Sylhet Engineering College, Sylhet

Department of Computer Science and Engineering

Automatic Text Summarization Using Supervised Machine Learning Approach for Bengali News Documents

Submitted by

(Mir Md. Mahedi Hasan, 2014331551, 4/2, CSE

&

Md. Tanvir, 2014331539, 4/2, CSE)

Supervised by

Summit Haque

Lecturer, Department of Computer Science and Engineering

Shahjalal University of Science and Technology

2nd November, 2019

CERTIFICATION

This thesis titled, "Automatic Text Summarization Using Supervised Machine Learning Approach for Bengali News Documents", submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in 2nd November, 2019.

Group Members:

Mir Md. Mahedi Hasan

Md. Tanvir

Supervisor:

_____

Summit Haque

Lecturer, Department of Computer Science and Engineering

Shahjalal University of Science and Technology

Qualification Form of Bachelor Degree

Students Name : Mir Md. Mahedi Hasan
Md. Tanvir

Thesis Title : Automatic Text Summarization Using Supervised Machine Learning Approach for Bengali News Documents

This is to certify that the thesis submitted by the student named above in November, 2019. It is qualified and approved by the following persons and committee.

Head of the Dept.	**External, Exam. Committee**	**Supervisor**
Abdur Rouf Associate Professor, Dept. of CSE, Sylhet Engineering College	Dr Mohammad Reza Selim Professor, Dept. of CSE, Shahjalal University of Science and Technology, Sylhet	Summit Haque Lecturer, Dept. of CSE, Shahjalal University of Science and Technology

Abstract

Automatic text summarization is the process of compressing large text into a shorter summary text which contains as much as possible the important informations of the original text. With the blessings of technology information is vastly stored in the cloud instead of hard copy documents or compact disk today. Hence summarization has become a basic need to keep these vast information in short and concise way and to find out the needed information fast and easily from these vast information. Though Bengali is one of the most used language in the world there is a lack of proper summarization system and a very few works have done on this language. Hence, in this paper, we present an approach to design an automatic text summarizer for Bengali text that generates a summary by extracting sentences from the source document. It takes care of a single document summarization based on supervised learning approaches. Each sentence in the document is represented by a set of features such as- topic feature, sentence position feature, term frequency inverse document frequency (Tf-idf). After each sentence is given feature values then three supervised machine learning model is used to train the summarizer. The used machine learning models are SVM, naïve bayes and extra tree classifier. After training these models a feature importance vector is extracted for ranking the sentences to generate a summary. Top ranked sentences are then included in the final summary. This experiment was performed on Bengali news articles of different category such as sports, bollywood, politics and crimes. The performance of the proposed method is compared with the human generated summaries. The result of the experiment shows a satisfactory output comparing to the human generated summaries.

Keywords: Bengali Text Summarization; Supervised Machine Learning; SVM; Naïve Bayes; Extra Tree; Sentence Extraction; Summary Generation

Table of Content

Page

List of Tables

List of Figures

Chapter 1
Introduction

The fastest growth of the Internet has resulted in boundless amount of information in natural language that has become increasingly more difficult to access effectively. Lots of people are using internet to keep their data secret and safe from the attack of unauthorized personnel today. We know that World Wide Web in 21st century makes a huge revolution in technological field or to be more specific in the Information Technology, and the more of this revolution are waiting to be happened. People are being more dependent on Internet these days, the production of informational data are increasing rapidly and to maintain and analysis of these data are becoming a biggest need at present. To maintain and use these data for future analysis Data Mining approach has become a buzzword in these decades. Data Mining refers to an approach which examines large scale of pre-existing databases in order to generate new information out of that. A small field of data mining is text mining. There are many types of data used day to day life and for educational, research and business purpose. Out of these data textual data is a type of data which are being used more and more in these decades. For these continuously increasing amount of textual data finding important information is getting harder than before. Text mining is the process of retrieving information from these vast amount of text data. Text mining has a lots of application out of those text summarization is application to summarize textual information. Text summarization is one of the most challenging problems compared to other problems of text mining.

In this chapter we will discuss about text summarization and automatic summarization techniques. Also we will discuss about the classification of text summarization. Then the objective of our project will be described. After that, we will discuss about the steps of summarizing process with brief discussion about extractive text summarization technique. In the end we will discuss about the need of text summarization and current condition of Bengali Language.

1.1 Text Summarization

Text summarization is a process in which a given text is converted into a shorter form containing the important informations of that text keeping in mind. Text summarization for these vast growing information technology can be single document or multi-document texts. By

summarizing a text document we will get a summary which will give precise information about the original text so that we don't have go through the whole document. These saves our valuable time and we can get our needed information quite easily. There is a huge amount of textual data and it is increasing very fast day by day. Those data are unstructured and unorganized for that reason we can't get our information out of those data easily, sometimes it's impossible. To have a better outcome and better overview about those data is to figure out some technique to organize those data in a specific way. There is a significant need to reduce as much as possible of this text data to a shorter and focused summaries that capture the salient details. So we can traverse it more efficiently as well as check whether the larger documents contain the information that we are trying to find out. So now we know the importance of summarization but then the problem appears which is we cannot summarize all the text data manually. That's when we are facing a need of automated procedure to summarize our document. And that's the time when the idea of an automatic text summarizer is introduced. The author of a renowned book named "Automatic Text Summarization" published at 2014 provides 6 reasons to demonstrate the importance of text summarization.

1. Summaries reduce reading time [1].
2. When researching documents, summaries make the selection process easier [1].
3. Automatic summarization improves the effectiveness of indexing [1].
4. Automatic summarization algorithms are less biased than human summarizers [1].
5. Personalized summaries are useful in question-answering systems as they provide personalized information [1].
6. Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process [1].

1.2 Automatic Text Summarization

Text summarization ideas create a great impact on data scientists. Text summarization is very important to data scientists but summarizing text manually is a very difficult task so an automated summarizer is become a necessity to this field. An automatic text summarization is a process through which a coherent and concise version of a long text document is created. Humans are generally good at this types of works. They first read the whole content, understand the meaning of that content and then make a concise summary with the most possible overview of the content. The forever goal of an automatic text summarizer is to create such a human level

summary. The automatic process is not just generating words or phrases that capture the substances of the source document. It is more like a process of generating a meaningful new document that should be easily read. Overall we can say that automatic text summarization is a task of producing a concise, meaningful and fluent summary while preserving key information and overall meaning of the main document.

