Utilizing Optical Character Recognition to Summarize Bengali Text

***Abstract*—Optical Character Recognition (OCR) has proven to be a valuable tool for extracting textual information from images, converting physical documents into digital formats that can be easily edited, searched, and shared. However, OCR technology is not without its limitations, particularly when it comes to recognizing and processing non-Latin scripts such as Bengali. Over 230 million people speak Bengali, making it one of the most widely used languages in the world. Summarizing Bengali material is one area where OCR could be especially helpful, saving time and effort for scholars and journalists who are working with a big corpus of text. In recent years, there have been several attempts to develop OCR-based text summarization techniques for Bengali. One challenge it struggles with is the complex nature of the language, its rich vocabulary, its highly inflected grammar and the inclusion of a number of ligatures in Bengali Script and diacritics that can significantly complicate the OCR process. Overall, the development of OCR-based text summarization could save time and effort, while also providing a more objective and systematic approach to analyzing large volumes of text. It is expected that in the future, Bengali text summarizing methods based on OCR will become more advanced and precise as technology advances.**

***Index Terms*—OCR, Bengali, Corpus, Summarization, Ligatures, Data Preprocessing, LDA, RNN, Accuracy, BLEU**

I. Introduction

OCR, or Optical Character Recognition, is a technology that scans printed or handwritten text and converts it into a digital format that may then be modified, searched, and saved digitally. Digitizing text required a lot of time and effort before the advent of OCR technology. This was inefficient because of the time it took and the potential for mistakes that may have negative effects on accuracy and output.

The goal of developing optical character recognition technology was to speed up and improve the accuracy of the method of digitizing text. In order to detect and digitize printed or handwritten text, OCR software use complex algorithms to evaluate the forms and patterns of the letters. This facilitates the rapid digitization of vast quantities of paper records, allowing users to save time and ensure precision in their work. By automatically turning scanned documents into text that is machine-readable, OCR technology does away with the necessity for human data input. This allows documents to be scanned and translated into other languages through machine translation algorithms, saving time and improving accuracy compared to manual data entry and making printed materials obtainable to people with visual disabilities by turning them into digital text that can be examined aloud by screen readers. Despite the potential of the technology, it still has trouble with non-Latin scripts. The Bangla language has several significant historical and writings published in it. Enhancing education by facilitating the sharing of educational resources across regions and countries, the development of Bangla OCR technology can aid in the preservation of this cultural heritage by making it simpler to digitize and store historical and written works in digital format.

However, OCR technology has struggled to accurately detect and digitize Bengali text due to the complexity of Bengali script. The researchers that need to deal with Bengali texts have faced a significant challenge because of this: it is time-consuming and inefficient to search and analyze vast amounts of Bengali text. While Bangla OCR technology has come a long way since the 1980s, current OCR systems still fall short of ideal accuracy. As a consequence, there is a lot of focus on developing better Bangla OCR systems. However, with the tools at our disposal now, a full-fledged Bangla OCR is not only feasible, but its use can yield accurate results and reduce processing time.

II. Literature Review

One of most researchable topic in NLP is text summarization. OCR (Optical Character Recognition) technology can convert printed text or scanned images into machine-readable digital text. Because this digitization automates the extraction of relevant information from documents, it reduces the requirement of manual summarization efforts and it is mostly used in text summarization process.

Many studies have been conducted in this field for various languages. To avoid ambiguity or contradiction, the process of calculating similarity via a basic pattern matching algorithm requires the possession of an exhaustive set of patterns for each meaning. The difficulty and time-consuming nature of manually compiling is tedious task. Recent natural language processing applications highlight the need for an efficient method to obtain the similarity between very textual data or sentences [1].The utilization of a lexical dictionary to evaluate the similarity of a set of words collected from several sentences being compared is one extended form of word co-occurrence methods. Sentence similarity can be calculated simply by adding the similarity values of all word pairs [2]. There is also a model that uses a reinforcement technique for abstractive text summarization and employs convolutional sequence learning. The central concept of this work is textual learning in context.

