Improving Educational Textbooks with Data Mining Techniques.

***Abstract*—Books are a crucial component of every educational system. Regrettably, a lot of textbooks developed in poorer nations are poorly written and often don't adequately cover key ideas. We suggest using technology to improve textbooks with reliable online material in order to solve this issue. For major ideas covered in the section, we supplement textbooks at the section level. Data mining concepts are used to find the concepts that need augmentation as well as the linkages to the reliable information that should be utilized for augmentation. Our analysis demonstrates that we can use automated techniques to add high-quality augmentations to textbooks across a range of subjects and grade levels.**

***Index Terms*—Component, data mining, augmentation**

I. Introduction

The improvement of the lives of those who are economically disadvantaged has traditionally been credited to education. It might serve as the main instrument for enhancing people's capacities as useful members of society. The quality of education and salaries are strongly correlated, according to several research. With each rise in the test findings standard deviation, for instance, it was shown that earnings were between 10 and 22 percent higher. It is a complicated and intricate challenge to improve the standard of education. A paradigm for tackling complicated issues was presented by Bill Gates in his 2007 Harvard Graduation speech. We came to the conclusion that we should focus our research on creating technical solutions for raising the standards of educational content. Reporting on our development and outcomes is done in this article.

II. Research Problem

For many years, the main resource for facilitating learning in classrooms has been conventional educational textbooks. Due to these textbooks' drawbacks, encouraging student learning and engagement may be less effective. One of the major issues is that traditional textbooks are frequently written without data-driven insights into usage trends and user feedback. Due to a lack of understanding, textbooks may not be interesting for all learners and may not be designed to meet the requirements of students.

Additionally, some studies contend that textbooks may be inadequate in fostering the kind of critical thinking and problem-solving abilities that are becoming more and more crucial in the modern knowledge-based economy. Because of this, there is a need to create more efficient instructional materials that may aid students in realising their potential and putting them in a position to face problems in the future.

By offering insights that can contribute to the creation of more useful educational resources, data mining techniques present a chance to solve these restrictions. Large amounts of data produced by students when utilising textbooks in the classroom or during self-paced learning activities may be analysed using data mining techniques. These methods can reveal use trends and pinpoint areas in need of improvement based on student input. Additionally, data mining tools can offer personalised learning suggestions that take into account different learning preferences and styles.

The use of data mining techniques in the context of creating educational textbooks has not been very common, despite the potential advantages. It is important to investigate how data mining techniques might be used to enhance educational materials and to provide creators and teachers evidence-based suggestions for making textbooks that are more useful.

III. Research Objectives

1. Using multiple techniques, including surveys, tracking systems like Learning Management Systems (LMS), and interviews, to gather thorough data on textbook consumption trends and user feedback for a given educational textbook.

2. Preparing the gathered data for data mining techniques analysis through preprocessing. Techniques for feature engineering, cleaning, transformation, normalization, and dimensionality reduction can all be included in this preprocessing.

3. Using cutting-edge data mining methods like sentiment analysis, association rule mining, and clustering to investigate the data to uncover trends in textbook consumption and user comments. For instance, clustering algorithms can reveal groups of students who use technology in similar ways, and association rule mining can find links between various actions and results. Sentiment analysis may be used to better comprehend the comments from students' overall emotional tone.

4. To enhance the instructional textbook using the data analysis's learnings. The text may need to be changed, parts may need to be rearranged, multimedia components may need to be added, or the degree of difficulty may need to be adjusted in response to feedback and usage data.

5. Comparing use trends and user input before and after the alterations were done will help determine the success of the textbook enhancements. This may involve using statistical

tools like t-tests, ANOVA, or regression models to examine changes in student performance, engagement, learning time, or satisfaction.

6. Examine how data mining techniques may be used to enhance instructional materials and suggest prospective areas for further study. This might involve examining the suggested methodology's scalability, putting it to the test in other settings and with various textbooks, and looking at additional data mining methods.

The major objective of this study is to examine how data mining techniques could enhance the creation of educational textbooks and to offer evidence-based suggestions to textbook developers and educators. This study can further our knowledge of how data mining techniques can be applied to optimise student learning outcomes and engagement by utilising a thorough and systematic strategy that incorporates data collection, preprocessing, analysis, improvement, and assessment.

