AUTOMATIC TEXT SUMMARIZATION USING ADVANCED SENTENCE RANKING SCHEMES

A Thesis Report
Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering

Submitted By

Naznin Ara Mitu	14.02.04.016
Sk. Tasbiha	14.02.04.025
Sarah Nuzhat Khan	14.02.04.064
Md. Siam Ansary	14.02.04.104

Supervised by
Mohammad Moinul Hoque
Associate Professor
Department of Computer Science and Engineering

AHSANULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY
DHAKA, BANGLADESH
DECEMBER 2018

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of Mohammad Moinul Hoque, Associate Professor, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE4100: Project & Thesis I and CSE4250: Project & Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

………………………..
Naznin Ara Mitu

………………………..
Sk. Tasbiha

………………………..
Sarah Nuzhat Khan

………………………..
Md. Siam Ansary

Countersigned by

Mohammad Moinul Hoque
Associate Professor, Department of CSE, AUST
…………………………………………...………………....

ACKNOWLEDGEMENT

ABSTRACT

Text summarization produces a concise amount of data out of document(s) which may represent the core information contained in those documents. Summaries should have all the important information and no crucial data should be left out. In recent times much research is being done in automatic text summarization. Having a detailed overview of the state-of-the-art for automatic text summarization, we have found that, this area is still very much open and there are many scopes for improvements. We have developed a system which uses combination of features to rank sentences across a document. While doing so, we have used a graph based approach with minimal text sentiment analysis and also considered the presence of key phrases. Using graph, we have found the interconnection between sentences and thus most significant sentence were identified. Neutral sentence were detected for inclusion in the final summery as well. Topic based key phrases which can also emphasize a sentence were also implemented. Combination of multiple features lead us to the identification process of the significant sentences which were finally included in the summary as per their sentence scores. We have done extractive summarization and compared them against some of the existing methods. Our method has worked quite well against the compared systems and can be used as an alternative approach towards the text summarization problem.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

Chapter 1

Introduction

1.1 Objective

A summary gives the overall sense of the document. Hence in different works, instead of going through the full document, from summary important information can be gained which makes the overall procedure more comfortable and less resources are needed in the long run. We have tried to explore a new method for the generation of summary. We need to consider all the major topics present in the document. If the summary contains sentences from all the major topics present in the document that has a better chance of giving better perspective of the document. It is typical of human summaries, too. When humans summarize a document, they include most of the topics present in the document in their summaries. Our summary generation approach is extractive, i.e., the summaries contain exact sentences present in the document.

Residing on the idea that a good summary must cover all major information present in the document, the first step towards text summarization is naturally to identify the important parts of texts. To accomplish this task, we need to extract the sentences which convey the significant information. Hence, sentences need to be scored which can be done in many approaches. A sentence that is hugely connected to other sentences that may convey the culmination of information is important. Some textual portions may carry notable emotional data. Therefore, sentiment analysis may play a vital part in identifying sentences that ought to be included in summary. Key-phrases in different topics emphasize on material of sentence which may lead to be included in summary.

The generated summaries needs to be evaluated to determine the effectiveness and efficiency of approaches. There are different dataset available for research purposes for evaluation causes.

In this chapter, we have discussed about text summarization and basic ideas surrounding it. In the later chapters, more have been scrutinized. In the second chapter, discussion has been done about state of the art and existing research issues and problems of the topic. In the third chapter, evaluation measures have been explored. In fourth chapter, a new approach for automatic summarization has been proposed and its implementation has been detailed. In fifth chapter, the experimental results and gained improvements have been analyzed. In sixth chapter, conclusion has been drawn going through the difficulties and

scopes of developments.

1.2   What is Text Summarization

Text summarization is a way to condense the large amount of information into a concise form by the process of selection of important information and discarding unimportant and redundant information. With the amount of textual information present in the world wide web the area of text summarization is becoming very important.