The convolutional neural network consolidates the entire process [3]. Instead of an encoder and decoder model, extractive summarization of text uses a decoder, which can generate a natural-sounding summary out of a lengthy text sequence, making Wikipedia-style summaries possible. A large, parallel dataset is not a problem for this technique, and neither is the generation of a multi-document summary [4].

III. OCR and Summarization Approaches

1) OCR: For the purpose of extracting Bangla text from images we will be using Bengali dataset that has Bengali images and their corresponding texts. Later we will use the pytesseract model using Bengali dataset that will be able to recognize Bengali texts from images fed to the model.

2) Summarisation: There are three major approaches for summarisation - extractive based, abstractive based and keyword based approach [5]. We will be working on abstractive summarisation approach [6] which has the following steps:

1) Data Collection
2) Data Preprocessing
3) Data Splitting
4) Model Selection
5) Fine-tuning
6) Summarization
7) Evaluation

IV. DATA COLLECTION AND PREPROCESSING

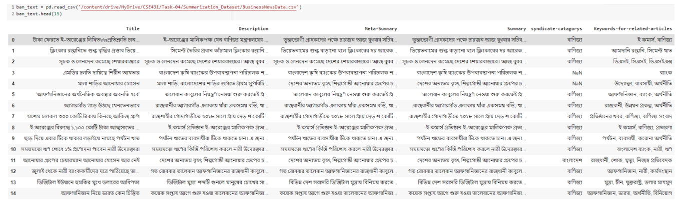
A. Data Collection

For OCR: To apply OCR technique we made our own dataset by taking around more than 300 pictures of text paragraphs written in Bengali language. Alongside we also collected the text data of those pictures and stored them to check the accuracy of OCR. .

For Summarization: To train a model for the purpose of summarization we collected around 23422 texts from the Prothom Alo Newspaper and their subsequent summarized texts. The texts were categorized into 7 types including-

1) Business News Data
2) Entertainment News Data
3) Opinion News Data
4) Politics News Data
5) Sports News Data
6) World News Data
7) Bangladesh News Data

Subsequent csv files were made for each type for data preprocessing. Each csv file had 6 columns- Title, Description, Meta-Summary, Summary, Syndicate-catagorys, Keywords-for-related-articles. An illustration of which is shown below:



```
1 ban_text = pd.read_csv('/content/drive/MyDrive/CSE431/Task-04/Summarization_Dataset/BusinessNewsData.csv')
2 ban_text.head(15)
```

