English Textbook Content Comprehension with LDA to improve the acquisition skills of learners

Abstract

In Bangladesh, rural students exhibit inadequate proficiency in English language. Pupils cannot fully grasp the context of the National Curriculum and Textbook Board (NCTB) provided textbooks. In this research study, an unsupervised topic modeling LDA approach is proposed to comprehend the context of NCTB's English book. Exploratory analysis is shown to depict significant keywords related to subtle topics in context. It identifies latent topics within lessons that uncover coherent themes from textual data. Extensive analysis is conducted to visualize high impact keywords, co-occurrence patterns, and correlations between extracted topics. It is anticipated that it will improve the curriculum provided by English textbook content synthesis and the acquisition skills of learners. A prototype mobile app is developed that incorporates topic modeling and extracted keywords. Furthermore, a qualitative research survey is undertaken to evaluate its effectiveness on end-users (course instructors at Bangladesh's higher secondary school). The challenges and future potential of LDA extracted content integrated into mobile apps in the learning process are explored. After collecting feedback, word clouds were used to analyze the participants' recommended terms, and the LIWC approach was used to estimate overall sentiment. The LIWC score showed positive sentiment, and the survey process enticed the participants, demonstrating that learners are eager to use the NLP technology driven topic modeling approach in teaching and learning, and there are tremendous opportunities.

Keywords: Natural Language Processing (NLP), Topic Modeling, Latent Dirichlet Allocation (LDA), Exploratory Analysis, Textbook Learning, Coherence

1 Introduction

In Bangladesh, there is a in effective acquisition and synthesis skills in the English language from the National Curriculum and Textbook Board (NCTB) curriculum provided textbook [32, 38]. especially in rural areas, in National

Board examinations like SSC and HSC, most of the students get poor marks in English subjects [1, 15]. It is anticipated that students have a lack of understanding of context. Topic modeling can play a significant role in context understanding for the curriculum provided in the English textbook [5, 8, 12, 28, 37]. Topic modeling is machine learning technique in the Natural Language Processing field that offers a promising unsupervised approach to identifying latent topics within provided documents [13, 18, 25, 26, 34]. It can help identify the main themes, concepts, and topics within the textbook's content, enabling instructors to tailor the learning experience to individual students. LDA is one of the prominent algorithms that can be used for topic modeling [9]. It provides coherent topics, dominant keywords, and a latent combination of features that characterize similarities between topics. In this research, LDA extracted keywords are rearranged and incorporated into a mobile app to observe the user experience. Across all English learning applications via digital media, about 55% exercise activities involved vocabulary learning [16, 20, 27] and other exercises include quizzes, flash cards, and games [36, 42] for enhancing learners' comprehension and self-checking ability [7, 17, 20, 23]. Therefore, it can be assumed that if LDA extracted vocabulary is learned, it will improve contextual understanding and enhance synthesis knowledge of the subtle meanings of correlated topics. There are numerous English learning apps that are prevailing [10, 23, 29]. One caveat is of these apps are based on the curriculum boards provided Textbook for learning English; hence, it could not attract a large number of pupils in Bangladesh who are mostly dependent on NCBI textbooks. To grasp the English language knowledge from curriculum provided in textbooks, a novel approach Latent Dirichlet Allocation (LDA) based unsupervised Topic modeling using the textbook corpus is adopted, and exploratory analysis is demonstrated in this study. Our anticipation is through this way, students can be able to interpret meaningful information, which facilitates their understanding of the correlated topics and important keywords related to those topics and leads them to understand the subtle meaning of the textbook context.

1.1 Research Overview

To detect underlying themes or topic keywords within a textbook corpus, the unsupervised probabilistic topic modeling technique LDA is used in this research [18] (see section 3.3 for details). Topic modeling of LDA doesn't directly account for student engagement, such as learning tasks in mobile apps [33]. Hence, a prototype app is developed and a qualitative survey is conducted to observe the instructor's sentiment impact similar to research paper in [44]. Qualitative survey research is undertaken to evaluate the effectiveness of unsupervised topic modeling LDA Bangladesh's National Curriculum Textbook Board (NCTB) provided English textbook for higher secondary school education. This study seeks to ascertain if instructors can teach and students can learn English better if a mobile app is introduced, which includes NLP's

LDA driven topic modeling applied extracted keywords and analysis. A prototype mobile app is developed to incorporate the topic-modeling extracted keywords into the app. This article presents the key findings and insights from the survey, shedding light on the perspectives of learners, especially instructors. In the survey questions, it was indicated whether the students, teachers, and instructors would find it acceptable and appreciated if textbook information were made available through a mobile app and presented in an interactive format. To demonstrate the mobile app idea during the interrogation survey session, a prototype is also prepared. Participants were asked for suggestions on how to make the app better and about any shortcomings. After collecting feedback, word clouds were used (see 3.4.2) to analyze the frequency of the participants' recommended terms, and the LIWC approach was used to estimate overall sentiment (see 3.4.1). The main contributions of this research study are:

- Unsupervised topic modeling technique LDA based keyword extraction from textbooks for the Bangladeshi setting affects the efficacy of education.
- Along with theoretical analysis with LDA, a prototype demonstration qualitative survey conducted to realize its potential and credibility.
- English Textbook content comprehension facilitates empowerment via education of mass population, which has direct impact on sustainability.

The following sections are organized as follows:. Background literature studied in Section 2 illustrates LDA for dominant keyword determination, followed by methodology. Survey response visualization and analysis are depicted in (subsection 3.4). The following section, 4 demonstrated the experiments and contains analysis separated into several subsections (4.1, 4.2 and 4.3). Then results discussion, limitations, and conclusion are illustrated in sections 5, 6 and 8 respectively.

2 Literature Review

Topic modeling has been used for semantic search, ontology exploration, classification, and dominant keyword searching in many research studies. For a curriculum-based textbook study, it could be an option that is not revealed from rigorous searches in online repositories, which we addressed in this research study. This research model involves infusing textbook features into LDA based topic modeling to discover a set of topic-words from the provided document for textbook comprehension and understanding context. Similar approaches and implications are described below:

2.1 Textbook Content extraction with LDA

According to Krishna Raj [30] the human brain has a propensity to forget a number of facts regarding the events in book. The LDA model is able to scan through large quantities of text in the book and extract intriguing key concepts and terms. Thereby, a book can be adapted by learners with the help

of the LDA model, greatly refining the learning process. A similar approach was proposed by Rani et al. for Hindi books and stories, a topic modeling text summarizing approach proposed in 2021 [31]. By incorporating linguistic features into LDA-based topic modeling, the suggested model can find a set of topic words from the given material to understand context. Educational content-based topic modeling for an intelligent system to develop a tutoring system is proposed in [37] by researcher Stefan Slater in 2017. To understand the linguistic content of mathematical problems, a personalized learning system is suggested that makes use of correlated topic modeling in natural language processing, an approach that can extract important keywords. A variety of significant and useful contents are explored in the context of addressing mathematical difficulties. They demonstrated that topic modeling is a useful method for personalized learning systems. For key term detection within articles, LDA-based topic modeling has been used in many research studies in which dominant keywords reveal future research trends or the most prominent topics. An investigation by Julio Guerra et al. (2013) demonstrated how the LDA model can be utilized for online content linking for any major subject, such as elementary algebra [12]. It can also be used to simplify context comprehension and content modeling for collections of reference books on the same topic. They concluded that the recommendation provided by LDA topic modeling for online educational systems is promising.

