**SAFE READ: A Keyword Based Classification System For Diverse Content, Recommending Age-Appropriate Reads**

By

**Mahmuda Afrin Tuli**

Roll: 1907019

&

**Sonjoy Roy**

Roll:1907073



**Supervisor:**

Farhan Sadaf

Lecturer

Dept. of Computer Science and Engineering Signature

Khulna University of Engineering & Technology

**Department of Computer Science and Engineering**

**Khulna University of Engineering & Technology**

**Khulna 9203, Bangladesh**

**November 2023**

**Acknowledgements**

First and foremost, we give thanks to Allah, the Almighty, who created us and will always be with us. We have completed our System with the mercy of Allah Almighty. I am grateful to Farhan Sadaf, Lecturer, Department of Computer Science and Engineering, Khulna University of Engineering and Technology, for guiding and directing us through the completion of this project. He made materials and tools available to us that will aid in the improvement of our system project work. He was a well-behaved, well-disciplined individual who assisted us in becoming more responsible. Next, I want to express my gratitude to all of my teachers who have shared their knowledge with us, which has aided us in completing our system project. Also, a big thank you to my classmates and seniors for sharing their knowledge and always giving support, which helped us finish this System project.

**Authors**

**Abstract**

The widespread and increasing availability of Bangla content increases the importance of the safety of under aged individuals from abusive content. Our system offers a keyword-based classification for political, religious, and abusive contents, as well as a recommendation system for each category. We have developed a significant collection of Bangla text that falls into three categories: abusive (3477 keywords), political (494 keywords), and religious (1411 keywords). Of these, only the religious and political datasets are publicly available for usage as abusive datasets might be exploited to create improper content. This dataset is used to match keyword with the given content and calculate word frequency of each category which is further used for recommendation. By employing this system, users, especially parents and guardians, can confidently navigate through a plethora of reading options and identify materials suitable for specific age brackets. Our project correctly predicted the class 90% of the time, and the recommendation algorithm relied heavily on this classification. This tailored approach not only enhances the overall user experience but also serves as a protective measure against exposing individuals, particularly minors, to content that may be inappropriate or unsuitable for their maturity level.

**Contents**

|  |  |  |
| --- | --- | --- |
|  |  | **Page** |
| Acknowledgement |  | ii |
| Abstract |  | iii |
| Contents |  | iv |
| List of Tables |  | vi |
| List of Figures |  | vii |

|  |  |  |
| --- | --- | --- |
| **1** | **Introduction** | 1 |
|  | * 1. Background   2. Problem Statement | 1  1 |
|  | 1.2 Objectives | 1 |
|  | 1.3 Scopes | 2 |
|  | 1.4 Unfamiliarity of the Problem | 2 |
|  | 1.5 Project Planning | 2 |
|  |  |  |
| **2** | **Related Works** | 3 |
|  | 2.1 Existing Solutions | 3 |
|  | 2.2 Limitation in Existing Solutions | 4 |
|  |  |  |
| **3** | **System Design** | 5 |
|  | 3.1 Dataset Preparation | 6 |
|  | 3.2 Class Recognition  3.2.1 Text Preprocessing  3.2.2 Feature Extraction  3.2.3 Classification  3.3 Recommendation | 6  6  10  12  15 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| **4** | **Project Implementation** | 16 |
|  | 4.1 Results  4.1.1 Dataset  4.1.2 Class Recognition  4.1.3 Recommendation  4.2 Tools / Platform used | 16  16  18  19  21 |
|  | 4.2.1 Google Colab | 21 |
|  | 4.2.2 Streamlit | 22 |
|  | * 1. Graphical User Interface   2. Use Case Diagram   3. Morality or Ethical Issues   4. Socio-economic Impact and Sustainability   5. Financial Analysis and Budget | 22  26  26  26  27 |
|  |  |  |
| **5** | **Conclusions** | 28 |
|  | 5.1 Challenges  5.2 Conclusion | 27  28 |
|  | 5.3 Future Study | 28 |
|  |  |  |
|  |  |  |
| **6** | **References** | 29 |
|  | Appendices |  |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Description** | **Page** |
| Table 3.1: | Example of Text Cleaning | 7 |

Table 3.2: Example of Stopwords Removal 7

Table 3.3:List of Stopwords in Bangla Language 8

Table 3.4: Example of Stemming 9

Table 3.5:Example of Word Tokenization 9

Table 3.6: Example of Bangla Word N-Gram 11

Table 3.7:Example of Matched Word after Applying Unigram. 12

Table 3.8: Example of Matched Word after Applying Bigram 13

and Trigram.

Table 4.1: The Books Name from Which Keywords Were Collected 16

Table 4.2: Total Count of Keywords in Each Category. 16

Table 4.3: Religious Keywords 17

Table 4.4:Political Keywords 17

Table 4.5: Abusive Keywords 18

Table 4.6: Class Recognition for Different Bangle Books 18

Table 4.7: Class Recommendation for Different Bangle Books 19

Table 4.8: Financial analysis and budget of the project 27

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Description** | **Page** |
| Figure 1.1: | Gantt Chart of Safe Read System Project | 3 |
| Figure 3.1:  Figure 3.2:  Figure 3.3:  Figure 4.1:  Figure 4.2:  Figure 4.3:  Figure 4.4:  Figure 4.5:  Figure 4.6:  Figure 4.7:  Figure 4.8:  Figure 4.9: | Workflow Diagram of Dataset Preparation  Work Flow Diagram of Class Recognition  Flow Chart for Recommendation algorithm  Initial User Interface  Select the Input file (.txt .docx .pdf)  File has been uploaded. Now click the button to convert to digital text  Total Count of Religious Keyword, Political Keyword and  Abusive Keyword  Frequently arrived words in Religious Keyword of Uploaded File  Frequently arrived words in Political Keyword of Uploaded File  Frequently arrived words in Abusive Keyword of Uploaded File  Pie Chart for Showing the Percentage of Each Category  Use Case Diagram of This Project. | 6  14  15  22  22  23  23  24  24  25  25  26 |

1. Introduction
   1. Background

The purpose of this project is to perform a keyword-based classification on different Bangla contents. The Keyword Based Classifier is a simple classifier that searches for repeating string sequences within a given file, in order to perform document classification. There are no websites that can classify and recommend safe Bangla content for under aged individuals. To protect our minors from abusive content we developed our system as a keyword-based classifier, which classifies contents into three categories: political, religious, and abusive. Based on the intensity of the words of different categories we set a threshold for recommending contents for minors.

* 1. Problem Statement

Text categorization is an active research area of text mining where the documents are classified with supervised, unsupervised or semi-supervised knowledge. Traditionally, this task is solved manually, but such manual classification is expensive to scale and also labor intensive. Besides, Sentiment analysis or opinion mining has been quite popular and has led to building of better products [1], but still review analysis can go wrong sometimes. If we classify the Bangla content by evaluating the words of a content, that will be more efficient for classifying and recommending for different ages.