	Title	Description	Meta-Summary	Summary	syndicate-catagorys	Keywords-for-related-articles
0	টাকা ফেরতে ই-অরেঞ্জের লিখিত\r\nপ্রতিশ্রুতি চান...	ই-অরেঞ্জের মালিকপক্ষ যেন বাণিজ্য মন্ত্রণালয়ের ...	ভুক্তভোগী গ্রাহকদের পক্ষে চারজন আজ বুধবার সচিব...	ভুক্তভোগী গ্রাহকদের পক্ষে চারজন আজ বুধবার সচিব...	বাণিজ্য	ই কমার্স, বাণিজ্য
1	ক্লিংকার রপ্তানিতে শুল্ক বৃদ্ধির প্রস্তাব ভিয়ে...	সিমেন্ট তৈরির প্রধান কাঁচামাল ক্লিংকার রপ্তানি...	ভিয়েতনামের শুল্ক বাড়ানো হলে ক্লিংকারের দর আরেক...	ভিয়েতনামের শুল্ক বাড়ানো হলে ক্লিংকারের দর আরেক...	বাণিজ্য	আমদানি রপ্তানি, সিমেন্ট খাত
2	সূচক ও লেনদেন কমেছে শেয়ারবাজারে	সূচক ও লেনদেন কমেছে দেশের শেয়ারবাজারে। আজ বুধব...	সূচক ও লেনদেন কমেছে দেশের শেয়ারবাজারে। আজ বুধব...	সূচক ও লেনদেন কমেছে দেশের শেয়ারবাজারে। আজ বুধব...	বাণিজ্য	ডিএসই, সিএসই, ডিএসইএক্স
3	এমডির চলতি দায়িত্বে শিরীন আখতার	বাংলাদেশ কৃষি ব্যাংকের উপব্যবস্থাপনা পরিচালক শ...	বাংলাদেশ কৃষি ব্যাংকের উপব্যবস্থাপনা পরিচালক শ...	বাংলাদেশ কৃষি ব্যাংকের উপব্যবস্থাপনা পরিচালক শ...	NaN	ব্যাংক
4	মালা শাড়ির আনোয়ার হোসেন	মালা শাড়ি, বাংলাদেশের শাড়ির জগতে প্রথম সুপরিচি...	দেশের অন্যতম বৃহৎ শিল্পগোষ্ঠী আনোয়ার গ্রুপের চ...	দেশের অন্যতম বৃহৎ শিল্পগোষ্ঠী আনোয়ার গ্রুপের চ...	NaN	উদ্যোক্তা, ব্যবসায়ী, অর্থনীতি
5	'আফগানিস্তানের অর্থনৈতিক অবস্থার অবনতি হবে'	তালেবান কাবুলের নিয়ন্ত্রণ নেওয়া শুরু করতেই প্র...	তালেবান কাবুলের নিয়ন্ত্রণ নেওয়া শুরু করতেই প্র...	তালেবান কাবুলের নিয়ন্ত্রণ নেওয়া শুরু করতেই প্র...	বাণিজ্য	আফগানিস্তান, ব্যাংক, অর্থনীতি
6	আগারগাঁও গড়ে উঠছে যেনতেনভাবে	রাজধানীর আগারগাঁও এলাকায় যাঁরা একসময় বস্তি, যা...	রাজধানীর আগারগাঁও এলাকায় যাঁরা একসময় বস্তি, যা...	রাজধানীর আগারগাঁও এলাকায় যাঁরা একসময় বস্তি, যা...	বাণিজ্য	রাজধানী, উন্নয়ন প্রকল্প, অর্থনীতি
7	হাশেম চালকল ৩০০ কোটি টাকায় কিনছে আকিজ গ্রুপ	রাজশাহীর গোদাগাড়ীতে ২০১৮ সালে প্রায় দেড় শ কোটি...	রাজশাহীর গোদাগাড়ীতে ২০১৮ সালে প্রায় দেড় শ কোটি...	রাজশাহীর গোদাগাড়ীতে ২০১৮ সালে প্রায় দেড় শ কোটি...	বাণিজ্য	প্রতিষ্ঠানের খবর, বাণিজ্য, বাণিজ্য সংবাদ