2.2 LDA for Dominant Keywords Determination

In 2019, Wafa Shafqat [35] proposed an architecture model for better understanding of crowdfunding comments posted by the investors to understand their motive to classify whether comments are scam or legitimate comments. Deep neural network language modeling, either LSTM or RNN-encoded embedding vectors, is fed into a LDA based topic modeling model to understand the context of discussion trends. Compared to simple Neural Networks (NNs) and non-LDA based approaches, this technique performs better at understanding crowdfunding comments. The capability of LDA-based topic modeling to detect the research trend of Bengali news published in web was analyzed by Kazi Masudul Alam in 2020 [4]. Their research demonstrates that using an appropriate corpus and labeled LDA is an effective combination model for predicting news topics efficiently. LDA labels key terms, making articles easier to read. Another research article by J. Lee published in 2022 conducted an experiment on the research trends of "COVID-19 and sports [21]. It used LDA and explored latent knowledge connectivity dimensions and structures in the articles. Rahul Gupta uses the application of LDA in 2022 to analyze research patterns of "Applied Intelligence" in 3269 articles published between 1991 and 2021 [14]. In this research, BoW and TF-IDF embedding are used for LDA based topic modeling.

2.3 LDA based Sentiment Analysis

Sentiment analysis has been a key research area in the NLP based research domain, where LDA has been applied to determine significant features, and those features contribute to segregating sentiments and providing recommendations. LDA based topic modeling has been used in sentiment analysis task. In 2021 Y. Cho et. al. published a research study of LDA-based topic modeling for sentiment analysis using the topic/document/sentence (TDS) model [11]. This article proposes a novel TDS approach that combines LDA-based topic modeling for sentiment analysis within documents. H. Wang et. al. examined Chinese people's public perceptions about Omicron variants on social media, Sina Weibo. Social media's 121,632 Omicron related data posts were analyzed using LDA-based topic modeling and sentiment analysis [41]. From topic analysis, they realized omicron's impact, infection situation, pandemic prevention, and control geographically. Hence, it is actually revealed that LDA based topic modeling can be used for understanding subtle topics and exploring various facts. Hence, from the above literature review, we can infer that LDA could help analyze the content of the textbook and identify the main topics covered. This information could then be used to enhance the learning experience.

3 Methodology

The methodology involves pre-processing textual data, training the LDA model on the pre-processed text, and subsequently interpreting and visualizing the generated topics. Data is collected from NCBI's English Textbook for class 9 of a higher secondary school. NLPs data mining approach is applied to segregate into subsequent lessons [19, 22]. Then textual data is pre-processed to remove noise, text is standardized, followed by the application of LDA to identify underlying topics, and then coherence measurements are applied. The research overview is depicted as follows:

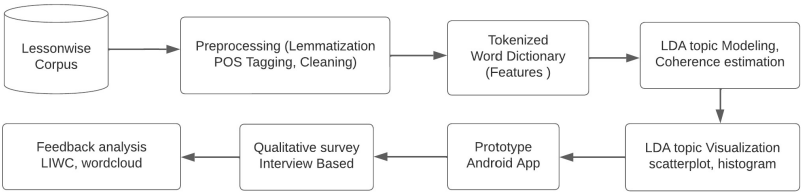


Fig. 1: Research workflow diagram

LDA algorithm automatically discovers latent topics within the documents based on word co-occurrences. Each topic is represented by a set of words; interpret these words to understand the main concepts associated with each

topic. Extensive exploratory analysis is conducted to visualize the topic modeling outputs. Analyze the topics generated by the model, which require manual review and adjustment to ensure the topics make sense.

3.1 Data Pre-Processing

Text is converted to Lowercased and Normalized to ensure consistent preprocessing [19].

(i) **Data cleaning**: unwanted characters, punctuation, and special characters are removed, and stop words (such as "and," "the," "is," etc) are removed. The Spacy library's English word model and NLTK's stopword list are used together. Also, words less than two characters are removed, such as: I, Hi, Oh, etc. Hence, noise is removed, and irrelevant characters, symbols, or data artifacts that have been introduced during data collection or scraping from a pdf file to a text file are separated. Hence, we found a clean corpus.

(ii) **Lemmatization**: Root words are collected words, and their dictionary form (lemma) is extracted using NLTK's WordNetLemmatizer package. Stemming reduces words to their base or root form, is not used since sometimes it changes the actual words.

(iii) **Part-of-Speech Tagging**: Spacy's English model 'en_core_web_sm' is used to extract interested words (such as noun, verb, and adjective) and excluded (CCONJ, AUX, DET, INTJ, PART etc which are Coordinating Conjunction, Auxiliary, Determinator, Interjection, Particle, etc) thereby collecting tokens for only those that are not punctuation, conjunction, symbol, etc.

3.2 Topic Modeling and Feature Extraction

Different techniques have been developed to perform topic modeling in the unsupervised topic modeling domain of Natural Language Processing (NLP), each with its own strengths and limitations [3, 40, 43]. Apart from LDA, Mallet LDA, Structural Topic Model (STM), Hierarchical Dirichlet Process (HDP), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA) etc are also prevailing and can be considered for comparative research studies. Prevailing LDA based data mining techniques are applied for feature extraction.