1.3 Classification of Text Summarization

Text summarization is a process of transforming one or more large text documents into a shorter text document that can represent the main concept of the original text document. A summarized document helps in understanding the main idea of a document instead of reading the whole document. This saves us a lot of time and we can find our desired information easily. Automatic text summarization is that kind of process which generates summaries with the help of a computer program. Text summarization has three main steps. These steps are identifying the topic, interpretation, and last step is summary generation [2]. Automatic text summarization can be classified based on different criteria. The different dimension of text summarization can be generally categorized based on its input type, purpose and output types [3]. There is another type of classification based on uses of external resources.

According to input type there are two types of summarization process. Which are-

i. Single Document: Single document summarization produces summary of single input document. The processing relatively easier. Many of the early summarization system dealt with single document summarization [3].

ii. Multi Document: Multi document summarization produces summary of multiple documents. These multiple inputs are often documents discussing the same topic [3].

According to the purpose there are three types of summarization technique. They are-

i. Generic: Generic summarization purpose is to summarize all texts regardless of its topic or domain; i.e., generic summaries make no assumptions about domain of its source information and view all documents as homogenous texts. The majority of the work that has been done revolves around generic summarization [3].

ii. Domain Specific: There have also been developments of summarization systems which are based on various domain of interest. Such that, summarizing finance articles, biomedical

documents, weather news, terrorist events and many more. Often, this type of summarization requires domain specific knowledge bases to assist its sentence selection process [3].
iii. Query based: Query-based summary contains only information about which are queried by the user. The queries are generally questions and keywords or phrases that are probably related to a particular subject. For example, snippets produced by search engines are an example of query-based application [3].

According to the output types there are two types of process. Those are-
i. Abstraction based: Abstraction based summarization forms an abstract of the original text document by using interpretation procedure and generate a summary that express the same idea in more concise way.
ii. Extraction based: Extractive summaries are generated based on sentence extraction from the original text documents. Extractive summaries contains importance sentences which are directly selected from the original document. Most of the previous and present summarization systems that have been developed are based on extractive summarization. Because abstraction based summarization is very complex task, extraction based summarization is getting more importance to produce summary. In this approach each of the document sentence are given some score based on different criteria and then top most ranked sentences are selected for the final summary. It uses different types of Natural Language Process (NLP) to retrieve information.

And lastly according to uses of external resources there are two different types of summarization techniques-
i. Knowledge poor: This type of summarizer do not use external corpus like Wikipedia and WordNet.
ii. Knowledge rich: This type of summarizer uses external corpus like Wikipedia and WordNet.

1.4 Objectives

The main objective of this project is to provide summaries of a single document based on extraction based summarization technique for different input documents and analyze the results to come up into a fair and gentle conclusion. Another objective is to make a summarizer for Bengali language because there are few Bengali summarizer for this language. As Bengali is

one of the most spoken and written language in the world summarization in Bengali has become a basic need these decade.

1.5 Fundamental Steps for Text Summarization

The fundamental steps for text summarization are given below-

i. Topic identification: The first step of summarization is topic identification. The algorithm that designed based on several aspects to point out some ways to at least have an idea about which topic we are working on. The techniques of topic identification include position, headline, tagline, word frequency, cue phrases.

ii. Interpretation: In this step the whole document is processed to produce a new content that is summarized with almost maximum amount of information from the main data. The most similar lines or sentences are find out to form a summarized document.

iii. Summary generation: After interpretation the automated system then generate summary in this step. The final summary then stored as a new document.

1.6 Extractive Text Summarization

Extractive text summarization is the process of selection of phrases and sentences from the source document to make up the new summary document. This technique involves ranking the connection of phrases in order to choose only those most relevant to the meaning of the source. Extractive summarization methods work by point out important sections of the text and generating them as in the original text. Most of the current automated text summarization systems use extraction based methods to produce a summary. Sentence extraction techniques are mainly used to produce extractive summaries. Extractive methods assign some numerical measure of a sentence for the summary called sentence scoring and then select the best sentences to form document summary based on the compression rate. In the extractive method, compression ratio is an important factor used to express the ratio between the length of the summary and the source text. If the compression rate increases then the summary length will also be increased and more insignificant content will be appeared in the summary. When the compression rate is being reduced the summary length will be decreased and more information will be lost. The acceptable compression rate is 5-30%. There are some common methods of extractive text summarization. These are as follows-

i. Statistical approach

ii. Graph based sentence similarity approach
iii. Machine learning approach
iv. Cluster based method
v. Centroid based summarization
vi. Text summarization with neural networks
vii. Fuzzy logic based method.

In our project, we will use machine learning approach to extract summary sentences based on some statistical feature which will give weights to the sentences of the document. We will use some supervised learning algorithm to train our model and compare their output and accuracy to find out which algorithm is more suitable for generating our desired summary. We will use three supervised learning algorithm and they are support vector machine (SVM), Naïve Bayes Classifier, ExtraTrees. Each of the sentences of the original document is given some feature weights to train these machine learning models. After training the model, summary generation will be the next step. In this step the summary will be generated according to the compression rate.

1.7 Why Bengali?

English text summarization had a severe research output and currently these developed summarizers' works more precisely. In Bengali language, this research is not up to the standard and as a result there is no decent summarizer at all which can be applied in Bangla text processing despite the content and users. A Bengali text summarizer has become obligatory and very demanding nowadays. A huge number of Internet, official or personal users will be benefited by using this automated Bengali text summarizer.

Chapter 2

Related Works

The very first work has done by Luhn [4] in automated summarization based on text extraction. It calculates measures of significance to regulate which sentences of an article may best serve as the auto-abstract. The "significance" factor of a sentence is obtained from an analysis of its words. It is here suggested that the frequency of word occurrence in an article enhances a useful measurement of word significance. It is further suggested that the relative position within a sentence of words having given values of significance enhances a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a mix of these two measurements.

Dragomir R. Radev et al. [5] have introduced a multi-document summarizer, MEAD, which generated summaries by employing cluster centroids generated by topic detection and tracking system. It considered two techniques, a centroid-based summarizer, and an evaluation scheme on the grounds of sentence usefulness and subsumption. The assessment was put through to single and also multiple document summaries. In the end, they developed about two user studies that test the models of multi-document summarization.