There is an issue of dealing with various types of content such as docx, pdf, and txt files, as well as written text. These need to be extracted and stored in a data frame for further use. Converting pdf files to data frames can be challenging as it requires a significant amount of time to complete.

* 1. Objectives

The fundamental purposes of our system is given below:

1. Review existing methods for classifying different types of Bangla content.

2. Apply our proposed preprocessing and highly modified dataset to classify the content.

3. Implement uni-gram, bi-gram, and tri-gram based comparison techniques as discussed.

4. Apply the proposed comparison algorithm to the Bangla dataset.

5. Develop an end-to-end system for online Bangla content classification and recommendation.

6. Design user-friendly web applications for optimal user experience.

7. Expand the principles of online keyword-based classification techniques to real-world scenarios.

8. Review the challenges we faced.

* 1. Scope

Our system, this project, is an innovative keyword-based classification system that recommends age-appropriate reads and ensures a safe reading environment for users of all ages. It's designed to provide a comprehensive and user-friendly approach to content categorization. Our primary focus is on addressing the need for a reliable and efficient content classification system that can help users find books that are suitable for their age and interests. By implementing our innovative system, we aim to make the reading experience safe and enjoyable for everyone. Our system is designed to address the need for a comprehensive and user-friendly approach to content categorization. The primary focus is on ensuring a safe reading environment for users of all ages by implementing an innovative keyword-based classification system.

* 1. Unfamiliarity of the Problem

This project's main attraction lies in its keyword-based classification system. Bangla content may contain a variety of words, including abusive, political, religious, and neutral ones. However, our system is designed to filter out politically/religiously/abusively charged content only. For keyword-based classification, simple comparison algorithms work effectively, especially when the dataset has enough key-words of each category.

* 1. Project Planning

The project planning consists of requirement analysis, study of techniques used, Improving UI/UX, Implementation of the whole system and other sections. The relevant details are shown in the following Figure 1.1.

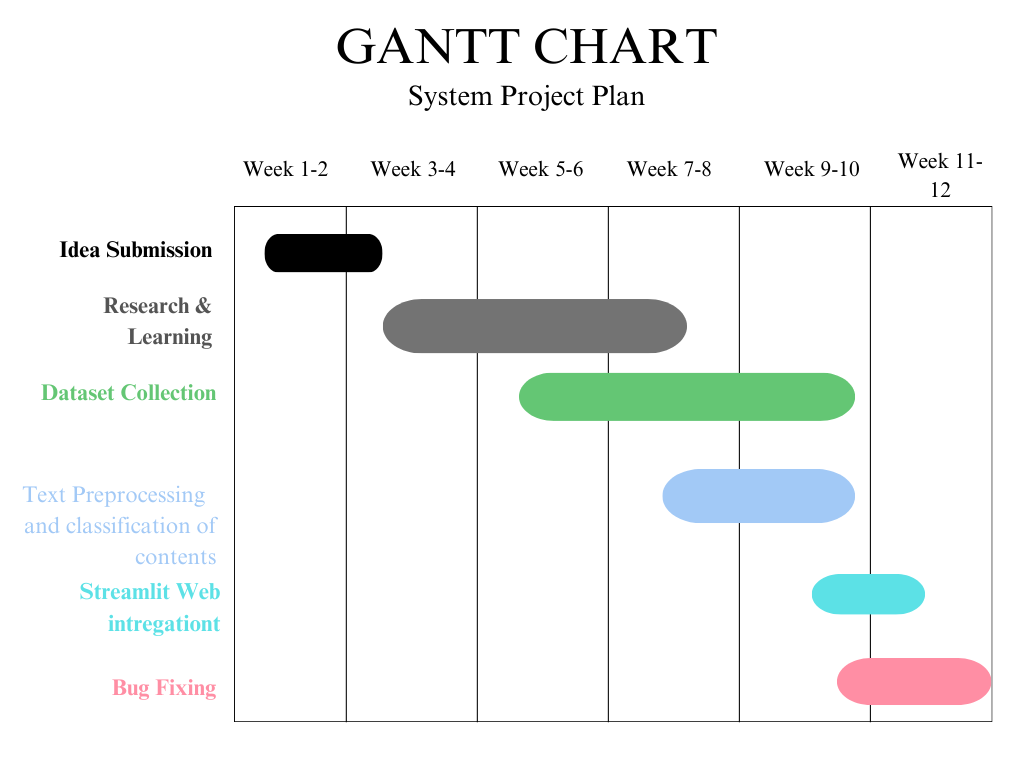


Figure 1.1: Gantt chart of Safe Read System Project.

1. Related Works
   1. Existing Solutions

There a Numerous research-based works have been conducted for the English language. A literature review shows that many established supervised learning algorithms have been implemented for text or document categorization. Efficient English text classification using selected machine learning techniques [11]. The most frequently used techniques are K-Nearest Neighbor (KNN), Naive Bayes (NB), N-Grams, Decision Tree (DT), Neural Network (NNet), and Support Vector Machines (SVM). NB is the most frequently used approach for text classification due to its simplicity in implementation and computation. In a study, NB was used to categorize the content on a website, which resulted in an accuracy of 80% [12]. EDA, Python script for preprocessing Bangla text dataset. Includes Bangla stop-words, Siyam Rupali font for Bengali language, and text categorization for five categories: sports, international, entertainment, state, economy, and education [12].

* 1. Limitation in Existing Solutions

A popular approach for recommending books involves analyzing the reader's review and assuming the sentiment of the book before recommending it to an audience. However, this process is time-consuming and not always reliable. Since everyone has different thoughts and opinions, relying solely on the review system may not be the best option.

Additionally, the classification method by sentences is not always accurate. Our system introduces a new and effective approach for categorizing books based on keywords. For example, political books will contain more political keywords while religious books will have more religious keywords. This keyword-based categorization system adds a new dimension for recommending books or content and can help users find exactly what they are looking for and what content they should recommend to their children.

1. System Design

The system was designed initially with in mind the fact that it can be scalable and adaptable by the common people. This system can be applied in our e-book websites for classifying books and also adaptable for future generation. It can also be used for blocking or removing abusive content from any social media site. The system was designed by following these three steps:

1. Dataset Preparation
2. Class Recognition
3. Recommendation
   1. Dataset Preparation

The dataset was prepared manually. To be more specific, keywords were collected from three categories of books, namely religious, political, and abusive. Firstly, we give a pdf file of the book as input. The “pdf2image” library is used to convert the PDF file into a list of PIL Images. Then the concurrent.futures, ThreadPoolExecutor [14] is employed to process the images in parallel for reducing the time for Bangla pdf to text conversion. And compiles the results into a Pandas Data Frame. The script efficiently handles the conversion and text extraction tasks in parallel, enhancing overall performance [5]. The final Data Frame contains page numbers and their corresponding extracted text. Finally, keywords were manually extracted from the text file and, if deemed valid, were stored in an MS Excel file and further downloaded in .csv format.



