BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Bengali Text Summarization Using TextRank, Fuzzy C-means and Aggregated Scoring Techniques

AUTHORS
Alvee Rahman
Fahim Md Rafiq
Ramkrishna Saha
Ruhit Rafian

SUPERVISOR
Mr. Hossain Arif
Assistant Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

We would like to dedicate this thesis to our loving parents …

Declaration

We, hereby declare that this thesis report is based on our own work and research and this report has not been submitted anywhere for any other degree or professional qualifications. The contents in this report have been prepared by us for our final undergraduate thesis and any other materials of work by other researchers have been acknowledged and referenced in the reference section.

Authors:

Alvee Rahman
Student ID: 15101036

Fahim Md Rafiq
Student ID:15101056

Ramkrishna Saha
Student ID: 15101024

Ruhit Rafian
Student ID:14201028

Supervisor:

Hossain Arif
Assistant Professor, Department of Computer Science and Engineering
BRAC UNIVERSITY

December 2018

The thesis titled “Bengali Text Summarization Using TextRank, Fuzzy C-means and Aggregated Scoring Techniques”

Submitted by:

Alvee Rahman **:** 15101036

Fahim Md Rafiq **:** 15101056

Ramkrishna Saha **:** 15101024

Ruhit Rafian **:** 14201028

of Academic Year Fall 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of BSc in Computer Science and Engineering

1. _____

Md. Abdul Mottalib

Professor & Chairperson

Department of CSE

BRAC University

2. _____

Hossain Arif

Assistant Professor

Supervisor

BRAC University

Acknowledgement

We take this opportunity to express our sincere gratitude to our supervisor Mr. Hossain Arif for his immense support during the period of our undergraduate thesis. His valuable guidance and immense knowledge in this field helped and motivated us in writing this thesis.

We would also like to thank all the faculty members of the Department of Computer Science and Engineering, BRAC University, for helping us with all the necessary support.

Abstract

In this world, it is very difficult and time consuming for humans to summarize large documents, reports, news and research articles. Multiple text summarization techniques play vital roles in picking the important points and sentences thus reducing the time and effort required to read a whole article. Numerous summarization techniques have been applied to the English language but comparatively work on Bengali text summarization is still limited. Furthermore, in our country, Bangladesh, all summarization is mainly done by humans. Keeping that in mind we aim to find a simple way of summarizing Bengali texts with the technology at hand. Text summarization can be of two types, either abstractive or extractive. In this paper we will use extractive text summarization to summarize Bengali passages, using Fuzzy C-Means, TextRank and Aggregate Sentence Scoring methodologies. We have also done a comparative study, among the 3 methodologies we have used and aim to find the most precise methodology for Bengali text summarization.

Table of Contents

List of Figures

List of Equations

List of Tables

Nomenclature

Acronyms / Abbreviation

FCM Fuzzy C-Means

NLP Natural Language Processing

NLTK Natural Language Toolkit

PCA Principal Component Analysis

PHP Hypertext Preprocessor

ROUGE Recall-Oriented Understudy for Gisting Evaluation

TF-IDF Term Frequency - Inverse Document Frequency

Chapter 1

Overview

1.1 Introduction

In an era where everyone needs to be updated every second, but no one has the adequate time to read and stay informed. So the world needs real-time automatic text summarization to help them stay informed with the least time consumed. Text summarization can also be used to skim through large Bengali documents and then deciding which one to read if it seems interesting enough. With the advancement in technologies in Bangladesh, the language Bengali is increasingly being used in almost all online platforms, hence the need of Bengali Text Summarization. Here in this proposed system, the objective is to take in articles written in Bengali and convert them into a shorter version, preserving the true meaning of the article. Text Summarization is primarily divided into two major sections: Extractive and Abstractive. In the Extractive approach, the system simply omits the sentences that possess the least weight in the true meaning of the given text, and generate a shorter and more precise version of the passage [11, 12]. In the abstractive approach, a summary of the original text is built keeping the same meaning and the theme intact. The summary built will be much like a one written by a human [10]. The main key points of a text are identified and then understandable sentences are constructed in a concise manner. This paper namely discusses the extractive approach which has been widely used over the years for summarization purpose.

News articles have been manually fed into the system which were collected from the national daily “The Daily Prothom Alo”, the system then processes the data before it can be summarized; in preprocessing, the system tokenizes the extract, and removes the stopwords from the extract so that they have no influence in the summary generation. After the removal of stop words, the system stems the words to their root forms, so that all the words generated from a common root is considered as a single unit.

The system primarily focuses on Fuzzy C-Means Clustering Algorithm [19] to generate an optimal summary. Along with FCM, TextRank [24] and Aggregate Sentence Scoring [7,12,13,14] has also been implemented to provide a comparative study at the end. For a uniform and accurate evaluation for the comparative study, the system uses the ROUGE [32] scoring method and later calculates F-Measure to provide an understandable illustrative comparative study.

1.2 Thesis Orientation

- Chapter 2 contains the Literature review which talks about previous works done on English and Bengali language. It also contains information about the algorithms used in the system.
- Chapter 3 talks about the Proposed Model which consist of the System Workflow, Dataset, Preprocessing, Sentence Scoring and the Classifying methodologies used.
- Chapter 4 shows the Evaluation, Results, Comparative study and Analysis.
- Chapter 5 finally concludes the paper and talks about the future scopes of FCM in Bengali Text Summarization.

1.3 Motivation

Bangladesh is currently going through a digital revolution, every day we are having newer innovations in our country, which in return requires a better support system. Like the rest of the world, Bangladesh also has most of its services delivered digitally to its subscribers. Monthly subscriptions of printed newspapers have declined dramatically in recent times, and more people have become dependant upon digital content. Such change in lifestyle, require us to have a system that will summarize all our documents in fractions of a second so that we are always informed about everything we intend to know, but with sparing much time of our fast-paced life.

1.4 Objective

The objective of this paper is to propose a system which would use two new algorithms to generate Bengali text summarization. The two algorithms proposed are as follows: 1) Fuzzy C-Means, 2) TextRank. Since work done on Bengali text summarization is limited, we hope our research will shine a new light on the subject and will open more doors for further research.

1.5 Challenges

Implementing a text summarization for the language Bengali was not as straightforward as it is for the more global language English. Numerous summarization projects has been carried out on English, and this lead to the availability of easily accessible packages and libraries which conducts the preprocessing of the test data in seconds. For Bengali, no such library could be found, hence codes had to be written from the very scratch to make the system a success.

Another challenge that the system had to overcome was that, in Bengali text words are generally for in their root (dictionary) form; Bengali language syntax tend to alter the words to match it with the context of the sentence which is very different from the English language syntax. In English, the maximum extent to which a sentence can be altered is adding a suffix/

prefix and this simple alteration can be easily taken care with the lemmatization method available in the Natural Language Processing Toolkit (NLTK) written in Python. For Bengali, the system had to be installed with a seperate stemming class, which converts every word into their very root format.

Chapter 2

Literature Review

Automatic text summarization was first introduced by Luhn [9] in 1958 where he proposed the idea of calculating the word frequencies in sentences and later using those to score sentences, ultimately selecting the highest ranked sentences for the summary. In recent years there have been numerous approaches toward automatic text summarization. Some of the approaches included abstract summarization techniques [10] such as structure-based approach and semantic-based approach. Among the extractive summarization techniques [11,12] cluster-based method, summarization with a neural network, graph-based method, latent semantic analysis (LSA) method, fuzzy logic based, query-based method are some of the most effective and popular ones. Text summarization has been an important application of natural language processing till date and although researchers have explored various methods of text summarization in English language, very few have been done in other natural languages like Bengali.

Despite the fact that Bengali text summarization is not as widely popular as English Text summarization, it did get an ample importance for being an emerging field of research in recent decades. First work in the Bengali text summarization field was done by Islam et al. in [1] in the year 2004. They proposed a keyword search-based technique for multiple documents where their corpus-based search engine searches the keyword in multiple documents and then makes a summary of the relevant documents. Later on, it was followed by Uddin and Khan [2] who implemented a summarizer using Java where they used location method, cue method, title, and term frequency to rank the sentences. The first 40% higher ranked sentences from a given text was given as the output. More methods such as TF*IDF, positional value, and sentence length was used by Sarkar [3] for summarizing Bengali news documents. His idea was to generate the main gist of a news article in order to aid the reader with an idea of the whole article. He used 30 Bengali documents and created a reference summary for each for evaluation purpose.

Efat et al. in [4] did a similar research using word frequency, cue words, sentence positional value and skeleton of the document for sentence scoring purposes. Their work showed an 83.57% match with human-generated summaries but their system's accuracy highly depended on the usage of keywords throughout the document.

Furthermore, a more sophisticated approach was taken by Das and Bandyopadhyay in [5] where they made a topic based opinion summarization system. Their system does two tasks: 1) Finding the theme of the document. 2) Finding the summary of the document. It does the first part by finding out the sentiment information in a document by following a topic-sentiment model which uses clustering model such as K-means and uses a theme relational graph technique for finding out the document level summary. Their theme detection technique generated 83.60% precision while summarization system generated 72.15%.

Another work was done by Sarkar in [6] who only used the TF*IDF model along with positional value and sentence length to generate a summary of a single document. He only used a single reference summary generated from LEAD baseline for evaluation which undermines the accuracy of the summarizer. Sarkar further continued his research [3] and used other systems such as System3, Baseline system2 to generate reference summaries for comparison.

A more rigorous research was done by Abjuar et al. in [7] where they used the following for word analysis and scoring: frequency, numeric value identification, repeated word distance, and cue words. For sentence analysis and scoring, they used the summation of frequent words, sentence length, sentence position, uniform sentences, imitation sentences, the skeleton of a document, frequent word percentile, prime sentences, aggregate similarities, and final gist analysis. They tested their system with 3 different Bengali texts and compared them with a human-generated summary.

Akter et al. in their paper [8] used a different approach in selecting sentences for generating a summary. They used K-means clustering after sentence ranking to choose the best and worst n-sentences for generating the summary. This shone a different light on selecting the sentences for generating the summary as worst scored sentences were not used in generating summaries before for Bengali text summarization. A newer approach was taken by Haque et al. [17] in the sense that they replaced pronoun by a corresponding noun. Furthermore, they always included the first sentence of the document in the summary generated. For the summary generation, they used popular methods for the sentence ranking and used one-third top-ranked sentences in the final summary and evaluated it using F-measure.

Out of the handful of brilliant groundbreaking researching, that have been done on Bengali Text Summarization, none have used the concepts of Fuzzy C-Means. However, implementation of the FCM algorithm, have been in talks for English language processing for quite some time now. In [18] Patil et. al. have proposed a text miner which is based on the Fuzzy C-means algorithm. Document clustering is a very important part of text mining and has two parts, namely hard clustering and soft clustering. In hard clustering, a data point belongs to only one cluster. Whereas, in the case of soft clustering, a data point may belong to multiple clusters. Each data point is associated with a membership function, which expresses the degree of its membership to a specific cluster. After the sentence clustering, it has been seen that clustering done using the Fuzzy C-Means algorithm outperforms the traditional K-means algorithm.

Fuzzy C-means (FCM) algorithm is a clustering algorithm based on fuzzy logic. There are different types of fuzzy clustering algorithms- fuzzy c-means, fuzzy-k nearest neighbour etc. However fuzzy c-means is the most widely used and popular algorithm. The FCM algorithm was developed by Dunn[19] in 1973 and was later modified by Bezdek [20] in 1981.

The idea behind the Fuzzy C-Means algorithm was used in various works related to natural language processing including [21] and [22].

Another process of generating text summarization is the highly accepted TextRank Algorithm. TextRank [24] is a graph based unsupervised algorithm derived from the PageRank algorithm [25]. PageRank was primarily introduced to rank the web pages which appear in online search results. In order to rank the web pages, probabilities of a user visiting a page is considered and a score is calculated. With these probability values, a matrix is initialized and the values update iteratively, ultimately creating a set of ranked web pages. TextRank is very similar to PageRank, except for the fact that instead of web pages, sentences are used. The sentences are converted to vector representations and then similarity scores are calculated. Using these scores, a graph is constructed and the top-ranked sentences are selected. In [26] Li et. al. have used the Wikipedia knowledge base to construct a modified TextRank model which extracts keywords for short texts. The main idea of the model was to treat each Wikipedia entry as independent concepts so that the semantic information of a word could be demonstrated in terms of the distribution of the word over the Wikipedia concepts. Upon using the classic TextRank algorithm the keywords extracted would just show the importance of the words over a single article, but in the aforementioned system, the importance of words is affected based on their availability on other Wikipedia articles. The results of their system show that their system performs better than the classical method and the common TF-IDF method.