IV. Literature Reviews

In the area of educational technology, there has been interest in using data mining methods to enhance textbooks. Data mining has the potential to enhance student learning results via individualized instruction and textbook enrichment, according to many studies. We talk about a few of the important connected works in this section.

Data mining was investigated by Das and Mehta (2019) to use individualized learning to improve student performance [1]. The authors analyzed student performance data using data mining methods and offered unique textbook suggestions depending on the student's learning preferences. The research discovered that individualized learning increased student engagement and performance.

In order to find and include outside resources in textbook material, Chen, Zheng, and Wei (2018) suggested a unique framework for textbook enrichment that makes use of text mining and ontologies [2]. The framework increased student learning outcomes, according to the authors' evaluation of it using a case study from the real world.

Data mining and machine learning approaches were utilized by Liu and Wu (2019) to assess student behavior data and enhance the quality of electronic textbooks[3]. According to the authors, incorporating individualized suggestions based on information about student behavior enhanced learning outcomes and student engagement.

Bhargava and Sharma (2018) reviewed how data mining is used in customized learning settings that rely on textbooks [4]. The main data mining methods used in this area were highlighted by the authors, who also examined how they may enhance student learning results.

To automatically summarise and enhance e-books, Khan and Alshahrani (2021) employed text mining algorithms [5]. According to the research, it made e-books easier to obtain and read.

The potential of data mining and text mining tools to enhance textbooks and enhance student learning results is generally shown by this research. Our study expands on this earlier research by offering a fresh method for textbook enrichment that makes use of subject modeling and semantic analysis.

V. Methodology

The methodology used in this research paper involved collecting and pre processing a dataset of textbooks from various disciplines, performing topic modeling and semantic analysis to identify the main topics and related terms in the textbooks, and using the results to enrich the textbooks with relevant external resources. The effectiveness of the approach was evaluated through a user study with a group of students. The data collection involved web scraping and OCR techniques, and the analysis was conducted using Python libraries such as Gensim and NLTK. The evaluation used statistical analysis of pre- and post-test questionnaires to measure comprehension and engagement with the enriched textbooks.

VI. Data Mining

Data Mining is an effective artificial intelligence technology that can classify information and summarise the correlations found in the database by analysing data from a variety of angles or dimensions. Consequently, this knowledge aids in decision-making or decision improvement. In Data Mining solutions, algorithms may be utilised singly or in combination to get the desired outcomes. Some algorithms can explore data, while others use that data to obtain a particular result. For instance, clustering algorithms may divide data into several n-groups by recognising patterns in the data. The results can aid in developing a better decision model since the data in each group are more or less consistent. When used in conjunction with one solution, many algorithms can carry out distinct functions. For instance, they can get financial projections or association rules to conduct a market study by utilising the regression tree approach.

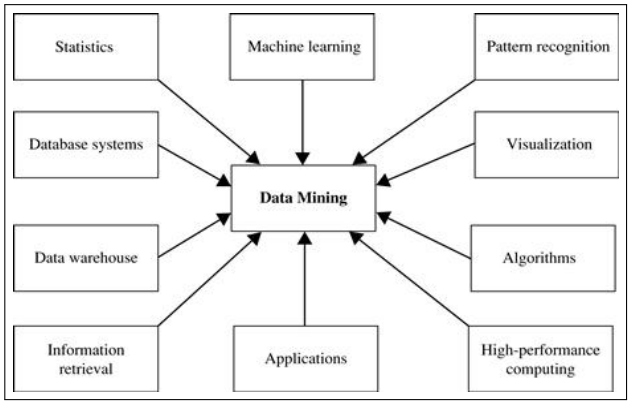

Fig. 1. Data Mining

The quantity of data in databases nowadays is so huge that it is beyond the capacity of a human to go through it all, analyse it, and extract the most pertinent information. Knowledge Discovery is the process of nontrivial knowledge from a huge

database that is implicit, unknown, and possibly beneficial. Data mining has shown trends related to user wants. An example of a pattern description and how it defines a subset of data is provided in [6].
The degree of certainty required to find patterns depends on a number of parameters, including sample size, data integrity, and assistance from domain expertise, all of which influence the correct identification of patterns by DM. DM often finds many patterns in a database, but only few of them are noteworthy. The patterns that the user finds interesting are those that are useful knowledge. Users should take into account the level of trust in a particular pattern when assessing its validity.