Edmundson [6] suggested some new methods in automatic extraction. The automatic extraction system was based on allocating to text sentences numerical weights. For computational understandability the sentence weights were taken as sums of the weights of four basic characteristics. The four fundamental methods are called Cue, Key, Title, and Location. The Cue method is based on the assumption that the probable relevance of a sentence is affected by the presence of reasonable words. The Cue dictionary consists of three sub dictionaries: Bonus words, that are positively relevant; Stigma words, that are negatively relevant; and Null that are positively relevant. Key method is homogeneous to Luhn (high-frequency content words are positively linked). Title method says that sentences containing words that are in the title or headings has higher score value. Location method rank sentences based on their position in the document. Sentences which are under headings have higher score. Sentences which are appeared near beginning or end of the document and/or paragraphs have higher score. Linear combination of four features: a1C + a2K + a3T +a4L where a1, a2, a3, and a4 are the

parameters which are positive integers for the Cue, Key, Title, and Location weights, respectively.

Kamal Sarkar [7] proposed a system that produced extractive summaries for Bengali news documents. The proposed approach has three key features to present each sentence as a numerical value, those features are TF*IDF, sentence position and sentence length. His approach conclude extraction of candidate sentences from the document sentences and ranking those extracted candidate sentences based on their feature weights. Based on this ranked sentences top few sentences were chosen for generating the summary document.

Ansamma John [8] proposed a supervised learning approach of multi-document summarization. He used a random forest classifier to train his model. This classifier is trained using feature score and summary information of each sentence of the document set. Feature scores of sentences of multiple documents which will be summarized are given as the test document for the classifier. From the output of the classifier, sentences that belonging to the summary class, a predetermined size summary is generated using Maximal Marginal Relevance. The experiments are done using the DUC 2002 dataset and its corresponding summary sentences.

Neto [9] has proposed an automatic text summarizer based on Machine Learning approach. He has presented a summarization procedure based on the application of trainable Machine Learning algorithms which employs a set of features extracted directly from the original text. These features are of two types: statistical - based on the frequency of some elements in the text; and linguistic - extracted from a simplified argumentative structure of the text. Two Machine Learning algorithms have been used to train on these features: one of them is C4.5 and the other is naïve bayes.

Efat [10] has proposed an extraction based summarization system which summarizes single Bengali news documents at a time. The proposed summarizer used a sentence scoring and ranking based summarization method. The sentences are ranked based on four basic features: Frequency, Position value, Cue words, Skeleton of the document. After ranking the document sentences top ranked sentences are then selected to form the output summary.

Chapter 3

Proposed Method

The goal of an automatic text summarization is to select the most important sentences from the original text of the Bengali news documents. Our proposed method uses various statistical features to find the most important sentences. Also it uses supervised learning models to find out the feature importance by training those models with these features. Extractive summarization can be considered as a binary classification problem. From the given features of a sentence, a machine learning model will judge how likely the sentence is important to be in the summary. Supervised models used to give better result than the statistical models. In our work we trained three models to see which model gives better outcome on the produced data set and then select one model to find the feature importance to generate the summary from the test documents. Our proposed method has four major steps which are- (1) preprocessing (2) feature scoring (3) train model (4) summary generation.

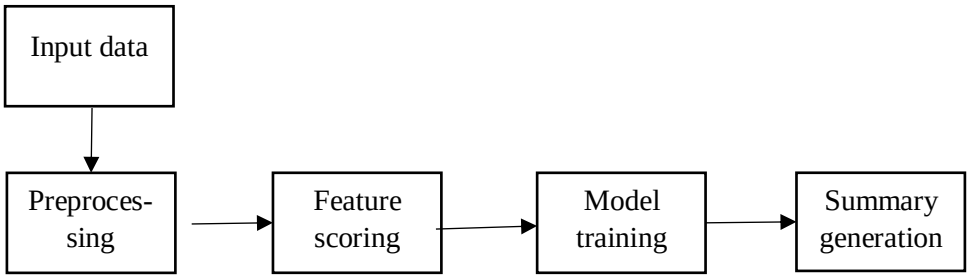

Fig. 1. Steps of the proposed text summarization technique.

3.1 Input Data

Input data is the prerequisite to perform any kind of research work. In our work we used an online provided Bengali news document datasets for summarization task. We collected this dataset from this online site (http://www.bnlpc.org/research.php). There are two datasets for the evaluation of the Bengali text summarization system. Each dataset contains 100 Bangla news document and three human generated model summary for each document. Two groups of scholars each contains three human judges generated these summaries. For our work we had to modify a few documents because those documents didn't come up with any punctuation marks inside. We tried to keep the rest of the dataset absolutely unchanged and for our work purpose we combined these two datasets into one data sets and divided into two parts one for training purpose and another for testing. Out of 200 news documents we selected the first 190 for train and test data for the used models and other 10 documents for generating our summaries and evaluation with human generated summary.

3.2 Preprocessing

In the preprocessing step of our proposed method the input text is first split into sentences. Then these sentences are further broken into words this process is called word tokenization. Tokenization can be two types- one is sentence tokenization and another is word tokenization. For English text there is a tool called NLTK (Natural Language Toolkit). Using this tool word and sentence tokenization can be performed easily. In our proposed method we could only use word tokenization using this tool, sentence tokenization is performed manually because NLTK doesn't support Bengali language for tokenization purpose. After tokenization stop words are removed using a Bengali stop word list created manually. After the removal of stop words stemming is performed to the words which contains the suffix from a given suffix list. This process is also done manually. So the preprocessing step can be divided into three steps- (1) tokenization (2) stop words removal (3) stemming.

3.2.1 Tokenization

Tokenization is the process of classifying sections of a string of input characters. After strings are being tokenized they have sent to the next process. Two types of tokenization is used for text summarization systems- one is sentence tokenization and the other is word tokenization.

Sentence tokenization: For tokenizing sentences from a text manually we have used the punctuation marks in Bengali language. When a sentence finds a punctuation marks it will separate into two different sentences. This process has done for all the documents.