3.3 Latent Dirichlet Allocation (LDA)

LDA model [13, 18, 25, 34] considers documents as mixes of topics, and each topic is a distribution over words. The objective is to derive the hidden topic assignments and the topic-word distributions that most effectively describe the observed documents. The goal of LDA is to uncover these latent topics from a collection of documents without needing any prior labeling or categorization of the content. An expression for the joint distribution of the LDA model is

described below:

$$P(\theta_d, z, w \| \alpha, \beta) = P(\theta_d \| \alpha) \prod_{n=1}^{N} P(z_{d,n} \| \theta_d) P(w_{d,n} \| z_{d,n}, \beta) \quad (1)$$

Where $w_{d,n}$ the n_{th} word in document d, $z_{d,n}$ the topic assigned to the n_{th} word in document d, α, β are the Dirichlet LDA model parameters. controls per-document topic distribution and per-topic word distribution. θ_d represents the topic distribution. $P(\theta_d \| \alpha)$ Dirichlet distribution representing the document-topic distribution, $P(z_{d,n} \| \theta_d)$ is the word topic assignment for the n_{th} word in document d, $P(w_{d,n} \| z_{d,n}, \beta)$ is the distribution representing the observed word given a topic. An algorithm for the LDA model is explain below:

We have chosen LDA as the baseline statistical topic modeling tool. An abstract of this model is depicted in figure 2

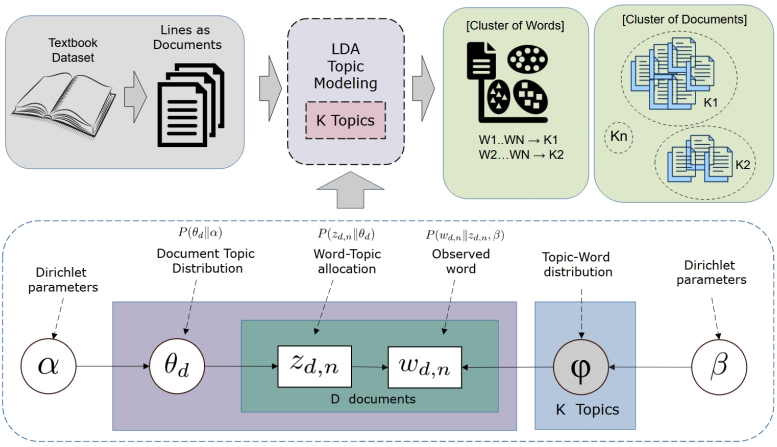


Fig. 2: Abstract LDA model workflow diagram

3.3.1 Comparative analysis of Topic modeling

While some variations of LDA, Mallet LDA is considered for large corpus processing and analysis [3, 6, 40]. It focuses on scalability. If large corpus needs to be analyzed, Mallet LDA might be more suitable. LDA in general can still be efficiently applied to moderately sized corpora. Analyzing topics within the context of metadata, STM could be a better fit. The Hierarchical Dirichlet Process (HDP) can be useful when we cannot guess the number of topics in advance. In [6] LSI, NMF, and LDA are compared in terms of coherence and similarity measures for the social media dataset, and in their analysis, NMF is observed to be the most effective measure. However, as a baseline model, LDA is often considered one of the most prominent choices. In this study, the

textbook corpus is divided into lessons, which are a mixture of topics, and using LDA, it is expected to be determined which words in the lesson belong to Lesson's topics. LDA produces interpretative results for exploratory topic analysis. The identified topics are represented as distributions over words, making it easy to assign meaningful labels to topics. Provided by most of the libraries and tools, making it easy to implement and can be integrated into existing workflows. Hence, LDA serves as a solid baseline for topic modeling tasks. However, how many topics are ideal needs to be determined, and topic modeling quality also needs to be measured.

LDA vs PCA: Mathematical Perspective

LDA sometimes compared directly with PCA as a dimensionality reduction tool that identifies the principal components capturing maximum variance in the data. **Covariance Matrix:**

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)^T \tag{2}$$

PCA represents data using the principal components, reducing dimensionality using Eigenvalue.

Dimensionality Reduction:

where Y is the reduced-dimensional representation, X is the original data, and V contains the principal components.

Variance Preservation:

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{D}\lambda_i} \tag{3}$$

PCA is effective for data compression and noise reduction widely used in scenarios where reducing the number of features is crucial, such as in image analysis or signal processing. LDA provides interpretable topics that can be useful for understanding the thematic structure of a document collection.

3.3.2 Optimal Topics with Coherence

The coherence score measures how coherent or interpreted the words in that topic and estimates the number of topic clusters [24]. The coherence score assesses the quality of the topics produced by LDA and ensures that the topics generated are statistically significant. Coherence C_{topic} can be expressed as follows:

$$C_{topic} = \sum_{i=1}^{N}\frac{1}{N(N-1)}\sum_{j=1}^{i}PMI(w_i, w_j) \tag{4}$$

Where, $PMI\,(w_i, w_j)$ represent pointwise mutual information statistical association between two words occurring together. PMI score indicates that the two words are more closely related within a topic. $PMI\,(w_i, w_j)$ can expressed

as

$$PMI\,(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{5}$$

where $P\,(w_i, w_j)$ is joint probability of occurrence of words w_i and w_j. To calculate the coherence score, the gensim library provides range of options such as $u_{mass}, c_v, c_{uci}, c_{npmi}$. u_{mass} and c_vThese two methods are most popular. For given topic with words $\{w_1, w_2, w_3, \ldots.., w_n\}$ a fixed context window size is provided (default size 10 words) then coherence score is calculated using an equation $\sum_{j=1}^{i} PMI\,(w_i, w_j)$ which provides negative coherence score. c_v can be expressed as

$$c_v = \frac{1}{N(N-1)} \sum_{j=1}^{i} similarity\,(w_i, w_j) \tag{6}$$

in which $similarity\,(w_i, w_j)$ represent the pairwise similarity between terms based on $PMI\,(w_i, w_j)$ scores. c_v provides a positive coherence score. Higher coherence values (higher than 0.5) indicate that the topics are moderately coherent and representative of meaningful themes within the text data.

3.4 Visualize participants response

3.4.1 Linguistic Inquiry and Word Count (LIWC)

In this research, LIWC is used to ascertain the general sentiment of the responses given by the survey participants (see section 4.4.6). Linguistic Inquiry and Word Count (LIWC), is a text analysis tool to measure psychological or emotional characteristics [2, 39]. It aims to quantify sentiment by examining the frequencies of different linguistic terms within given text based on predefined dictionary of words associated with various categories. Let's assume text as a sequence of words: $\{w_1, w_2, \ldots., w_n\}$ and M different linguistic categories: $\{C_1, C_2, \ldots., C_m\}$. Proportion of words in each category $P[i] = \frac{w[i]}{T}$ where T is the total number of Text. Now, a matrix $X[i, j]$ can be formed, where w_i represents the frequency of the word in the linguistic category C_j. LIWC vector containing the proportions of words in each linguistic category can be expressed as $P_{total}[j] = \frac{\sum_{i=1}^{T} X[i,j]}{T}$.

3.4.2 Word cloud

LIWC involves linguistic analysis using mathematical expressions, but using word clouds, survey answers can be visualized vividly in an interpretable interactive format. Word cloud consider a set of words $\{w_1, w_2, \ldots., w_n\}$ extracted from text document and associated frequencies $\{f_1, f_2, \ldots., f_n\}$, s_i represent the proportional size of the word in the cloud can be expressed as $s_i = \frac{f_i}{\sum_{j=1}^{n} f_j}$ where normalized frequency f_i. In sections 4.4.6 figure 10a and 10b we can see the participants responses most frequent terms.