Figure 3.1: Workflow Diagram of Dataset Preparation.

* 1. Class Recognition
     1. Text Preprocessing

Text preprocessing is an indispensable step in the text classification task, particularly in preparing the text data to be effectively utilized as input for classification. In the context of Bangla text classification, several commonly employed preprocessing techniques include:

Text Cleaning

In the realm of text classification, be it in Bengali or any other language, the significance of eliminating punctuation and emoticons cannot be overstated. This is because these elements generally do not add much to the semantic meaning of a sentence. By removing them, the quality and relevance of the textual data can be enhanced, and unnecessary noise can be reduced, allowing the classifier to concentrate on the core textual content.

**Table 3.1:** Example of Text Cleaning.

|  |  |
| --- | --- |
| **Raw Text** | **Clean Text** |
| /n/n রাজনীতির সম্পর্কে আলোচনা করা যাক, এবং ধর্মীয় সহিত সকল দৃষ্টিভঙ্গি মনোনিবেশ করা গুরুত্বপূর্ণ ,শালার বেটাকে এবং কুত্তার বাচ্চাকে মার ধরে  ??। 🗳️🙏 | রাজনীতির সম্পর্কে আলোচনা করা যাক এবং ধর্মীয় সহিত সকল দৃষ্টিভঙ্গি মনোনিবেশ করা গুরুত্বপূর্ণ শালার বেটাকে এবং কুত্তার বাচ্চাকে মার ধরে |

**Stopwords Removal**

In the Bangla language, there exist common words, such as "এবং", "অথবা"," অতএব", which are referred to as stopwords. These words do not contribute significantly to the meaning of a sentence or phrase and are often excluded during text analysis. To facilitate this process, a comprehensive list of 736 Bangla stopwords [6] was created. During text analysis, each word in the text data is compared to the stopword list, and if a word is found to be a stopword, it is removed from the text data. This ensures that the resulting analysis is more accurate and meaningful.

**Table 3.2:** Example of Stopwords Removal.

|  |  |
| --- | --- |
| **Cleaned Text** | **Text After** **Stopwords Removal** |
| রাজনীতির সম্পর্কে আলোচনা **করা** যাক **এবং** ধর্মীয় সহিত সকল দৃষ্টিভঙ্গি মনোনিবেশ **করা** গুরুত্বপূর্ণ শালার বেটাকে **এবং** কুত্তার বাচ্চাকে মার **ধরে** | রাজনীতির সম্পর্কে আলোচনা যাক ধর্মীয় সকল দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ শালার বেটাকে কুত্তার বাচ্চাকে মার |

**Table 3.3:** List of Stopwords in Bangla Language.

['অই', 'অগত্যা', 'অত: পর', 'অতএব', 'অথচ', 'অথবা', 'অধিক', 'অধীনে', 'অধ্যায়', 'অনুগ্রহ', 'অনুভূত', 'অনুযায়ী', 'অনুরূপ', 'অনুসন্ধান', 'অনুসরণ', 'অনুসারে', 'অনুসৃত', 'অনেক', 'অনেকে', 'অনেকেই', 'অন্তত', 'অন্য', 'অন্যত্র', 'অন্যভাবে', 'অন্যান্য', 'অপেক্ষাকৃতভাবে', 'অবধি', 'অবশ্য', 'অবশ্যই', 'অবস্থা', 'অবিলম্বে', 'অভ্যন্তরস্থ', 'অর্জিত', 'অর্থাত', 'অসদৃশ', 'অসম্ভাব্য', 'আইন', 'আউট', 'আক্রান্ত', 'আগামী', 'আগে', 'আগেই', 'আগ্রহী', 'আছে', 'আজ', 'আট', 'আদেশ', 'আদ্যভাগে', 'আন্দাজ', 'আপনার', 'আপনি', 'আবার', 'আমরা', 'আমাকে', 'আমাদিগের', 'আমাদের', 'আমার', 'আমি', 'আর', 'আরও', 'আশি', 'আশু', 'আসা', 'আসে', 'ই', 'ইচ্ছা', 'ইচ্ছাপূর্বক', 'ইতিমধ্যে', 'ইতোমধ্যে', 'ইত্যাদি', 'ইশারা', 'ইহা', 'ইহাতে', 'উক্তি', 'উচিত', 'উচ্চ', 'উঠা', 'উত্তম', 'উত্তর', 'উনি', 'উপর', 'উপরে', 'উপলব্ধ', 'উপায়', 'উভয়', 'উল্লেখ', 'উল্লেখযোগ্যভাবে', 'উহার', 'ঊর্ধ্বতন', 'এ', 'এপর্যন্ত', 'এঁদের', 'এঁরা', 'এই', 'এইগুলো', 'এইভাবে', 'এক', 'একই', 'একটি', 'একদা', 'একবার', 'একভাবে', 'একরকম', 'একসঙ্গে', 'একা', 'একে', 'এক্', 'এখন', 'এখনও', 'এখনো', 'এখানে', 'এখানেই', 'এছাড়াও', 'এটা', 'এটাই', 'এটি', 'এত', 'এতটাই', 'এতদ্বারা', 'এতে', 'এদিকে', 'এদের', 'এপর্যন্ত', 'এবং', 'এবার', 'এমন', 'এমনকি', 'এমনকী', 'এমনি', 'এর', 'এরকম', 'এরা', 'এল', 'এলাকায়', 'এলাকার', 'এস', 'এসে', 'ঐ', 'ও', 'ওঁদের', 'ওঁর', 'ওঁরা', 'ওই', 'ওকে', 'ওখানে', 'ওদের', 'ওর', 'ওরা', 'ওহে', 'কক্ষ', 'কখন', 'কখনও', 'কত', 'কবে', 'কম', 'কমনে', 'কয়েক', 'কয়েকটি', 'করছে', 'করছেন', 'করতে', 'করবে', 'করবেন', 'করলে', 'করলেন', 'করলো', 'করা', 'করাই', 'করাত', 'করার', 'করায়', 'করি', 'করিতে', 'করিয়া', 'করিয়ে', 'করে', 'করেই', 'করেছিল', 'করেছিলেন', 'করেছে', 'করেছেন', 'করেন', 'কর্তব্য', 'কাউকে', 'কাছ', 'কাছাকাছি', 'কাছে', 'কাজ', 'কাজে', 'কারও', 'কারণ', 'কারণসমূহ', 'কারো', 'কি', 'কিংবা', 'কিছু', 'কিছুই', 'কিছুটা', 'কিছুনা', 'কিনা', 'কিন্তু', 'কিভাবে', 'কী', 'কূপ', 'কে', 'কেউ', 'কেউই', 'কেউনা', 'কেখা', 'কেন', 'কেবল', 'কেবা', 'কেস', 'কেহ', 'কোটি', 'কোথা', 'কোথাও', 'কোথায়', 'কোন', 'কোনও', 'কোনো', 'ক্রম', 'ক্ষেত্রে', 'কয়েক', 'কয়েকটি', 'খুঁজছেন', 'খুব', 'খোলা', 'খোলে', 'গড়', 'গত', 'গিয়ে', 'গিয়েছিলাম', 'গিয়েছে', 'গিয়ে', 'গিয়েছে', 'গুরুত্ব', 'গুলি', 'গেছে', 'গেল', 'গেলে', 'গোটা', 'গোষ্ঠীবদ্ধ', 'গ্রহণ', 'গ্রুপ', 'ঘর', 'ঘোষণা', 'চলে', 'চান', 'চায়', 'চার', 'চালা', 'চালান', 'চালু', 'চায়', 'চেয়ে', 'চেয়েছিলেন', 'চেষ্টা', 'চেয়ে', 'ছয়', 'ছাড়া', 'ছাড়াছাড়ি', 'ছাড়া', 'ছাড়াও', 'ছিল', 'ছিলেন', 'ছোট', 'জন', 'জনকে', 'জনাব', 'জনাবা', 'জনের', 'জন্য', 'জানতাম', 'জানতে', 'জানা', 'জানানো', 'জানায়', 'জানিয়ে', 'জানিয়েছে', 'জানে', 'জায়গা', 'জিজ্ঞাসা', 'জিজ্ঞেস', 'জিনিস', 'জে']