8	ই-অরেঞ্জের বিরুদ্ধে ১,১০০ কোটি টাকা আত্মসাতের ...	ই-কমার্স প্রতিষ্ঠান ই-অরেঞ্জের মালিকপক্ষ প্রতা...	ই-কমার্স প্রতিষ্ঠান ই-অরেঞ্জের মালিকপক্ষ প্রতা...	ই-কমার্স প্রতিষ্ঠান ই-অরেঞ্জের মালিকপক্ষ প্রতা...	বাণিজ্য	ই কমার্স, বাণিজ্য, প্রতারণা
9	ছাড় দিয়ে এবার টিকে থাকার লড়াইয়ে নামছে পর্যটন খাত	পর্যটন খাতের ব্যবসায়ীরা টিকে থাকতে চান। এ জন্য...	পর্যটন খাতের ব্যবসায়ীরা টিকে থাকতে চান। এ জন্য...	পর্যটন খাতের ব্যবসায়ীরা টিকে থাকতে চান। এ জন্য...	বাণিজ্য	পর্যটন, ব্যবসায়ী, করোনা অর্থনীতি
10	সময়মতো ঋণ শোধে ১% প্রণোদনা পাবেন নারী উদ্যোক্তারা	সময়মতো ঋণের কিস্তি পরিশোধ করলে নারী উদ্যোক্তার...	সময়মতো ঋণের কিস্তি পরিশোধ করলে নারী উদ্যোক্তার...	সময়মতো ঋণের কিস্তি পরিশোধ করলে নারী উদ্যোক্তার...	বাণিজ্য	বাংলাদেশ ব্যাংক, নারী, ঋণ
11	আনোয়ার গ্রুপের চেয়ারম্যান আনোয়ার হোসেন আর নেই	দেশের অন্যতম বৃহৎ শিল্পগোষ্ঠী আনোয়ার গ্রুপের চ...	দেশের অন্যতম বৃহৎ শিল্পগোষ্ঠী আনোয়ার গ্রুপের চ...	দেশের অন্যতম বৃহৎ শিল্পগোষ্ঠী আনোয়ার গ্রুপের চ...	বাংলাদেশ	রাজধানী, শোক, মৃত্যু, নিজস্ব প্রতিবেদক
12	জুলাই থেকে নারী ব্যাংককর্মীদের ঘরে পাঠিয়েছে তা...	গত রোববার তালেবান আফগানিস্তানের রাজধানী কাবুলে...	গত রোববার তালেবান আফগানিস্তানের রাজধানী কাবুলে...	গত রোববার তালেবান আফগানিস্তানের রাজধানী কাবুলে...	বাণিজ্য	আফগানিস্তান, নারী, কর্মসংস্থান
13	ডিজিটাল ইউয়ানে হুমকির মুখে ডলারের আধিপত্য	'ডিজিটাল মুদ্রা' শব্দটি শুনলে মানুষের চোখের সা...	বিভিন্ন দেশ সরাসরি ডিজিটাল মুদ্রায় বিনিময় করতে...	বিভিন্ন দেশ সরাসরি ডিজিটাল মুদ্রায় বিনিময় করতে...	বাণিজ্য	মুদ্রা, চীন, যুক্তরাষ্ট্র, ডলার মাহমুদ
14	আফগানিস্তান নিয়ে ভারত কেন চিন্তিত	কয়েক সপ্তাহ আগে শুরু হওয়া তালেবানের আফগানিস্তা...	কয়েক সপ্তাহ আগে শুরু হওয়া তালেবানের আফগানিস্তা...	কয়েক সপ্তাহ আগে শুরু হওয়া তালেবানের আফগানিস্তা...	বাণিজ্য	আফগানিস্তান, ভারত, অর্থনীতি, বিনিয়োগ