4 Experiment and analysis

In this research study, for the dataset we have collected, Bangladesh NCBI provided an English textbook for classes 9–10. Here, the whole book is segregated into lessons, and we wanted to explore the important topics within the content. Similar topic words remain together. Therefore, assumptions are that it helps students understand the words, sentences, and context of the book. The ideal number of topics are determined using the coherence score [24].

4.1 Coherence measurements

To measure coherence in the context of LDA, following steps are followed:

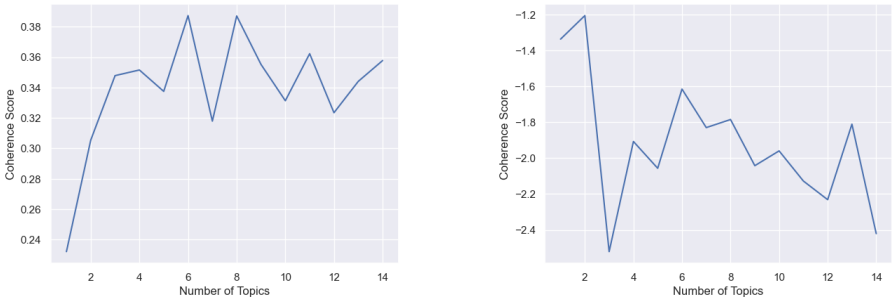
(i) Cleaned document samples are prepared using Python's NLP data mining techniques explain in detailed in data reprocessing section. Prepared $T(d_i)$ set of tokens in Documents d_i for i^{th} document samples in corpus D.

(ii) Doc to BOW corpus dictionary is prepared with a Doc2Bow vector. This vector x_d can be represented as where $n(w_i, d)$ denotes the count of words w_i for the document d.

(iii) Trained LDA Model: During the training phase, gensim's MulticoreLDA model with four CPU worker thread is set. Doc2Bow dictionary is applied, along with 20 iterations that are invoked. The rest of the parameters for LDA model training were the default parameter settings of gensim library.

(iv) Calculate Coherence: To calculate the coherence score for each LDA model for n number of topics, step 3 is iterated for $n = 15$ times.

(v) Iteration result coherence score for n number of topics are saved in a list and plotted using Seaborn.



(a) Coherence for c_v approach (b) Coherence for u_{mass} approach

Fig. 3: Coherence score to estimate optimal number of Topics

From the chart we can see that six topics are dominant in our provided corpus. The chart shown at the left shows the coherence score for u_{mass} and

the right chart represents the score for c_v for multiple iterations. Using 6 topics we can see the output of corresponding topic and top 10 words in a topic.

Topics	Dominant Keywords and Weights
Topic 01	'energy' 0.060, 'source' 0.029, 'renewable' 0.018, 'water'0.016, 'use' 0.013, 'gas' 0.013, 'produce' 0.013, 'green' 0.013, 'warm' 0.013, 'cause' 0.012
Topic 02	'pastime' 0.024, 'computer' 0.024, 'social' 0.023, 'user' 0.022, 'network' 0.020, 'student' 0.019, 'class' 0.017, 'change' 0.016, 'book' 0.015, 'survey' 0.013
Topic 03	'mother' 0.083, 'buy' 0.021, 'love' 0.018, 'child' 0.014, 'worker' 0.014, 'begin' 0.013, 'cultural' 0.012, 'observe' 0.012, 'thing' 0.012, 'language' 0.011
Topic 04	'life' 0.016, 'Bangladesh' 0.016, 'family' 0.015, 'home' 0.014, 'root' 0.014, 'language' 0.014, 'country' 0.013, 'Pakistan' 0.010, 'war' 0.010, 'man' 0.009
Topic 05	'country' 0.031, 'river' 0.022, 'India' 0.022, 'land' 0.021, 'boat' 0.015, 'small' 0.015, 'population' 0.015, 'lake' 0.013, 'group' 0.012, 'house' 0.011
Topic 06	'job' 0.064, 'English' 0.023, 'learn' 0.021, 'teacher' 0.017, 'use' 0.016, 'dream' 0.016, 'think' 0.016, 'thing' 0.015, 'school' 0.014, 'education' 0.013

4.2 Dominant topic determination

In LDA models, each document is composed of multiple topics. But typically, some specific topics are dominant. The following experiment extracts this dominant topic for each sentence and shows the relative weight of the topic and the keywords. It estimated which document belongs predominantly to which topic. How frequently the words have appeared in the documents and the weights of each keyword in the same chart, words that occur in multiple topics, and the ones whose relative frequency is greater than the weight.

4.2.1 Relative Importance measurement

Word frequency $n(w_j, d_i)$ in each document D is measured in equation 8 as below which identifies the most frequent words within each document and across the entire corpus.

$$D = \sum_{d_i \in D} \begin{cases} 1, n\,(w_i, d_i) > 0 \\ 0, n\,(w_i, d_i) = 0 \end{cases} \tag{7}$$

we can visualize relative importance of any keywords in terms of frequency and plotted inclined with LDA provided weights (figure 4).

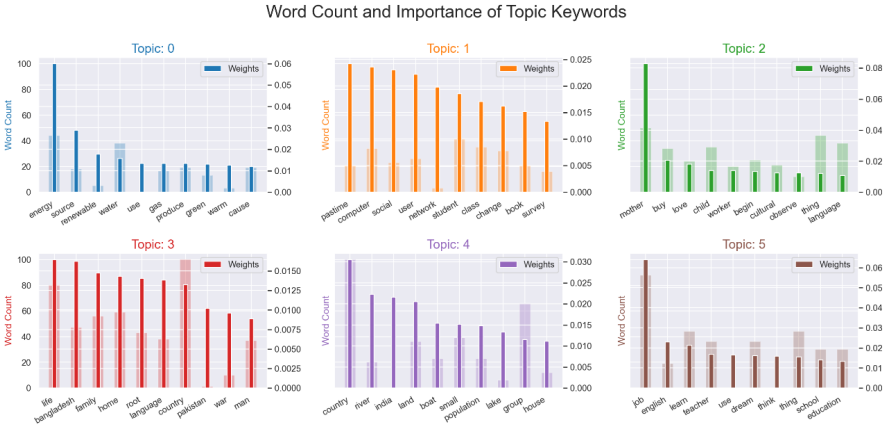


Fig. 4: Word frequency and its relative importance

4.3 Topic-Term Matrix Visualization

Visualizing the topics and their relationships in a topic model The Python library PyLDAvis is used to provide an interactive web-based interface to explore and analyze the LDA results of topic modeling. PyLDAvis itself abstracts away much of the underlying mathematical complexity and provides a user-friendly way to generate visualizations and interactively explore topics and their relationships. Key components of distance among topics and salient terms are explained below:

4.3.1 Inter-Topic Distance Map

Distance among topics refers to the measurement of similarity between topics in a high dimensional space matrix provided by the LDA model. PyLDAvis library is used to conserve dimensionality reduction using PCA and for calculating distance between topics metric like Euclidean distance or Cosine Similarity.

