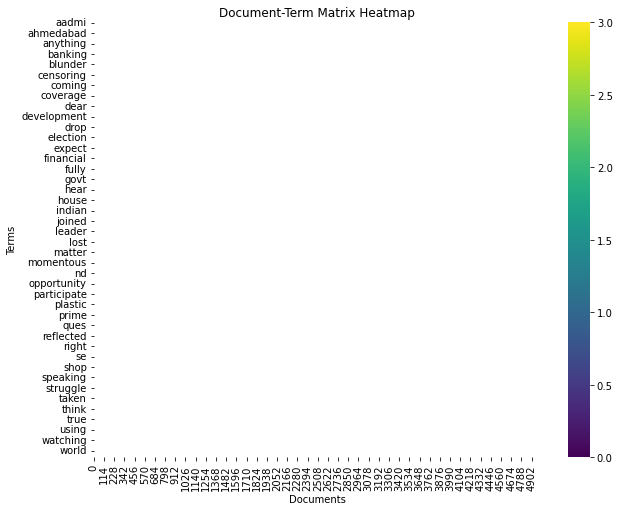


Fig 1: Heatmap

The results of visualization in the form of word cloud was done using the library ‘WordCloud’. The results of the visualization are

A close-up of words

Description automatically generated

Fig 2 : Word Cloud Visualization of TOPIC 1

A close-up of words

Description automatically generated

Fig 3 : Word Cloud Visualization of TOPIC 2

A close-up of words

Description automatically generated

Fig 4 : Word Cloud Visualization of TOPIC 3

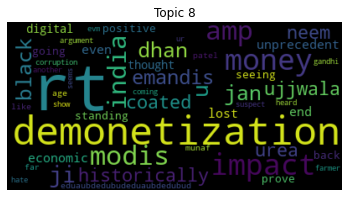


Fig 5 : Word Cloud Visualization of TOPIC 8

A close-up of words

Description automatically generated

Fig 6 : Word Cloud Visualization of TOPIC 9

Also, the visualization was done with ‘pyLDAvis’. It is a useful library that provides an interactive web-based visualization that allows users to explore the relationships between topics, the prevalence of topics in the corpus, and the terms associated with each topic. The results of the visualization are:

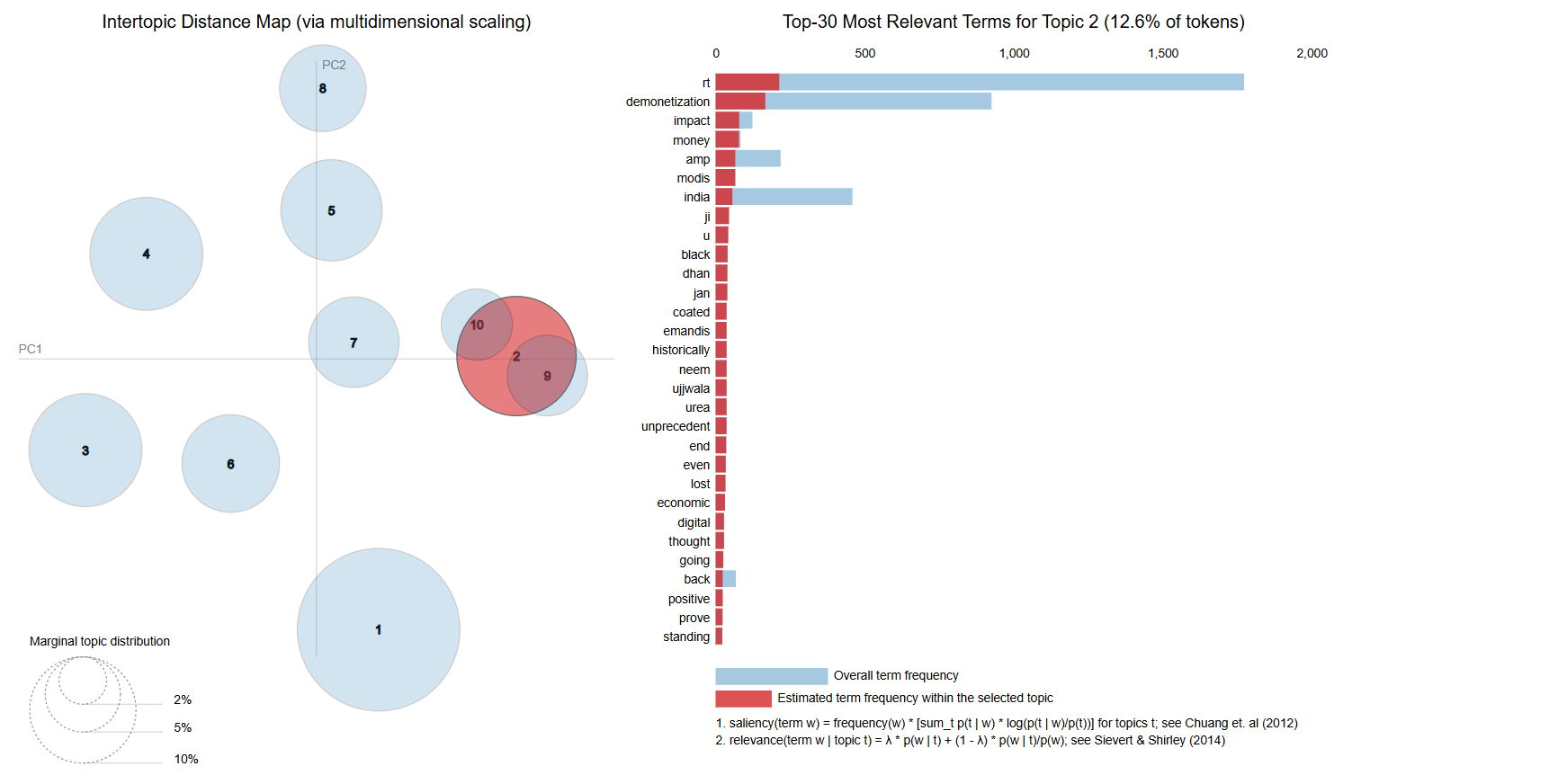


Fig 7 : pyLDAvis visualization 1

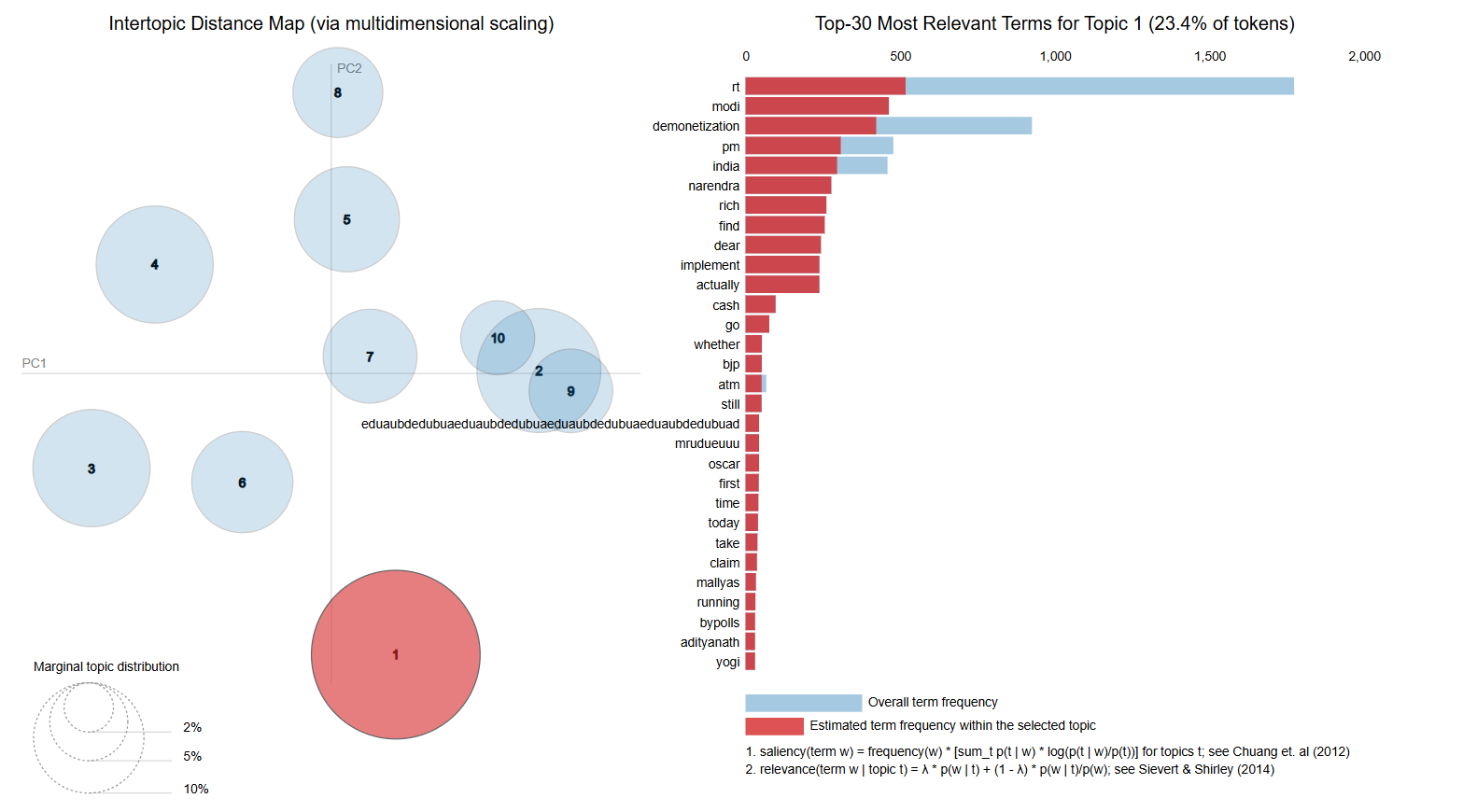


Fig 8 : pyLDAvis visualization 2

A screenshot of a computer screen

Description automatically generated

Fig 9 : pyLDAvis visualization 3

From the analysis of the above shown visualization results, it is observed that most of the people are in support of the demonetization act in India.

While some of the people think it as a scam (as seen in topic 2), many are in support of Modi, PM of India and the goverment(topic 3, topic 5, topic 8, topic 9).

People also believe that the demonetization of currency can restrict and put an end to existing fake currencies (topic 6, topic 8).

# Conclusion

Demonetization in India refers to the government's decision on November 8, 2016, to render the 500 and 1000 rupee notes invalid, which made up 86% of the cash in circulation. Prime Minister Narendra Modi announced the move with the objectives of combating black money, counterfeit currency, and corruption. Citizens were allowed until the end of December 2016 to deposit the invalidated notes into banks and exchange them for new 500 and 2000 rupee notes [2], [3].

Finding out public reaction in this context can provide useful insights. Hence, Topic Modeling was done on the dataset available in Twitter using LDA.

People mostly were in support of the government’s decision to demonetization. Although there were mixed reactions, some against the decision , but majority of the people welcomed the decision. They believed that the demonetization of currency can restrict and put an end to existing fake currencies and also the black currencies.

Since, the tweets are mostly from Indians, some of the stop words that are available in Indian languages are seen frequently in visualized report. Such stop words can be omitted during pre-processing of data in future work.

# References

[1] A. Rathee, "Demonetization in India Twitter Data," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/arathee2/demonetization-in-india-twitter-data. [Accessed: May 21, 2024].

[1] A. Shukla and R. Kumar, "Modi Government’s Demonetisation Move - Objectives and Impact," The Economic Times, 2016. [Online]. Available: https://economictimes.indiatimes.com/news/politics-and-nation/demonetisation-how-the-numbers-stack-up/articleshow/56495722.cms. [Accessed: 29-May-2024].

[2] P. Mukherjee, "Demonetisation in India: Impact and Aftermath," The Hindu Business Line, 2016. [Online]. Available: https://www.thehindubusinessline.com/opinion/demonetisation-in-india-impact-and-aftermath/article9395448.ece. [Accessed: 29-May-2024]