Fig. 1. Dataset of Summarization

B. Data Preprocessing

For OCR: In our dataset, we discovered a wide variety of images that required preprocessing before the pytesseract model could be applied to extract texts. We used normalization, skew correction, image scaling, noise removal, thresholding, binarization, skeletonization, thinning, etc as preprocessing techniques for this purpose. To begin, we flipped the pictures into monochrome gray. The picture quality was then improved by eradicating any noise or distortions that might compromise OCR readability. The grayscale picture was then thresholded to create a binary one. The picture is then transformed into a monochrome version, with the words appearing in black on a white backdrop. Next, we rotate the image to straighten out the text lines if they are not already perfectly horizontal. After that, we got rid of the last of the image's imperfections and noise, such as lines and dots. To get the desired level of text thickness in the photos, we next used Thinning and Skeletonization. The next step we took to increase OCR precision was to standardize the size and orientation of the fragmented letters or phrases.

For Summarization: We applied multifarious preprocessing techniques on the collected dataset in csv format to ensure that it was in a suitable format for training our model. For each of the 7 csv files we checked for duplicate and missing/null values [7]. A few duplicate texts were found which might have been mistakenly taken during data collection. We removed the rows with duplicate values. Meanwhile for the null values we used an imputation technique. We got null values for the 'syndicate-catagorys' column of the csv file only. As the dataset was already divided, it was easy to replace the null values with its subsequent category using .fillna() function of the pandas library. Finally after preprocessing, all the refined datasets were combined together into 1 csv file to train our model.

V. Methodology

First we will be cleaning and translating the raw data into a structure useful for testing and training the summarization model constitutes this stage. Here, we'll go over the process of preparing a dataset for text summarization and the many phases required.

The first phase is data cleaning, in which irrelevant information like HTML elements, punctuation, extra spaces and stop words is eliminated. Stop words are those that appear frequently in a language but contribute nothing to a summary; hence, it is important to get rid of them. Since the machine cannot understand a text in its short form, we must define the word's full meaning before adding contractions. The dataset is made

more manageable by this method so that it may be thoroughly evaluated.

Tokenizing sentences is the next phase, which entails parsing the text into its component sentences. This is a stage since it helps the summarization model comprehend the organization of the text and the connections between individual phrases. Tokenizing a Bangla phrase effectively requires knowledge of the language's morphology, making it more difficult than in English. Tokenization, the process of parsing phrases into their constituent words, is the third phase. This is a vital stage since it teaches the summarization model the significance of each word and their connections to one another. We will use 'bnlp toolkit' for the purpose of tokenization of our Bengali texts [8].

Thirdly, the regular expression is used to remove the rare character or unwanted character to remove from texts. Removing spaces, English characters, punctuation form text, Bengali digit from text is the prime objective of using regular expression in our research.

Data normalization, the fourth phase, involves reformatting the text into an easily analyzed format. In the case of Bangla, this means adapting the text to a uniform script like Unicode. This is a vital stage since it guarantees that the summarization method has a proper grasp of the text.

Data representation is the fifth stage, and it entails expressing the text in a way that the summarization model is able to understand. Bag of Words, TF-IDF, and Word Embeddings are all useful tools for doing this phase, which entails transforming the text into numerical values.

Initially we divide our summarisation dataset into two parts-train and test data. We use 80% of the dataset for training our models and the rest of 20% data is used to test our models.

Now, selecting the best model is essential since it affects the final quality of the summary. To effectively summarize the most crucial points of the input text, a decent model should be able to extract the relevant semantic and contextual information. While there are a variety of models for text summarization, not all of them work well with the Bangla language.

Pointer-generator networks:
Using a combination of extractive and abstractive methods, pointer-generator networks are a type of hybrid model. In order to produce a summary, they employ a sequence-to-sequence architecture and equip their model with a pointer mechanism for lifting individual words from the source text [9].

Latent Dirichlet Allocation (LDA):
The LDA model of a corpus is a probabilistic generative model. The documents are modeled as stochastically mixed across latent themes, with each topic being defined as the distribution over words. One of the most often used approaches to topic modeling is latent Dirichlet allocation (LDA), which was initially proposed by Blei, Ng, and Jordan in 2003 [1]. LDA uses the probabilities of words to symbolize concepts. Word chances from LDA show that the most probable words in each topic may provide a good overview of the subject at hand. The most popular approach to topic modeling is Latent Dirichlet Allocation (LDA), an uncontrolled generative probabilistic technique for modeling a corpus. It is assumed in LDA that all documents have a similar Dirichlet prior for their topic distributions and that each document may be represented as a probability distribution over latent topics. The LDA model represents each latent subject as a probability distribution over words, with all topic word distributions having the same Dirichlet prior. Latent variables and hyper parameter values may be inferred by maximizing the probability of the observed data, denoted by the letter D. It may be written as, $p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(Z_d n|\theta_d)p(w_{dn}|Z_{dn},\beta))d\theta_d$