4.3.2 Salient Terms or dominant keywords

Salient terms in a topic are words W that are most strongly associated with a specific topic. The mathematical expression for finding salient terms w for a topic t involves the extraction of top n words that pose the highest probability scores for topic t in the topic-term matrix $P[t, w]$.

The top 30 most salient terms are shown at right in the bar chart histogram, and the left figure shows inter-topic distance, their size, etc. (see figure 5). The PCA dimensionality reduction technique is applied here to embed the LDA result into a 2D plain scale. Projecting the data in lower-dimensional subspace by computing eigenvalues reduced the circle overlapping. Topics that are closer together in the map are more similar in terms of the distribution of words.

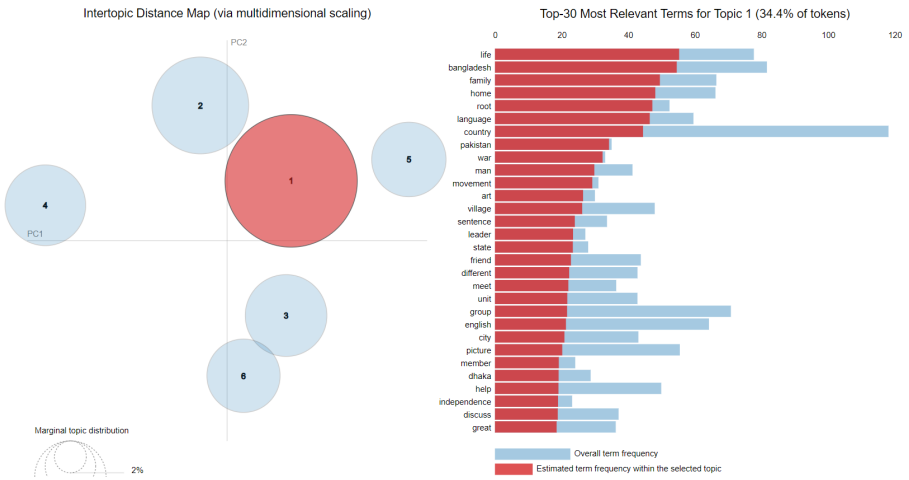


Fig. 5: Topic model co-occurrence visualization with dominant keywords'

4.4 Qualitative survey

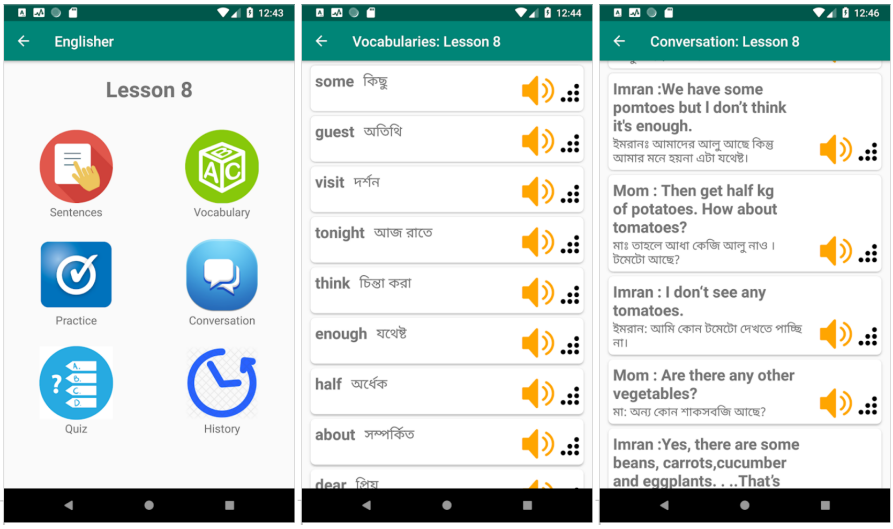
In the survey questions, it was indicated whether the students, teachers, and government organizations would find it acceptable and appreciated if textbook information were made available through a mobile app and presented in interactive format. To demonstrate the mobile app idea during the interrogation survey session, a prototype app, Englisher, is prepared (see demo in Section 4.4.1). Participants were asked for suggestions on how to make the app better and to specify shortcomings. Presumably It provides an insight into teacher's emotions about the inclusion of mobile technology in the higher secondary English education system.

4.4.1 Englisher Mobile App

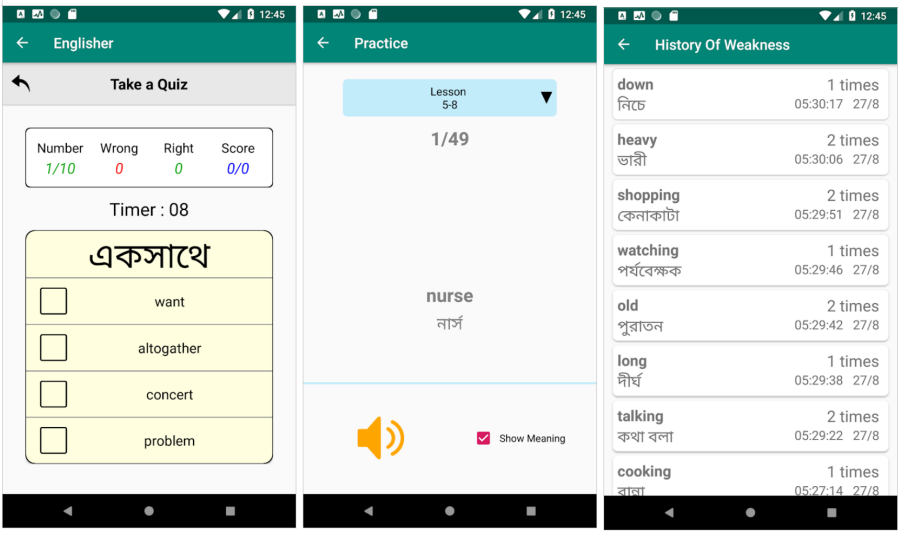
A mobile application (Englisher) is being developed with content from the NCTB's English Textbook for class 9. The extracted keywords are organized into lessons, and furthermore, a quiz is introduced as an exercise. Each sentence's and word's Bengali meanings are provided in accordance with the lesson. Students can take quizzes, and their results are recorded in history so that history can be reviewed and performance can be improved with more practice in the future.

