Analyzing Review on Education System by Topic Modelling

[1]Jannat Ara Tasnim
Department of Science and Technology
American International University-Bangladesh
Dhaka, Bangladesh
jannatara2399@gmail.com

[2]Nabiha Tahsin
Department of Science and Technology
American International University-Bangladesh
Dhaka, Bangladesh
tahsin.n611@gmail.com

Abstract— **Education plays a vital role in social development. It's also important for analyzing discussions and can provide valuable insights into prevailing keys, challenges as well as areas for improvement. This research employs Latent Dirichlet Allocation (LDA), a topic modeling technique to extract and categorize dominant topics from a collection of education-related texts. The study identifies four key topics, each reflecting different facets of educational discourse. Through visualizations we have to analyze the distribution of these topics across various documents. It reveals significant trends in education discourse. The findings highlight the diverse aspects of education from academic literature and institutional discussions to policy transformations. These insights can help educators, researchers, and policymakers make informed decisions regarding educational reforms and curriculum development. The primary objectives of this study are to uncover hidden patterns in education discourse, track the evolution of key themes over time, and assess how various stakeholders—such as educators, researchers, and policymakers—shape and respond to these discussions. By leveraging data-driven insights, this study aims to facilitate more evidence-based decision-making in educational reforms, curriculum development, and policy formulation.**

Keywords: *Topic Modelling, LDA, Education System, Textual Data, Preprocessing*

I. Introduction

Despite being a vital pillar of national growth, the education system often experiences frequent and inconsistent reforms that fall short of addressing its underlying issues. In Bangladesh, exam forms, curricular frameworks, and educational and learning techniques have all undergone several modifications sometimes as a result of temporary governmental changes rather than long-term strategic planning. Expected changes have not occurred despite the establishment of many education committees, leaving important problems including gender inclusion, job disparities among graduates, curriculum instability, and a lack of teacher autonomy unsolved. There has been a loop of experimenting without significant advancement due to the lack of an organized and evidence-based approach to policymaking. Addressing these concerns requires a systematic method for analyzing education policies, identifying key themes in policy discourse and monitoring their development over time.

A recent article highlights a noteworthy shift in academic focus, highlighting the increasing focus on specialized educational routes that better meet the needs of the business. highlighting the increasing focus on specialized educational routes that better meet the needs of the business. In this dynamic context, data science and its subfield, topic modeling, offer powerful tools to analyze and interpret these trends. Topic modeling is a text mining approach used in natural language processing (NLP) that uses unsupervised learning on large text collections to generate a summary set of words that are generated from those documents and reflect the collection's overall principal set of themes [1]. Topic modelling find frequently occurring words or phrases in a text sample and classify them under many themes.

This study examines the discourse around education policy in Bangladesh using the probabilistic topic modeling technique Latent Dirichlet Allocation (LDA). LDA makes the assumption that every document is made up of several themes, each of which is represented by a word distribution. By iteratively allocating words to topic clusters and adjusting the probability distribution to best reflect the data's structure, the algorithm finds these subjects [2]. By doing this, LDA uncovers primary patterns in conversations about education, which helps academics and policymakers gain deeper insights into policy trends, emerging challenges, and areas requiring intervention. Tokenization, lemmatization, stopwords removal, and TF-IDF weighting are examples of data preparation techniques that improve the quality of topic extraction by lowering noise in the dataset and are crucial to the efficacy of LDA [1]. By pulling these techniques, this study can enhance the accuracy of topic identification, enabling a more data-driven and evidence-based approach to educational analysis and decision-making.

The need for a comprehensive, data-driven approach to education policy analysis is particularly relevant in the context of ongoing discussions about Bangladesh's education system. Therefore, topic modeling may help educators and policymakers find hidden patterns and developing themes in huge volumes of textual data, including curricular material, academic papers, and student feedback. Also, topic modelling can be used to uncover gaps in current curricula, predict future educational requirements, and have a greater understanding of the changing priorities within educational systems [3]. This method not only improves our comprehension of contemporary educational trends but also gives decision-makers the tools they need to create more responsive and future-ready educational frameworks. The findings in this study contribute to the growing field of computational social science and policy analytics, demonstrating how machine learning techniques can enhance our understanding of complex policy challenges and support more informed decision-making in education reform.