Before any summarization techniques can be applied on any sort of text, a proper preprocessing is required, similar preprocessing techniques were discussed in [7, 8], where they conducted tokenization, stop word removal and stemming; making the text ready for mathematical analysis. Akter et. al. in [8] used the concept of TF*IDF for word scoring and incorporated cue/skeleton words concepts into their sentence scoring mechanism which further improve their system's accuracy. Works in [7, 12, 13, 14] brought in the idea of sentence scoring, by taking the sentence length into consideration, [7, 13, 14] took it one step further and also altered the sentence priority based on the sentence position within the text extract. Krishnaveni et. al. in [14], implemented the idea of topic scoring, where sentences containing the words that are present in the topic sentence is given a higher priority.

Having problems in manipulating data for having an unfavourable number of dimensions is not a new problem in this field of research. Nonetheless, it was quite remarkably handled by Tian et al, in their work [16] where they discussed the method of how Principal Component Analysis (PCA) can be used to reduce the dimension of the data, to make complex data more susceptible for manipulation and visualization. PCA mainly works by taking in multidimensional data, and then analyzes their standardized form and determines the predefined number of Principal Components for the given data, based upon the variation of the data points in any chosen dimension.

In this paper, a comparative study and a thorough analysis is performed, between the techniques TextRank, Aggregate Scoring method and FCM integrated with PCA to summarize Bengali Text and news articles by the extractive method into concise and meaningful texts.

Chapter 3

Proposed Model

3.1 System Workflow

The system proposed in this paper uses three popular text summarization methodologies to summarize Bengali text documents and provides a comparative study on the outputs generated. Figure 3.1 below represents a detailed experimental workflow of the system proposed in this paper.

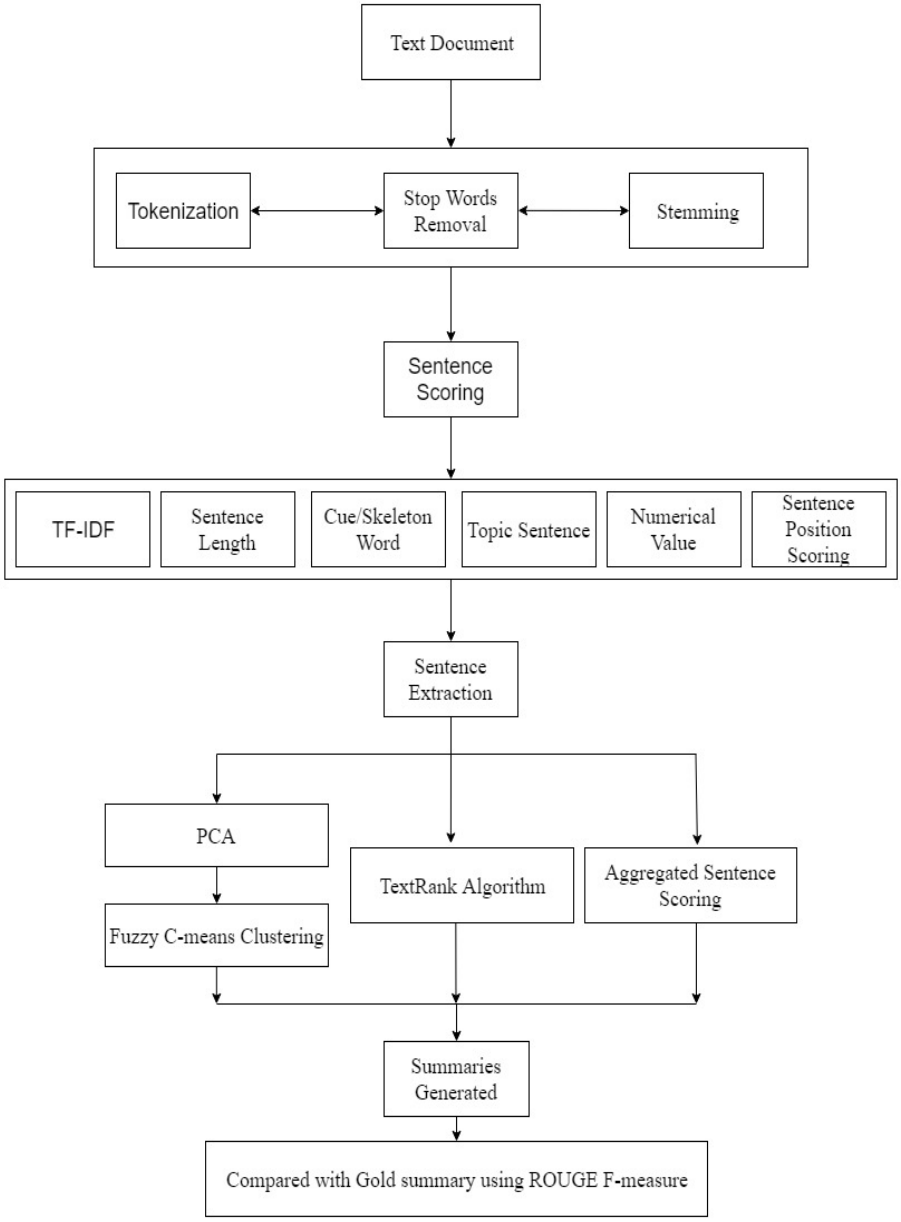

Figure 3.1 Workflow of the system

At the very primary stage Bengali text article is fed into the system; once the file is read it undergoes the immersive preprocessing procedures to prepare the text document for scoring. During preprocessing, the system removes the stopwords present, splits the text into paragraphs, sentences and later into words (tokenization). The system also stems the words to its root version so that in circumstance a word is not misinterpreted as different words in case of occurrences at multiple instances in different forms of the root version.

After the preprocessing is done the system moves on to the feature extraction part which is the scoring mechanism of the sentences in order to generate the Extractive summary

of the input text. For the scoring mechanism, the system is equipped with 6 different scoring techniques. The scores used for the system are TF-IDF, Numerical Value, Sentence length, Cue/Skeleton word, Topic Sentence and Sentence Position Scoring.

Upon the successful extraction of the features, the system produces a 6-dimensional array upon which Principal Component Analysis (PCA) is performed to reduce the 6-dimensional data to a 2-dimensional data. The 2-dimensional data is then subjected to Fuzzy C-Means (FCM), to classify the sentences into 2 clusters, later the cluster having a greater F-measure value is printed as the output summary.

Apart from FCM, the system also uses TextRank and Aggregate Scoring techniques to generate 2 more summaries for each article. TextRank is a form of summary generation that is derived from PageRank where sentence similarity is used to find the most important sentences. Next, the system finds the Aggregate Scores of the sentences and create a third summary using the most important sentences from the set. The F-Measure is calculated for each of the summaries, comparing it with the Gold Summary (Human Generated Summary) that is manually fed into the system, and a comparative study is conducted exhibiting the classifying methodology with the maximum accuracy.

3.2.1 Dataset

Dataset has been taken in from an online repository [31], which is basically a group of texts with their human-generated summary. Each instance of data has a full-sized text, along with three human-generated summaries. The human-generated summary is named “Gold Summary”. The Gold Summary also follows the extractive method to summarization, to keep consistency between it and output summary.

News articles from different national daily newspapers including “The Daily Prothom Alo” and “Kaler Kantho” have also been considered as texts for text summarization. In Section 3.3 the test articles given are two news articles taken from the Daily Prothom Alo website.

/

3.2.2 Preprocessing

Bengali language processing have not been as widely popular as English language processing, hence Bengali does not have any sort of libraries like NLTK that is readily available for

English. For making the system to perform in a way that has been proposed initially, a proper preprocessing had to be done. Preprocessing methodologies that have been implemented in this system, tend to make the text readable by the machine, and help the scoring mechanism to perform precisely. Most of the codes used in the preprocessing has been solely written for this system, due to the unavailability of any pre-designed library. A detailed description of the system's preprocessing methodologies has been discussed below:

1. Stemming

In Bengali, a certain root word can be manipulated in multiple ways to make it best suited with the sentence and the context it is used for. For example, the word 'কাজ' can be used as 'কাজের', 'কাজটি' etc, but all of these words originate from the same root word which is 'কাজ', hence to make the system's scoring mechanism more accurate and relevant, a stemming mechanism is incorporated in preprocessing, which simply converts all the words to their very root version. If the following words are taken as an example,'কাজের', 'কাজটি' etc will all be converted to 'কাজ'. So that each time the words come up, the system's scoring mechanism will recognise them and treat the words as the same word as the root word. A rule-based generic Bengali stemmer as implemented in [23] has been used which converts a Bengali word into its stemmed form. The following table, Table 3.1 demonstrates what happens when words used in sentences are stemmed into their root version.

Table 3.1: Stemming output

Original Word	Stemmed Word
তেলের	তেল
যুক্তরাষ্ট্রে	যুক্তরাষ্ট্র
সালের	সাল
অক্টোবরের	অক্টোবর
নেমে	নাম
অক্টোবরেই	অক্টোবর
কারণে	কারণ
মাসের	মাস

2. Stopword Removal

Words that hold any contribution for expressing the meaning of a sentence and have very little meaning themselves are called stop words. Bengali sentences are often filled with numerous stopwords. Bengali language and its grammar are designed as such that one has to use stopword to make it complete. Words such as '**অবশ্য**', '**এই**', '**কয়েক**' are merely few of words from the enormous list of stopwords that has been installed in the system. All such words are detected by the system and are removed before the scoring starts. If the stopwords are not removed then they tend to take up of a lot of computational resources and as these words are likely to be repeated, they appear to be scored higher than the actual meaningful words and eventually contribute to generating inaccurate summaries. Table 3.2 below, exhibits the sentence structure upon removal of the stopwords.

Table 3.2: Representation of stopword removal

Sentences with stopword	Sentences after stopword is removed
অথচ গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল।	গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল।
শুক্রবার এই তেলের দাম কমেছে ৫ দশমিক ৫ শতাংশ।	শুক্রবার তেলের দাম কমেছে ৫ দশমিক ৫ শতাংশ।

3. Paragraph Splitting

An article contains multiple paragraphs and this system is built such that, it separates each paragraph into objects which contains necessary information about the paragraph it's referencing. Furthermore, each paragraph is processed iteratively to rank its topic and concluding sentences.

4. Sentence Splitting

The position of the sentence within a paragraph is also one of the many attributes the system takes into account in order to generate the most accurate summary. A separate sentence class is written, where instances of each sentence are created and the numerical value of each sentence feature is stored within.

5. Tokenization

Tokenization is the process of splitting each sentence into separate words. In order to check for occurrences of words in the sentences and to increment the scores for any positive matches, tokenization had to be done.

3.3 Feature Extraction

3.3.1 TF-IDF Scoring

TF-IDF stands for Term Frequency-Inverse Document Frequency; this score represents the importance or significance of a specific word in the entire document. TF is defined by the number of occurrences of the word w, in a total number of sentences S.

$$TF - IDF(w) = Frequency(w) * log\left(\frac{Number\ of\ Sentences}{Frequency(w)}\right) \quad 3.1$$

Each sentence in the text then receives a score based on the TF-IDF of each word in the sentence.

$$Score(s_1) = \sum_{w=0}^{n} TF - IDF(w_i) \quad 3.2$$

After scores for each of the sentences are generated, each score is normalized so as to make the system's scores compatible with the clustering algorithms. The normalization algorithm simply counts the maximum score of a sentence to be 1, and all other scores are normalized in relation to the maximum score.

$$Normalized\ Score(s_1) = \frac{Score(s_1)}{Max(Scores(s_1))} \quad 3.3$$

3.3.2 Numeric value based Sentence Scoring

A sentence containing any numerical values are generally considered more important as a number can add a lot of value to a summary. Thus, the text is scanned for the presence of the Bengali numerals "০, ১, ২, ৩, ৪, ৫, ৬, ৭, ৮, and ৯"; if a numeral is found, regardless of combinations, the score of the sentence is incremented. Then, the score is calculated using:

$$Numerical\ Score(s_2) = \frac{Numeral\ Count(s_2)}{Length(s_2)} \quad 3.4$$

These scores do not require any normalization as 0≤Numerical Score(s) <1

3.3.3 Sentence Length based Scoring

Length based scoring relies on the length of each sentence, where a sentence is scored based on how its length compares to the average length of sentences in the article. This method of scoring relies on the fact that sentences that are too long or too short do not hold a lot of significance. Sentences that are too short usually contain anecdotes, exclamations, or quotes that do not reflect importance. On the other hand, sentences that are too long are considered vague, bringing unwanted information to the reader.

$$Score(s_3) = \frac{Length(s_3)}{Average\ Length\ of\ Sentence\ in\ text} \quad 3.5$$

After scoring each sentence the values are normalized to meet the algorithm's required ranges of 0-1. To do this, a piecewise function was used

$$NormalizeScore = \begin{cases} ratio & if\ ratio \leq 0 \\ 2 - ratio & if\ ratio \leq 2 \\ 0 & ratio > 2 \end{cases} \quad 3.6$$