Word tokenization: For tokenizing words for an entire document we used natural language toolkit (NLTK). This toolkit automatically tokenize words and store them to a list. This list is sent to further processing.

3.2.2 Stop Words Removal

Stop words are those words which are most commonly used in every sentences. This words are worthless and do not have any value to the sentences. Having these stop words can increase the sentences weight unnecessarily. So we have to remove these stop words from the input documents. We analyzed that Bengali documents contains 25-30% or more stop words. Example of some of the stop words in Bengali are- “হতে”, “থেকে”, “হয়”, “তবে”, “যা”, “হবে”, “যদিও”, “হোক” etc.

3.2.3 Stemming

In information retrieval stemming is the process of finding the root word of a derived word. In English word for example there are three forms of word “go” which are “going”, “goes”, “gone”. Stemming will reduce these strings to its root form “go”. This process is necessary otherwise similar words will count as different words. For this reason we have used stemming for our method.

The input and corresponding output for the preprocessing step in our work is shown in fig. 2 and fig. 3 respectively.

মানুষের মনে বিচার বিভাগ নিয়ে আস্থার সংকট তৈরি হয় এমন কোনো বক্তব্য দেওয়া থেকে রাজনীতিক, মন্ত্রী-সাংসদসহ সবাইকে বিরত থাকার আহ্বান জানিয়েছেন প্রধান বিচারপতি সুরেন্দ্র কুমার সিনহা।
সম্প্রতি বিচার বিভাগ নিয়ে বিএনপির ভারপ্রাপ্ত মহাসচিব মির্জা ফখরুল ইসলাম আলমগীরের এক বক্তব্যের বিষয়ে তাঁর লিখিত ব্যাখ্যা পাওয়ার পর গতকাল সোমবার প্রধান বিচারপতি এ কথা বলেন। একই সঙ্গে বিচার বিভাগ নিয়ে বক্তব্য দেওয়ার ক্ষেত্রে মির্জা ফখরুলকে সতর্ক করে দিয়েছেন আপিল বিভাগ।
প্রধান বিচারপতির নেতৃত্বাধীন আপিল বিভাগের পাঁচ সদস্যের বেঞ্চে গতকাল মির্জা ফখরুলের বক্তব্যের ব্যাখ্যা জমা দেন তাঁর আইনজীবী জয়নুল আবেদীন। সঙ্গে ছিলেন সগীর হোসেন। রাষ্ট্রপক্ষে ছিলেন অতিরিক্ত অ্যাটর্নি জেনারেল মুরাদ রেজা।
পরে মির্জা ফখরুলের আইনজীবীর উদ্দেশে প্রধান বিচারপতি বলেন, \"আপনারা রাজনীতিবিদরা বিচার বিভাগকে সরকারের অংশ বলেন, এটা ঠিক না। বিচার বিভাগ রাষ্ট্রের একটি অঙ্গ। সংবিধান মানতে হলে বিচার বিভাগকে মানতে হবে। বিচার বিভাগের ওপর আস্থা হারায় এমন কিছু করবেন না। আইনজীবীরা বলেন, সরকারের মন্ত্রীও বলেন, আপনারা বলেন—তাহলে বিচার বিভাগ যাবে কোথায়?\"
গত ৭ ফেব্রুয়ারি সিলেটে এক অনুষ্ঠানে মির্জা ফখরুল অভিযোগ করেন, সরকার বিচার বিভাগকে নিয়ন্ত্রণ করার চেষ্টা করছে।
মির্জা ফখরুলের ওই বক্তব্যের বিষয়ে ১৮ ফেব্রুয়ারি লিখিত ব্যাখ্যা চান আপিল বিভাগ।
গতকাল একই সঙ্গে রাজধানীর পল্টন থানায় নাশকতার তিন মামলায় মির্জা ফখরুলের জামিনের মেয়াদ ১৫ দিন বাড়ান আপিল বিভাগ।
গত বছরের ২৪ নভেম্বর পল্টন থানায় নাশকতার তিন মামলায় মির্জা ফখরুলকে তিন মাসের জামিন দেন হাইকোর্ট।
পরে ওই জামিন অব্যাহত রাখার জন্য আপিল বিভাগে যান মির্জা ফখরুল।

Fig. 2. Sample input text

[[মানুষ', বিচ', বিভাগ', আস্থ', সংকট', তৈরি', রাজনীতিক', মন্ত্রী-সাংসদসহ', সবাই', বিরত', থাক', আহ্বান', জানিয়', প্রধান', বিচারপতি', সুরেন্দ্র', কুম', সিনহা'], [বিচ', বিভাগ', বিএনপির', ভারপ্রাপ্ত', মহাসচিব', মির্জা', ফখরুল', ইসলাম', আলমগীর', এক', বক্তব্য', বিষয়', লিখিত', ব্যাখ্যা', পাওয়', গতকাল', সোমব', প্রধান', বিচারপতি', কথা'], [বিচ', বিভাগ', মির্জা', ফখরুল', সতর্ক', আপিল', বিভাগ'], [প্রধান', বিচারপতির', নেতৃত্বাধীন', আপিল', বিভাগ', পাঁচ', সদস্য', বেঞ্চ', গতকাল', মির্জা', ফখরুল', বক্তব্য', ব্যাখ্যা', জমা', আইনজীবী', জয়নুল', আবেদীন'], [সগীর', হোস'], [রাষ্ট্রপক্ষ', অতিরিক্ত', অ্যাটর্নি', জেনার', মুরাদ', রেজা'], [মির্জা', ফখরুল', আইনজীবীর', উদ্দেশ', প্রধান', বিচারপতি', আপনা', রাজনীতিবিদ', বিচ', বিভাগ', সরকার', অংশ'], [বিচ', বিভাগ', রাষ্ট্র', অঙ্গ'], [সংবিধান', মান', বিচ', বিভাগ', মান'], [বিচ', বিভাগ', ওপর', আস্থা', হারায়'], [আইনজীবী', সরকার', মন্ত্রী', আপনা', বলেন—তাহল', বিচ', বিভাগ', কোথায়'], [গত', ৭', ফেব্রুয়ারি', সিলেট', এক', অনুষ্ঠান', মির্জা', ফখরুল', অভিযোগ', সরক', বিচ', বিভাগ', নিয়ন্ত্রণ'], [মির্জা', ফখরুল', বক্তব্য', বিষয়', ১৮', ফেব্রুয়ারি', লিখিত', ব্যাখ্যা', আপিল', বিভাগ'], [গতকাল', রাজধানীর', পল্টন', থানায়', নাশকত', তিন', মামলায়', মির্জা', ফখরুল', জামিন', মেয়াদ', ১৫', বাড়ান', আপিল', বিভাগ'], [গত', বছর', ২৪', নভেম্বর', পল্টন', থানায়', নাশকত', তিন', মামলায়', মির্জা', ফখরুল', তিন', মাস', জামিন', হাইকোর্ট'], [জামিন', অব্যাহত', রাখ', আপিল', বিভাগ', মির্জা', ফখরুল']]