4.4.2 Survey Planning

The survey was conducted over a period of four weeks with 50 high schools in Dhaka and Bogura districts of Bangladesh. It encompasses only English subject areas Teachers who teach in high schools from grade six to grade ten and teaches regularly in the school. A questionnaire was distributed to teachers, allowing us to gather answers.



(a) Lesson wise exercise



(b) Quiz with Vocabulary

Fig. 6: Englisher Mobile app for Learning LDA based topic model words

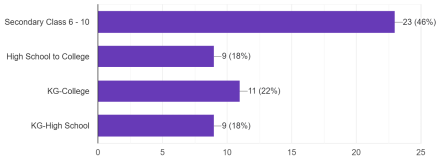
4.4.3 Participants

During the survey, standard participants were chosen emphasizing infrastructure quality, teaching experience, class size, etc. At the beginning, 100 institutions were selected. Then half of them were excluded since the infrastructure's overall quality and condition were not above average. Among the

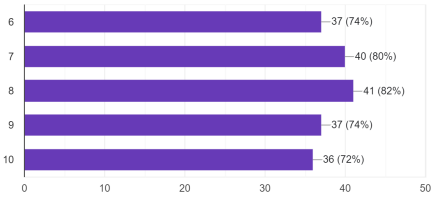
chosen samples, 76% were good and 26% considered average institutions. Privately held 45%, 32% partially government, and 22% are government institutes. Over 1000 students study in almost 40% of these institutions, and a sizable number of pupils are present in each section and class. 38% class have a size greater than 50. So, we can presume that the participating teachers have quite a bit of experience teaching a sufficient number of students.

4.4.4 Survey Results

We have done extensive analysis with the survey data collected. In our data collection, the highest priority is given to the secondary class student teachers who teach between 6 and 10th grade, about 46%. High school, KG college, and KG high school. Details about the statistics are depicted in the following figures. Adjacent chart explains the percentage of teachers who teach in which class. Hence, from these two figures, we can get a vivid image of collected dataset resources about the participating teachers.



(a) Participating institution category and size

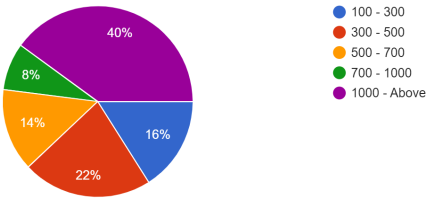


(b) Participating instructors' class/Label

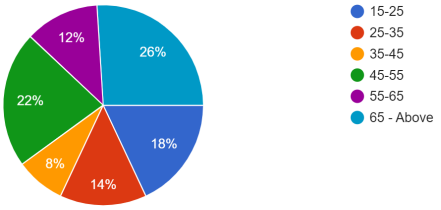
The following graphs give an overview of the English teaching experiences of the teachers as well as the general consensus regarding the use of digital content and mobile apps in everyday teaching and learning. Almost 62% of teachers have been teaching for more than 8 to 10 years, and some of them have been teaching for decades in higher secondary education. 32% of teachers have three to eight years of experience, while just 6% are new to the profession. Around 83.7% of teachers said the language of instruction during their graduation was English, and their major was also English. Very few teachers 13.3% graduation major is something other than English, yet teachers teaching English in secondary schools probably have sufficient English language proficiency.

4.4.5 Analyzing survey Facts

More than half of teachers, or 58%, have no prior experience utilizing mobile apps or technology for teaching, but 90% of them agree, and more than 45% strongly agree, that it encourages pupils to engage actively in their learning. However, they (almost 60%) also hold the opinion that a notebook cannot be

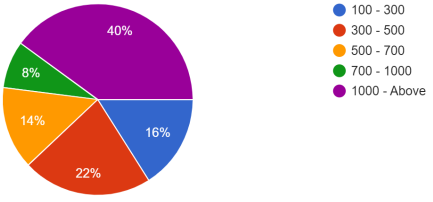


(a) Participating institutions' number of students

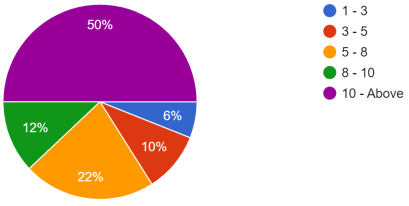


(b) Participating instructors' class size

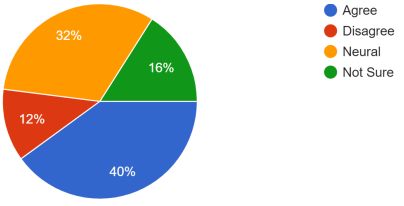
completely replaced, despite the fact that mobile apps may solve many problems and provide technological support for teaching and learning. Promisingly optimistic approximately 40%, although thinking that the notebook-based content memorizing learning method can be replaced, feel that mobile app-based learning can replace it permanently.



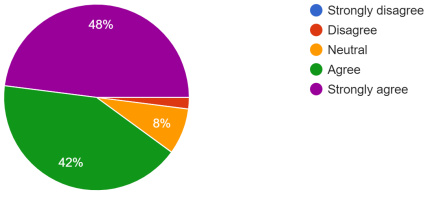
(a) Participating institutions' number of students



(b) Participating instructors' Teaching experience



(c) App could help students in replacement of guide book



(d) App could motivate students

This study proposes the Englisher mobile app and presents it to the participating teachers to gather their insightful feedback. 92% of teachers reported that they would use this type of mobile app for teaching if it were made available after using the trial version of the offered customized Englisher app. Teachers anticipate that 80% of students will utilize this app during class. 86% of respondents believed it may help students' English proficiency, and 98% agreed that the government should support this kind of innovation in the education sector.