II. Literature Review

Topic modeling is a powerful tool used in data science and research to extract hidden patterns and thematic structures from large text collections. It also enhances search engines by organizing results according to topics, improving user experience. Topic modeling is a crucial tool in academic and scientific research, aiding in literature reviews, trend analyses, and citation analysis. It is also used in healthcare and biomedical research to analyze electronic health records, doctor's notes, and patient reviews. In business intelligence and market research, it helps organizations understand customer preferences, track industry trends, and assess employee sentiment. Legal and policy research uses topic modeling for document classification, regulatory compliance, and policy analysis. In finance and economics, it predicts stock market movements, detects fraud, and forecasts trends using techniques like Latent Dirichlet Allocation and Latent Semantic Analysis.

. Tong and Zhang (2016) explored the application of LDA in text mining by analyzing Wikipedia articles and Twitter data. Their study showed that LDA improves information retrieval and recommendation systems by constructing document-topic and user-topic models. The findings highlight the importance of data preprocessing and model training in maximizing LDA's effectiveness [1].

In qualitative research, topic modeling has been increasingly used to identify themes in large text corpora. While the details of Topic Modelling for Qualitative Studies (2015) are not fully available, existing literature suggests that LDA enhances qualitative analysis by automating pattern detection, reducing subjectivity, and increasing analytical rigor. Compared to traditional manual coding, topic modeling allows researchers to efficiently process large textual datasets, making it valuable in disciplines such as social sciences and humanities [2].

Another significant application of LDA is in sentiment analysis on social networks. Naskar et al. (2016) proposed a Sent_LDA model that integrates topic modeling with sentiment analysis to analyze Twitter conversations. Their framework associates identified topics with emotions using the ANEW lexical dictionary and Russell's model of affect. Additionally, they explore user interactions to detect sentiment-based communities. Their study demonstrates that sentiment-aware topic modeling provides deeper insights into user discourse and opinion dynamics in social media [3].

These studies collectively emphasize the versatility of LDA in improving text analysis. While these studies highlight LDA's strengths, they also have notable limitations. Tong and Zhang (2016) show LDA's effectiveness in text mining but overlook issues like hyperparameter sensitivity, topic coherence, and the assumption of a fixed number of topics. In qualitative research, LDA automates theme detection but still requires human interpretation, making results partially subjective. It also struggles with short texts due to insufficient word co-occurrence. Naskar et al. (2016) enhance LDA with sentiment analysis (Sent_LDA), but their approach depends on external lexical resources, which may not generalize well. Additionally, sarcasm and evolving language complicate sentiment detection, and the model's computational demands can hinder real-time analysis. Despite these flaws, whether for enhancing search algorithms, automating qualitative research, or integrating sentiment analysis, LDA has proven to be a crucial tool for extracting meaningful insights from large-scale textual data.

III. Methodology

The method used in this study is topic modelling for finding topic and patterns in an article for analyzing review on education system in Bangladesh. For that, some step-by-step processes were followed throughout the experiment. The steps are followed by-

i. Data Collection

First of all, a news article, "Education system needs intensive review" was taken from a Bangladeshi online news website Prothom Alo English.

In a new R file, firstly, all necessary packages for this study were installed. Then, web scrapping was done by giving the URL of this article as input for taking the data for preparation.

ii. Data Preprocessing

After web scrapping, data preprocessing was started. The input data were converted from uppercase to lowercase, punctuations and numbers were removed. Stopwords like "the," "is," "and" were also removed to focus on meaningful Words. Then, lemmatization and stemming were done. Lemmatization reduces words to their base form (e.g., "running" → "run"). Stemming removes suffixes to main words. After that, texts were split into individual words or tokens which is called tokenization. Then, this entire preprocessed text was placed in a new dataset.

iii. Document Term Matrix Preparation

Putting the preprocessed text in a corpus, a Document Term Matrix was created. A Document-Term Matrix (DTM) represents text data in a structured format where rows represent documents, columns represent words (terms), values represent the count of each word in a document. There is sparsity in DTM which means most entries are zero since a document contains only a subset of all possible terms.

iv. Calculation of TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is calculated to weigh words based on their importance in a document while reducing the influence of commonly used words. In DTM, there are types of values which are- Term Frequency (TF) and Inverse Document Frequency (IDF) which is altogether TF-IDF.