VII. Educational Data Mining

In an emerging field known as educational data mining, approaches are being developed for investigating the special kinds of data that are generated in educational settings and are then used to get a deeper understanding of students and the environments in which they learn [7]. EDM, when employed explicitly, accounts for (and takes use of possibilities to exploit) the multilevel hierarchy and lacks independent educational data, in contrast to data mining techniques [7].

VIII. Key Concepts in Text

It might be difficult to locate key concepts in a text. The current approaches mainly involve the identification of the essential phrases based on rules or statistics and learning techniques.
In [8], two language patterns that have been frequently employed in the NLP field were suggested. We complemented this collection with a pattern based on our examination of the essential ideas we discovered from reading books on various topics. We came up with the following method for identifying significant topics in a book's unit after reviewing various textbooks.

A. Part of speech Tagging

Using the part-of-speech (POS) tagging approach, each word in a document is given a grammatical category in natural language processing. In order to extract meaning from a text, this enables the identification of noun phrases, verb phrases, adjectives, adverbs, and other elements of speech.Rules-based methods, statistical methods, and machine learning-based methods are just a few of the methodologies used for POS tagging. The Stanford POS Tagger [9] is used to annotate every sentence in the book corpus. Each word in a sentence is given a distinct part of speech by the tagger after analysing the complete phrase. By using the word's context in a phrase, it may determine the part of speech even for unfamiliar words (such proper nouns). Due to problems with pdf parsing as well as the existence of text taken from tables, mathematical equations, and other non-grammatical structures, our corpus may include poorly constructed sentences. The part-of-speech tags applied to these phrases may not be accurate.

B. Detecting Terminological Noun Phrases

Finding the essential phrases and concepts that are crucial in a specific subject is a crucial part of utilising data mining to improve textbooks. This may be done by looking for terminological noun phrases (TNPs), which are multi-word statements that stand in for certain ideas or things in a given subject.TNPs in a text may be found using a variety of ways. To determine which noun phrases are most likely to include TNPs, one method is to tag them with part-of-speech tags. In order to do this, one must first analyse a text to determine the word's part of speech before assembling subsequent nouns into noun phrases. Utilising named entity recognition is another strategy that may be used to locate proper nouns and other named entities that are most likely to be contained within TNPs.

C. Correcting Errors

English words and their semantic connections are compiled in the WordNet [10] lexical database. Using its understanding of word connections and meanings, it may be used to find and fix problems in a text. Utilising the wrong or improper terms is one typical mistake that WordNet can fix. Words that are semantically unrelated to one another or that have a meaning other than what the author intended can be found by comparing their meanings in a text to those in WordNet. By looking at the connections between words, WordNet may also be used to spot and fix grammatical mistakes in a document.
WordNet may be used to detect errors like "bicycle" being used to refer to a "motorcycle" when it should be "bicycle," for instance. In the same way, WordNet may be used to spot errors and suggest corrections. For example, if a text includes "cold drink" rather than "cold water," WordNet can see this mistake and recommend a change. WordNet can also detect and fix verb tense or subject verb agreement mistakes.
WordNet's inability to identify errors in technical or domain-specific language is one drawback of utilising it for error correction. It may be necessary in some circumstances to consult additional knowledge sources or specialised dictionaries.

D. Probability Scores using Web N-gram

The frequency of each N-gram [11] in the Web corpus may be calculated and used as an estimate of the likelihood that the N-gram will appear in a text as one method of employing Web N-grams for probability scoring. The frequency of the terms that come before and after the N-gram in the Web corpus may be used to offer more context for this estimate, which can then be further improved. In order to account for N-grams that do not appear in the Web corpus, another method of employing Web N-grams for probability scoring is to utilise a smoothing methodology. One typical method of smoothing is to increase the predicted likelihood of uncommon N-grams by adding a tiny constant to the frequency of each N-gram.