3.3.4 Cue/Skeleton Word Scoring

The words that itself does not hold, much significance but when used in a sentence; provides a truer picture of the context is called the Cue/Skeleton words. Sentences containing the Cue words such as 'কারন', 'যেহেতু', 'অতএব' etc. are likely to hold a greater importance in holding the gist of the actual text extract. In [9, 14] the cue feature was discussed where sentences containing the words that are in the list of predefined list of cue word are given a higher importance. In the system, all the words in their tokenized form are crossed checked

with the list of cue words, and if a positive match is found the sentence containing the Cue word is given a higher score.

for every cue word i present in sentence j

$$Score(s_4) \text{ for sentence } j = Score(s_4) \text{ for sentence } j + 1 \quad (3.7)$$

$$Normalized\ Score(s_4) = \frac{Score(s_4)}{Max(Scores(s_4))} \quad (3.8)$$

3.3.5 Topic Sentence Scoring

For any given text extract, the first sentence of the extract, and the first sentence of each subsequent paragraph are supposed to contain the words that are more relevant to the subject of the text and are more like to give the overview of the context. A concept as such was discussed in [17]. In the system, the words present in the Topic Sentences are matched with the other sentences of any given paragraph, and the sentences containing the words of the topic sentence are given a higher priority. This is done by flagging a word that is present in one of the topic sentences, then checking for the flagged word in the other sentences. For example, if the word "আবহাওয়া" exists in the topic sentence, and then also in sentence S, the score is incremented by 1.

$$Score(s_5) = Sum(Flagged\ words) \quad (3.9)$$

$$Normalized\ Score(s_5) = \frac{Score(s_5)}{Max(Scores(s_5))} \quad (3.10)$$

3.3.6 Sentence Position based scoring

The first line of a paragraph, which is the topic sentence, usually highlights and sums up what the whole article is about. This is also done by the concluding sentence in the paragraph. So, a scoring system is used where the paragraphs are iterated through, and the topic sentence and concluding sentences are ranked highest. Also, the immediate line after the topic sentence and the line before the conclusion sentence usually contains important information as well. Keeping that in mind the first 10% and last 10% sentences of each paragraph are scored more

than the rest. The following function shows the selection process where S (I) is the positional sentence scoring of each sentence.

$$Score(s_5) = \begin{cases} S(i) = 1 & if\ position\ of\ sentence\ i \leq 0.1 * total\ no.of\ sentences\ in\ a\ paragraph \\ S(i) = 1 & if\ position\ of\ sentence\ i \geq 0.9 * total\ no.of\ sentences\ in\ a\ paragraph \\ S(i) = 0 & otherwise \end{cases} \quad 3.11$$

Table 3.3: Score table of the score for test Article 1

Sentence Number	TF-IDF	Numerical Value-Based	Sentence Length Based	Cue/Skeleton word	Topic Sentence	Sentence Position Based
1	0.173	0	0	0	0.2727	1
2	0.2228	0.125	0	0	0.0909	1
3	0.5143	0.5714	1	0.5	0.2727	1
4	0.2831	0.25	0	0	0.3636	1
5	0.697	0	1	0.5	0.3182	1
6	0.4604	0.1667	1	0	0.2727	0
7	0.1995	0	0	0	0.1364	1
8	0.6223	0	1	0	0.8182	1
9	1	0	0	0	1	1
10	0.597	0	1	0	0.1818	1
11	0.7138	0	1	0	0.2727	1
12	0.2002	0	0	0	0.4091	1
13	0.4094	0.2222	0	0	0.0909	1
14	0.4323	0.5455	1	0	0.1818	1
15	0.6274	0.2	1	0	0.0909	0
16	0.3514	0.1111	0	0	0.0455	1
17	0.3239	0.1111	0	0	0.0455	1
18	0.5531	0	1	0	0.5909	1
19	0.7598	0	1	0	0.2727	1
20	0.3419	0	0	0	0.1364	1
21	0.4277	0	0	0.5	0.4091	1
22	0.2534	0	0	0	0	1
23	0.2298	0	0	0	0	0
24	0.1088	0	0	0	0.0455	0
25	0.4223	0	1	0.5	0.1818	1
26	0.7632	0	1	1	0.8636	1
27	0.276	0	0	0.5	0.1364	1
28	0.574	0.1176	1	0	0.0909	0
29	0.2095	0	0	0	0.1364	1
30	0.2175	0	0	0	0.3636	1
31	0.2601	0.1111	0	0	0.0455	1
32	0.1796	0	0	0	0	0
33	0.1929	0	0	0	0.0909	0
34	0.7351	0.375	1	0	0.2273	1
35	0.6635	0	1	0	0.5909	1

3.4 Classifying Methodologies

3.4.1 Fuzzy C-Means Clustering

Principal Component Analysis (PCA)

The system that has been developed for the Text Summarization, is designed to have 6 features i.e. the data generated is 6 dimensional, now the Fuzzy C-means algorithm implementation this system uses 2-dimensional data; thus the need of Principal Component Analysis.

One of the very basic functions of the Principal Component Analysis is that it reduces data dimensions making it feasible for data visualization [15, 16]. The algorithm initially takes in multidimensional data, then it standardizes the data and finally deduces the principal components based on the variation of the data. For the system developed, 6 features/columns were converted into 2 features/columns of data; Principal Component 1 (PC1) and Principal Component 2 (PC2). PC1 demonstrates the direction when there is the most variation in the input dataset, and PC2 demonstrates the same where the second most variation occurs. Upon generation of the 2 Principal Components, Fuzzy C-means algorithm is implemented on the 2-dimensional data, to generate the clusters based on which the summary is determined. Table 3.4 below shows the PC1 and PC2 values generated for each sentence for the test Article 1.

Table 3.4: PCA Score table for Test Article 1

Sentence Number	Principal Component 1	Principal Component 2
1	-1.12037	-1.04909
2	-1.27738	-0.20211
3	1.830519	2.160945
4	-0.48953	0.00867
5	2.031563	-0.35554
6	0.052385	1.816605
7	-1.32648	-0.84062
8	2.128207	-0.64812
9	2.502514	-1.58377
10	0.777594	0.236368
11	1.265407	0.17756
12	-0.77416	-1.22567
13	-0.7288	0.344891
14	0.698672	2.583186
15	0.141872	2.322483
16	-1.04289	-0.12352
17	-1.11454	-0.13992
18	1.489062	-0.36811
19	1.385257	0.205004
20	-0.95546	-0.75566
21	0.536656	-1.54871

22	-1.46139	-0.61565
23	-2.1827	0.413302
24	-2.40611	0.276796
25	1.040501	-0.32662
26	4.023113	-1.54576
27	-0.40909	-1.25374
28	-0.05016	1.921285
29	-1.30042	-0.83465
30	-0.82093	-1.15104
31	-1.28076	-0.17799
32	-2.3135	0.383352
33	-2.09534	0.262795
34	1.469992	1.935296
35	1.776702	-0.30224

Fuzzy C-means Clustering:

The term Fuzzy set in mathematics refers to a set in which each element of the set has varying degrees of membership. In the traditional set theory, the membership of elements in a set are expressed in a binary fashion- an element either belongs to the set or it does not. However, in fuzzy set theory, the membership of elements are expressed with the help of membership function and the membership values varying in the interval [0, 1].

The Fuzzy C-means algorithm is a soft computing technique initially developed by Dunn [19]. This algorithm is based on the fuzzy set theory mentioned above. The idea of membership in fuzzy set theory is modified in the case of Fuzzy C-means algorithm and a membership matrix is formed known as the partition matrix, which contains the degree of memberships of elements across different clusters. In this proposed system, the number of clusters specified for clustering is set to be 2.

The Fuzzy C-means algorithm runs very similarly to the K-means algorithm. Firstly, the number of clusters required is needed to be specified. Next, an initial partition matrix is created and data points are randomly distributed over the clusters in a binary way. The algorithm converges, when the change between the membership values between two iterations is greater than ε - the specified error limit or the maximum number of iterations has reached.

The primary aim of the FCM is to minimize the objective function as (3.12) before the algorithm converges.

$$J = \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^{m} \left\| x_i - c_j \right\|^2 \qquad 3.12$$

Where *c* is the total number of clusters, *n* is the total number of data points. μ_{ij} stands for the degree of membership of x_i in the *jth* cluster and *m* is any real number greater than 1.

The FCM algorithm aims to partition a set of *n* data points $X = \{x_1, x_2, x_3 \ldots x_n\}$ into a set of specified different clusters. Fuzzy partitioning is done by the iterative optimization of the objective function (3.12) with the update of the membership values (3.14) and the calculation of the centroid of the clusters as (3.13) [28].

$$c_j = \frac{\sum_{i=1}^{n}(\mu_{ij}^{m} x_i)}{\sum_{i=1}^{n}(\mu_{ij}^{m})} \qquad 3.13$$

Where c_j is the d-dimension centre of the cluster.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \qquad 3.14$$

The FCM algorithm [27] has the following steps:

- *Initialize* $U = [u_{ij}]$ *matrix randomly,* $U^{(0)}$*, where* $U^{(0)}$ *is the initial partition matrix.*
- *At k-step: calculate the centers vectors* $C^{(k)} = [c_j]$ *with* $U^{(k)}$ *with equation (3.13)*
- *Update* $U^{(k)}$ *,* $U^{(k+1)}$ *with equation (3.14)*
- *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

This system uses the FCM algorithm in [29] for implementation.

FCM is a clustering model which was used in this system where it was fetched witch 2-dimensional data generated from the PCA model. The FCM model automatically calculated the optimum number of centres based on the given input data and then it iteratively found out to which centre the data points are the closest to, which ultimately led to creating the clusters. The optimum number of centres is found using the Fuzzy Partition Coefficient (FPC) value, greater the value greater is the accuracy of the number of centres best fitted according to the data which can be seen from the following graph which was generated from Test Article 1. An illustration of the centres with their FPC values is given below in Figure 3.2.

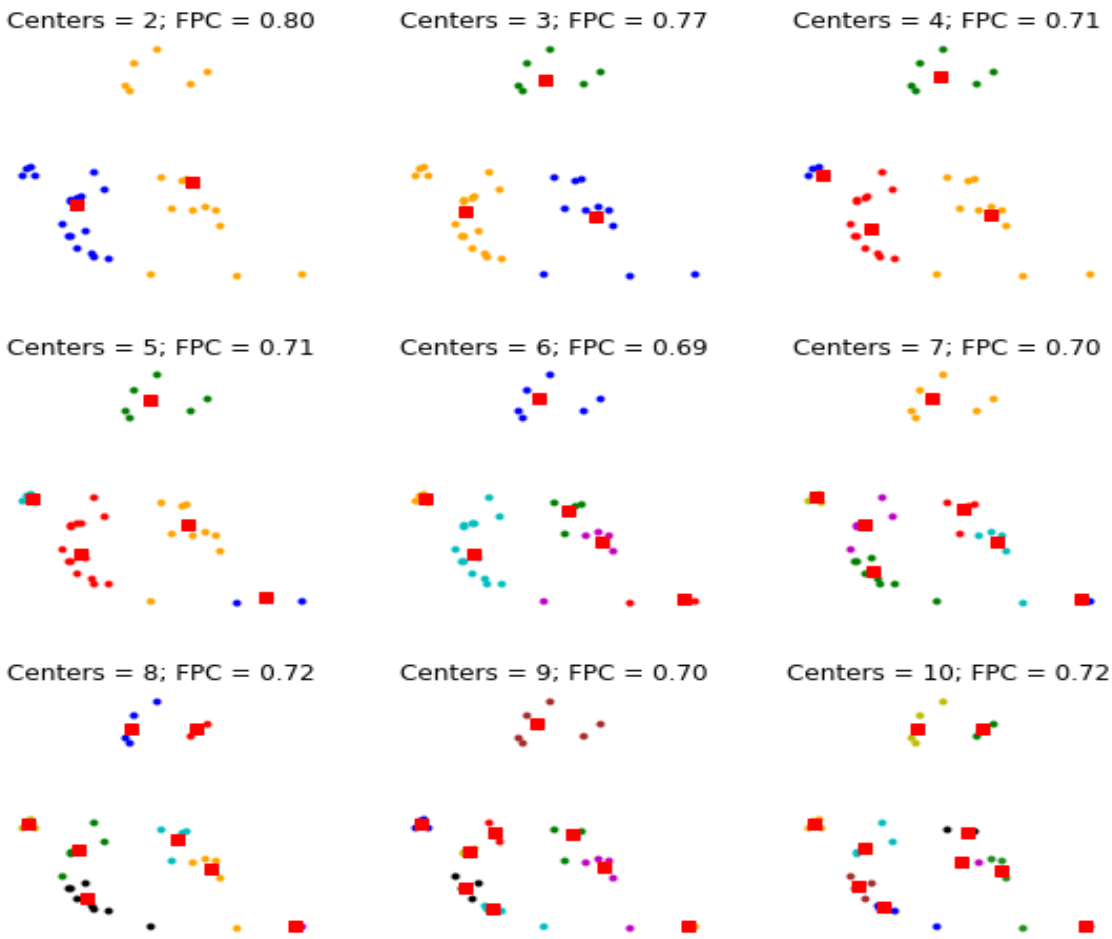

Figure 3.2. Number of centres and FPC values of FCM

Figure 3.2 shows the system's experiment with multiple centres ranging from 2 to 10. It can be seen that having 2 centres, the greatest FPC value was generated, which means it would give best results if 2 centres are initiated. The main objective of using FCM here was to find out which sentences should be chosen for the system's summary based on its processed score.