**Stemming**

This is the process of reducing words to their root form. The primary objective of stemming is to simplify each word in a text to its core meaning, thereby enabling words with similar meanings but different forms to be treated as identical words. This approach helps to reduce the dimensionality of the text data and enhance the performance of classification. We have used ‘BanglaStemmer’ from bangla-stemmer [7].

**Table 3.4:** Example of Stemming.

|  |  |
| --- | --- |
| **Text After Stopwords Removal** | **Clean Text** |
| রাজনীতির সম্পর্কে আলোচনা যাক ধর্মীয় সকল দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ শালার বেটাকে কুত্তার বাচ্চাকে মার | রাজনীতির সম্পর্ আলোচনা যাক ধর্মীয় সকল দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ শালা বেটা কুত্ত বাচ্চা মার |

Word Tokenization

Word tokenization is a fundamental process in natural language processing that involves breaking down a sequence of text into individual words or tokens. The main objective of word tokenization is to create a meaningful representation of the textual content by identifying and isolating individual words. In this context, a word token refers to a unit of

text that possesses semantic meaning and is treated as a discrete entity.

**Table 3.5:** Example of Word Tokenization.

|  |  |
| --- | --- |
| **Clean Text** | **Tokenized Text** |
| রাজনীতির সম্পর্ আলোচনা যাক ধর্মীয় সকল দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ শালা বেটা কুত্ত বাচ্চা মার | ['রাজনীতির', 'সম্পর্', 'আলোচনা', 'যাক', 'ধর্মীয়', 'সকল', 'দৃষ্টিভঙ্গি', 'মনোনিবেশ', 'গুরুত্বপূর্ণ', 'শালা', 'বেটা', 'কুত্ত', 'বাচ্চা', 'মার'] |

* + 1. Feature Extraction

Word N-Gram

N-grams refer to the contiguous sequences of word symbols, or tokens, present in a given document. In more technical terms, they can be defined as the neighboring sequences of items in a document. The Bengali language is known for its complex grammatical structures, conjugations, and compound words. N-grams play a crucial role in adapting to this complexity by providing a detailed representation of language patterns. In our study, we have utilized unigram, bigram, and trigram models to analyze the Bengali language.

**Table 3.6:** Example of Bangla Word N-Gram.

|  |  |  |  |
| --- | --- | --- | --- |
| **Clean Text** | **Unigram** | **Bigram** | **Trigram** |
| রাজনীতির সম্পর্ আলোচনা যাক ধর্মীয় সকল দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ শালা বেটা কুত্ত বাচ্চা মার | ['রাজনীতির', 'সম্পর্', 'আলোচনা', 'যাক', 'ধর্মীয়', 'সকল', 'দৃষ্টিভঙ্গি', 'মনোনিবেশ', 'গুরুত্বপূর্ণ', 'শালা', 'বেটা','কুত্ত', 'বাচ্চা', 'মার'] | ['রাজনীতির সম্পর্', 'সম্পর্ আলোচনা', 'আলোচনা যাক', 'যাক ধর্মীয়', 'ধর্মীয় সকল', 'সকল দৃষ্টিভঙ্গি', 'দৃষ্টিভঙ্গি মনোনিবেশ', 'মনোনিবেশ গুরুত্বপূর্ণ', 'গুরুত্বপূর্ণ শালা', 'শালা বেটা', 'বেটা কুত্ত', 'কুত্ত বাচ্চা', 'বাচ্চা মার'] | ['রাজনীতির সম্পর্ আলোচনা', 'সম্পর্ আলোচনা যাক', 'আলোচনা যাক ধর্মীয়', 'যাক ধর্মীয় সকল', 'ধর্মীয় সকল দৃষ্টিভঙ্গি', 'সকল দৃষ্টিভঙ্গি মনোনিবেশ', 'দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ', 'মনোনিবেশ গুরুত্বপূর্ণ শালা', 'গুরুত্বপূর্ণ শালা বেটা', 'শালা বেটা কুত্ত', 'বেটা কুত্ত বাচ্চা', 'কুত্ত বাচ্চা মার'] |

* + 1. Classification

We implemented a keyword-based classification approach. Initially, we conducted individual word matching with the political, religious and abusive keywords and calculated the overall count of matched keywords in each category. The category with the highest count of matched keywords was considered as the predicted category. To ensure more precise results, we performed this analysis on unigram, bigram, and trigram.

**Table 3.7:** Example of Matched Word After Applying Unigram.

|  |  |  |  |
| --- | --- | --- | --- |
| **Text after Applying Unigram** | **Political Keywords** | **Religious Keywords** | **Abusive Keywords** |
| ['রাজনীতির', 'সম্পর্', 'আলোচনা', 'যাক', 'ধর্মীয়', 'সকল', 'দৃষ্টিভঙ্গি', 'মনোনিবেশ', 'গুরুত্বপূর্ণ', 'শালা', 'বেটা','কুত্ত', 'বাচ্চা', 'মার'] | **রাজনীতির** | **ধর্মীয়** | **শালা,**  **কুত্ত** |

Here we used a simple mathematical equation for calculating the percentage of matched words. Initially all parameters are kept 0. Here for unigram, counting occurrences:

*total\_count\_religious=∑found\_keys\_religious.values()*

*total\_count\_political=∑found\_keys\_political.values()*

*total\_count\_abusive=∑found\_keys\_abusive.values()*

*total\_count\_match=total\_count\_religious + total\_count\_political + total\_count\_abusive*