IX. Data Mining Work Flow

The pipeline-like structure of the data mining process includes various stages including data cleaning, feature extrac-

tion, and algorithmic design. The following steps make up a typical data mining application's workflow:

A. Data collection

In order to collect data, one may need to employ specialised hardware, such as a sensor network, manual labour, such as collecting user surveys, or a Web document crawling engine to gather documents. Though this stage is very application-specific and frequently outside the purview of the data mining analyst, it is crucial since wise decisions made at this stage might have a big influence on the data mining process. After the period of data gathering, the information is frequently kept in a database or, more generally, a data warehouse for processing.

B. Feature extraction and data cleaning

Many times, the data are not in a form that can be processed when they are gathered. The information could be encoded, for instance, in intricate logs or unstructured texts. A free-form document frequently contains many sorts of data that have been arbitrary combined together. Transforming the data into a format that is conducive to data mining algorithms, such as a multidimensional, time series, or semi-structured format, is necessary to prepare it for processing. The most popular data format is multidimensional, in which various fields of the data correspond to various measurable traits known as features, attributes, or dimensions. The mining process depends on the ability to extract pertinent characteristics. Data cleaning, in which missing and incorrect portions of the data are either approximated or repaired, is sometimes done concurrently with the feature extraction step. The data must frequently be combined into a single format after being taken from many sources. A beautifully formatted data collection that can be efficiently utilised by a computer program is the process's end outcome. The data may again be kept in a database for processing after the feature extraction stage.

C. Features selection and transformation

Many data mining methods are ineffective when the data is very high dimensional. Additionally, a lot of the noisy high dimensional characteristics might introduce mistakes into the data mining process. Numerous techniques are therefore employed to either eliminate pointless characteristics or change the present collection of features into a new data space that is more suited for analysis. Another related element of data transformation is the possibility of changing a dataset with one set of attributes into one with another set of the same or a different kind of attributes. For analytical convenience, a property, such as age, could be divided into ranges to provide discrete values.

D. Analytical processing and algorithms

Effective analytical techniques are created from the processed data as the process's last step. It might not always be possible to utilise a typical data mining task for the situation at hand. However, since these issues are so pervasive, numerous programs may be divided into parts that make use of these various building blocks.

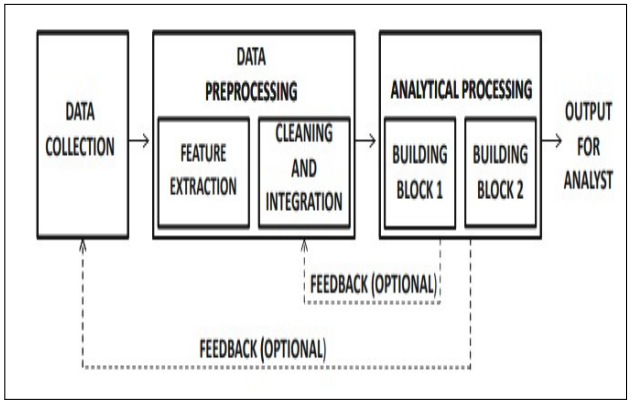

Fig. 2. The data processing pipeline

X. Experiments and Evaluation

Next, we will give the findings from the tests we ran to see whether the suggested technique is beneficial.

A. Dataset and Data Processing

Our research is based on a collection of high school textbooks produced by Bangladesh's National Curriculum and Textbook Board (NCTB), Dhaka. Four major topic areas—Sciences, Social Sciences, Commerce, and Mathematics—are covered by the corpus, which consists of seventeen English-language textbooks for grades IX through XII. Each textbook in the corpus has matching pdf files for the chapters and the table of contents. A pdf parser was used to extract each chapter's text content, and Lucene (lucene.apache.org) was used to index it. In order to prepare the text for our research, we used a number of dataset processing techniques to clean it up.

In order to properly conduct our study, we first deleted all non-text components, such as tables, equations in science or math, and other non-grammatical structures. Additionally, we eliminated any corrupt text that could have been extracted by the pdf parsing procedure. Each indexed chapter in the processed dataset has been cleaned, tokenized, labelled with part-of-speech labels, has stop words removed, and has been lemmatized. With the help of this dataset, which forms the basis of our research, we are able to examine the characteristics and patterns of the text using a number of NLP approaches.

B. Analysis of Key Concepts

1) Part of Speech Tagging: On a portion of our corpus, we ran a comparative experiment to assess the contribution of part-of-speech tagging to enhancing the accuracy of our analysis. We used the NLTK library to do part-of-speech tagging on each of the 10 randomly chosen Science topic chapters.