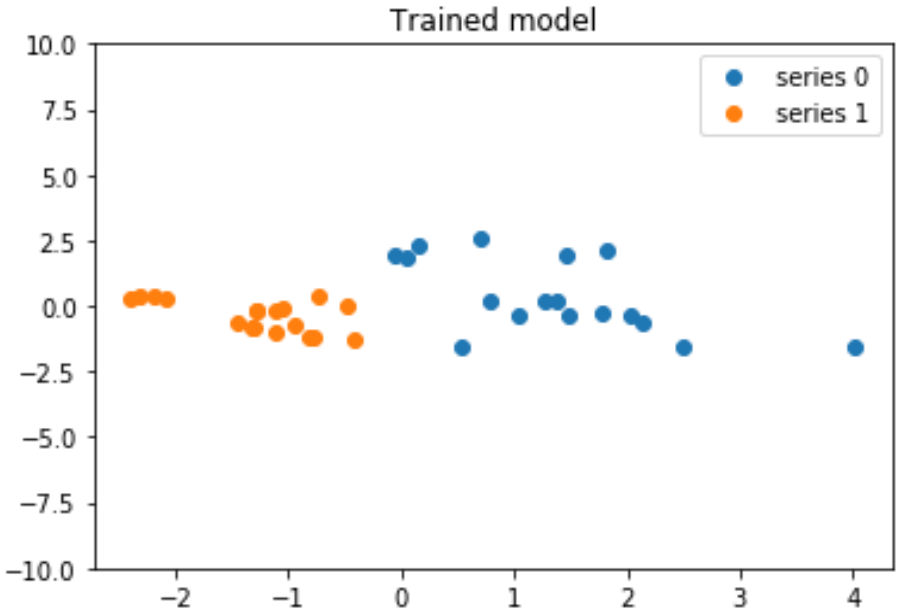

Figure 3.3. FCM clustering output

Table below demonstrates the list of sentences that have been clustered into 2 different clusters, which have generated using the Fuzzy C-Means.

Table 3.5: Sentence allocation in Clusters

Sentences in Cluster 1	Sentences in Cluster 2
2	0
4	1
5	3
7	6
8	11
9	12
10	15
13	16
14	19
17	21
18	22
20	23
24	26
25	28
27	29
33	30
34	31
	32

3.4.2 TextRank Algorithm

TextRank has been a popular summary generator for the English language. Hence it is used to generate another summary from the same article used before for comparison purpose. The

TextRank code was from a repository in GitHub [30] in which the TextRank algorithm is implemented using PHP. The article was first stemmed and all stop words were removed in order to attain a greater accuracy by the algorithm. The summary generated was output onto a text which was later used for comparison purpose with the other two algorithms used.

3.4.3 Aggregated Sentence Scoring

Aggregates Sentence Scoring is a very simple and straightforward approach to generate a summary based on the features obtained from analyzing a given text extract. It is the traditional method of summarization where all the scores of every sentence are added together, and all the scores are ranked based on the cumulative score of the sentences. A summary length ratio relative to the original text is usually predefined, and the extractive summary is computed by selecting the sentences that hold the highest scores among all the sentences.

For a clearer understanding, let us assume the system is fed with an original text comprising a total of 20 sentences, and the predefined accepted ratio is given to be 0.4 or 40%. After the aggregate score is computed and the sentences are ranked in a descending order, then only the top 40% is determined to be the summary of the original text; in the particular case, the summary will have a length of 8 sentences.

For every feature i in sentence j

$$Aggregate\ Score(s_j) = \sum_{w=0}^{i} Score(s_i) \qquad 3.15$$

Table 3.6: Sentences with their Aggregate Scores for Test Article 1

Sentence	Aggregate Score
এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, ‘তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে	4.627
২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল	3.858
কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে	3.515
সোসিয়েট জেনারেলের পণ্য গবেষণা বিভাগের প্রধান মাইকেল হেইগ বলেন, ছয় সপ্তাহ ধরে দাম যে হারে কমছে, তাতে বিনিয়োগকারীদের নাভিশ্বাস উঠে যাওয়ার জোগাড়	3.44
যুক্তরাষ্ট্রে শুক্রবার এক গ্যালন জ্বালানির দাম ছিল ২ দশমিক ৫৮ ডলার, যা এক মাস আগেও ছিল ২.৮৪ ডলার	3.337
এই পরিস্থিতিতে তেল উৎপাদনকারী দেশগুলো আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকের দিকে তাকিয়ে আছে	3.254
শুক্রবার ২০১৮ সালের মধ্যে ব্রেন্টের দাম সর্বনিম্ন ৫৯ ডলারে নেমে আসে	3.16
ইরানের ওপর নিষেধাজ্ঞা আসছে এই আশঙ্কায় সৌদি আরবসহ ওপেকভুক্ত দেশগুলো তেলের উৎপাদন বাড়িয়ে দেয়	3.144
এসব কারণে বৈশ্বিক অর্থনীতির চালিকা শক্তি জ্বালানি তেলের বাজার রমরমা হওয়ার	3.104

সম্ভাবনা নেই	
কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়	3.033
তেল খাতসংশ্লিষ্ট ব্যক্তিরা আশা করছেন, আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকে সৌদি আরবসহ অন্যরা তেলের সরবরাহ যথেষ্ট হারে কমাবে এবং তাতে বাজার কিছুটা সাশ্রয়ী হবে	3
আর সম্প্রতি তিনি যেভাবে সৌদি আরবের প্রশংসা করলেন, তাতে বিনিয়োগকারীদের মনে শঙ্কা, সৌদি আরব সম্ভবত উৎপাদন তেমন একটা কমাবে না	2.987
তবে দাম কমে যাওয়া সত্ত্বেও মার্কিন প্রেসিডেন্ট ডোনাল্ড ট্রাম্প ওপেকসহ সৌদি আরবকে উৎপাদন না কমাতে চাপ দিচ্ছেন	2.779
অন্যদিকে বৈশ্বিক আর্থিক বাজারে প্রবৃদ্ধি নিয়ে আবারও আশঙ্কা তৈরি হয়েছে	2.337
তেল কোম্পানির শেয়ারের দাম পড়ে যাওয়ায় শুক্রবার ডাও সূচকের মান ১৭৮ পয়েন্ট কমে যায়	1.918
তেলের দাম এভাবে কমার কারণে অনেকেই হতবুদ্ধ হয়ে গেছেন	1.912
এক মাস আগেই পর্যবেক্ষকেরা দিন গুনছিলেন, তেলের দাম কবে ব্যারেলপ্রতি ১০০ ডলারে উঠবে	1.9
অথচ গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল	1.897
সোসিয়েট জেনারেলের হিসাব মতে, চলতি প্রান্তিকে বড় বড় তহবিলের ক্ষতির পরিমাণ ৭৭০ কোটি ডলার ছাড়িয়ে গেছে	1.783

Chapter 4

Evaluation and Results

4.1 Rouge

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [32] is a metric system to compare machine generated summaries or translation against a reference summaries a.k.a Gold Summary. ROUGE tends to generate a metric value that determines the accuracy of the generated summary by generating a ratio of overlapping sentences.

For the evaluation of the system's summary generated from the 3 different methods, the ROUGE-2 measure was used. What it does is, it compares the summary generated by the system with the reference summary (Human-produced). It has two criteria for evaluation:
1) Recall and 2) Precision.

Recall finds out if the system summary has sentences which match with the reference summary or not. It uses the following formula for computation:

$$Recall = \frac{Number\ of\ overlapping\ Sentences}{Total\ number\ of\ Sentences\ in\ reference\ summary} \quad 4.1$$

A perfect score of 1 would mean the system summary matched fully with the reference summary. However the system summary might have useless and unnecessary information in addition to the information present in the reference summary, and still, recall would give a good score. A better way to see if in fact only the relevant information is present in the system summary or not is by using precision measure.

Precision measure finds out how much of the reference summary is actually present in the system summary by the following formula:

$$Precision = \frac{Number\ of\ overlapping\ Sentences}{Total\ number\ of\ Sentences\ in\ system\ summary} \quad 4.2$$

It simply finds out if the system summary is indeed relevant and concise or not.

Lastly, the F1 measure which is a measure of a test's accuracy is calculated using both recall and precision values. A score of 0 means the test yielded the worst result while 1 stands for the best. According to the system, a score 1 means the system summary matched exactly with the gold summary while 0 means the system summary is totally inaccurate. The F1 measure is calculated by the following formula:

$$F1 = 2 * \frac{precision * recall}{precison + recall} \quad 4.3$$

4.2 Comparative Study and Analysis

As represented by Table 4.1 and visualized by Figure 4.1 above, on both test cases, FCM yields a higher number of common sentences. This means that our FCM algorithm returns sentences that have a higher probability of carrying more importance from the input article.

	TextRank	**Aggregate Scoring**	**FCM**
Article 1	10	10	12
Article 2	7	9	10

Table 4.1: Number of common sentences in the summaries generated

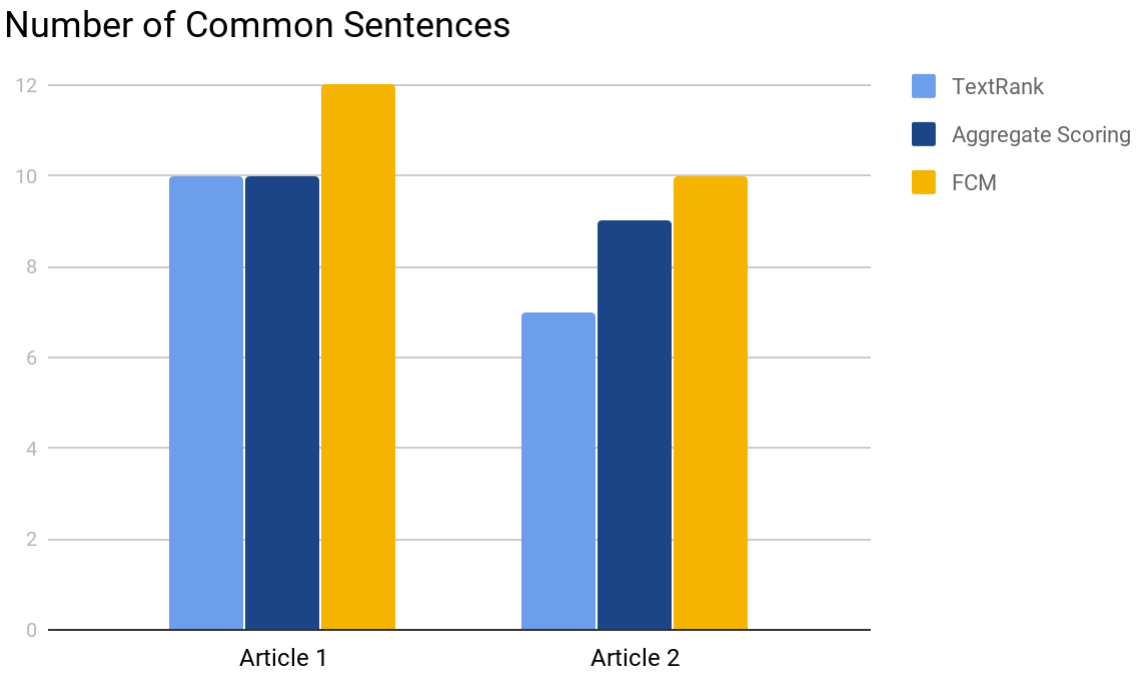

Figure 4.1: Bar chart comparing the number of common sentences in the summaries

Table 4.2: Comparison between F-number, Precision, and Recall for Test Article 1

	TextRank	Aggregate Scoring	FCM
F1 measure	0.625	0.5882352941	0.6857142857
Precision	0.7142857143	0.625	0.7058823529
Recall	0.5555555556	0.5555555556	0.6666666667

