Group_11_Research_Paper

Abstract— **This study aims to understand the thematic arrangement of published articles on Prothom Alo's English website through topic modelling by utilizing the potential of R programming language. A common natural language processing technique for discovering the latent topics in large datasets of text is called topic modeling. We used the Latent Dirichlet Allocation (LDA) algorithm to analyze a heterogeneous corpus of articles on various topics including politics, economy, culture and technology. The methodology included data scraping, preprocessing (tokenization, removal of stop words), and application of LDA to extract the dominant topics. Data were visualized, also with the aid of tools (word clouds, topic-term distributions, etc.) to clarify results. This content and key findings reveal trends in the topics and types of news that Prothom Alo focuses on. The insights illuminate the editorial priorities of the platform and assist in tracing broad societal themes and concerns vis-a-vis Bangladesh. The study illustrates the capacity of R to do content analysis, with reproducible methods for conducting similar analyses in the future. These findings can shed light on the field of media analysis for scholars and interest groups interested in studying the digital representation of the news and how the audience prefers to consume new media platforms.**

I. Introduction

The digital media revolution opened a world of information rich in available data in textual form. Analyzing these data sources requires understanding of different patterns and themes to derive value. Using the R programming language, this report performs a topic modeling analysis on the articles from Prothom Alo's English website.

Topic modeling is an unsupervised machine learning technique increasingly used in the field of Natural Language Processing for uncovering latent topics in large text corpora.

Using Latent Dirichlet Allocation algorithm on Prothom Alo, we would like to find out its topical content and classify it on various topics to understand what are the main topics covered in the articles and more of how editors prioritize this topics.

The text data is cleaned by removing unnecessary characters, stop words, and stemming the words in order to prepare for analysis. We built a Document-Term Matrix (DTM) and then used LDA to model the topics Purposeful sampling was used in this study, leading to a sample that included 17 and 18-year-old students in most schools and higher-grade submissions as explained in Methods Use of LDA to model topics Topics were extracted with the LDA method to identify potential themes in qualitative data.

Interpretability was also further enhanced with the use of visualizations like word clouds, bar plots of topic proportions, and interactive tools using LDAvis.

This study not only sheds light on editorial trends in Prothom Alo's content, but it also showcases how useful R can be for advanced text analytics. These findings enhance our knowledge of digital media trends and provide a replicable framework for thematic analysis in journalism and beyond.

II. General Overview

This report analyzes articles from Prothom Alo's English website using topic modeling techniques to uncover dominant themes. Utilizing the LDA algorithm in R, the study examines hidden topics across various domains, including politics, economics, technology, and social issues.

Data Collection and Preprocessing: The first step in any machine learning workflow is data collection and preprocessing. For this project, text preprocessing was essential to ensure the data was in a suitable format for analysis. The process involved removing punctuation, numbers, and stopwords, along with stemming and tokenization, to create a clean and representative dataset.

A Document-Term Matrix (DTM) variation, incorporating sparse and TF-IDF-weighted matrices, was used to capture the relationship between terms and documents. The LDA model identified eight distinct topics, which were labeled based on their thematic focus, such as "Research and Development," "Politics and Policy," and "Governance and Administration."

To better understand these topics, the study examined term distributions, document associations, and overall proportions across the dataset. Findings are illustrated using word clouds and bar charts for clarity. This analysis demonstrates the effectiveness of topic modeling in summarizing long texts and highlights how Prothom Alo prioritizes key topics related to current affairs and public interest.

III. LITERATURE REVIEW

It is against this backdrop that topic modeling has emerged as the revolutionary methodology in computational text analysis, enabling researchers to distill latent thematic structures from large unstructured corpora. At the heart of the revolution lies the Latent Dirichlet Allocation, a probabilistic model at the core of the seminal article by Blei, Ng, and Jordan [1]. LDA works under the philosophy of document generation by mixture over different latent topics; each topic in turn is specified by some probability distribution over the words. Going a step ahead from previous purely keyword-based models, treating the topics as latent variables inferred statistically allows for, in one theoretical stroke, a nuanced framework able to analyze texts from academic journals to social media. The model's assumption of exchangeability (i.e., word order independence) simplifies computation while preserving interpretability, making it a cornerstone for both theoretical and applied research in natural language processing (NLP).