Topic Dirichlet prior parameters and the overall distribution of words across themes are defined with respect to the Dirichlet prior distribution. T stands for the total number of themes, M for the total number of documents, and N for the overall size of the word. The topic distributions at the corpus level are represented by the Dirichlet multinomial pair (α, θ). Given (β, ϕ), this is the Dirichlet-multinomial combination for word-topic distributions. Document-level variables, sampled once every document, are denoted by d. Each word in a document is individually sampled for the zdn and wdn variables. LDA is a well-regarded method for distributing latent topics throughout a large corpus. Therefore, it can categorize a large body of patents into granular subject matter and display those patents using a variety of topic distributions. Using LDA, a vocabulary can be built from the terms in a collection of documents to unearth previously unknown relationships and themes. A topic is the distribution of probabilities across these words, and documents are seen as a collection of topics. After this, each document is interpreted as the distribution of probabilities over a collection of themes. The data may be seen as the result of a generative process, the parameters of which are the joint distribution of probabilities across the observable and concealed features [10].

Generative Model:
A summary $\hat{y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_m\}$ is anticipated based on the input text $x = \{w_1, w_2, ..., w_n\}$. The total amount of phrases in the initial document, denoted by *n*, and the number of characters in the projected summary, denoted by *m*, are shown below. With a unidirectional long short-term memory (LSTM) encoder, the input text x is encoded into a collection of hidden states with $h = \{h_1, ..., h_n\}$. An awareness-based LSTM decoder is used to determine the on-time step t, unobserved state s_t of the decoders, and context vector c_t, (See, Liu, and Manning, 2017). The implementation details are included in the supplementary material of this work (or (See, Liu, and Manning 2017)). All of generator G's settings go by the symbol. At each time step t, the decoder state s_t, the context

vector c_t, and a softmax layer are passed through to determine the likelihood of correctly detecting a word from the target set of words.

$$P_{vocab}(\hat{y}_t) = softmax(V'(V[s_t, c_t] + b) + b')$$

whereas V′, V, b, b′ could be addressed. To employ either word generators from a set vocabulary or pointer copying uncommon or unseen from the input sequence, we implement a switching pointer-generator network, which is comparable to the work of (See, Liu, and Manning 2017). The summary probability $P(\hat{y}_t)$ for each token $\hat{y}_t$ may then be calculated.

Discriminative Model:
The discriminator is a binary classifier that attempts to determine whether or not the input sequence was synthesized by a machine or authored by humans. Due to its impressive performance in text classification (Kim, 2014), we encode the sequence of inputs using a CNN. To acquire a wide variety of features, we apply a peak-over-time pooling technique to the output of many filters whose window widths vary. These combined characteristics are then sent to a fully linked softmax layer, the output of which is the likelihood that the data is unique [11].

Cluster-based summarization:
One component of the success of text summarization is the use of clustering. By grouping similar statements together, it's easier to choose the most informative ones and get rid of the rest. Clustering methods choose one or two phrases from comparable sentences according to chosen criteria and include those sentences as a summary if a news story keeps repeating the same idea using the same set of terms. If there are more sentences in the text, then more clusters are needed to provide a more in-depth overview of the content. A cluster size of about 7 would provide a one-third summary of a text of 50 sentences, while a cluster size of 10 would do the same for a text of 100 phrases [12].

Recurrent neural networks (RNNs):
Models of abstract summarization such as these have the ability to predict the subsequent letters in a sequence in order to produce summaries.