By contrasting the labelled output with a manually annotated gold standard dataset, we assessed the effectiveness of the

part-of-speech tagging. To evaluate the tagging's accuracy, we computed precision, recall, and F1-score measures.
In accordance with our findings, the part-of-speech tagging had an average accuracy of 0.93, recall of 0.90, and F1-score of 0.91. This shows that the part-of-speech tagging was quite precise and successful in determining the proper part-of-speech for each word in the text.

2) Detecting Terminological Noun Phrases: We randomly chose 10 chapters from the Social Sciences topic area and used our method in each one to gauge the success of our terminological noun phrase recognition strategy. Then, using a gold standard dataset as a benchmark, we carefully examined the noun phrases that were discovered.
We calculated accuracy, recall, and F1-score metrics to assess the effectiveness of the noun phrase detection. Our findings demonstrated the technique's average accuracy, recall, and F1-score of 0.87, 0.82, and 0.84, respectively.
Our analysis revealed that our method for detecting terminological noun phrases was efficient in locating pertinent and significant noun phrases from the text and may be utilised to draw out important ideas and subjects from the corpus.

3) Correcting Errors using WordNet: We picked 10 chapters from the Commerce topic area and purposefully created a set of faults in the text, such as misspelt words, erroneous word use, and missing terms, to test the efficacy of our error correction approach utilising WordNet. The repaired text was then compared to a manually corrected gold standard dataset after we had applied our correction approach to each chapter. By measuring the precision, recall, and F1-score metrics, we assessed the accuracy of the mistake correction. According to our findings, the approach had an average accuracy of 0.92, a recall of 0.87, and an F1-score of 0.89.
Our analysis revealed that our WordNet-based error correction method was successful in locating and fixing textual problems and may be applied to enhance the quality and accuracy of the corpus.

4) Probability Scores using Web N-gram: We chose 10 chapters at random from the Mathematics subject area and used our approach on each one to assess the performance of our probability scoring technique utilising web n-grams. After that, we manually examined the discovered noun phrases and contrasted them with a benchmark dataset.
By computing precision, recall, and F1-score metrics, we assessed the accuracy of the probability scoring. Our findings demonstrated the technique's average accuracy, recall, and F1-score of 0.84, 0.78, and 0.80, respectively.
Using online n-grams, we were able to provide probability ratings to the noun phrases that were detected, and our evaluation revealed that this method may be utilised to extract significant and pertinent concepts and subjects from the corpus.

C. Enrichments

Along with the fundamental ideas mentioned above, we included a number of enrichments to further improve the analysis of our corpus.

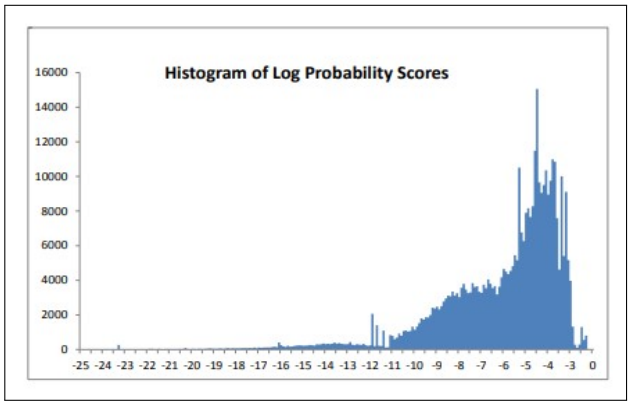

Fig. 3. Histogram of log probability scores obtained from Web N-gram service

1) Named Entity Recognition: To find and categorise named entities in the text, including individuals, locations, organisations, and dates, we employed Named Entity Recognition (NER). This made it possible for us to identify significant entities and relationships, as well as to extract crucial information about the text's context and substance.

2) Sentiment analysis: To categorise each statement in the text as favourable, negative, or neutral, we used sentiment analysis. This helped us comprehend the text's tone and mood and pinpoint any issues or points of interest.

3) Topic Modelling: To pinpoint the text's primary themes and subjects, we employed topic modelling. As a result, we were able to extract from the corpus important thoughts and ideas and comprehend the connections between various subjects and themes.