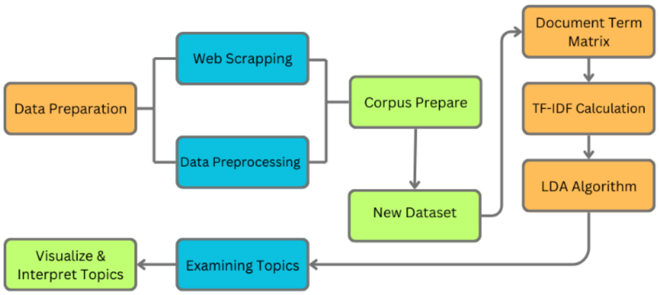

Fig. 1. Block Diagram of workflow

$$a_{ij} = TF_{ij} \times IDF_j$$

Where,

$$TF_{ij} = \frac{Number\ of\ times\ term\ j\ appears\ in\ document\ i}{Total\ number\ of\ terms\ in\ document\ i}$$

$$IDF_j = log \frac{D}{number\ of\ documents\ containing\ term\ j}$$

TF means the frequency of a word presented in the document whereas IDF means how rare the word is among all documents. TF-IDF is basically the calculation of a word with more importance for that document. The higher the TF-IDF value of a word, the better it is for that document.

v. *LDA Algorithm application*

After calculating TF-IDF, we use LDA (Latent Dirichlet Allocation) algorithm to identify latent topics. LDA mainly finds hidden topics from a collection of word. On this study of topic modelling analysis, four topics were assumed to be existed in the dataset. LDA follows an iterative process by initially assigning words randomly to topics. Here every word is given a probabilistic score corresponding to the most likely subject to which it could belong in LDA, where each document is a mixture of topic probabilities and each topic is a mixture of words. In this context, the probability distribution over words is as follows:

$$P(\omega_j) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

where $P(z_i = j)$ is the probability that the *j-th* topic was sampled for the *i-th* word, and $P(w_i|z_i = j)$ is the probability of word w_i of topic *j*. [4]

In this study, a beta matrix was taken to help identify which words are most relevant to each topic. Highest beta value indicating most relevant word for the topic depended on by extracting top terms per topic. Also, the gamma values included in this study stored the topic distribution or probability for each document. It was intended to show each topic proportion in each document. Lastly, the topics were examined and analyzed through visualization plots.

IV. ANALYSIS

After data preprocessing and creating a clean text corpus with meaningful words, DTM was generated and so do the DTM shape was found as 17 300 where 17 means number of documents or text paragraphs extracted and 300 means number of unique words used in the corpus. In this study, term frequencies in the first 5 documents showed high sparsity of 68% which means many words are rare in the dataset. Maximal term length of a word was 11. In this first five documents, the word “education” appeared frequently.

From the TF-IDF scores, the top 6 words are experiment, carry, regard, sometimes, examination, drop with tf = 1, idf = 6.24 and high tf_idf = 6.24. These words appear infrequently across documents but are important when present.

By analyzing, it can be assumed that in the beta matrix, words like "adopt," "carry," "certain" have high beta values for Topic 3. In bar plot visualization of top 10 words per topic of beta matrix, each color represents a different topic helping interpret topic themes based on word distribution.