Fig. 3. Sample output after preprocessing phase

3.3 Feature Scoring

After preprocessing documents are then used to compute the feature score of each sentence. We selected three different features of sentences for scoring which are topic feature, TF-IDF and sentence position feature. Every sentence of every documents of the dataset will compute these features and assign feature weights to each sentences. After completing scoring each sentence

feature weight final dataset will form and those data will be then labeled and used to train and test the machine learning models. The features selected for our proposed method are as follows:

3.3.1 Topic Feature

Each news documents of our experimental dataset has a title. Title of a document contains the important idea about the document. So when summarizing a document title feature can be an important feature. Usually titles of the documents are not include in the summary documents. So sentences with similar information as that of the title will consider as important sentences for the summary. This feature weight is calculated using cosine similarity measure for each sentence. To achieve this feature weight each sentence and the entire document is represented as a vector in the 'n' dimensional vector space where 'n' is the number of stemmed unique non-stop words in the document set. Let S and T represents a sentence vector and a title vector respectively for a document. Title feature represents as Tf of a sentence S is calculated using equation (1).

$$Tf = \frac{\sum_{k=1}^{n} S_k * T_k}{\sqrt{\sum_{k=1}^{n} {S_k}^2 * \sum_{k=1}^{n} {T_k}^2}} \quad (1)$$

3.3.2 TF-IDF Feature

TF-IDF short form of term frequency inverse document frequency is used to measure how important a word to a document in our case sentence. TF-IDF assigns each word in a document a weight based on its term frequency (tf) and inverse document frequency (idf). The words with higher weight scores will be considered to be more important. The TF-IDF score for each sentence in a document is calculated as follows:

$$W_{tfidf} = \sum_{w \in s} TF(w) * IDF(w) \quad (2)$$

Where:

TF(*w*): the number of times a word '*w*' occurs in the document,

IDF(*w*): *log(N/df(w))*, where *N* = total number of documents in our corpus and *df(w)* called document frequency of a word *w* indicates the number of documents where the word *w* occurs at least once.

3.3.3 Sentence Position Feature

In this method, a sentence score is calculated based on the position of a sentence in the input document. Sentence position weight is calculated by assigning higher values to the starting and ending sentences of a document. Sentences in the middle of a document are less important according to the position of sentence feature. The score of a sentence is calculated through this formula:

$$P = \frac{1}{\sqrt{i}} \qquad (3)$$

Where: *P* = the weight of a sentence S due to its position in the document
i = the position of the sentence in the document.

3.4 Labeling data

After each sentence of all the documents were given feature weights the next step is to give a class to all the sentences of the documents. As text summarization is considered a binary classification problem there will be only two class given to the sentences based on the summary sentences. A sentence class will be either in summary or not in summary. In the feature vector this classes are represents as 0 or 1, if a sentence is in the summary sentence then the class will be 1 and if a sentence is not in the summary sentences then the class of this sentence will be 0. This labeling process is done by checking if a document sentence is present in the human generated extractive summary. Then 0 or 1 value is assign as a class to each of the document sentences. After labeling the data final dataset is prepared which is in a vector representation of the feature values against each sentences and a class value. Then with this data a classifier model is trained.

3.5 Training Classifier

Text is considered as a binary classification problem. All sentences are classified into two class either in summary or not in summary. Then a supervised learning algorithm is applied to train from the given data. In our proposed method we trained 3 machine learning classifier such as- SVM (Support Vector Machine), Naïve Bayes Classifier, Extratrees (Extremely Randomized Trees). For train the preprocessed and featured weighted data we split the data into two section. One section is used for training the model and the other for testing the model to find how accurately our used model can classify the sentences according to their two classes. Out total dataset used for training phase is 190 documents. We used 80% of this collection for training purpose and the rest 20% used for testing the model. The work flow of the model training is given below:

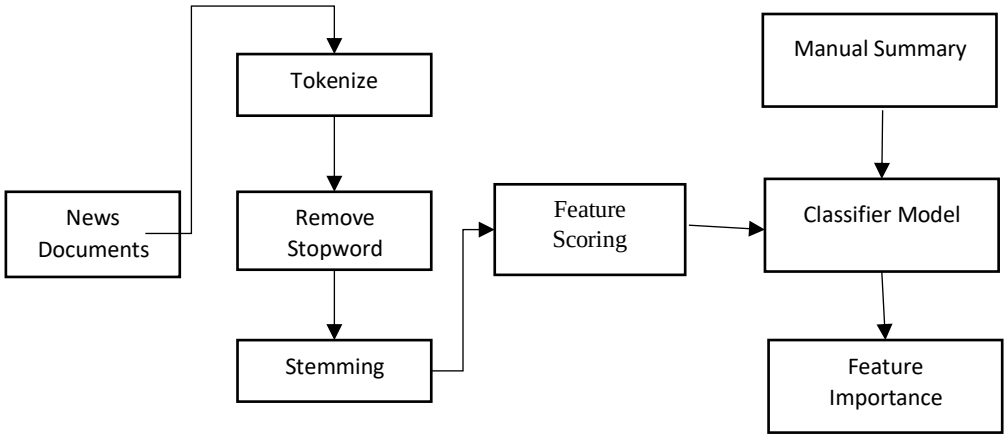

Fig. 4. Training Phase

Fig. 4 shows how a model is trained for out proposed method. After training the model feature importance is calculated from the trained model. This feature importance of a model is then applied to generate summary from the input document. Feature importance provides a score that indicates how important or valuable each feature was in the construction on our used model in that case extratrees model. The higher feature importance value indicates that the feature is more important to the data. Machine learning models provide this feature importance class to

calculate feature importance. After calculating the feature importance the next step is generating the summary.

