Bangla text summarizer using hybrid of gpt2-bengali and pegasus transformers

1st Shafin Mahmud Jalal
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
shafin.mahmud.jalal@g.bracu.ac.bd

2nd Md. Rezuwan Hassan
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.rezuwan.hassan@g.bracu.ac.bd

3th S.M Niaz Morshed
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sm.niaz.morshed@g.bracu.ac.bd

4th Masum Uddin Ahmed
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
masum.uddin.ahmed@g.bracu.ac.bd

5th Rufaida Tasin
Computer Science and Enginnering
BRAC University
Dhaka, Bangladesh
rudaida.tasin@g.bracu.ac.bd

6th Md Humaion Kabir Mehedi
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

7th Md Mustakin Alam
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

8th Annajiat Alim Rasel
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Automatic text summarization is needed to concisely extract a small subset of text portions from a large text where the isolated text may have sentences that are more significant compared to other sentences in the text. Although there have been a lot of approaches to English text summarization, very few works have been done on automatic Bengali text summarization. For the evaluation purpose, a dataset was formulated from the scratch with Bengali news documents from two reputed newspapers. The evaluation dataset was classified into four different classes with benchmark standard summary text, generated by a group of random human contributors for each of the documents. The current work presents a hybrid approach for dealing with the summarization process of Bengali text documents. The hybrid model is introduced with a goal to improve the overall accuracy of the summary text generation. The proposed model generates a summary text based on keyword scoring, sentiment analysis, and the interconnection of sentences. After conducting the evaluation on the existing dataset, the proposed system performs with an average of 0.77 Recall Score, 0.57 Precision Score, and 0.64 F-measure Score. Empirical verification with other similar systems shows that the proposed model can be used as an alternative system to address the Text Summarization problem of Bengali documents.**

***Index Terms*—Hybrid Bangla Document Summarization, Hybrid transformers, Sentence Scoring, Sentiment Analysis, Text Ranking**

I. Introduction

Text summarization requires a short, accurate, and fluent summary of a longer text document. From the summary, important information can be gained, making the overall procedure more comfortable, and fewer resources are needed. To discover relevant information faster from a huge number of text documents available online, automatic text summarization ideas have been found very significant. A few methods have been explored for the generation of summary from Bengali documents. If the summary contains sentences from the document's major topics, it has a better chance of giving a better perspective of the document. The summary generation approach of the proposed system is extractive, i.e., they contain sentences as it appears in the document. According to Sarkar [1], text summarization involves preprocessing, stemming, sentence ranking, and summary generation. The preprocessing step requires removal of stopwords, stemming and converting the input into a collection of sentences Uddin and Khan [2] described an extraction-based method for summarizing Bengali documents. Different features, such as location, term frequency, numerical data, etc., were used to rank the sentences. Based on the features, they have designed the Bengali summarizer and concluded that the summary size should be 40 percent of the actual content. Das and Bandyopadhyay [3] have summarized Bengali documents using sentiment information. They have tried to identify the sentiment information in a document and then aggregated that for generating the summary. Mihalcea [4] [5] has focused on text summarization based on graphs. A graph can be constructed considering the sentences as nodes and connecting them with edges. After that, edge weights may be measured by calculating the similarity between two nodes. Keeping the state-of-the-art in view, a hybrid Bangla Text Summarizer has been proposed in this work, which combines the following methods:

- Sentiment Scoring
- Keyword Ranking
- Text Ranking

In the proposed method, the top 40 percent of the actual document was considered as a generated summary based on the combined weighted score, which we describe in an upcoming section.

II. Literature Review

Explainable AI is a class of techniques or methods that facilitates us to comprehend and evaluate a machine learning model's reached conclusion. We may use it to improve the performance of our models, as well as to guide others to understand how and what made the model take a particular decision. Some of the techniques belonging to that class are SHAP, DeepSHAP, DeepLIFT, CXplain, LIME. We integrated the LIME method with our created model as it makes a model's reaching conclusion individually comprehensive. Interpretations for different models and classifiers are all endorsed by LIME.

A. Hasan et al. [**?**] compared sentiment analysers for twitter accounts using machine learning [**?**]. They examined the sentiment analysis algorithms TextBlob, SentiWordNet, and W-WSD using the Naive Bayes classifier. They demonstrated that W-WSD had the greatest accuracy (79 percent) on 6250 tweets. Additionally, they discovered that W-WSD achieved the maximum accuracy (62 percent) on 5000 tweets when employing an SVM classifier.

K. h. manguri et al. used the tweepy and TextBlob packages to analyze user data [**?**]. They gathered data from twitter using two hashtags: 'COVID-19' and 'coronavirus'. The data was acquired during one of the coronavirus's most active weeks, and they analyzed over 53 thousand tweets.

G. a. ruz et al. evaluated sentiment during critical occurrences using five classifiers citeruz2020sentiment[**?**]. They employed two Spanish datasets for the study: the 2010 Chilean earthquake and the Catalan independence referendum (2017). They concluded that their behavior was same regardless of language. SVM achieved the best accuracy on the first dataset, whereas Random forest achieved the highest accuracy on the second. Additionally, they said that TAN and BF TAN provide intriguing qualitative data that may be used to appreciate the major characteristics of an event's dynamic from a historical and sociological perspective.

K. sailunaz et al. created a recommendation system based on sentiment analysis of user tweets and responses [**?**]. They determined the agreement, sentiment, and emotion score of responses while calculating influence scores. Additionally, they compiled a list of tweet creators who had the same perspective on certain themes and expressed similar feelings and thoughts about those topics.

M. danilevsky et al. performed a study to examine the use of Explainable AI in natural language processing [**?**]. It summarized recent advances in XAI research in natural language processing as addressed at the main NLP conferences over the previous seven years. They discussed the major classification schemes for explanations, as well as the technical operations and explainability tools now available for developing explanations for model predictions.

S. m. mathews used Explainable AI to assess natural language processing on twitter data, tumor diagnosis on biological signals, and Windows pc malware detection using LIME [**?**]. They visualized their findings in an appropriate manner.

A. Waldis et al. created a framework for text features based on statistical information [**?**]. Their conclusion demonstrated their capacity to explain and forecast. They obtained superior results with the decision tree classifier than with the random forest and CNN classifiers.