In order to include the concept of time, the recurrent neural networks are a combination of feedforward neural networks by adding recurrent edges that span many time steps. It's possible that RNNs don't have any cycles at all, but recurrent edges, including self-connections, may generate cycles. Input activation at time t is provided to nodes along recurrent edges by both the current instance $x^{(t)}$ and the hidden nodes $h^{(t-1)}$ in the prior state of the network. Given the secret state $h^{(t)}$, the output $\hat{y}^{(t)}$ can be computed. Thus, these recurrent connections allow the input $x^{(t-1)}$ at time t-1 to affect the output $\hat{y}^{(t)}$ at time t. In a straightforward recurrent neural network, we can demonstrate all computations required for calculation at any given step on the forward path with just two equations:
$h^{(t)} = \sigma(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h)$
$\hat{y}^{(t)} = softmax(W^{yh}h^{(t)} + b_y)$

Here The weights that exist between the input data and hidden layers are represented by the matrix W_{hx}, while the weights across the layers that are hidden at successive time steps are represented by the matrix W_{hh}. The biases are vectors that each node uses to figure out its own offset. In this article, recurrent hidden layer networks make up the bulk of the models. However, the outputs of one state can be connected to the hidden layer of the next state in some suggested models, such as Jordan Networks. Others, including the sequence-to-sequence learning model proposed by Sutskever et al [13].

Thus after training the model with train data, we check its accuracy and if it seems to be usable we insert images to extract texts from them and then using pytesseract model our required text will be extracted. Later the applied model of summarisation will generate our summary in human readable form. An overview of the overall workflow for summarisation is as follow:

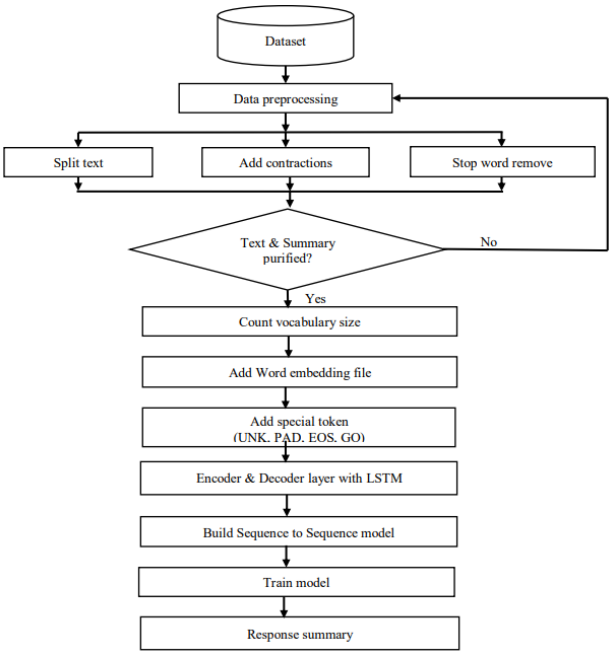


Fig. 2. Workflow of Summarization

VI. RESULTS AND ANALYSIS

Evaluation allows us to measure the performance and effectiveness of the generated summaries.

Firstly for OCR, we got around 94% accuracy owing to some images having extensive noise or due to having too thin or thick texts that could not be handled even after multiple scaling or preprocessing techniques [14].

To evaluate the quality of generated summaries, we commonly use metrics such as ROUGE and BLEU. BLEU compares the n-grams in the produced and reference summaries to determine how well they match, whereas ROUGE quantifies the overlap among the two summaries [15].

Although these metrics provide a quantitative measure of summary quality, they may not always align with human judgment. Therefore, it is important to conduct a qualitative evaluation of the summaries as well. This involves having human evaluators rate the summaries based on criteria such as coherence, informativeness, and readability.

In the case of our Bangla text summarization using the aforementioned models, the evaluation process would involve computing BLEU scores for the generated summaries compared to the reference summaries. Additionally, human evaluators could be used to assess the summaries based on criteria specific to Bangla language and culture. By combining both quantitative and qualitative evaluation metrics, we can gain a more comprehensive understanding of the effectiveness of the summarization system and identify areas for improvement.