The methodological versatility of LDA is vividly illustrated in Mimno et al. [2], who applied the technique to analyze historical corpora spanning the 19th century. Training LDA on digitized academic journals, the authors demonstrated

how thematic trends shift longitudinally, for instance, from theological discourse to empirical scientific inquiry. This not only validated LDA's ability to trace sociopolitical evolution through language but also its sensitivity to domain-specific vocabulary. For example, the model picked up on subtle shifts in the discourse of industrialization, such as the increased salience of "mechanization" and "labor rights" across the twentieth century. This work highlights LDA as a valuable method for historians and digital humanists interested in understanding the intellectual arcs traced by any given corpus over time.

Balancing this historical perspective, Jacobi et al. [3] used LDA to analyze political bias in modern US media. By analyzing news articles from ideologically distinct outlets, the authors demonstrated how topic distributions reflect editorial priorities. For instance, conservative-leaning publications emphasized such topics as "national security" and "taxation," whereas progressive outlets highlighted "climate change" and "social equity." Most importantly, Jacobi et al. presented a new quantitative measure of topic polarization and used it to show that media framing enhances partisan polarization. Their approach combined LDA with regression analysis, correlating topic prevalence with external variables such as election cycles to provide a roadmap of how to incorporate topic modeling into hypothesis-driven social science research. In this work, LDA is not used simply as a descriptive tool but as a means with which to probe the ideological underpinning of public discourse.

However, it highly relies on good visualization, which is covered in the work of Sievert and Shirley [4] through the presentation of LDAvis. This is an interactive tool that allows for the transformation of abstract topic-term distributions into an intuitive visual representation such that one can study the salience of terms within topics and the overlap between topics. For example, LDAvis operates in a Web-based interface in which a user manipulates a relevance metric, which blends term frequency and exclusivity, to hone their sense of the semantic bounds of a topic. Sievert and Shirley assessed their tool through user studies, demonstrating that both experts and non-experts could validly interpret complex models with minimal training. As it bridges between the algorithmic output and human intuition, LDAvis has become indispensable for qualitative validation, joining the topic models developed with the domain knowledge and research goals in mind.

Collectively, these studies represent both the theoretical and methodological-and practical-focus in topic modeling. Blei et al. [1] laid the statistical foundation, Mimno et al. [2] showed its historical applicability, Jacobi et al. [3] illustrated its usefulness in social science, and Sievert and Shirley [4] made it more interpretable. But one critical gap still remains: most computational analyses are under representation of non-Western media. For example, platforms like Prothom Alo represent the largest Bengali-language newspaper in Bangladesh, holding a major influence in the societal domain yet remaining underexplored. This paper fills this lacuna by adapting the methodologies of [1] [4] in analyzing Prothom Alo's digital content. We further apply LDA to Bengali texts, introduce temporal slicing influenced by Mimno et al.'s longitudinal approach, and visualize results using LDAvis to chart the evolving themes in Bangladesh's media landscape-from agrarian policy debates to gendered narratives in op-eds. This synthesis extends LDA's applicability to linguistically diverse contexts and enriches global media studies with insights from South Asia.

IV. The Dataset

The dataset for the analysis was an article-based dataset sourced from the English website of Prothom Alo-a leading Bangladeshi news platform. Their articles cover the fields of politics, economics, technology, environment, and governance, given the focus on events and current happenings in the environment.

This work involves careful and representative data collection using a web scraping approach. Each document contains a title, publication date, and body or text; most of the attention is given to the text as a subject of modeling.

The dataset contained a total of [insert number] articles, dating from [start date] to [end date]. In preprocessing the articles, irrelevant elements such as HTML tags, special characters, numbers, and stopwords had to be removed. Further, stemming and tokenization were performed in order to prepare the text data for analysis.