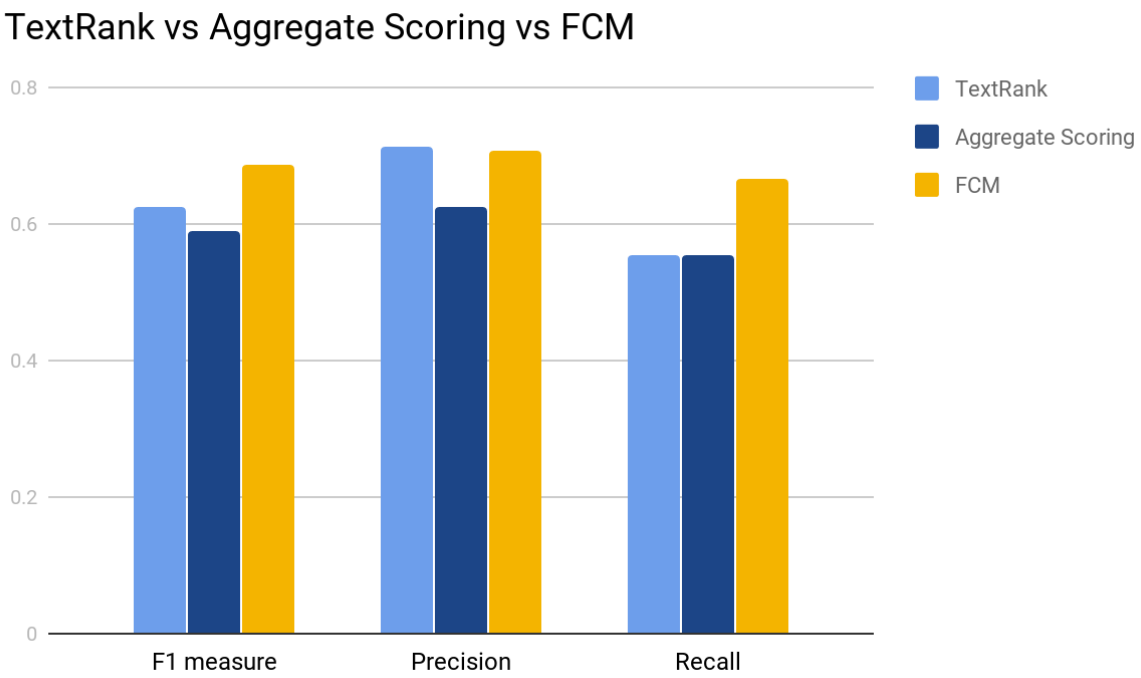

Figure 4.2: Bar Chart for Test Article 1

When judging the accuracy of summaries, we can look at two factors, the F1 Measure and Common Sentences. For the first article, we notice a higher F1 measure for FCM summary than both Aggregate scoring and TextRank. This is backed up by the fact that FCM summary generates more common sentences than both TextRank and Aggregate scoring. FCM generated 2 more relevant sentences than TextRank and Aggregate Scoring, which results in a higher F1 measure.

Table 4.3: Comparison between F-number, Precision and Recall for Test Article 2

	TextRank	Aggregate Scoring	FCM
F1 measure	0.35	0.5	0.606060606
Precision	0.304347826	0.473684210	0.625
Recall	0.411764705	0.529411764	0.588235294

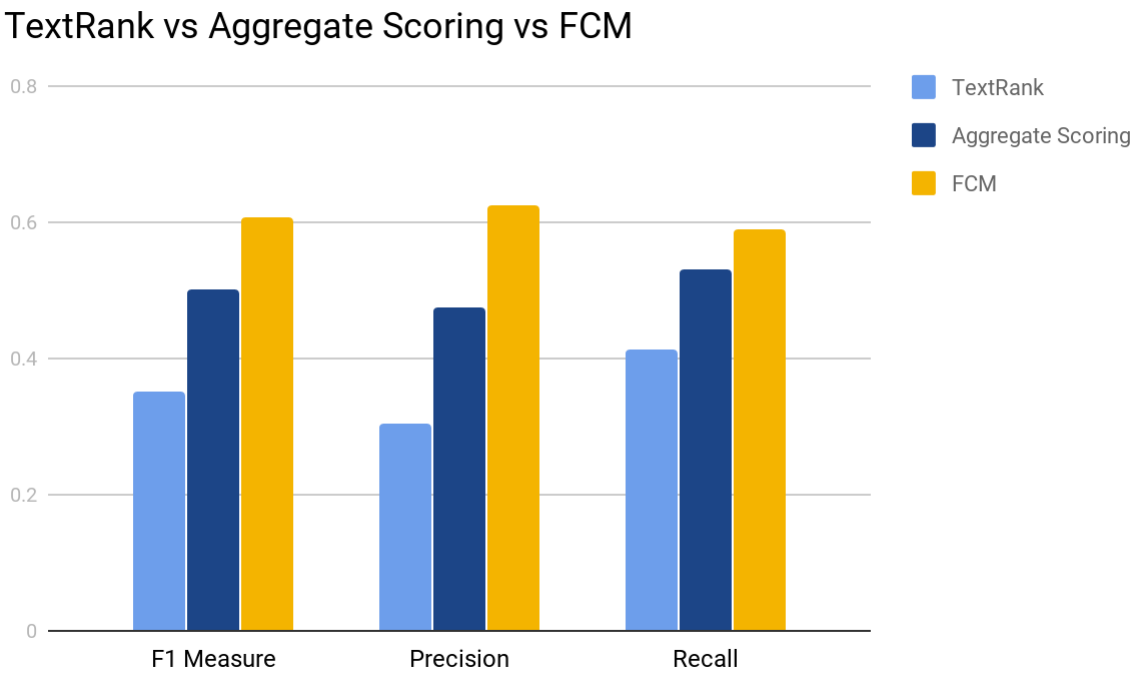

Figure 4.3: Bar chart for Test Article 2

In case of the second article, FCM consistently returns a higher F1 Measure, meaning that the summary generated by the FCM algorithm is more accurate and better retains information from the initial article. This is backed up by the number of sentences common between FCM summary and Gold summary being higher in both cases.

Table 4.4: Percentage increase in Summary Accuracy

	Percentage Increase from TextRank	**Percentage Increase from Aggregate Scoring**
Article 1	9.264	15.303
Article 2	53.565	19.178

4.3 Test Articles

Article 1: ‘৫০ ডলারে নেমে এল তেলের দাম’

অপরিশোধিত তেলের দাম আরও এক দফা কমল। গতকাল যুক্তরাষ্ট্রে তেলের দাম ৭ শতাংশ কমেছে। ২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল।

অথচ গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল। কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে। এক মাস আগেই পর্যবেক্ষকেরা দিন গুনছিলেন, তেলের দাম কবে ব্যারেলপ্রতি ১০০ ডলারে উঠবে। এখন তেলের এই পড়তি দাম দেখে তাঁদের কপালে চিন্তার ভাঁজ পড়েছে।

সোসিয়েট জেনারেলের পণ্য গবেষণা বিভাগের প্রধান মাইকেল হেইগ বলেন, ছয় সপ্তাহ ধরে দাম যে হারে কমছে, তাতে বিনিয়োগকারীদের নাভিশ্বাস উঠে যাওয়ার জোগাড়।

তেল খাতসংশ্লিষ্ট ব্যক্তিরা আশা করছেন, আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকে সৌদি আরবসহ অন্যরা তেলের সরবরাহ যথেষ্ট হারে কমাবে এবং তাতে বাজার কিছুটা সাশ্রয়ী হবে। তবে দাম কমে যাওয়া সত্ত্বেও মার্কিন প্রেসিডেন্ট ডোনাল্ড ট্রাম্প ওপেকসহ সৌদি আরবকে উৎপাদন না কমাতে চাপ দিচ্ছেন। আর সম্প্রতি তিনি যেভাবে সৌদি আরবের প্রশংসা করলেন, তাতে বিনিয়োগকারীদের মনে শঙ্কা, সৌদি আরব সম্ভবত উৎপাদন তেমন একটা কমাবে না।

তেলের দামের বৈশ্বিক মানদণ্ড হচ্ছে অপরিশোধিত ব্রেন্ট তেলের দাম। শুক্রবার এই তেলের দাম কমেছে ৫ দশমিক ৫ শতাংশ। শুক্রবার ২০১৮ সালের মধ্যে ব্রেন্টের দাম সর্বনিম্ন ৫৯ ডলারে নেমে আসে। তেল কোম্পানির শেয়ারের দাম পড়ে যাওয়ায় শুক্রবার ডাও সূচকের মান ১৭৮ পয়েন্ট কমে যায়। শেভরন ও কনোকোফিলিপসের শেয়ারের দাম ৩ শতাংশ পড়ে যায়। আর শেল উৎপাদক ইওজি রিসোর্সের দাম পড়েছে ৫ শতাংশ।

ইরানের ওপর নিষেধাজ্ঞা আসছে এই আশঙ্কায় সৌদি আরবসহ ওপেকভুক্ত দেশগুলো তেলের উৎপাদন বাড়িয়ে দেয়। কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়। এতে বাজারে তেলের দাম ক্রমেই কমতে কমতে এ জায়গায় এসে দাঁড়িয়েছে।

অন্যদিকে বৈশ্বিক আর্থিক বাজারে প্রবৃদ্ধি নিয়ে আবারও আশঙ্কা তৈরি হয়েছে। অর্থনীতিবিদেরা ইতিমধ্যে প্রবৃদ্ধির প্রাক্কলন কম করে ধরছেন। বিশ্বের তৃতীয় ও চতুর্থ অর্থনীতি ইতিমধ্যে সংকুচিত হচ্ছে। চীনের প্রবৃদ্ধিও কমছে। এসব কারণে বৈশ্বিক অর্থনীতির চালিকা শক্তি জ্বালানি তেলের বাজার রমরমা হওয়ার সম্ভাবনা নেই।

এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, 'তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে।' তেলের দাম এভাবে কমার কারণে অনেকেই হতবুদ্ধ হয়ে গেছেন। সোসিয়েট জেনারেলের হিসাব মতে, চলতি প্রান্তিকে বড় বড় তহবিলের ক্ষতির পরিমাণ ৭৭০ কোটি ডলার ছাড়িয়ে গেছে। পণ্যের বাজারের পরিস্থিতিও হতাশাজনক।

তবে তেলের পড়তি দাম ভোগ্যপণ্য ক্রেতাদের জন্য আশীর্বাদ হয়ে এসেছে। সামনে বড়দিন, তার আগে ২২ নভেম্বর থ্যাংকসগিভিং ডেও পালিত হয়েছে। এ উপলক্ষে মানুষ বেড়াতে যায়। তেলের দাম কম থাকায় মানুষের চলাফেরা বেড়েছে। যুক্তরাষ্ট্রে শুক্রবার এক গ্যালন জ্বালানির দাম ছিল ২ দশমিক ৫৮ ডলার, যা এক মাস আগেও ছিল ২.৮৪ ডলার।

এই পরিস্থিতিতে তেল উৎপাদনকারী দেশগুলো আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকের দিকে তাকিয়ে আছে।

Article 2: 'শিক্ষাবিপ্লবে মরিয়া চীন'

বিশ্বের অন্যতম অর্থনৈতিক পরাক্রমশালী রাষ্ট্র চীন। সামরিক শক্তি ও প্রতিরক্ষা, নতুন নতুন তথ্য ও যোগাযোগ প্রযুক্তির উপকরণ উদ্ভাবন এবং উন্নয়নে দেশটি যথেষ্ট এগিয়ে। সেই তুলনায় শিক্ষাক্ষেত্রে যেন এগিয়ে যেতে পারেনি।

যুক্তরাজ্যের সাপ্তাহিক প্রকাশনা টাইমস হায়ার এডুকেশন সাময়িকীর সর্বশেষ জরিপে বিশ্বের শীর্ষ ১০টি বিশ্ববিদ্যালয়ের মধ্যে চীনের কোনো বিশ্ববিদ্যালয়ের নাম নেই। বিষয়টি তাদের ভাবিয়ে তুলেছে। অবশ্য সাংহাই র ্যাঙ্কিংয়ে শীর্ষ ৫০০ বিশ্ববিদ্যালয়ের মধ্যে চীনেরই আছে ৪৫টি বিশ্ববিদ্যালয়। এতে অবশ্য তারা সন্তুষ্ট নয়। শিক্ষাক্ষেত্রে বিপ্লব ঘটিয়ে শীর্ষস্থান দখল করতে চীন এখন মরিয়া। আর এ জন্য দেশটির নামকরা সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়ের ওপর আরও বেশি জোর দিয়েছে চীনা কর্তৃপক্ষ।