Here we got different accuracy rates for different texts in our matrix = [0.72,0.81,0.9,0.51,0.6,0.33,.....,0.4,0.54] which gave our overall BLEU score to be 0.81 or 81%. We also generate the BLEU scores for N-grams of the summaries = [[0.4,0.5,0.2,0.1],[0.3,0.4,0.1,0.22],....,[0.5,0.2,0.1,0]]. These findings allow us to either modify the machine learning model to enhance the quality of the summaries or change the summary procedure to increase the BLEU score [16].

VII. Conclusion

Bengali text preprocessing is more challenging than that of other languages. This necessitates the development of a preprocessing library for Bengali text. We have completed our statistical analysis and sample comparisons. Moreover, no machine can guarantee a 100% reliable outcome. Every device has a restricted range of applications. The same holds true for our summarizer model. Our research led us to conclude that building our own Bangla WordNet and corpus would improve efficiency and accuracy. Then, and only then, will it have a chance at improved output. Our models may have underperformed because of the data source, especially the summary and article lengths [17]. The model might not grasp the data set's structure. Because our model was being run on Google Colab, it underperformed and frequently crashes whenever we inputted a large batch size due to the GPU memory limit being reached.Future work will hopefully include taking use of pre-trained models and fine-tuning them to carry out summarization [18]. In addition, we want to clean up our data by hand, article by article, to ensure the highest possible quality. This will aid our little model in its pattern-detection efforts, ultimately improving its overall performance.

References

[1] D. Michie, "Return of the imitation game," *Electron. Trans. Artif. Intell.*, vol. 5, pp. 203–221, 2001.

[2] J.-M. Torres-Moreno, "Automatic text summarization," 2014.

[3] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*, 2017.

[4] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. M. Shazeer, "Generating wikipedia by summarizing long sequences," *ArXiv*, vol. abs/1801.10198, 2018.

[5] P. Raundale and H. Shekhar, "Analytical study of text summarization techniques," *2021 Asian Conference on Innovation in Technology (ASIANCON)*, pp. 1–4, 2021.

[6] J. Yan and S. Zhou, "A text structure-based extractive and abstractive summarization method," *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 678–681, 2022.

[7] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," *ArXiv*, vol. abs/1908.08345, 2019.

[8] T. Islam, M. Hossain, and M. F. Arefin, "Comparative analysis of different text summarization techniques using enhanced tokenization," *2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6, 2021.

[9] N. Dhar, G. Saha, P. Bhattacharjee, A. Mallick, and M. S. Islam, "Pointer over attention: An improved bangla text summarization approach using hybrid pointer generator network," *2021 24th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, 2021.

[10] H. Jelodar, Y. Wang, C. Yuan, and X. Feng, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, pp. 15 169–15 211, 2017.

[11] S.-H. Liu, K.-Y. Chen, B. Chen, E.-E. Jan, H. Wang, H.-C. Yen, and W.-L. Hsu, "A margin-based discriminative modeling approach for extractive speech summarization," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pp. 1–6, 2014.

[12] P. K. Bhole and S. Ramdeobaba, "Single document text summarization using clustering approach implementing for news article," *international journal of engineering trends and technology*, vol. 15, pp. 364–368, 2014.

[13] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," *ArXiv*, vol. abs/1506.00019, 2015.

[14] K. Todorov and G. Colavizza, "An assessment of the impact of ocr noise on language models," in *International Conference on Agents and Artificial Intelligence*, 2022.

[15] Y. Graham, "Re-evaluating automatic summarization with bleu and 192 shades of rouge," in *Conference on Empirical Methods in Natural Language Processing*, 2015.

[16] D. Roy, S. Fakhoury, and V. Arnaoudova, "Reassessing automatic evaluation metrics for code summarization tasks," *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021.

[17] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," *ArXiv*, vol. abs/1808.10792, 2018.

[18] B. Salih and E. S. GUNAL, "The impact of features and preprocessing on automatic text summarization," *SCIENCE AND TECHNOLOGY*, vol. 25, no. 2, pp. 117–132, 2022.