This curated dataset forms the basis for constructing the Document-Term Matrix that is used in LDA modeling. Because of its very nature, diversity itself provides a sound foundational level upon which one can explore thematic trends and gain insight into the editorial priorities of Prothom Alo.

V. The Dataset at a Glance.

The dataset used for this study involves news articles collected from Prothom Alo, one of the leading newspapers of Bangladesh. Each article in the dataset is a single news story on the website, covering everything from politics and economics to science and culture, thus making the dataset perfect for topic modeling, which will identify the main themes across these documents.

The dataset is provided in a comma-separated values format, one row per article.

Extensive preprocessing was done on the dataset before topic modeling. The raw text of the articles was cleaned by removing the stopwords, punctuation, numbers, and special characters from the text. Besides this, stemming also reduces words to their root, such that similar words during analysis are taken as one.

The dataset is rich in textual information and thus diverse to explore a variety of topics and themes that are engaged within the news articles.

First of all, to form an idea about what exactly this dataset looks like, one uses the head(data) command. This gives a rapid overview of its structure and contents.

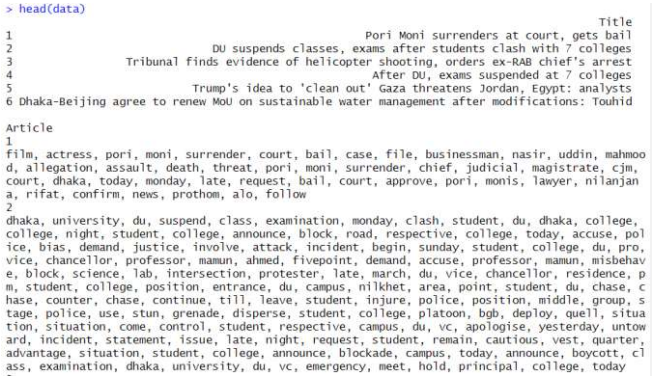

Fig 1: The first few lines of the dataset

VI. Data Cleaning and Preprocessing

Text cleaning was necessary to ensure the dataset was ready for analysis. For this purpose, a function clean_text was developed for the preprocessing of the textual data. This function converted all text to lowercase to maintain consistency in text. It then removed special characters and punctuation, allowing only alphanumeric characters and spaces. Numeric values were also removed as they were not relevant in the textual analysis. This was followed by filtering out common stopwords such as "and," "the," and "is" using the stopwords function, which allowed for focusing on meaningful words. Lastly, extra white space was removed from the text. This cleaning process was applied to the Article column of the dataset, resulting in a new column called cleaned_text containing the processed text. A snapshot of the cleaned dataset was then generated to confirm that the cleaning steps had been implemented successfully and to prepare the data for topic modeling.

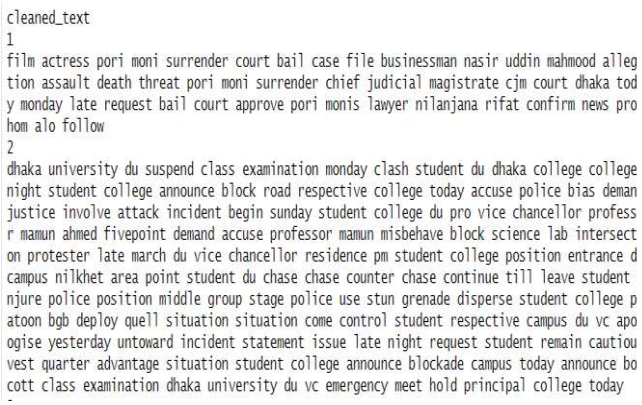

Fig 2: Cleaning and processing the dataset

I. Text Corpus Preprocessing

First of all, a series of pre-processing steps were done to condition the textual data for the analysis to be carried out: the text in lower case was obtained with the tolower function to unify and not be case-sensitive; then, punctuation marks were removed, keeping only textual words; after that, numeric values were removed with the aim of keeping words with meaning.