১৯১১ সালে বেইজিংয়ে প্রতিষ্ঠিত হয় সিংহুয়া বিশ্ববিদ্যালয়। শতাধিক বছরের পুরোনো এই বিশ্ববিদ্যালয় এখন গবেষণা, বিজ্ঞান, প্রযুক্তি, প্রকৌশল ও গণিত বিষয়ে চীনাদের গর্বের প্রতীক। পশ্চিমা গবেষণাভিত্তিক বিশ্ববিদ্যালয়গুলোর আদলে পরিচালিত হচ্ছে চীনের সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়। এই দুটি বিশ্ববিদ্যালয় পরস্পর প্রতিবেশী ও প্রতিযোগী, যা কিনা চীনের অক্সফোর্ড ও কেমব্রিজ হিসেবে খ্যাত। সিংহুয়া হচ্ছে প্রচলিত ও বাস্তবধর্মী বিশ্ববিদ্যালয়। এই বিশ্ববিদ্যালয়ে পড়াশোনা করেছেন চীনের বর্তমান প্রেসিডেন্ট সি চিন পিং, সাবেক প্রেসিডেন্ট হু জিনতাওসহ বিখ্যাত অনেকেই। আর পিকিং বিশ্ববিদ্যালয় হচ্ছে দেশটির কবি, দার্শনিক ও বিপ্লবীদের তীর্থস্থান। চীনের সাবেক শীর্ষ নেতা মাও সেতুং এই বিশ্ববিদ্যালয় গবেষণা করেছিলেন। ১৯৮৯ সালে তিয়েনআনমেন স্কয়ারে বিক্ষোভে বিশ্ববিদ্যালয়টি অগ্রণী ভূমিকা রাখে।

লন্ডনভিত্তিক সাময়িকী দ্য ইকোনমিস্টের প্রতিবেদন বলছে, ১৯৯৫ সাল থেকে চীনা কেন্দ্রীয় সরকার দেশটির বিশ্ববিদ্যালয়গুলো বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ে উন্নীত করতে লাখ লাখ ডলার ব্যয় অব্যাহত রেখেছে। এর আওতায় প্রথম ২১১টি প্রকল্প হাতে নেওয়া হয়েছে। একবিংশ শতাব্দীর চ্যালেঞ্জ মোকাবিলায় প্রায় ১০০টি প্রতিষ্ঠানকে প্রস্তুত করা হয়েছে। ২০১৫ সাল থেকে চালু করা হয়েছে ডবল ফার্স্ট ক্লাস প্ল্যান প্রকল্প। এর লক্ষ্য দ্রুততম সময়ে বিশ্ববিদ্যালয়গুলো থেকে একটি বিশ্ববিদ্যালয়কে বিশ্বমানে পরিণত করা ও প্রতিষ্ঠানের পরিসর বাড়ানো।

যেকোনো কিছুর পেছনে অর্থ হচ্ছে মূল চালিকা শক্তি। সেই অর্থ খরচ করতে প্রস্তুত চীন। অর্থায়ন প্রক্রিয়া বিশ্ববিদ্যালয়গুলোকে উৎকৃষ্ট মানের গবেষণায় অনুপ্রাণিত করে। চীনা বিশ্ববিদ্যালয়ের একাডেমিক গবেষণায় নিয়োজিত ব্যক্তিদের যথেষ্ট প্রণোদনারও ব্যবস্থা রয়েছে। প্রযুক্তি ও প্রতিযোগিতানির্ভর বিশ্বের বিভিন্ন দেশের সরকারও নীতিনির্ধারণে পরিবর্তন নিয়ে আসছে। বিশ্বমানের বিশ্ববিদ্যালয় প্রতিষ্ঠা, উন্নত গবেষণা, র‍্যাঙ্কিংয়ে অন্তর্ভুক্তি ও অগ্রগতির জন্য শিক্ষা খাতে চীন ছাড়াও ভারত, সিঙ্গাপুর, দক্ষিণ কোরিয়া, তাইওয়ান, ফ্রান্স, জার্মানি বিপুল অর্থ ব্যয় করছে। প্রতিবেশী দেশ ভারত তাদের ২০টি বিশ্ববিদ্যালয়কে বিশ্বমানে নেওয়ার ঘোষণা দিয়েছে। এমনকি নাইজেরিয়ার মতো দেশ ২০২০ সালের মধ্যে তাদের অন্তত দুটি বিশ্ববিদ্যালয়কে বিশ্বের শীর্ষ ২০০টির মধ্যে অন্তর্ভুক্তির লক্ষ্যমাত্রা নির্ধারণ করেছে।

সিংহুয়ার মেধাবী শিক্ষার্থীরা সেরা গবেষক হয়ে দেশের উন্নয়নে কাজ করেন কিংবা দেশের হয়ে বিদেশে গবেষণায় নিযুক্ত হন। ২০১৭ সালে সিংহুয়া বিশ্ববিদ্যালয় ১ হাজার ৩৮৫ জনকে ডক্টরেট উপাধি দিয়েছে। একেই সময় যুক্তরাষ্ট্রের ম্যাসাচুসেটস ইনস্টিটিউট অব টেকনোলজিতে (এমআইটি) ৬৫৪ জনকে ডক্টরেট দেওয়া হয়। অবশ্য এই সংখ্যা সিংহুয়া বিশ্ববিদ্যালয়ের সাফল্যের প্রধান কারণ নয়। সিংহুয়া বিশ্ববিদ্যালের ভাইস চেয়ারম্যান ইয়াং বিন বলেন, 'সিংহুয়ার সবচেয়ে গুরুত্বপূর্ণ উন্নয়ন ছিল ১৯৭৮ সালে, যখন ডেং জিয়াওপিং (প্রয়াত রাজনীতিক) বলেন, চীন বিপুলসংখ্যক শিক্ষার্থী বিদেশে পাঠাবে।' তিনি আরও বলেন, '১০ হাজার শিক্ষার্থী বিদেশে পাঠানো প্রয়োজন। আমাদের বৈজ্ঞানিক শিক্ষার স্তর উন্নত করার এটাই হচ্ছে অন্যতম প্রধান পথ।

4.4 Output Summaries

Gold Summary for Test Article 1

অপরিশোধিত তেলের দাম আরও এক দফা কমল। ২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল। অথচ গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল। কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে। এক মাস আগেই পর্যবেক্ষকেরা দিন গুনছিলেন, তেলের দাম কবে ব্যারেলপ্রতি ১০০ ডলারে উঠবে। তেলের দামের বৈশ্বিক মানদণ্ড হচ্ছে অপরিশোধিত ব্রেন্ট তেলের দাম। শুক্রবার এই তেলের দাম কমেছে ৫ দশমিক ৫ শতাংশ। শুক্রবার ২০১৮ সালের মধ্যে ব্রেন্টের দাম সর্বনিম্ন ৫৯ ডলারে নেমে আসে। তেল কোম্পানির শেয়ারের দাম পড়ে যাওয়ায় শুক্রবার ডাও সূচকের মান ১৭৮ পয়েন্ট কমে যায়। ইরানের ওপর নিষেধাজ্ঞা আসছে এই আশঙ্কায় সৌদি আরবসহ ওপেকভুক্ত দেশগুলো তেলের উৎপাদন বাড়িয়ে দেয়। কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়। এতে বাজারে তেলের দাম ক্রমেই কমতে কমতে এ জায়গায় এসে দাঁড়িয়েছে। অন্যদিকে বৈশ্বিক

আর্থিক বাজারে প্রবৃদ্ধি নিয়ে আবারও আশঙ্কা তৈরি হয়েছে। তবে তেলের পড়তি দাম ভোগ্যপণ্য ক্রেতাদের জন্য আশীর্বাদ হয়ে এসেছে। এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, ‘তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে। সোসিয়েট জেনারেলের হিসাব মতে, চলতি প্রান্তিকে বড় বড় তহবিলের ক্ষতির পরিমাণ ৭৭০ কোটি ডলার ছাড়িয়ে গেছে। যুক্তরাষ্ট্রে শুক্রবার এক গ্যালন জ্বালানির দাম ছিল ২ দশমিক ৫৮ ডলার, যা এক মাস আগেও ছিল ২.৮৪ ডলার। এই পরিস্থিতিতে তেল উৎপাদনকারী দেশগুলো আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকের দিকে তাকিয়ে আছে।

Summary Generated by Fuzzy C-Means (Article 1)

২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল। কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে। এক মাস আগেই পর্যবেক্ষকেরা দিন গুনছিলেন, তেলের দাম কবে ব্যারেলপ্রতি ১০০ ডলারে উঠবে।সোসিয়েট জেনারেলের পণ্য গবেষণা বিভাগের প্রধান মাইকেল হেইগ বলেন, ছয় সপ্তাহ ধরে দাম যে হারে কমছে, তাতে বিনিয়োগকারীদের নাভিশ্বাস উঠে যাওয়ার জোগাড়।তেল খাতসংশ্লিষ্ট ব্যক্তিরা আশা করছেন, আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকে সৌদি আরবসহ অন্যরা তেলের সরবরাহ যথেষ্ট হারে কমাবে এবং তাতে বাজার কিছুটা সাশ্রয়ী হবে। তবে দাম কমে যাওয়া সত্ত্বেও মার্কিন প্রেসিডেন্ট ডোনাল্ড ট্রাম্প ওপেকসহ সৌদি আরবকে উৎপাদন না কমাতে চাপ দিচ্ছেন। আর সম্প্রতি তিনি যেভাবে সৌদি আরবের প্রশংসা করলেন, তাতে বিনিয়োগকারীদের মনে শঙ্কা, সৌদি আরব সম্ভবত উৎপাদন তেমন একটা কমাবে না। শুক্রবার ২০১৮ সালের মধ্যে ব্রেন্টের দাম সর্বনিম্ন ৫৯ ডলারে নেমে আসে। তেল কোম্পানির শেয়ারের দাম পড়ে যাওয়ায় শুক্রবার ডাও সূচকের মান ১৭৮ পয়েন্ট কমে যায়।ইরানের ওপর নিষেধাজ্ঞা আসছে এই আশঙ্কায় সৌদি আরবসহ ওপেকভুক্ত দেশগুলো তেলের উৎপাদন বাড়িয়ে দেয়। কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়।অন্যদিকে বৈশ্বিক আর্থিক বাজারে প্রবৃদ্ধি নিয়ে আবারও আশঙ্কা তৈরি হয়েছে। এসব কারণে বৈশ্বিক অর্থনীতির চালিকা শক্তি জ্বালানি তেলের বাজার রমরমা হওয়ার সম্ভাবনা নেই। এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, ‘তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে। সোসিয়েট জেনারেলের হিসাব মতে, চলতি প্রান্তিকে বড় বড় তহবিলের ক্ষতির পরিমাণ ৭৭০ কোটি ডলার ছাড়িয়ে গেছে। যুক্তরাষ্ট্রে শুক্রবার এক গ্যালন জ্বালানির দাম ছিল ২ দশমিক ৫৮ ডলার, যা এক মাস আগেও ছিল ২.৮৪ ডলার।এই পরিস্থিতিতে তেল উৎপাদনকারী দেশগুলো আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকের দিকে তাকিয়ে আছে।

Summary Generated by TextRank Algorithm (Article 1)

অপরিশোধিত তেলের দাম আরও এক দফা কমল। গতকাল যুক্তরাষ্ট্রে তেলের দাম ৭ শতাংশ কমেছে। ২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল। অথচ গত অক্টোবরেই তেলের দাম ব্যারেলপ্রতি ৭৬ ডলারে উঠেছিল। কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে। এক মাস আগেই পর্যবেক্ষকেরা দিন গুনছিলেন, তেলের দাম কবে ব্যারেলপ্রতি ১০০ ডলারে উঠবে। এখন তেলের এই পড়তি দাম দেখে তাঁদের কপালে চিন্তার ভাঁজ পড়েছে। তেলের দামের বৈশ্বিক মানদণ্ড হচ্ছে অপরিশোধিত ব্রেন্ট তেলের দাম। শুক্রবার এই তেলের দাম কমেছে ৫ দশমিক ৫ শতাংশ। শেভরন ও কনোকোফিলিপসের শেয়ারের দাম ৩ শতাংশ পড়ে যায়। কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়। এতে বাজারে তেলের দাম ক্রমেই কমতে কমতে এ জায়গায় এসে দাঁড়িয়েছে। এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, ‘তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে.’ তেলের দাম এভাবে কমার কারণে অনেকেই হতবুদ্ধ হয়ে গেছেন। তবে তেলের পড়তি দাম ভোগ্যপণ্য ক্রেতাদের জন্য আশীর্বাদ হয়ে এসেছে।