Table 1. Data training example

Sentence	Topic feature	TF-IDF	Sentence position feature	Summary
1	0.214286	0.140703	1	Yes
2	0.0	0.159316	0.904762	No
3	0.142857	.305215	0.809524	No
4	0.0	0.302812	0.714286	No
Etc.	0.111111	0.151662	0.428571	Yes

Table 1 explains the final dataset we used for training our model. Table 1 shows that every sentence are the sample data , topic feature, tf-idf and sentence position are attributes and the summary columns are the attribute target whose two attributes are Yes or NO. Attribute Yes means sentence appears in the summary while No means the sentence doesn't appear in the summary.

3.6 Summary Generation

In this phase a summary is generated from the input document using feature importance values from the training phase of our model. For generating summary we have to go through all the preprocessing and feature scoring methods we used to train our models. Summary generation phase is shown in fig. 5.

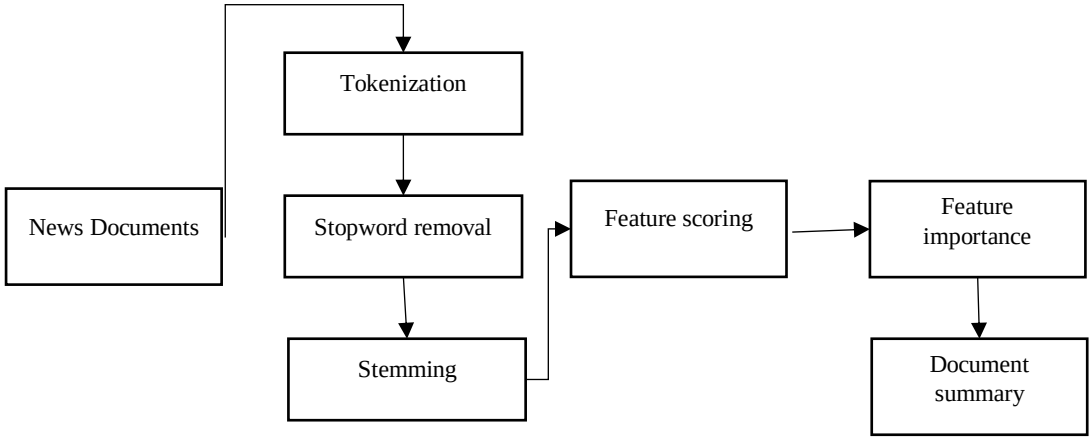

Fig. 5. Summary generation phase

After tokenization, stopword removal and stemming the input data then scored according to their feature weights. After that, each feature score will multiply by the feature importance score for the same feature to each of the sentences. After multiplying all the three features their summation is taken for sentence ranking process. After ranking of sentences top ranked sentences are selected according to the compression rate. The final summary is generate in this process.

Chapter 4

Experiments and Result

In this chapter we discuss about the experiment we conducted for generating summary of the Bengali news documents. For implementing the above discussed approach we used python-3 as programming language and a compatible environments. We used the encoding format UTF-8 for out experiments of Bengali text. Feature importance is used to rank the sentences which we produced in the training phase of our extratrees model.

4.1 Data Analysis

As we discuss in the previous chapter the source of our data, we will discuss the further analysis of our data set to this section. Sentences from the dataset are extracted and represented as a vector in vector space model after preprocessing. We then extracted similarity between sentences in the document with the title sentence which is called topic feature, then term frequency inverse document frequency is calculated for each sentences and sentence position feature is given to these sentences. Our data set contains news articles from different categories of news namely Bollywood, politics, sports, crime etc.

Table 2. Characteristics of dataset

No. of articles collected	190 Bengali new articles
Domain categories	Sports, Crime, Bollywood, Politics
Average Sentences in each documents	10-20
Average Words	250-350

To generate the tagged corpus, the sentences are classified manually into two classes namely Yes or No based on the human generated summary. If a sentence is presented in the human generated summary then that sentence is tagged with Yes and if the sentence is not present then tagged with No. In the vector representation of the sentences class of the sentences assigned as 1 or 0.

4.2 Evaluation

For evaluation of our used models confusion matrix is generated which contains information about the actual and predicted classification done by classifier. We used three different binary classifier model which classifies each sentence into one of the two classes either true (sentence is in the summary) or false (sentence not in summary). Binary classifier can give four possible classification for each instance: a true positive (actually in summary and predicted as in summary), a true negative (actually not in summary and predicted not in summary), a false positive (actually not in summary but predicted as in summary) or false negative (actually in summary but predicted as not in summary). From these performance metrics accuracy, precision and recall can be calculated using (4).

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative} \quad (4)$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (5)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (6)$$

Accuracy generated from naïve bayes classifier is 66%.

Table 3. Confusion matrix for naïve bayes classifier

Class Label	Precision	Recall	F1-score
insummary	0.67	0.97	0.79
notinsummary	0.62	0.09	0.16
accuracy			0.66
macro avg	0.65	0.53	0.48
weighted avg	0.65	0.66	0.57

Accuracy of the SVM classifier is 65%.

Table 4. Confusion matrix for SVM classifier

Class Label	Precision	Recall	F1-score
insummary	0.65	1.00	0.79
notinsummary	0.00	0.00	0.00
accuracy			0.65
macro avg	0.33	0.50	0.39
weighted avg	0.42	0.65	0.51

Accuracy of the ExtraTree classifier is 70%

Table 5. Confusion matrix for ExtraTree classifier

Class Label	Precision	recall	F1-score
insummary	0.71	0.92	0.80
notinsummary	0.67	0.30	0.42
accuracy			0.70
macro avg	0.69	0.61	0.61
weighted avg	0.70	0.70	0.67

From table 3, 4, 5 we can see that extratree performs better for our dataset. So we can use this classifier to calculate feature importance for ranking the sentences.