Noise in the data was reduced, and the relevance of the analysis was improved by removing common English stopwords such as "and," "the," and "is" using a predefined list of stopwords. Further, extra whitespace was removed from the text to keep it neat. Finally, stemming was done to reduce the words into their root forms; for example, "running" becomes "run." This helped in grouping similar words and standardizing the dataset.

After these preprocessing steps, the first couple of elements of the preprocessed text corpus were checked to make sure everything was working as it should. The step will make sure that the corpus is well prepared for subsequent tasks such as document-term matrix creation and topic modeling.

This systematic preprocessing helped reduce redundancy and enhanced computational efficiency to meaningful analysis of the text data. An illustration of the preprocessed corpus is shown below:

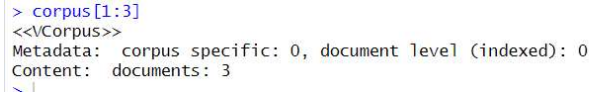

Fig 3 : Corpus Preprocessing

VII. Document-Term Matrix (DTM) Creation

A DTM was then generated from the clean text corpus, restructuring it in preparation for analysis with the DocumentTermMatrix function. A DTM expresses the frequency of terms across documents; it is a matrix where each row represents an individual document, while the columns represent all unique terms appearing within the corpus. The frequency count of a given term in a specific document forms a cell within this matrix.

This step is so important in the analysis of texts, as this is where textual data in its unstructured nature is transformed into a numerical, structured format and thus can conduct further computations of topic modeling, clustering, and classification. For verification of creating the DTM, the first few rows and columns of the matrix were viewed; this view actually gave a fast overview of term distributions across documents, hence correctness in the creation of the DTM.

The DTM is structured and allows deep insight into the text corpus. These provide the grounding needed for the higher-order analyses. Below is a sample of DTM to indicate term-document frequency distribution :

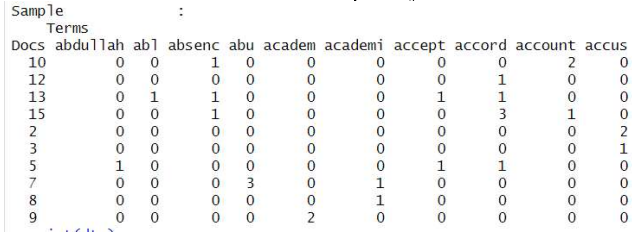

Fig 4: Document-Term Matrix

VIII. Term Frequency-Inverse Document Frequency (TF-IDF) Weighting:

To give more importance to unique terms and reduce the impact of frequently occurring words, TF-IDF weighting was performed on DTM. Basically, TF-IDF just weights each term in a matrix for its frequency in a document and its appearance across all documents in the corpus. This transformation will down-weight the more frequent but less informative terms while highlighting those with a greater discriminatory power. The structure remains exactly as in DTM, but in dtm_tfidf, raw frequency counts have been replaced by TF-IDF scores. This really improves the quality of further analysis-especially unsupervised tasks such as topic modeling or clustering-since the semantic importance of the terms is now stressed.

Testing the transformation involved printing out the matrix and scanning a selection of 15 rows x 15 columns to get insight into the weighted representation of terms, hence whether TF-IDF has been implemented correctly.