For bigram, counting occurrences:

*total\_count\_religious=total\_count\_religious+ ∑found\_keysB\_religious.values()*

*total\_count\_political= total\_count\_political +∑found\_keysB\_political.values()*

*total\_count\_abusive= total\_count\_abusive + ∑found\_keysB\_abusive.values()* *total\_count\_match=∑found\_keysB\_religious.values()+∑found\_keysB\_political.values()+∑found\_keysB\_abusive.values()*

*total\_word\_in\_input=total\_word\_in\_input− total\_count\_match ×2+ total\_count\_match*

For trigram, counting occurrences:

*total\_count\_religious=total\_count\_religious+ ∑found\_keysT\_religious.values()*

*total\_count\_political= total\_count\_political +∑found\_keysT\_political.values()*

*total\_count\_abusive= total\_count\_abusive + ∑found\_keysT\_abusive.values()*

*total\_count\_match=∑found\_keysT\_religious.values()+∑found\_keysT\_political.values()+∑found\_keysT\_abusive.values()*

*total\_word\_in\_input=total\_word\_in\_input− total\_count\_match ×3+ total\_count\_match*

Now percentage count:

**Table 3.8:** Example of Matched Word After Applying Bigram and Trigram.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bigrams** | **Trigrams** | **Political Keywords** | **Religious Keywords** | **Abusive Keywords** |
| ['রাজনীতির সম্পর্', 'সম্পর্ আলোচনা', 'আলোচনা যাক', 'যাক ধর্মীয়', 'ধর্মীয় সকল', 'সকল দৃষ্টিভঙ্গি', 'দৃষ্টিভঙ্গি মনোনিবেশ', 'মনোনিবেশ গুরুত্বপূর্ণ', 'গুরুত্বপূর্ণ শালা', 'শালা বেটা', 'বেটা কুত্ত', 'কুত্ত বাচ্চা', 'বাচ্চা মার'] | ['রাজনীতির সম্পর্ আলোচনা', 'সম্পর্ আলোচনা যাক', 'আলোচনা যাক ধর্মীয়', 'যাক ধর্মীয় সকল', 'ধর্মীয় সকল দৃষ্টিভঙ্গি', 'সকল দৃষ্টিভঙ্গি মনোনিবেশ', 'দৃষ্টিভঙ্গি মনোনিবেশ গুরুত্বপূর্ণ', 'মনোনিবেশ গুরুত্বপূর্ণ শালা', 'গুরুত্বপূর্ণ শালা বেটা', 'শালা বেটা কুত্ত', 'বেটা কুত্ত বাচ্চা', 'কুত্ত বাচ্চা মার'] |  |  | শালা বেটা,  কুত্ত বাচ্চা |

Based on the data presented in the table 3.7 and table 3.8, it can be observed that the total count for abusive keywords is the highest among all the categories. Therefore, it can be concluded that the predicted category is Abusive.

The system contains several steps of working procedure. As working with lots of data or sensitive information, the flow of the system needs to be understood efficiently. As such the movement of data is shown in the work flow diagram, Figure 3.2.

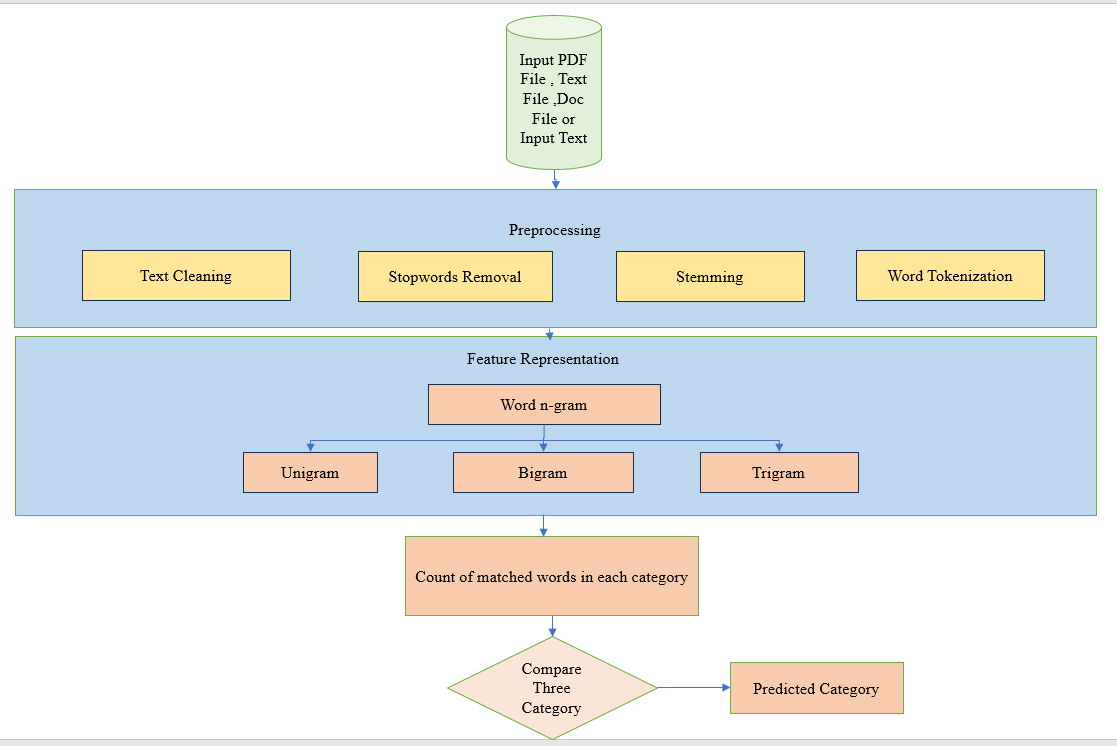


Figure 3.2: Work Flow Diagram of Class Recognition.

* 1. Recommendation

The system utilizes a recommendation algorithm to suggest appropriate content for individuals of varying ages. This step involves comparing the numerical values obtained from the classification result with a predetermined threshold, which then leads to a decision being made. Here is the Recommendation Sudo code for decision making:

***if predicted category is 'Religious':***

***write message (“This book is suitable for all.”)***

***elseif predicted category is 'Political' and abusive-key <= 0.01:***

***write message ("This book is suitable for 8 to 10 years old children.")***

***elseif predicted category is 'Political' and abusive-key <= 0.03:***

***write message ("This book is suitable for 10 to 15 years old children.")***

***elseif predicted category is 'Political' and abusive-key > 0.03:***

***write message ("This book is suitable for adults.")***

***elseif predicted category is 'Abusive':***

***write message (“It is an adult content. This book is not suitable for underaged individuals”)***

***else:***

***write message (“This book is suitable for all.”)***



Figure 3.3: Flow Chart for Recommendation algorithm.