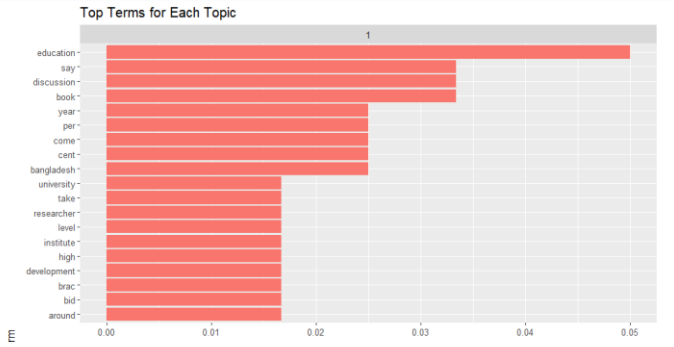

Fig. 2. Top Terms in Topic 1

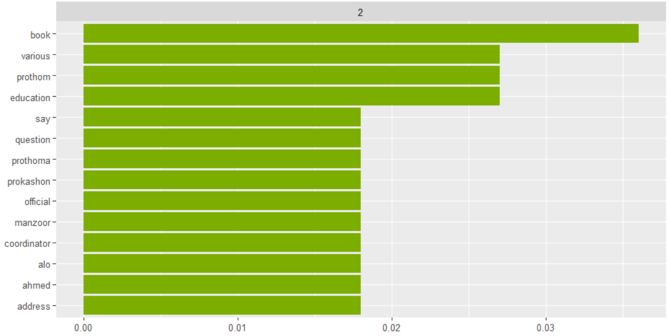

Fig. 3. Top Terms in Topic 2

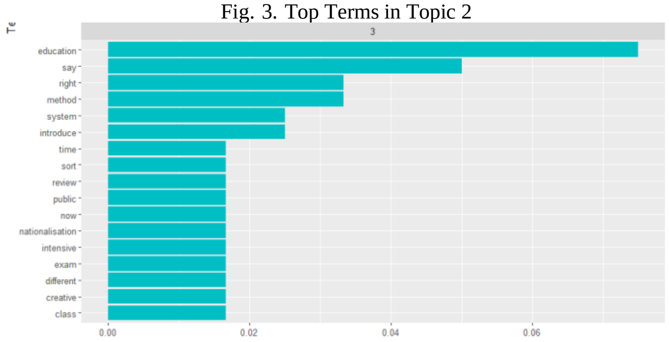

Fig. 4. Top Terms in Topic 3

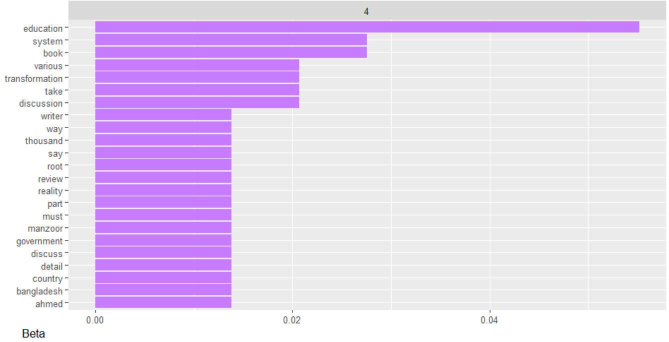

Fig. 5. Top Terms in Topic 4

Topic 1 contains words like "discussion," "researcher," "Bangladesh," and "university", indicating it focuses on education-related research and institutions. Topic 2 includes terms like "book," "prothom," and "prokashon," which suggests it covers publishing, books, and possibly news media related to education. Topic 3 emphasizes "method," "system," "right," and "introduce," pointing toward educational policies, teaching methods, and reforms. Topic 4 includes "transformation," "government," "review," and "policy," indicating a focus on systematic changes in education, possibly policy reforms.

For documents probabilities, gamma values indicate the probability of a document belonging to a specific topic. Document 2 & 7 have very high gamma values (0.999 & 0.998) for Topic 1, meaning they are strongly associated with that topic.

Other documents have lower gamma values, indicating mixed topic memberships.