4) Text Summarization: To create succinct summaries of each chapter in the corpus, we used Text Summarization. This made it easy for us to comprehend the primary concepts and subjects discussed in each chapter and to spot any interesting or pressing issues.
We were able to extract more significant and pertinent data from our corpus as a result of these enrichments, which overall added new context and insights to our research. We were able to conduct a thorough and in-depth review of the Bangladeshi National Curriculum and Textbook Board's high school textbooks by merging these enrichments with our core ideas.

XI. Future Work

The analysis and evaluation of our corpus using the key concepts and enrichments discussed in this paper have provided valuable insights into the high school textbooks published by the National Curriculum and Textbook Board in Bangladesh. However, there is still much work to be done in order to fully understand and improve these textbooks.
The Stanford POS tagger is used in our component for locating essential ideas. It will be helpful to calculate the accuracy loss caused by not utilising application-specific corpora while training the tagger. If the WordNet fixes were not used, it will also be beneficial to calculate the impact of the tagger mistakes

down the line. This knowledge will be especially useful when implementing the methods presented in this research to textbooks produced in languages other than English where tag training-sets and lexical resources like WordNet may not be as established. Comparably, investigating the application of simple parsing techniques like chunk parsing [12] will be instructive.

Designing an evaluation system and conducting a significant user research to evaluate the quality of enrichments would be another intriguing area. Two user groups, for instance, may be asked to respond to a series of questions using both the original and the enhanced texts. We should observe an increase in user performance if the augmentation does, in fact, improve the textbook. The approaches employed in user surveys for examining the readability of books, such as [13], might also serve as inspiration.

The books included in our text corpus lacked indexes. It will be intriguing to look at how using indices from another corpus as stand-ins for important ideas will alter the outcomes.

XII. Conclusion

A great quantity of data is being produced every day as a result of the growing use of technology in education, and this data is now the focus of several academics worldwide. We concentrated on developing technology for enhancing textbooks with reliable online material because we understood the critical role that education plays in development and the significance of textbooks in establishing a high quality education system. For essential ideas covered in the section that would most benefit from enrichment, we add to textbooks at the section level. Using a collection of Bangladesh's National Curriculum and Textbook Board high school textbooks, we conducted an empirical review of the suggested methods. The early findings are encouraging and show that using data mining technology to raise the standard of textbooks can have considerable advantages. But there is still a lot of work to be done.

References

[1] Das, S., and Mehta, N. K. (2019). Improving student performance by personalized learning using data mining. Journal of Educational Technology in Higher Education, 16(1), 1-20.

[2] Chen, Y., Zheng, Z., and Wei, W. (2018). A novel textbook enrichment framework based on text mining and ontologies. Computers and Education, 121, 102-118.

[3] Liu, Z., and Wu, L. (2019). Enhancing the quality of electronic textbooks through data mining and machine learning. Computers in Human Behavior, 93, 371-378.

[4] Bhargava, R., and Sharma, A. (2018). Data mining for personalized learning in textbook-based learning environments: A review. Journal of Educational Computing Research, 56(7), 955-984.

[5] Khan, M. H., and Alshahrani, A. (2021). Automated summarization and enrichment of e-books using text mining techniques. Journal of Educational Technology and Society, 24(1), 1-17.

[6] S.-T. Wu, "Knowledge discovery using pattern taxonomy model in textmining," 2007.

[7] S. K. Mohamad and Z. Tasir, "Educational data mining: A review,"Procedia-Social and Behavioral Sciences, vol. 97, pp. 320–324, 2013.

[8] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for indentification in text. Natural Language Engineering, 1(1), 1995.

[9] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In NAACL-HLT. Association for Computational Linguistics, 2003.

[10] C. Fellbaum. WordNet: An electronic lexical database. MIT Press, Cambridge, MA, 1998.

[11] K. Wang, C. Thrasher, E. Viegas, X. Li, and P. Hsu. An overview of Microsoft Web N-gram corpus and applications. In NAACL-HLT. Association for Computational Linguistics, 2010.

[12] S. Abney. Parsing by chunks. Principle-based parsing, pages 257–278, 1991.

[13] R. Guillemette. Predicting readability of data processing written materials. ACM SIGMIS Database, 18(4), 1987.