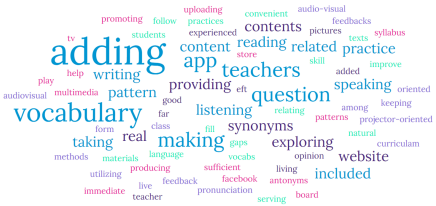
Questions	**yes**	**No**
Do you use digital content for teaching or digital medium for teaching and learning?	84%	16%
Have you ever used Internet or Mobile app to teach students or asked students to find learning materials from internet or Mobile App?	76%	24%
Education during graduation was English and English was used for learning	83.70%	16.30%
Customized mobile app for Learning and Teaching English		
Do you think topic model based mobile app-based learning can improve English proficiency of students?	86%	14%
Do you think Govt should promote these types of innovation for education sector?	98%	2%

4.4.6 Qualitative Survey Sentiment

Traditional LIWC Dimension	**Answer Text**	**Standard Commercial Language**	**Answer Text**	**Standard for Formal Language**	**Answer Text**	**Standard for story language**
Positive Tone	2.54	3.96	3.91	2.33	3.22	2.18
Negative Tone	0	1.1	0	1.38	0	1.75
Social Words	2.54	6.87	5.65	6.54	4.08	10.5
Cognitive Processes	13.56	9.35	18.26	7.95	15.88	8.7
Allure	2.54	7.79	3.04	3.58	2.79	5.48
Moralization	0	0.2	0	0.3	0	0.21

From this LIWC table, a higher proportion of words related to positive emotions indicates a positive emotional tone in the text in the answer to the questions related to "How this app can be improved" and "How English learning can be improved using a mobile app". LIWC is applied to three different categories: "commercial writing", "Formal language", and "story language", and in all the categories, the answer text showed highly positive sentiment from the survey user. Though respondents had a mix of optimism and skepticism regarding the use of mobile apps in teaching and learning,. During the interrogation session, their tone was positive and enticed participants.

(a) word cloud for the question regarding app improvement



(b) word cloud for the question regarding English learning app improvement

The word cloud is generated from the answers provided to the question "how the app can be improved" and followed by a more generalized question "how English learning can be improved using an app". The participants narrated a variety of viewpoints on the questions. From the word cloud, it is inferred that adding graphical content would enhance the apps' usefulness and make them more visually appealing to users. More practice resources for listening and exercise would be helpful. Another suggestion is to include synonym and syllabus-related instances as well as audiovisual engagement with the app. Add additional vocabulary and involve more experienced teachers who have greater experience in digital learning and teaching.

5 Results Discussion and outcomes

The results demonstrate that LDA-based topic modeling significantly enhances content comprehension by providing concise summaries of the learning material. Readers can grasp the main ideas and connections between topics, aiding in retention and knowledge acquisition. From the qualitative survey, it was revealed that LDA based topic modeling approaches based on extracted keywords within mobile apps seem effective as they provide more contextual meaning to the learners. Most of the participating teachers are enthusiastic about topic modeling-based contextual resource learning related to technology incorporated into pedagogy. Participants appreciated the cutting edge NLP's learning resources available through mobile devices. Teachers admitted that available digital resources facilitated a deeper understanding of topics and catered to different learning styles, nurturing a more engaging learning environment. This will positively impact student motivation and overall engagement and, hence, boost overall learning. Some crucial suggestions were to improve the graphics of the app so that it becomes interactive and guardian involvement can be introduced. Based on the survey results, it is revealed that the potential for digital mobile-based learning in schools is immense. The government should take initiatives to incorporate it into the course curriculum syllabus and could impose an obligation to adopt mobile app-based learning teaching in the school.

Apps need to be improved by including collaborative form of learning.Additionally, the interactive interface receives positive feedback for its

user-friendly design and utility in assistingreaders' navigation through the textbook.The recommendation is to make it specific to NCTB books only for particular classes. This approach is also our goal, considering NCTB Books. Including interactive e-books, dictionaries, educational apps, and multimedia content.

6 Limitations

LDA could play a role in understanding the topics covered in an English textbook and potentially aiding in content customization and topic relevance for personalized learning. In a personalized learning context, the goal is to tailor the educational experience to the individual needs and preferences of each learner. This involves understanding the learner's strengths, weaknesses, interests, and learning style. While LDA could be useful in some aspects of this process, it might not directly address all the requirements of personalized learning for an English textbook. In this research study, we showed that LDA based topic modeling could be a solution to enhance the context understanding of the learners. However, from the survey, it was revealed that the app was not sufficient. Learner oriented topic-document distribution to identify which topics are most relevant to a specific student can be provided. The app can provide additional explanations, examples, or resources to cater to their individual learning style. Assessments and exercises focused on the topics that need reinforcement for each student. Monitor their progress and adjust the learning path accordingly. Analyze students' performance, engagement, and feedback to refine the topic modeling process and its integration into the learning environment. More sophisticated approaches, such as adaptive learning systems and AI-based tutoring, might be needed to truly personalize the learning experience in a comprehensive manner.

7 Special Remarks

For data privacy and security issues, many teachers were reluctant to provide their social website addresses to the surveyor. Among all the participants, only 24% attendees provided their social media addresses to use them publicly for research purposes.

8 Conclusion

By employing topic modeling in a personalized learning context, educators can create a more engaging and effective learning experience. This approach allows for enhanced understanding and retention of the textbook context. The school survey with the prototype app reaffirmed its potential in learning experiences. LDA based topic modeling leverages learning experience to improve interpretation and knowledge acquisition. The synthesis of existing research sheds light on the potential of topic modeling to improve textbook context

comprehension and the knowledge retention of learners. It was revealed that teachers and instructors would find it acceptable and appreciated if textbook information were presented using NLP technology driven algorithms like LDA topic modeling in mobile apps. The study concludes that apps seem effective as they provide a personal and learner-centered learning opportunity ubiquitously. Reveal to the user as complementary essential material to learn English textbooks quickly and effectively. The survey's findings show that teachers are eager to use NLP provided extracted keywords technology in teaching and learning. There are tremendous opportunities; however, apps need to be improved by including collaborative form of learning.

References

[1] Bangladesh Education Statistics 2021.

[2] Welcome to LIWC-22 introducing liwc-22 a new set of text analysis tools at your fingertips.

[3] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, page 102131, 2022.

[4] Kazi Masudul Alam, Md. Tanvir Hossain Hemel, S.M. Muhaiminul Islam, and Aysha Akther. Bangla News Trend Observation using LDA Based Topic Modeling. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6, December 2020.

[5] Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174:27–42, June 2019.

[6] Ashjan Alotaibi and Najwa Altwaijry. A comparison of topic modeling algorithms on visual social media networks. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)*, pages 26–31, 2022.

[7] Matthew L. Bernacki, Jeffrey A. Greene, and Helen Crompton. Mobile technology, learning, and achievement: Advances in understanding and measuring the role of mobile technology in education. *Contemporary Educational Psychology*, 60:101827, January 2020.

[8] Steven Bethard, Soumya Ghosh, James H. Martin, and Tamara Sumner. Topic model methods for automatically identifying out-of-scope resources. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 19–28, Austin TX USA, June 2009. ACM.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Xiaojun Chen. Evaluating Language-learning Mobile Apps for Second-language Learners. *Journal of Educational Technology Development and Exchange*, 9(2), December 2016.