Aggregated Summary for Article 1

২০১৭ সালের অক্টোবরের পর এই প্রথম তেলের দাম ব্যারেলপ্রতি ৫০ দশমিক ৪২ ডলারে নেমে এল। কিন্তু অতি সরবরাহ নিয়ে শঙ্কা, চাহিদা পড়ে যাওয়া—এসব কারণে এক মাসের মধ্যে তেলের দাম এতটা কমে গেছে।সোসিয়েট জেনারেলের পণ্য গবেষণা বিভাগের প্রধান মাইকেল হেইগ বলেন, ছয় সপ্তাহ ধরে দাম যে হারে কমছে, তাতে বিনিয়োগকারীদের নাভিশ্বাস উঠে যাওয়ার জোগাড়।তেল খাতসংশ্লিষ্ট ব্যক্তিরা আশা করছেন, আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকে সৌদি আরবসহ অন্যরা তেলের সরবরাহ যথেষ্ট হারে কমাবে এবং তাতে বাজার কিছুটা সাশ্রয়ী হবে। তবে দাম কমে যাওয়া সত্ত্বেও মার্কিন প্রেসিডেন্ট ডোনাল্ড ট্রাম্প ওপেকসহ সৌদি আরবকে উৎপাদন না কমাতে চাপ দিচ্ছেন। আর সম্প্রতি তিনি যেভাবে সৌদি আরবের প্রশংসা করলেন, তাতে বিনিয়োগকারীদের মনে শঙ্কা, সৌদি আরব সম্ভবত উৎপাদন তেমন একটা কমাবে না। শুক্রবার ২০১৮ সালের মধ্যে ব্রেন্টের দাম সর্বনিম্ন ৫৯ ডলারে নেমে আসে। তেল কোম্পানির শেয়ারের দাম পড়ে যাওয়ায় শুক্রবার ডাও সূচকের মান ১৭৮ পয়েন্ট কমে যায়।ইরানের ওপর নিষেধাজ্ঞা আসছে এই আশঙ্কায় সৌদি আরবসহ ওপেকভুক্ত দেশগুলো তেলের উৎপাদন বাড়িয়ে দেয়। কিন্তু যুক্তরাষ্ট্র এরপর ভারত, চীনসহ বেশ কয়েকটি দেশকে ইরান থেকে তেল কেনার বেলায় ছাড় দিলে বাজারে তেলের সরবরাহ অনেকটা বেড়ে যায়।অন্যদিকে বৈশ্বিক আর্থিক বাজারে প্রবৃদ্ধি নিয়ে আবারও আশঙ্কা তৈরি হয়েছে। এসব কারণে বৈশ্বিক অর্থনীতির চালিকা শক্তি জ্বালানি তেলের বাজার রমরমা হওয়ার সম্ভাবনা নেই। এফএক্সটিএমের বিশ্লেষক লুকমান ওতুনুগা বলেন, ‘তেলের সরবরাহ একদিকে বাড়ছে, অন্যদিকে চাহিদা কমছে—এই দুই কারণে তেলের বাজারে বিপর্যয় নেমে আসছে।’ তেলের দাম এভাবে কমার কারণে অনেকেই হতবুদ্ধ হয়ে গেছেন। যুক্তরাষ্ট্রে শুক্রবার এক গ্যালন জ্বালানির দাম ছিল ২ দশমিক ৫৮ ডলার, যা এক মাস আগেও ছিল ২.৮৪ ডলার।এই পরিস্থিতিতে তেল উৎপাদনকারী দেশগুলো আগামী মাসে ভিয়েনায় ওপেক ও সহযোগী দেশগুলোর বৈঠকের দিকে তাকিয়ে আছে।

Gold Summary Article 2

বিশ্বের অন্যতম অর্থনৈতিক পরাক্রমশালী রাষ্ট্র চীন। সামরিক শক্তি ও প্রতিরক্ষা, নতুন নতুন তথ্য ও যোগাযোগ প্রযুক্তির উপকরণ উদ্ভাবন এবং উন্নয়নে দেশটি যথেষ্ট এগিয়ে। সেই তুলনায় শিক্ষাক্ষেত্রে যেন এগিয়ে যেতে পারেনি।যুক্তরাজ্যের সাপ্তাহিক প্রকাশনা টাইমস হায়ার এডুকেশন সাময়িকীর সর্বশেষ জরিপে বিশ্বের শীর্ষ ১০টি বিশ্ববিদ্যালয়ের মধ্যে চীনের কোনো বিশ্ববিদ্যালয়ের নাম নেই। অবশ্য সাংহাই র‍্যাঙ্কিংয়ে শীর্ষ ৫০০ বিশ্ববিদ্যালয়ের মধ্যে চীনেরই আছে ৪৫টি বিশ্ববিদ্যালয়। আর এ জন্য দেশটির নামকরা সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়ের ওপর আরও বেশি জোর দিয়েছে চীনা কর্তৃপক্ষ।পশ্চিমা গবেষণাভিত্তিক বিশ্ববিদ্যালয়গুলোর আদলে পরিচালিত হচ্ছে চীনের সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়। এই দুটি বিশ্ববিদ্যালয় পরস্পর প্রতিবেশী ও প্রতিযোগী, যা কিনা চীনের অক্সফোর্ড ও কেমব্রিজ হিসেবে খ্যাত। লন্ডনভিত্তিক সাময়িকী দ্য ইকোনমিস্টের প্রতিবেদন বলছে, ১৯৯৫ সাল থেকে চীনা কেন্দ্রীয় সরকার দেশটির বিশ্ববিদ্যালয়গুলো বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ে উন্নীত করতে লাখ লাখ ডলার ব্যয় অব্যাহত রেখেছে। ২০১৫ সাল থেকে চালু করা হয়েছে ডবল ফার্স্ট ক্লাস প্ল্যান প্রকল্প। এর লক্ষ্য দ্রুততম সময়ে বিশ্ববিদ্যালয়গুলো থেকে একটি বিশ্ববিদ্যালয়কে বিশ্বমানে পরিণত করা ও প্রতিষ্ঠানের পরিসর বাড়ানো।যেকোনো কিছুর পেছনে অর্থ হচ্ছে মূল চালিকা শক্তি। অর্থায়ন প্রক্রিয়া বিশ্ববিদ্যালয়গুলোকে উৎকৃষ্ট মানের গবেষণায় অনুপ্রাণিত করে। চীনা বিশ্ববিদ্যালয়ের একাডেমিক গবেষণায় নিয়োজিত ব্যক্তিদের যথেষ্ট প্রণোদনারও ব্যবস্থা রয়েছে। সিংহুয়ার মেধাবী শিক্ষার্থীরা সেরা গবেষক হয়ে দেশের উন্নয়নে কাজ করেন কিংবা দেশের হয়ে বিদেশে গবেষণায় নিযুক্ত হন। ২০১৭ সালে সিংহুয়া বিশ্ববিদ্যালয় ১ হাজার ৩৮৫ জনকে ডক্টরেট উপাধি দিয়েছে। একেই সময় যুক্তরাষ্ট্রের ম্যাসাচুসেটস ইনস্টিটিউট অব টেকনোলজিতে (এমআইটি) ৬৫৪ জনকে ডক্টরেট দেওয়া হয়।

Summary Generated by FCM (Article 2)

যুক্তরাজ্যের সাপ্তাহিক প্রকাশনা টাইমস হায়ার এডুকেশন সাময়িকীর সর্বশেষ জরিপে বিশ্বের শীর্ষ ১০টি বিশ্ববিদ্যালয়ের মধ্যে চীনের কোনো বিশ্ববিদ্যালয়ের নাম নেই। অবশ্য সাংহাই র‍্যাঙ্কিংয়ে শীর্ষ ৫০০ বিশ্ববিদ্যালয়ের মধ্যে চীনেরই আছে ৪৫টি বিশ্ববিদ্যালয়। আর এ জন্য দেশটির নামকরা সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়ের ওপর আরও বেশি জোর দিয়েছে চীনা কর্তৃপক্ষ। পশ্চিমা গবেষণাভিত্তিক বিশ্ববিদ্যালয়গুলোর আদলে পরিচালিত হচ্ছে চীনের সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়। ১৯৮৯ সালে তিয়েনআনমেন স্কয়ারে বিক্ষোভে বিশ্ববিদ্যালয়টি অগ্রণী ভূমিকা রাখে।লন্ডনভিত্তিক সাময়িকী দ্য ইকোনমিস্টের প্রতিবেদন বলছে, ১৯৯৫ সাল থেকে চীনা কেন্দ্রীয় সরকার দেশটির বিশ্ববিদ্যালয়গুলো বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ে উন্নীত করতে লাখ লাখ ডলার ব্যয় অব্যাহত রেখেছে। এর আওতায় প্রথম ২১১টি প্রকল্প হাতে নেওয়া হয়েছে। একবিংশ শতাব্দীর চ্যালেঞ্জ মোকাবিলায় প্রায় ১০০টি প্রতিষ্ঠানকে প্রস্তুত করা হয়েছে। ২০১৫ সাল থেকে চালু করা হয়েছে ডবল ফার্স্ট ক্লাস প্ল্যান প্রকল্প। এর লক্ষ্য দ্রুততম সময়ে বিশ্ববিদ্যালয়গুলো থেকে একটি বিশ্ববিদ্যালয়কে বিশ্বমানে পরিণত করা ও প্রতিষ্ঠানের পরিসর বাড়ানো। প্রতিবেশী দেশ ভারত তাদের ২০টি বিশ্ববিদ্যালয়কে বিশ্বমানে নেওয়ার ঘোষণা দিয়েছে। এমনকি নাইজেরিয়ার মতো দেশ ২০২০ সালের মধ্যে তাদের অন্তত দুটি বিশ্ববিদ্যালয়কে বিশ্বের শীর্ষ ২০০টির মধ্যে অন্তর্ভুক্তির লক্ষ্যমাত্রা নির্ধারণ করেছে।সিংহুয়ার মেধাবী শিক্ষার্থীরা সেরা গবেষক হয়ে দেশের উন্নয়নে কাজ করেন কিংবা দেশের হয়ে বিদেশে গবেষণায় নিযুক্ত হন। ২০১৭ সালে সিংহুয়া বিশ্ববিদ্যালয় ১ হাজার ৩৮৫ জনকে ডক্টরেট উপাধি দিয়েছে। একেই সময় যুক্তরাষ্ট্রের ম্যাসাচুসেটস ইনস্টিটিউট অব টেকনোলজিতে (এমআইটি) ৬৫৪ জনকে ডক্টরেট দেওয়া হয়। সিংহুয়া বিশ্ববিদ্যালের ভাইস চেয়ারম্যান ইয়াং বিন বলেন, ‘সিংহুয়ার সবচেয়ে গুরুত্বপূর্ণ উন্নয়ন ছিল ১৯৭৮ সালে, যখন ডেং জিয়াওপিং (প্রয়াত রাজনীতিক) বলেন, চীন বিপুলসংখ্যক শিক্ষার্থী বিদেশে পাঠাবে।

Summary Generated by TextRank Algorithm (Article 2)