1. Project Implementation
   1. Results

4.1.1 Dataset

We have created three distinct datasets categorized as religious, political, and abusive. These datasets were compiled by extracting keywords from four different types of books, which are outlined in the table 4.1.

**Table 4.1:** The Books Name from Which Keywords Were Collected.

|  |  |  |
| --- | --- | --- |
| **Book Name** | **Author Name** | **Book Type** |
| রাজনীতির অভিধান | সৌরেন্দ্রমোহন গঙ্গোপাধ্যায় | Political |
| অপরাধ-জগতের ভাষা ও শব্দকোষ | ভক্তি প্রসাদ মল্লিক | Abusive |
| গালি অভিধান | আব্দুল মান্নান স্বপন | Abusive |
| যার যা ধর্ম | মুহাম্মদ হাবিবুর রহমান | Religious |

**Table 4.2:** Total Count of Keywords in Each Category.

|  |  |
| --- | --- |
| **Category Name** | **Total Keywords** |
| Religious | 1411 |
| Political | 494 |
| Abusive | 3477 |

These are the set of tables that represent snapshots of the dataset we are currently working with. These tables provide a glimpse into the data at different points in time and serve as a reference for our analysis and decision-making processes.

**Table 4.3:** Religious Keywords.

|  |  |  |
| --- | --- | --- |
| **Category** | **Keyword** | **Label** |
| Religious | অংশুমান | 0 |
| Religious | অক্ষৌহিণী | 0 |
| Religious | অগ্নি | 0 |
| Religious | হাঁসিয়া | 0 |
| Religious | অঙ্গদ | 0 |
| Religious | অঙ্গিরা | 0 |
| Religious | অগস্ত্য যাত্রা | 0 |
| Religious | অচিন্ত্য-ভেদ-অভেদ | 0 |
| Religious | অজ | 0 |

**Table 4.4:** Political Keywords.

|  |  |  |
| --- | --- | --- |
| **Category** | **Keyword** | **Label** |
| Political | অকালি দল | 1 |
| Political | অছিবাদ | 1 |
| Political | অধিকার | 1 |
| Political | অধিকারী | 1 |
| Political | গঙ্গাধর | 1 |
| Political | মোরেশ্বর | 1 |
| Political | অনুশীলন সমিতি | 1 |
| Political | প্রতিষ্ঠান | 1 |
| Political | অন্তর্দলীয় গণতন্ত্র | 1 |

**Table 4.5:** Abusive Keywords.

|  |  |  |
| --- | --- | --- |
| **Category** | **Keyword** | **Label** |
| Abusive | অঁড়কঁক | 2 |
| Abusive | অউরর পোয়া | 2 |
| Abusive | অক্কুম | 2 |
| Abusive | অকর্মণ্য | 2 |
| Abusive | অকর্মা | 2 |
| Abusive | অকর্মার ধাড়ি | 2 |
| Abusive | অক্দমা | 2 |
| Abusive | অকাইঠা | 2 |

* + 1. Class Recognition

**Table 4.6:** Class Recognition for Different Bangle Books.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Book No.** | **Book Title** | **Actual Class** | **Predicted** | |
| 01 | সহীহ-বুখারী(পার্ট-০৫) -ইসলামবাংলা | Religious | R | 3.21% |
| P | 2.93% |
| A | 0.37% |
| 02 | অ্যারিস্টটল এর পলিটিকস - সরদার ফজলুল করীম | Political | R | 0.86% |
| P | 7.96% |
| A | 0.59% |
| 03 | আল কুরআন ও তার সুমহান মর্যাদা – আব্দুল আযীয ইবন আব্দুল্লাহ | Religious | R | 5.98% |
| P | 0.44% |
| A | 0.14% |
| 04 | খুন - আহমেদ ফারুক | Abusive | R | 2.67% |
| P | 0.59% |
| A | 2.89% |
| 05 | সাতই মার্চের ভাষণ | Political | R | 0.75% |
| P | 3.01% |
| A | 0.13% |
| 06 | সন্মোহন - সংগীত বান্দাপাধায়ে | Abusive | R | 2.05% |
| P | 0.17% |
| A | 2.29% |
| 07 | রামায়নঃ যুদ্ধ কাণ্ড ৪র্থ খণ্ড | Religious | R | 2.72% |
| P | 0.76% |
| A | 0.24% |

* + 1. Recommendation

**Table 4.7:** Class Recommendation for Different Bangle Books

|  |  |  |  |
| --- | --- | --- | --- |
| **Book No.** | **Book Title** | **Actual Recommendation** | **Predicted Recommendation** |
| 01 | সহীহ-বুখারী(পার্ট-০৫) -ইসলামবাংলা | This book is suitable for all. | This book is suitable for all. |
| 02 | অ্যারিস্টটল এর পলিটিকস - সরদার ফজলুল করীম | This book is suitable for 8 to 10 years old children. | This book is suitable for 8 to 10 years old children. |
| 03 | আল কুরআন ও তার সুমহান মর্যাদা – আব্দুল আযীয ইবন আব্দুল্লাহ | Religious | This book is suitable for all. |
| 04 | খুন - আহমেদ ফারুক | It is an adult content. This book is not suitable for underaged individuals. | It is an adult content. This book is not suitable for underaged individuals. |
| 05 | সাতই মার্চের ভাষণ | This book is suitable for 8 to 10 years old children. | This book is suitable for 8 to 10 years old children. |
| 06 | সন্মোহন - সংগীত বান্দাপাধায়ে | It is an adult content.This book is not suitable for underaged individuals | It is an adult content.This book is not suitable for underaged individuals |
| 07 | রামায়নঃ যুদ্ধ কাণ্ড ৪র্থ খণ্ড | This book is suitable for all. | This book is suitable for all. |
|  |

* 1. Tools

A wide variety of tools were used to complete the system. Each tool played a significant role in the completion of the project.

* + 1. Google Colab

Google Colab, or Colaboratory [15] is a free cloud-based platform by Google for writing and executing Python code in collaborative Jupyter Notebooks. In terms of GPU specifications, it offers variable GPU memory (12GB/16GB), a variable GPU memory clock (0.82GHz/1.59GHz), and performance ranging from 4.1 to 8.1 TFLOPS, all contingent on the specific GPU assigned during runtime. Google Colab also supports mixed precision based on the assigned GPU. Users should consult the official Google Colab documentation for real-time information on specifications during usage.

* + 1. Streamlit

Streamlit [10] is a Python library designed for easily creating interactive web applications for data science and machine learning. It simplifies the development process with minimal code, provides a range of widgets for interactivity, integrates well with popular data science libraries, supports rapid prototyping, and allows easy sharing and deployment of applications.

* 1. Graphical User Interface (GUI)

At first, we developed a web application which converts user’s raw input handwriting image to digital text. We used an open-source Python library called Streamlit which makes it easy to create and share custom web apps for machine learning and data science applications. Figures below demonstrates interfaces of this application and how to use it. To run the web application, change the current directory to project directory and run the following command:

*streamlit run app.py*

Where app.py is the streamlit filename. This will run streamlit’s server in local machine where the web application is hosted.