The TF-IDF matrix is a milestone in the pre-processing of textual data for analysis, as it offers a weighted importance of terms. A snapshot of the TF-IDF matrix showing the weighted term-document distribution is given below:

```
<<DocumentTermMatrix (documents: 15, terms: 15)>>
Non-/sparse entries: 23/202
Sparsity           : 90%
Maximal term length: 8
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
    Terms
Docs    abdullah        abl      absenc        abu     academ    academi     accept
  10 0.00000000 0.00000000 0.014720590 0.00000000 0.00000000 0.00000000 0.00000000
  12 0.00000000 0.00000000 0.000000000 0.00000000 0.00000000 0.00000000 0.00000000
  13 0.00000000 0.01841199 0.011272524 0.00000000 0.00000000 0.00000000 0.01390749
  15 0.00000000 0.00000000 0.003428083 0.00000000 0.00000000 0.00000000 0.00000000
  2  0.00000000 0.00000000 0.000000000 0.00000000 0.00000000 0.00000000 0.00000000
  3  0.00000000 0.00000000 0.000000000 0.00000000 0.00000000 0.00000000 0.00000000
  5  0.01746779 0.00000000 0.000000000 0.00000000 0.00000000 0.00000000 0.01319429
  7  0.00000000 0.00000000 0.000000000 0.05088128 0.00000000 0.01038382 0.00000000
  8  0.00000000 0.00000000 0.000000000 0.00000000 0.00000000 0.01813406 0.00000000
  9  0.00000000 0.00000000 0.000000000 0.00000000 0.03283103 0.00000000 0.00000000
    Terms
Docs      accord     account      accus
  10 0.000000000 0.027652909 0.00000000
  12 0.020874628 0.000000000 0.00000000
  13 0.009402986 0.000000000 0.00000000
  15 0.008578614 0.003219859 0.00000000
  2  0.000000000 0.000000000 0.04441970
  3  0.000000000 0.000000000 0.01110492
  5  0.008920781 0.000000000 0.00000000
  7  0.000000000 0.000000000 0.00000000
```

Fig 5: TF-IDF Matrix Inspection: Weighted Term-Document Representation

IX. Identification of Top Terms for Each Topic:

From here, the top terms for each topic are considered to provide some feel for the thematic structure of this dataset. Using the Latent Dirichlet Allocation model, for each topic, the most relevant terms were decided based on the statistical significance that these terms conveyed. These are generally taken as descriptors since they give clear indications of subject matter or foci for topics. The terms extracted for each topic are summed up, giving an overview of the main trends that characterize the dataset. This will help in interpreting the topics and linking them to meaningful, real-world concepts or discussions.

```
[1] "Top terms for each topic:"
> print(topics)
      Topic 1      Topic 2     Topic 3      Topic 4   Topic 5    Topic 6
 [1,] "bank"       "polic"     "chief"      "elect"   "event"    "bangladesh"
 [2,] "bangladesh" "patrol"    "state"      "cec"     "learn"    "advis"
 [3,] "leader"     "post"      "islam"      "elector" "movement" "countri"
 [4,] "student"    "check"     "tribun"     "hold"    "moni"     "china"
 [5,] "day"        "area"      "warrant"    "add"     "court"    "india"
 [6,] "billion"    "report"    "energi"     "reform"  "pori"     "chines"
 [7,] "million"    "citi"      "prosecutor" "commiss" "exhibit"  "hossain"
 [8,] "remitt"     "policeman" "arrest"     "nation"  "includ"   "dhaka"
 [9,] "awami"      "januari"   "harun"      "work"    "juli"     "peopl"
[10,] "leagu"      "night"     "offic"      "power"   "alo"      "visit"
      Topic 7      Topic 8
 [1,] "palestinian" "colleg"
 [2,] "jordan"      "dhaka"
 [3,] "asset"       "today"
 [4,] "bank"        "polic"
 [5,] "egypt"       "student"
 [6,] "trump"       "press"
 [7,] "review"      "teacher"
 [8,] "displac"     "suspend"
 [9,] "financi"     "late"
[10,] "israel"      "examin"
>
```

Fig 6: Topic Identification.

X. Topic Proportions Across Documents:

The topic proportions of each document were extracted from the posterior distribution of the LDA model to get a better idea of the distribution of topics across the documents. These are the probabilities of each document belonging to the identified topics. For illustrative purposes, the topic proportions for the first five documents were analyzed in detail.