যুক্তরাজ্যের সাপ্তাহিক প্রকাশনা টাইমস হায়ার এডুকেশন সাময়িকীর সর্বশেষ জরিপে বিশ্বের শীর্ষ ১০টি বিশ্ববিদ্যালয়ের মধ্যে চীনের কোনো বিশ্ববিদ্যালয়ের নাম নেই। অবশ্য সাংহাই র‍্যাঙ্কিংয়ে শীর্ষ ৫০০ বিশ্ববিদ্যালয়ের মধ্যে চীনেরই আছে ৪৫টি বিশ্ববিদ্যালয়। আর এ জন্য দেশটির নামকরা সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়ের ওপর আরও বেশি জোর দিয়েছে চীনা কর্তৃপক্ষ। ১৯১১ সালে বেইজিংয়ে প্রতিষ্ঠিত হয় সিংহুয়া বিশ্ববিদ্যালয়। শতাধিক বছরের পুরোনো এই বিশ্ববিদ্যালয় এখন গবেষণা, বিজ্ঞান, প্রযুক্তি, প্রকৌশল ও গণিত বিষয়ে চীনাদের গর্বের প্রতীক। পশ্চিমা গবেষণাভিত্তিক বিশ্ববিদ্যালয়গুলোর আদলে পরিচালিত হচ্ছে চীনের সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়। এই দুটি বিশ্ববিদ্যালয় পরস্পর প্রতিবেশী ও প্রতিযোগী, যা কিনা চীনের অক্সফোর্ড ও কেমব্রিজ হিসেবে খ্যাত। সিংহুয়া হচ্ছে প্রচলিত ও বাস্তবধর্মী বিশ্ববিদ্যালয়। আর পিকিং বিশ্ববিদ্যালয় হচ্ছে দেশটির কবি, দার্শনিক ও বিপ্লবীদের তীর্থস্থান। চীনের সাবেক শীর্ষ নেতা মাও সেতুং এই বিশ্ববিদ্যালয় গবেষণা করেছিলেন। লন্ডনভিত্তিক সাময়িকী দ্য ইকোনমিস্টের প্রতিবেদন বলছে, ১৯৯৫ সাল থেকে চীনা কেন্দ্রীয় সরকার দেশটির বিশ্ববিদ্যালয়গুলো বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ে উন্নীত করতে লাখ লাখ ডলার ব্যয় অব্যাহত রেখেছে। চীনা বিশ্ববিদ্যালয়ের একাডেমিক গবেষণায় নিয়োজিত ব্যক্তিদের যথেষ্ট প্রণোদনারও ব্যবস্থা রয়েছে। প্রযুক্তি ও প্রতিযোগিতানির্ভর বিশ্বের বিভিন্ন দেশের সরকারও নীতিনির্ধারণে পরিবর্তন নিয়ে আসছে। বিশ্বমানের বিশ্ববিদ্যালয় প্রতিষ্ঠা, উন্নত গবেষণা, র‍্যাঙ্কিংয়ে অন্তর্ভুক্তি ও অগ্রগতির জন্য শিক্ষা খাতে চীন ছাড়াও ভারত, সিঙ্গাপুর, দক্ষিণ কোরিয়া, তাইওয়ান, ফ্রান্স, জার্মানি বিপুল অর্থ ব্যয় করছে। এমনকি নাইজেরিয়ার মতো দেশ ২০২০ সালের মধ্যে তাদের অন্তত দুটি বিশ্ববিদ্যালয়কে বিশ্বের শীর্ষ ২০০টির মধ্যে অন্তর্ভুক্তির লক্ষ্যমাত্রা নির্ধারণ করেছে। ২০১৭ সালে সিংহুয়া বিশ্ববিদ্যালয় ১ হাজার ৩৮৫ জনকে ডক্টরেট উপাধি দিয়েছে। অবশ্য এই সংখ্যা সিংহুয়া বিশ্ববিদ্যালয়ের সাফল্যের প্রধান কারণ নয়। সিংহুয়া বিশ্ববিদ্যালের ভাইস চেয়ারম্যান ইয়াং বিন বলেন, ‘সিংহুয়ার সবচেয়ে গুরুত্বপূর্ণ উন্নয়ন ছিল ১৯৭৮ সালে, যখন ডেং জিয়াওপিং (প্রয়াত রাজনীতিক) বলেন, চীন বিপুলসংখ্যক শিক্ষার্থী বিদেশে পাঠাবে।’ তিনি আরও বলেন, ‘১০ হাজার শিক্ষার্থী বিদেশে পাঠানো প্রয়োজন। ৪০ বছর ধরে সিংহুয়া এবং দেশের অন্য শীর্ষ বিশ্ববিদ্যালয়গুলো তাদের কৃতিত্ব ধরে রেখেছে। এসব বিশ্ববিদ্যালয়ের

প্রতি আকর্ষণ বাড়াতে সরকারও অতিরিক্ত সম্পদ ও প্রয়োজনীয় উপকরণ সরবরাহ করে আসছে। বিশ্বের শীর্ষ জার্নালগুলো ইংরেজি ভাষায় লিখিত ও প্রকাশিত হয়, যা চীনা বিজ্ঞানীদের জন্য প্রতিবন্ধকতা তৈরি করে। মোদ্দা কথা হচ্ছে, শিক্ষাক্ষেত্রে বিপ্লব ঘটাতে এবং বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ের কাতারে চীনের বিশ্ববিদ্যালয়কে অন্তর্ভুক্তি করতে যা করা প্রয়োজন, তা–ই তারা করে যাচ্ছে। অক্সফোর্ড বিশ্ববিদ্যালয়ের অধ্যাপক ও সিংহুয়া জার্নাল অব এডুকেশনের সম্পাদকীয় পরিষদের সদস্য সিমন মার্জিনসন বলেন, আগামী পাঁচ বছর কিংবা এর কম সময়ের মধ্যে নাম্বার ওয়ান বিশ্ববিদ্যালয় হবে সিংহুয়া।

Aggregate Summary (Article 2)

যুক্তরাজ্যের সাপ্তাহিক প্রকাশনা টাইমস হায়ার এডুকেশন সাময়িকীর সর্বশেষ জরিপে বিশ্বের শীর্ষ ১০টি বিশ্ববিদ্যালয়ের মধ্যে চীনের কোনো বিশ্ববিদ্যালয়ের নাম নেই। পশ্চিমা গবেষণাভিত্তিক বিশ্ববিদ্যালয়গুলোর আদলে পরিচালিত হচ্ছে চীনের সিংহুয়া বিশ্ববিদ্যালয় ও পিকিং বিশ্ববিদ্যালয়।লন্ডনভিত্তিক সাময়িকী দ্য ইকোনমিস্টের প্রতিবেদন বলছে, ১৯৯৫ সাল থেকে চীনা কেন্দ্রীয় সরকার দেশটির বিশ্ববিদ্যালয়গুলো বিশ্বের শীর্ষ বিশ্ববিদ্যালয়ে উন্নীত করতে লাখ লাখ ডলার ব্যয় অব্যাহত রেখেছে। ২০১৫ সাল থেকে চালু করা হয়েছে ডবল ফার্স্ট ক্লাস প্ল্যান প্রকল্প। এর লক্ষ্য দ্রুততম সময়ে বিশ্ববিদ্যালয়গুলো থেকে একটি বিশ্ববিদ্যালয়কে বিশ্বমানে পরিণত করা ও প্রতিষ্ঠানের পরিসর বাড়ানো। এমনকি নাইজেরিয়ার মতো দেশ ২০২০ সালের মধ্যে তাদের অন্তত দুটি বিশ্ববিদ্যালয়কে বিশ্বের শীর্ষ ২০০টির মধ্যে অন্তর্ভুক্তির লক্ষ্যমাত্রা নির্ধারণ করেছে।সিংহুয়ার মেধাবী শিক্ষার্থীরা সেরা গবেষক হয়ে দেশের উন্নয়নে কাজ করেন কিংবা দেশের হয়ে বিদেশে গবেষণায় নিযুক্ত হন। ২০১৭ সালে সিংহুয়া বিশ্ববিদ্যালয় ১ হাজার ৩৮৫ জনকে ডক্টরেট উপাধি দিয়েছে। একেই সময় যুক্তরাষ্ট্রের ম্যাসাচুসেটস ইনস্টিটিউট অব টেকনোলজিতে (এমআইটি) ৬৫৪ জনকে ডক্টরেট দেওয়া হয়।

Chapter 5

Conclusion

As the world progresses in this Information Technology era, research in the Bengali language becomes more and more important. A text summarization system holds significance because of the importance of saving time, effort and also data. Text summarization can have two schools of thoughts: extractive summarization, and abstractive summarization. While the output from the abstractive method of summarization is more natural and coherent, it needs more processing and the complexity of the program is too high. As such, the extractive method to summarization provides a greater trade-off value due to the lower computational requirements. In this paper, an FCM based algorithm is used in conjunction with 6 sentence scoring methods to find the most important sentences. For the purpose of a comparative study, a TextRank algorithm was used to generate a summary along with an Aggregate scoring algorithm. TextRank algorithm simply uses a similarity measure to find the most useful sentences in an article. The Aggregate Scoring algorithm also uses the 6 scoring methods and the sum of all 6 scores from these methods are added to calculate aggregate scores for each sentence. These scores are then finally sorted in descending order, and then the top scoring sentences are then printed in original order to form an extracted summary. An FCM based algorithm tends to return higher F-number as well as a higher number of relevant sentences (sentences that are also found in the Gold summary).

In the future, an FCM based algorithm in conjunction to word-based scoring as well as more sentence based scoring methods can be used for further improvements in extractive summarization techniques. FCM can also be modified and implemented in abstractive text summarization for better human like summaries. Furthermore, we can also implement Automatic Text Sumarization by the help of a web plugin or mobile application which would automatically scrape data from websites and generate a real time text summary.

References

[1] Islam, M. T., & Al Masum, S. M. (2004, December). Bhasa: A corpus-based information retrieval and summariser for bengali text. In *Proceedings of the 7th International Conference on Computer and Information Technology*.

[2] Uddin, M. N., & Khan, S. A. (2007, December). A study on text summarization techniques and implement few of them for Bangla language. In *Computer and information technology, 2007. iccit 2007. 10th international conference on* (pp. 1-4). IEEE.

[3] Sarkar, K. (2012, August). An approach to summarizing Bengali news documents. In *proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 857-862). ACM.

[4] Efat, M. I. A., Ibrahim, M., & Kayesh, H. (2013, May). Automated Bangla text summarization by sentence scoring and ranking. In *Informatics, Electronics & Vision (ICIEV), 2013 International Conference on* (pp. 1-5). IEEE.

[5] Das, A., & Bandyopadhyay, S. (2010, August). Topic-based Bengali opinion summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 232-240). Association for Computational Linguistics.

[6] Sarkar, K. (2012). Bengali text summarization by sentence extraction. *arXiv preprint arXiv:1201.2240*.

[7] Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017, July). A heuristic approach of text summarization for Bengali documentation. In *Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on* (pp. 1-8). IEEE.

[8] Akter, S., Asa, A. S., Uddin, M. P., Hossain, M. D., Roy, S. K., & Afjal, M. I. (2017, February). An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In *Imaging, Vision & Pattern Recognition (icIVPR), 2017 IEEE International Conference on*(pp. 1-6). IEEE.

[9] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.

[10] Moratanch, N., & Chitrakala, S. (2016, March). A survey on abstractive text summarization. In *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on* (pp. 1-7). IEEE.

[11] Andhale, N., & Bewoor, L. A. (2016, August). An overview of text summarization techniques. In *Computing Communication Control and automation (ICCUBEA), 2016 International Conference on* (pp. 1-7). IEEE.

[12] Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. In *Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on* (pp. 1-6). IEEE.

[13] Krishnaveni, P., & Balasundaram, S. R. (2017, July). Automatic text summarization by local scoring and ranking for improving coherence. In *Computing Methodologies and Communication (ICCMC), 2017 International Conference on*(pp. 59-64). IEEE.

[14] Vijay, S., Rai, V., Gupta, S., Vijayvargia, A., & Sharma, D. M. (2017, December). Extractive text summarisation in hindi. In *Asian Language Processing (IALP), 2017 International Conference on* (pp. 318-321). IEEE.

[15] Galarnyk, M. "PCA Using Python (Scikit-Learn) – Towards Data Science." *Towards Data Science*, Towards Data Science,5 Dec. 2017, towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60.

[16] Tian, S. (2017). A hybrid debris flow hazard degree analysis model based on PCA and SFLA-FCM. *Revista de la Facultad de Ingeniería*, *31*(9).

[17] Haque, M., Pervin, S., & Begum, Z. (2017). An Innovative Approach of Bangla Text Summarization by Introducing Pronoun Replacement and Improved Sentence Ranking. *Journal of Information Processing Systems*, *13*(4).

[18] Patil, D. B., & Dongre, Y. V. (2015). A fuzzy approach for text mining. *IJ Mathematical Sciences and Computing*, *4*, 34-43.

[19] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.

[20] Bezdek, J. C. (1981). Objective function clustering. In *Pattern recognition with fuzzy objective function algorithms* (pp. 43-93). Springer, Boston, MA.

[21] Witte, R., & Bergler, S. (2007). Fuzzy clustering for topic analysis and summarization of document collections. In *Advances in Artificial Intelligence* (pp. 476-488). Springer, Berlin, Heidelberg.

[22] Pole, K. R., & Mote, V. R. (2017, October). Improvised fuzzy clustering using name entity recognition and natural language processing. In *Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on*(pp. 123-126). IEEE.

[23] Kamal, R. (2014). rafi-kamal/Bangla-Stemmer. [online] GitHub. Available at: https://github.com/rafi-kamal/Bangla-Stemmer.

[24] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[25] Langville, A. N., & Meyer, C. D. (2011). *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

[26] Li, W., & Zhao, J. (2016, July). TextRank algorithm by exploiting Wikipedia for short text keywords extraction. In *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)* (pp. 683-686). IEEE.

[27] Ross, T. J. (2005). *Fuzzy logic with engineering applications*. John Wiley & Sons.

[28] (n.d.). Retrieved from https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html

[29] Josh Warner, Jason Sexauer, scikit-fuzzy, twmeggs, Alexandre M. S., Aishwarya Unnikrishnan, … Himanshu Mishra. (2017, October 6). JDWarner/scikit-fuzzy: Scikit-Fuzzy 0.3.1 (Version 0.3.1). Zenodo. doi:10.5281/zenodo.1002946

[30] DavidBelicza. (2018, October 08). DavidBelicza/PHP-Science-TextRank. Retrieved from https://github.com/DavidBelicza/PHP-Science-TextRank

[31] Research work on Bangla NLP. (n.d.). Retrieved from http://www.bnlpc.org/research.php

[32] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.