Figure 4.1: Initial user interface.

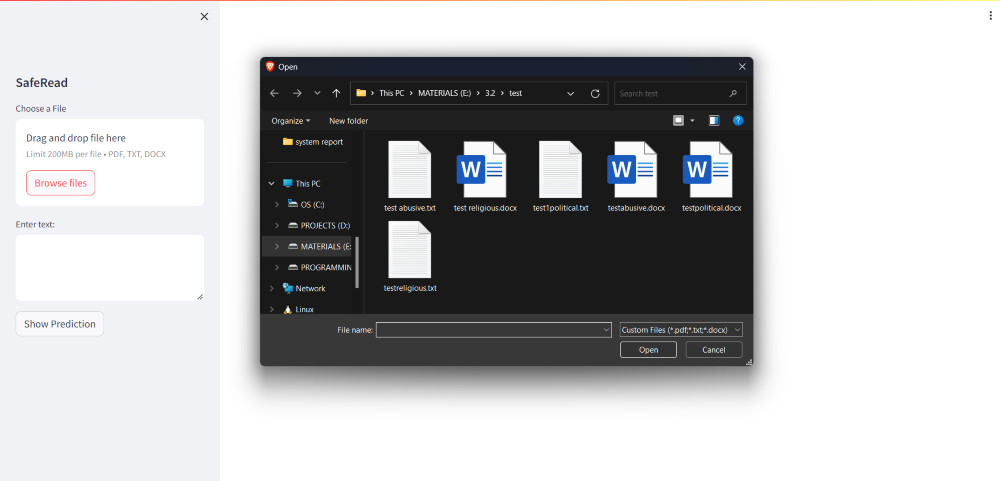


Figure 4.2: Select the Input file (.txt .docx .pdf).



Figure 4.3: File has been uploaded. Now click the button to convert to digital text.

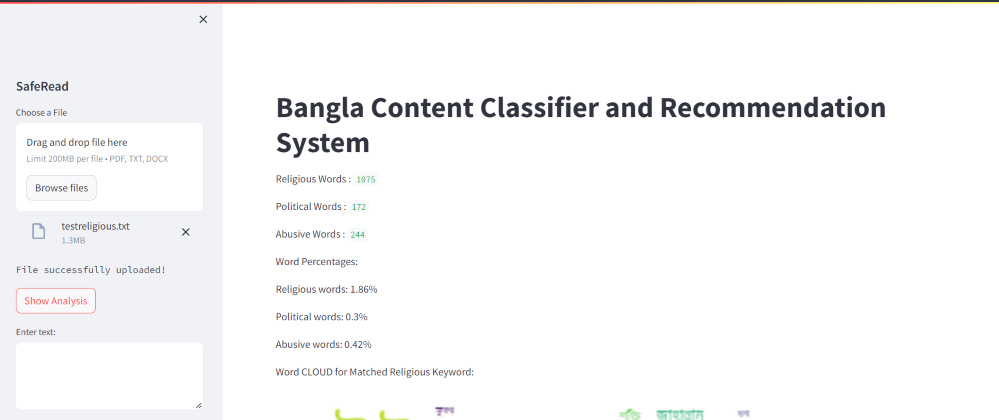


Figure 4.4: Total Count of Religious Keyword, Political Keyword and Abusive Keyword.



Figure 4.5: Frequently arrived words in Religious Keyword of Uploaded File.



Figure 4.6: Frequently arrived words in Political Keyword of Uploaded File.

Figure 4.7: Frequently Arrived Words in Abusive Keyword of Uploaded File.

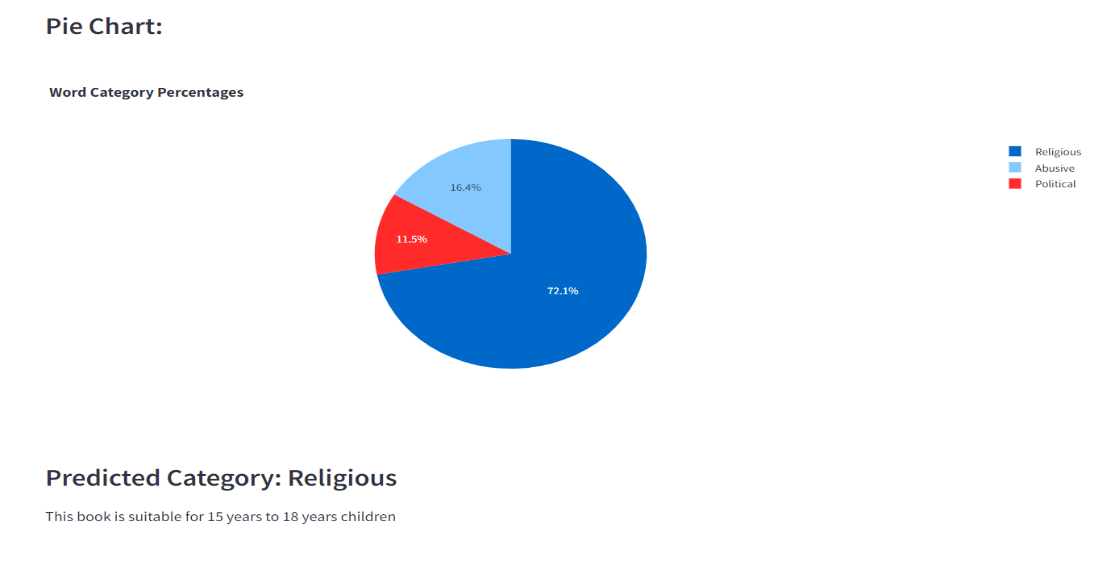


Figure 4.8: Pie Chart for Showing the Percentage of Each Category.

Based on the data presented in the pie chart above, it is evident that the highest percentage belongs to the religious category. Therefore, it can be concluded that the predicted category is religious.

* 1. Use Case Diagram

The system consists of a wide variety of roles. The perspective of these roles is exhibited in the Figure 4.9.