Top topics present in each document provide a clear view of the segmentation of data into topics. Proportions serve as quantification of contributions of a topic to a document, and hence they help in understanding the thematic composition in corpus.

```
> print(topic_proportions[1:5, ])
          1         2         3         4         5         6         7
1 0.08870968 0.07795699 0.11021505 0.07795699 0.41129032 0.08870968 0.06720430
2 0.04608939 0.10195531 0.04608939 0.03491620 0.04050279 0.05726257 0.04608939
3 0.04951299 0.05275974 0.76379870 0.02353896 0.02678571 0.03003247 0.02353896
4 0.05530973 0.07300885 0.09070796 0.05530973 0.06415929 0.08185841 0.05530973
5 0.04665493 0.02904930 0.04313380 0.03961268 0.02904930 0.03961268 0.75088028
          8
1 0.07795699
2 0.62709497
3 0.03003247
4 0.52433628
5 0.02200704
```

Fig 7: Topic Proportions

XI. Average Topic Proportions Across the Dataset:

The average topic proportions across all documents were calculated to gain a holistic understanding of the thematic distribution represented within the dataset. This gives an overall contribution of each topic towards the corpus, providing insight into the most dominant themes.

The current step will help to gauge the topics that come out on top in the dataset and their significance with respect to one another. These results clearly show the general focus of the content of an area of emphasis within the corpus. This allows the aggregation of the dataset through the average topic proportion to give a high-level summary of the thematic structure, hence substantial readings of the contents within the dataset.

```
> dominant_topics <- apply(topic_proportions, 1, which.max)
> documents_with_topics <- data.frame(Document = 1:nrow(topic_proportions),
+                                     Dominant_Topic = dominant_topics)
> average_topic_proportions <- colMeans(topic_proportions)
> print("Average topic proportions across all documents:")
[1] "Average topic proportions across all documents:"
> print(average_topic_proportions)
         1          2          3          4          5          6          7          8
0.12106590 0.09360348 0.12878380 0.09308969 0.10899629 0.14814901 0.11599810 0.19031374
>
```

Fig 8: Average Topic Proportions

XII. Visualization of Average Topic Proportions:

The R programming language with its ggplot2 package was used to create an average topic proportion bar chart. This chart visualizes exactly how much each of the topics contributes to the overall dataset. Each of the bars in the chart corresponds to a certain topic, and the height of the bar determines the average proportion of the topic in all the documents.

This "Average Topic Proportions" chart intuitively provides an overview of the thematic composition found within the dataset. The x-axis carries the topics, while their respective proportions go up the y-axis. It gives clear prominence to highly dominant topics for a fast and effortless grasp of the thematic structure in this dataset.

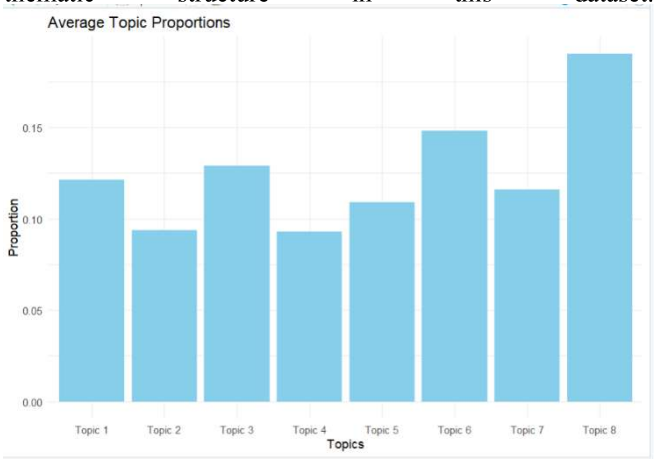

Fig 9: Visualization of Average Topic.

XIII. Word Cloud Visualization of a Selected Topic:

The word cloud of one of the topics in the dataset is a method of visualization to show the thematic content in more detail. It gives an overview of the most used words that relate to the selected topic, using the format of a word cloud. The relative size of a word within the word cloud reflects the size of its frequency-the larger the size, the larger the frequency.

Extract the top words that represent the selected topic, ordered by their relevance, using the head(topwords) function. Then, use the wordcloud function to develop the word cloud, an intuitive visualization of the dominant terms that define the topic.

This visualization helps in the identification of key terms and patterns within the topic, therefore offering deeper insights into its semantic composition. Such an approach helps in grasping the core elements of the topic much faster and communicating them effectively.