4.3 Results

We have tested our proposed method for several input Bengali news documents. Our method successfully generate summary which are close to the human generated summary. Some of the sample input and output are given below to show our work-

Sample Input 1:

Title:
কাঠগড়ায় পুলিশ Text: চা-দোকানি বাবুল মাতবর (৫০) সবকিছুর ঊর্ধ্বে চলে গেছেন। তাঁর কাছে কেউ আর চাঁদা চাইতে পারবে না। চাঁদা না পেয়ে কেউ লাথি মেরে জ্বলন্ত চুলার ওপর ফেলে দিতে পারবে না। কারণ, পুড়ে অঙ্গার বাবুল গতকাল বৃহস্পতিবার দুপুরে ঢাকা মেডিকেল কলেজ হাসপাতালের বার্ন ইউনিটে মারা গেছেন।
বাবুলের ওপর নিষ্ঠুরতার সঙ্গে জড়িত পুলিশ ও পুলিশের তথ্যদাতাদের (সোর্স) বিচারের দাবিতে গতকাল দুপুরে মিরপুর ১ নম্বর গুদারাঘাটের সড়ক অবরোধ করে বিক্ষোভ করেছেন এলাকাবাসী। সংসদে ৩০০ বিধিতে স্বরাষ্ট্রমন্ত্রীর বিবৃতি দাবি করেছেন একজন সাংসদ।
ভোলা সদর উপজেলার মুন্সিচরের বাবুল মাতবর মিরপুর ১ নম্বরের গুদারাঘাটের ফুটপাতে চায়ের দোকান চালাতেন।
গত বুধবার রাত সাড়ে নয়টার দিকে শাহ আলী থানার চার পুলিশ সদস্যের উপস্থিতিতে তাঁদের সোর্স বাবুলের কাছ চাঁদা দাবি করেন।
চাঁদা দিতে রাজি না হওয়ায় বাবুলকে লাথি মেরে কেরোসিন তেলের জ্বলন্ত চুলার ওপর ফেলে দেন সোর্স দেলোয়ার।

Sample Output 1:

চা-দোকানি বাবুল মাতবর (৫০) সবকিছুর ঊর্ধ্বে চলে গেছেন।
তাঁর কাছে কেউ আর চাঁদা চাইতে পারবে না।
চাঁদা দিতে রাজি না হওয়ায় বাবুলকে লাথি মেরে কেরোসিন তেলের জ্বলন্ত চুলার ওপর ফেলে দেন সোর্স দেলোয়ার।

Human generated summary for sample input 1:

চা-দোকানি বাবুল মাতবর (৫০) সবকিছুর ঊর্ধ্বে চলে গেছেন।
কারণ, পুড়ে অঙ্গার বাবুল গতকাল বৃহস্পতিবার দুপুরে ঢাকা মেডিকেল কলেজ হাসপাতালের বার্ন ইউনিটে মারা গেছেন।
চাঁদা দিতে রাজি না হওয়ায় বাবুলকে লাথি মেরে কেরোসিন তেলের জ্বলন্ত চুলার ওপর ফেলে দেন সোর্স দেলোয়ার।

Analysis: Comparing our proposed methods output and human generated output we can see that both of the summary almost same except out method couldn't predict the second sentence correctly for this 3 line summary.

Sample Input 2:

Title:
সাংবাদিক ফরহাদ খাঁ দম্পতি হত্যার রায় দ্রুত কার্যকরের দাবি Text: সাংবাদিক ফরহাদ খাঁ দম্পতি হত্যাকারীদের ফাঁসির রায় দ্রুত কার্যকরের দাবি জানিয়েছে সাংবাদিক ফরহাদ খাঁ স্মৃতি ফাউন্ডেশন। গতকাল শনিবার বেলা ১২টায় জাতীয় প্রেসক্লাবের সামনে আয়োজিত এক মানববন্ধন কর্মসূচি থেকে এ দাবি করা হয়।
২০১১ সালের ২৭ জানুয়ারি রাতে রাজধানীর নিজ বাসায় ফরহাদ খাঁ দম্পতিকে হত্যা করা হয়। ২০১২ সালে বিচারিক আদালত হত্যার দায়ে দুজনকে ফাঁসির দণ্ড দেন।
মানববন্ধনে বাংলাদেশ ফেডারেল সাংবাদিক ইউনিয়নের মহাসচিব ওমর ফারুক, ঢাকা সাংবাদিক ইউনিয়নের সাবেক সাধারণ সম্পাদক শাবান মাহমুদ, সাংবাদিক মৃণাল চক্রবর্তী, শাকিল আহমেদ চৌধুরী, সাংবাদিক ফরহাদ খাঁ স্মৃতি ফাউন্ডেশনের সদস্যসচিব হাসান টগর, যুগ্ম আহ্বায়ক আতিকুল ইসলাম প্রমুখ বক্তব্য দেন।

Sample Output 2:

সাংবাদিক ফরহাদ খাঁ দম্পতি হত্যাকারীদের ফাঁসির রায় দ্রুত কার্যকরের দাবি জানিয়েছে সাংবাদিক ফরহাদ খাঁ স্মৃতি ফাউন্ডেশন।

Human generated summary for sample input 2:

সাংবাদিক ফরহাদ খাঁ দম্পতি হত্যাকারীদের ফাঁসির রায় দ্রুত কার্যকরের দাবি জানিয়েছে সাংবাদিক ফরহাদ খাঁ স্মৃতি ফাউন্ডেশন। গতকাল শনিবার বেলা ১২টায় জাতীয় প্রেসক্লাবের সামনে আয়োজিত এক মানববন্ধন কর্মসূচি থেকে এ দাবি করা হয়।

Analysis: From sample 2 input out method gives one line summary which compared to human generated summary is a good summary.