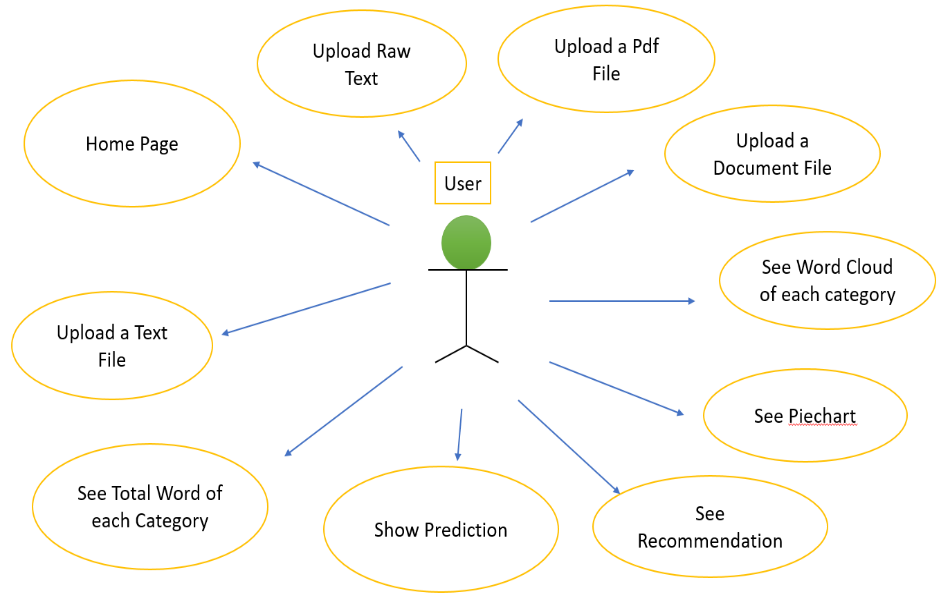


Figure 4.9: Use Case Diagram of This Project.

4.5 Morality or Ethical Issues

The project developed took a wide variety of sources. The dataset was gathered from various books that contained keywords from different categories. To accomplish this, we learned about Bangla pdf-to-text conversion [3], preprocessing techniques [1, 2], n-grams [8], and comparison algorithms [9] from official documentation and research papers. For the web application, we followed Streamlit documentation to ensure seamless productivity and a smooth user experience, and designed the user interface to be easily understandable for everyone. The political and religious datasets are available to the public, while the abusive dictionary is private due to its potential for misuse, such as generating negative comments on social media or making others feel uncomfortable by using inappropriate language

* 1. Socio-economic Impact and Sustainability

We have been very vigilant in monitoring the social standards while categorizing our content. In accordance with our social morality standards, we have recommended certain contents to different age groups. For instance, religious content is recommended for all users, whereas political content may not be suitable for 8-9-year-old children. Moreover, abusive or adult content cannot be recommended to minors. Our dataset contains key-words from different generations and literature. So it can be used for future projects such as social media post classification, book categorization or generating passages by these keywords.

* 1. Financial Analysis and Budget

Financial analysis and budget for this project involves the cost of development, equipment, implementation, maintenance and other relevant expenses. A general measurement of the budget has been estimated in following Table 4.2 The overall budget is characterized by minimal external expenditures, with majority of costs being self-funded, utilization of available software and technologies.

**Table 4.8:** Financial Analysis and Budget of the Project.

|  |  |
| --- | --- |
| **Types of Tools** | **Budget (Taka)** |
| Hardware (Laptop) | Personal |
| Software (Google colab, Streamlit, Ms Excel ) | Free version |
| Data collection and processing | 1,000 – 3,000 |
| Miscellaneous | 1,000 – 1,500 |
| Total: | 2,000 – 4,500 |

1. Conclusion
   1. Challenges

We faced several challenges when we started developing our system due to very limited resources. The first obstacle was collecting a dataset of numerous Bangla words. Our dataset includes abusive/slang words from different regions and cultures, religious conceptual words, and characters, as well as political words which required us to go through history. Secondly, we encountered an optimization problem while converting a Bangla PDF to a data frame using Pytesseract [3]. Although reliable, the process was time-consuming and required extensive resources. Thirdly, stemming was not effective during the text preprocessing stage in the Bangla language. However, it wasn't a necessary step for our system.

* 1. Conclusion

The development of the online keyword-based Bangla content classifier system is a significant advancement in protecting children in Bangladesh and other Bangla-speaking countries from inappropriate content. It also saves time by classifying content much faster than manual classification. Moreover, the rich dataset used in the system can be utilized for various content generation purposes. The pdf-to-text conversion feature, optimized by multithreading [5], saves time without compromising the quality of Bangla content conversion. Overall, this system offers numerous optimization opportunities to solve various problems, as described.

* 1. Future Study

In the future, parallel computing will play a crucial role in finding the optimal model for converting PDFs to text. To achieve better output, we need to expand our dataset. We plan to include an example of each keyword for sentence-based analysis, which will involve word encoding and word embedding techniques such as TF-IDF, Word2Vec, GloVe etc. Improving the stemmer is also a priority**.** Sentence-based approach has the potential to enhance the system's understanding of semantic relationships and contextual similarities among words in book descriptions or content, resulting in more accurate results. To capture sequential dependencies in book descriptions, we will use machine learning and DNN based models such as LSTMs. Additionally, we aim to develop and employ comprehensive evaluation metrics to assess the performance of the enhanced system.

# References

[1] Rahman, M. H., Islam, M. S., Jowel, M. M. U., Hasan, M. M., & Latif, M. S. (2021, July). Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[2] Sarker, S. (2021). Bnlp: Natural language processing toolkit for bengali language. *arXiv preprint arXiv:2102.00405*.

[3] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.

[4] Redoy, A. F. K., & Hossain, M. A. Depression Detection From Social Media Textual Data Using Natural Language Processing and Machine Learning Techniques. Bhowmik, N. R., Arifuzzaman, M., & Mondal, M. R. H. (2022).

[5] PAUDEL, P., SHAH, R., GC, R., & KHADKA, S. (2023). SHRUTI-A NEPALI-BABGALI BOOK READER.

[6] Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, *13*, 100123.

[7] Hoque, M. N., & Seddiqui, M. H. (2015, December). Bangla parts-of-speech tagging using Bangla stemmer and rule based analyzer. In *2015 18th International Conference on Computer and Information Technology (ICCIT)* (pp. 440-444). IEEE.

[8] Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization.

In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and*

*information retrieval* (Vol. 161175, p. 14).

[9] Summerfield, M. (2010). Programming in Python 3: a complete introduction to the

Python language. Addison-Wesley Professional.

[10] <https://discuss.streamlit.io/t/hyperlink-a-column-based-on-another-column/46325>

[11] Luo, X. (2021). Efficient English text classification using selected machine learning

techniques. *Alexandria Engineering Journal*, *60*(3), 3401-3409.

[12] Ababneh, J. (2019). Application of Naïve Bayes, decision tree, and K-nearest

neighbors for automated text classification. *Modern Applied Science*, *13*(11), 31.

[13] Bhowmik, N. R., Arifuzzaman, M., & Mondal, M. R. H. (2022). Sentiment analysis

on Bangla text using extended lexicon dictionary and deep learning

algorithms. *Array*, *13*, 100123.

[14] Richie-Halford, A., & Rokem, A. (2018). Cloudknot: A python library to run your

existing code on aws batch. In *Proceedings of the 17th python in science*

*conference* (pp. 8-14).

[15] <https://colab.research.google.com/drive/1JmyGb5b7N2txvy9RISJZXSXOQ7>

efE4tj?usp=sharing