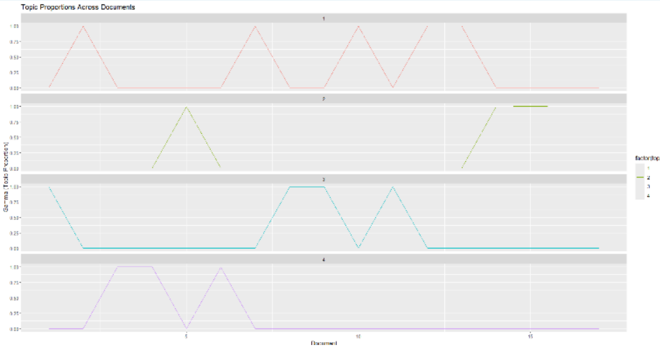

Fig. 6. Topic Proportions Across Documents

The variation in topic proportions across documents indicates that different parts of the dataset focus on different themes. Topic 1 has sharp peaks, meaning some documents are heavily centered on research and institutions, while others are not. Topic 2 is dominant in specific documents, likely sections discussing books and publications. Topic 3 appears in waves, meaning educational methods and policies are discussed in certain parts of the dataset but not consistently throughout. Topic 4 follows a fluctuating pattern, indicating education transformation and policy discussions appear intermittently across different documents.

V. RESULT & DISCUSSION

This study performs topic modeling on text data, likely related to education and publications. The graphs reveal important insights about the structure of the data and topic distribution. The word "education" appears in almost all topics, suggesting that the entire dataset is centered around educational discussions. In plots of top terms per topic, the presence of "Bangladesh" in Topic 1 and Topic 4 suggests a country-specific focus. "Book" appears in multiple topics, reinforcing the role of academic literature in education discussions. The different distributions of words show diverse discussions within education, ranging from teaching methods to publications and policy reforms. In topic proportions across documents, the variation in topic proportions across documents indicates that different parts of the dataset focus on different themes. No single topic dominates all documents, suggesting a well-balanced dataset covering multiple education-related themes. Some documents are almost entirely composed of one topic, while others contain a mix of topics, indicating a diverse but structured discussion. The high proportion of some topics in specific sections suggests that the dataset may include articles or reports with distinct focuses, such as education policies vs. academic publishing.

In summary, the analysis interprets that the dataset is well-structured and contains multiple dimensions of education discussions. Different documents emphasize different aspects of education, from policies and teaching methods to books and publications. The LDA model has successfully grouped words into meaningful topics, making it easier to understand the key themes present in the text. The results suggest that discussions on education are diverse and multi-dimensional, with different documents emphasizing distinct aspects of the broader topic. The results could be useful for education researchers, policymakers, or publishers.

VI. CONCLUSION

This study successfully applies topic modeling using LDA to extract meaningful insights from education-related texts. The analysis reveals a well-balanced unsupervised dataset covering multiple themes, including academic research, publishing, teaching methods, and policy discussions. The topic proportion visualization indicates that different documents emphasize different aspects of education. It suggests a structured yet diverse discourse. The presence of education system transformations and institutional discussions signifies an ongoing evolution in educational practices and policies. These findings can aid stakeholders in education. It includes policymakers, researchers, and institutions, in understanding key trends and making data-driven decisions. Future research can enhance this study by incorporating larger datasets, refining preprocessing techniques. It can also integrate sentiment analysis to measure opinions on education reforms.

REFERENCES

[1] Tong, Zhou & Zhang, Haiyi. (2016). A Text Mining Research Based on LDA Topic Modelling. Computer Science & Information Technology. 6. 201-210. 10.5121/csit.2016.60616.

[2] Nikolenko, Sergey & Koltsov, Sergei & Koltsova, Olessia. (2015). Topic modelling for qualitative studies. Journal of Information Science. 43. 10.1177/0165551515617393.

[3] Naskar, Debashis & Mokeddem, Sid Ahmed & Rebollo, M. & Onaindia, Eva. (2016). Sentiment Analysis in Social Networks through Topic Modeling.

[4] Cifuentes, J., Olarte, F. A macro perspective of the perceptions of the education system via topic modelling analysis. *Multimed Tools Appl* **82**, 1783–1820 (2023). https://doi.org/10.1007/s11042-022-13202-6