Sample Input 3:

Title:
মজুত চাল নিয়ে বিপাকে সরকার Text: এখন বাজারে মোটা চালের কেজি ৩০ থেকে ৩২ টাকা। আর খাদ্য অধিদপ্তর খোলাবাজারে বিক্রি করছে ২০ টাকায়। কিন্তু এই চালের মান পড়ে যাওয়ায় সহজে বিক্রি হচ্ছে না। তাই ৩২ টাকা কেজি দরে কেনা এই চাল এখন ১৫ টাকায় বিক্রির পরিকল্পনা করছে খাদ্য মন্ত্রণালয়।
এদিকে আমনের ৮৯ হাজার টন চাল সংগ্রহ বাকি আছে। আগামী মে থেকে ১০-১২ লাখ টন বোরো সংগ্রহ শুরু হবে। এ জন্য গুদাম খালি করতে হবে। কিন্তু সরকারের খাদ্য বিক্রির অন্যতম উপায় খোলাবাজারে চাল (ওএমএস) বিক্রি প্রায় বন্ধ হয়ে আছে। সাতটি সামাজিক নিরাপত্তা কর্মসূচিতে বরাদ্দ করা খাদ্যের বণ্টন ও বিক্রিও চলছে শ্লথগতিতে। গত ১৩ নভেম্বর খাদ্য অধিদপ্তর থেকে মন্ত্রণালয়ে পাঠানো প্রতিবেদন অনুযায়ী, গুদামে থাকা চালের বড় অংশের মান ক্রমশ কমছে। ওই প্রতিবেদনে দ্রুত চাল খালাসের নির্দেশনা চাওয়া হয়েছে। কিন্তু এখনো সরকারি গুদামে মান পড়ে যাওয়া এক লাখ পাঁচ হাজার টন চাল রয়ে গেছে।
গত বুধবার ত্রাণ মন্ত্রণালয়বিষয়ক সংসদীয় স্থায়ী কমিটির সভায় সাংসদেরা টিআর ও কাবিখায় চাল-গম চান না বলে জানিয়ে দিয়েছেন। তাঁরা সরকারের কাছে সামাজিক নিরাপত্তা কর্মসূচির জন্য নগদ অর্থ চেয়েছেন।

Sample Output 3:

এখন বাজারে মোটা চালের কেজি ৩০ থেকে ৩২ টাকা। আর খাদ্য অধিদপ্তর খোলাবাজারে বিক্রি করছে ২০ টাকায়। কিন্তু এই চালের মান পড়ে যাওয়ায় সহজে বিক্রি হচ্ছে না। এ জন্য গুদাম খালি করতে হবে।
তাঁরা সরকারের কাছে সামাজিক নিরাপত্তা কর্মসূচির জন্য নগদ অর্থ চেয়েছেন।

Human generated summary for sample input 3:

এখন বাজারে মোটা চালের কেজি ৩০ থেকে ৩২ টাকা। আর খাদ্য অধিদপ্তর খোলাবাজারে বিক্রি করছে ২০ টাকায়। কিন্তু এই চালের মান পড়ে যাওয়ায় সহজে বিক্রি হচ্ছে না। তাই ৩২ টাকা কেজি দরে কেনা এই চাল এখন ১৫ টাকায় বিক্রির পরিকল্পনা করছে খাদ্য মন্ত্রণালয়। গত বুধবার ত্রাণ মন্ত্রণালয়বিষয়ক সংসদীয় স্থায়ী কমিটির সভায় সাংসদেরা টিআর ও কাবিখায় চাল-গম চান না বলে জানিয়ে দিয়েছেন।

Analysis: In our sample input 3 we can see that the machine generated summary has first 3 sentences from the model summary exactly same out of 5 sentences. Other two sentences are different for our machine generated summary. So analyzing these sample input and output we can say that out proposed method perform better than the previously applied statistical approaches. From the experiment result we have seen that our machine generated summary is 50-60% similar to the human generated summary.

Chapter 5

Conclusion

In the whole project we have tried to implement supervised learning approach to summarize a single text document in Bengali language. We have analyzed the samples and calculated the numbers. From the analysis we have come to a decision that to make things more efficient and accurate we need to design our own Bangla WordNet, stemmer and corpus. And after that we will be able to get better results from our project.

In spite of having many challenges to complete our project because of the constraints of the resources out method performs satisfactory. Our prosed work is based on learning based approach in which we have trained 3 classifier such as SVM, naïve bayes and extratree. Out of these three classifier extratree performs better than the other two. We have used these classifier to train on our dataset and generate a feature importance vector so that we can apply this vector weights for ranking our sentences to give them class according to their importance in the summary class.

Chapter 6

Future Works

Our proposed method performs well though the performance of this approach can be improved. By introducing more features like cue words, bigrams, trigrams, named entity recognition etc. Also we have to design our own Bangla WordNet, Stemmer and corpus to improve performance of many other works include ours. More labeled data can be added to this project to improve its performance. We have used three supervised learning model, there are many other models out there, training those models may give better results. This method can be also extended to work on multiple documents summarization. Also the same technique can be applied on domains other than news and later we can study the effects of various domain characteristics on the suggested features and overall performance of the technique.

References

[1] J.-M. Torres-Moreno, Automatic Text Summarization, Great Britain: ISTE Ltd, 2014.

[2] N. S. M. J. Yazan Alaya AL-Khassawneh, "Improving Triangle-Graph Based Text Summarization," *Indian Journal of Science and Technology,* vol. 10, no. 8, pp. 1-15, 2017.

[3] Y. a. G. O. S. a. B. H. a. N. H. C. a. S. P. Jaya Kumar, "A review on automatic text summarization approaches," *Journal of Computer Science ,* vol. 12, no. 4, pp. 178-190, 2016.

[4] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development,* vol. 2, no. 2, pp. 159 - 165, 1958.

[5] H. J. M. S. D. T. Dragomir R. Radev, "Centroid-based summarization of multiple documents," *Information Processing and Management,* vol. 40, no. 6, p. 919–938, 2004.

[6] H. P. EDMUNDSON, "New Methods in Automatic Extracting," *Journal of the Association for Computing Machinery,* vol. 16, no. 2, pp. 264-285, 1969.

[7] K. Sarkar, "An Approach to Summarizing Bengali News Documents," *Proceedings of the International Conference on Advances in Computing, Communications and Informatics - ICACCI,* 2012.

[8] D. M. W. Ansamma John, "RANDOM FOREST CLASSIFIER BASED MULTI-DOCUMENT SUMMARIZATION SYSTEM," *IEEE Recent Advances in Intelligent Computational Systems (RAICS),* 2013.

[9] A. A. F. a. C. A. A. K. Joel Larocca Neto, "Automatic Text Summarization Using," *Lecture Notes in Computer Science ,* vol. 2507, pp. 205-215, 2002.

[10] M. I. ,. H. K. Md. Iftekharul Alam Efat, "Automated Bangla Text Summarization by Sentence," *International Conference on Informatics, Electronics and Vision (ICIEV),* 2013.