[11] Akhmedov Farkhod, Akmalbek Abdusalomov, Fazliddin Makhmudov, and Young Im Cho. LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. *Applied Sciences*, 11(23):11091, January 2021. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

[12] Julio Guerra, Sergey Sosnovsky, and Peter Brusilovsky. When One Textbook Is Not Enough: Linking Multiple Textbooks Using Probabilistic Topic Models. In Davinia Hernández-Leo, Tobias Ley, Ralf Klamma, and Andreas Harrer, editors, *Scaling up Learning for Sustained Impact*, Lecture Notes in Computer Science, pages 125–138, Berlin, Heidelberg, 2013. Springer.

[13] Aakansha Gupta and Rahul Katarya. PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning. *Computers in Biology and Medicine*, 138:104920, November 2021.

[14] Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, and Thoudam Doren Singh. Prediction of research trends using LDA based topic modeling. *Global Transitions Proceedings*, 3(1):298–304, June 2022.

[15] Wasim Bin Habib and Tuhin Shubhra Adhikary. English, maths drag results down again, May 2018.

[16] Yungwei Hao, Kathryn S. Lee, Szu-Ting Chen, and Sin Chie Sim. An evaluative study of a mobile application for middle school students struggling with English vocabulary learning. *Computers in Human Behavior*, 95:208–216, June 2019.

[17] Safarova Fotima Isamiddinovna. Mobile Applications As A Modern Means Of Learning English. pages 1–5, November 2019.

[18] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, June 2019.

[19] Anne Kao and Steve R. Poteet. *Natural Language Processing and Text Mining*. Springer Science & Business Media, March 2007. Google-Books-ID: CVtxFWbKT7wC.

[20] Blanka Klímová and Aleš Berger. Evaluation of the Use of Mobile Application in Learning English Vocabulary and Phrases – A Case Study. In Tianyong Hao, Wei Chen, Haoran Xie, Wanvimol Nadee, and Rynson Lau, editors, *Emerging Technologies for Education*, Lecture Notes in Computer Science, pages 3–11, Cham, 2018. Springer International Publishing.

[21] Jea Woog Lee, YoungBin Kim, and Doug Hyun Han. LDA-based topic modeling for COVID-19-related sports research trends. *Frontiers in Psychology*, 13, 2022.

[22] Philip M. McCarthy and Chutima Boonthum-Denecke, editors. *Applied Natural Language Processing: Identification, Investigation and Resolution*. IGI Global, 2012.

[23] Rastislav Metruk. The Use of Smartphone English Language Learning Apps in the Process of Learning English: Slovak EFL Students' Perspectives. *Sustainability*, 13(15):8205, January 2021.

[24] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

[25] Sergey Nikolenko. SVD-LDA: Topic Modeling for Full-Text Recommender Systems. In Obdulia Pichardo Lagunas, Oscar Herrera Alcántara, and Gustavo Arroyo Figueroa, editors, *Advances in Artificial Intelligence and Its Applications*, volume 9414, pages 67–79. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[26] Sergey Nikolenko. SVD-LDA: Topic Modeling for Full-Text Recommender Systems. In Obdulia Pichardo Lagunas, Oscar Herrera Alcántara, and Gustavo Arroyo Figueroa, editors, *Advances in Artificial Intelligence and Its Applications*, volume 9414, pages 67–79. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.

[27] Petra Poláková and Blanka Klímová. Mobile Technology and Generation Z in the English Language Classroom—A Preliminary Study. *Education Sciences*, 9(3):203, September 2019. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[28] Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. Short text topic modeling techniques, applications, and performance: a survey. *IEEE Transactions on Knowledge and Data Engineering*,

34(3):1427–1445, 2020.

[29] Karmila Rafiqah M. Rafiq, Harwati Hashim, and Melor Md Yunus. Sustaining Education with Mobile Learning for English for Specific Purposes (ESP): A Systematic Review (2012–2021). *Sustainability*, 13(17):9768, January 2021. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[30] Krishna Raj P M and Jagadeesh Sai D. Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques. *Materials Today: Proceedings*, 51:576–584, January 2022.

[31] Ruby Rani and D. K. Lobiyal. An extractive text summarization approach using tagged-LDA based topic modeling. *Multimedia Tools and Applications*, 80(3):3275–3305, January 2021.

[32] Star Report. Schools of 0, May 2018.

[33] Jacobijn Sandberg, Marinus Maris, and Kaspar de Geus. Mobile English learning: An evidence-based study with fifth graders. *Computers & Education*, 57(1):1334–1347, August 2011.

[34] M. Selvi, K. Thangaramya, M. S. Saranya, K. Kulothungan, S. Ganapathy, and A. Kannan. Classification of Medical Dataset Along with Topic Modeling Using LDA. In Vijay Nath and Jyotsna Kumar Mandal, editors, *Nanoelectronics, Circuits and Communication Systems*, Lecture Notes in Electrical Engineering, pages 1–11, Singapore, 2019. Springer.

[35] Wafa Shafqat and Yung-Cheol Byun. Topic predictions and optimized recommendation mechanism based on integrated topic modeling and deep neural networks in crowdfunding platforms. *Applied Sciences*, 9(24):5496, 2019.

[36] Mitchell Shortt, Shantanu Tilak, Irina Kuznetcova, Bethany Martens, and Babatunde Akinkuolie. Gamification in mobile-assisted language learning: a systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*, 36(3):517–554, March 2023. Publisher: Routledge _eprint: https://doi.org/10.1080/09588221.2021.1933540.

[37] Stefan Slater, Ryan Baker, Ma. Victoria Almeda, Alex Bowers, and Neil Heffernan. Using correlational topic modeling for automated topic identification in intelligent tutoring systems. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 393–397, Vancouver British Columbia Canada, March 2017. ACM.

[38] Bangladesh Bureau of Educational Information and Statistics. *Bangladesh Educational Statistics 2016*. Bangladesh Bureau of Educational Information and Statistics, first edition edition, January 2017.

[39] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[40] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[41] Han Wang, Kun Sun, and Yuwei Wang. Exploring the Chinese Public's Perception of Omicron Variants on Social Media: LDA-Based Topic Modeling and Sentiment Analysis. *International Journal of Environmental Research and Public Health*, 19(14):8377, January 2022. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

[42] Zhihong Xu, Zhuo Chen, Lauren Eutsler, Zihan Geng, and Ashlynn Kogut. A scoping review of digital game-based technology on English language learning. *Educational Technology Research and Development*, 68(3):877–904, June 2020.

[43] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval: 31th European Conference on IR Research, ECIR 2009, Toulouse, France, April 6-9, 2009. Proceedings 31*, pages 29–41. Springer, 2009.

[44] Huseyin oz. An Investigation of Preservice English Teachers' Perceptions of Mobile Assisted Language Learning. In *English Language Teaching, ERIC*, 8(2): 22–34, 2015.