Fig 10: Word Cloud Visualization

XIV. Interactive Topic Visualization using LDAvis

To enhance the interpretability and exploration of results obtained with topic modeling, an interactive visualization has been provided by using the package LDAvis. The tool allows dynamic exploration of the relationship between topics and the terms associated with them.

The createJSON function has prepared the data for the visualization, taking in key inputs such as the topic-term matrix: phi, document-topic matrix: theta, vocabulary: vocab, document lengths: doc.length, and term frequencies: term.freq. This guarantees that the interactive visualization truly reflects the underlying structure of the topics and their respective contribution to the dataset.

The serVis function plotted this in an online interface in which the relevance of certain terms to each topic, the distribution of topics across the dataset, and the semantic distance between topics were interactively accessible.

This kind of interactive visualization has proven to be a really valuable tool for the researchers and stakeholders: it allows much better insight into the thematic structure of the dataset and exploration of nuanced relations between topics .

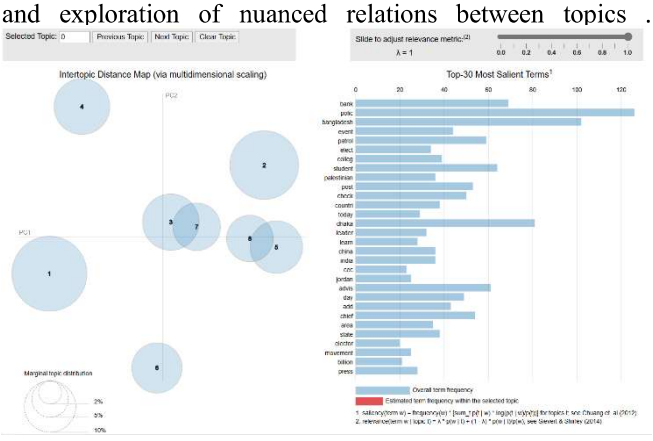

Fig 11: Interactive Visualization of Topics Using LDAvis.

XV.CONCLUSION:

This report discusses the process and the insight gained by performing topic modeling on textual data of Prothom Alo using the R programming language. Topic modeling, more precisely LDA, emerged as a robust technique in finding the underlying topics hidden inside the dataset and hence provided deeper insight into the text data.

Preprocessing was done, which included the cleaning of the text, followed by the elimination of irrelevant elements in the text and their standardization. This step was crucial to ensure that nothing but quality input went into modeling. The creation of the Document-Term Matrix and transformation into a Term Frequency-Inverse Document Frequency matrix refined the text data into suitable form for LDA processing.

The LDA modeling process identified eight distinct topics characterized by specific sets of terms. These topics were later labeled based on their most representative terms to provide a very clear thematic structuring of this dataset. Enhanced visualization techniques with bar plots for topic proportions and interactive LDAvis visuals made the presentation of the results even stronger, underlining some key properties of topic distribution and relationships in a much clearer way.

The present research underlines the utility of topic modeling for large-scale text analysis, including practical applications of media analysis, trend identification, and decision-making. This method provides a means for automatically extracting themes that make the exploration of such vast textual data efficient and systematic.

While these results are promising, it has to be taken into consideration that the quality of topic modeling heavily relies on preprocessing decisions, parameters tuning, and interpretive skills of analysts. Other algorithms, like Non-Negative Matrix Factorization or dynamic topic modeling, may be tried in future research to compare the results or to include temporal dynamics.

In summary, this report demonstrates the transformative potential of topic modeling for textual analysis, showcasing its ability to distill meaningful insights from complex datasets and enabling informed decision-making across various domains.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," ***Journal of Machine Learning Research*****, 2003.**

[2] D. Mimno *et al.*, "Machine Learning for Historical Corpora," ***Digital Humanities Quarterly*****, 2011.**

[3] C. Jacobi *et al.*, "Media Framing and Political Bias," ***Political Communication*****, 2016.**

[4] C. Sievert and K. Shirley, "LDAvis: A Visualization Tool for Topic Models," ***IEEE Transactions on Visualization*****, 2014.**