The search engines do a remarkable job in searching through a heap of information to dish out the most relevant information the user is looking for. Even the information picked by search engines with a great precision is of a daunting amount. It is very time consuming to read through entire length of the document. Invariably a decision for a certain task has to be made in certain time frame and reading through all the documents is simply impossible. The process speeds up considerably when the essence i.e., the summary of the document is also available. The technology of automatic text summarization is critical when dealing with problems like that.

There has been a rapid increase in the amount of researches done in the field of automatic text summarization. It is also increasingly being exploited in the commercial sector, in the telecommunications industry, data mining of text databases, in filters for web based information retrieval, and in word processing tools. In addition to the traditional focus of automatic indexing and automatic abstracting to support Information Retrieval, researchers are investigating the challenging problems, including multilingual summarization. As the information overload problem grows, and the information becomes more and more accessible, the field of automatic text summarization can be expected to become more and more important, and the new applications for the text summarization will surface.

1.3   Categories of Text Summarization

Text Summarization approaches are mainly of two categories. They are:

1. Extractive Summarization
2. Abstractive Summarization

1.3.1   Extractive Summarization

The extractive summarization is the one where the exact sentences present in the document

are used as summaries. The extractive summarization is simpler and is the general practice among the automatic text summarization researchers at the present time. Extractive summarization process involves giving scores to sentences using some method and then using the sentences that achieve highest scores as summaries. As the exact sentence present in the document is used the semantic factor can be ignored which results in generation of less calculation intensive summarization procedure. This kind of summary is generally completely unsupervised and language independent too. Although this kind of summary does its job in conveying the essential information it may not be necessarily smooth or fluent. Sometimes there can be almost no connection between adjacent sentences in the summary resulting in the text lacking in readability. In this research, only extractive summarization has been studied.

1.3.2   Abstractive Summarization

The abstractive summarization is the process in which the abstract of the document is created. The abstract can contain the words and phrases not present in the original document. The abstractive summarization procedure is a very complicated process as the semantics of sentences has to be dealt with. Several other factors such as word sense, grammatical structure has to be taken into consideration before creating a useful abstractive summary. The abstractive summary that has all the characteristics of a good summary is the ultimate goal of automatic text summarization. Abstractive summary is important because as the textual structure present in the summary can vary significantly from the original text the copyright of the original text won't be infringed and the intellectual rights for the summary can be owned by the summarizing party.

Abstractive summarization approach can be further extended to two methods. They are as below:
- Structure based approach
- Semantic based approach

Structure based approach uses prior knowledge. It encodes most vital data from the document(s) through psychological feature schemes like templates, extraction rules, alternative structures such as tree, ontology, lead and body rule, graph based structure etc. Semantic based approach employs the linguistic illustration of document(s) to feed into Natural Language Generation system. This technique specializes in identifying noun phrases and verb phrases by processing linguistic data.

1.4   Modes of Summarization Systems

- Generic summarization
- Query based summarization

Generic summarization consists in extracting text spans relevant to the main topics of a whole document, while Query based summarization abstracts the information relevant to a given query.

1.5   Stages of Text Summarization

Andhale and Bewoor [1] accumulated that text summarization process work in three steps which are analysis, transformation and synthesis. Hovy [2] amassed that text summarization has three distinct phases which are topic identification, interpretation or topic fusion and summary generation. We can come to decision that the whole process may be of two steps which are topic identification and summary generation. Summary generation may be extended. It may have analysis, transformation and synthesis sub processes.

We can represent the idea in the following way:

- Topic Identification
- Summary generation
  - Analysis
  - Transformation
  - Synthesis

To perform topic identification stage most of the systems employ several independent modules according to Hovy [2]. Each module assigns a score to each unit of input, then a combination module combines the scores for each unit to assign a single integrated score to it. Finally the system returns then highest scoring units, according to the summary length requested by the user.

Summary generation phase leads to the desired summary through various sub-phases. The different phases can be of extractive or abstractive approach.

At first analysis is done based on important features and techniques. Then, transformation is done. Sentences may be reduced in size to make the summary condensed or fusion may

be implemented. Later, in the synapsis phase, the summary is presented.

The analysis phase can have many manners such as machine learning approach, ranking based on positional criteria, cue phrase, word and phrase frequency etc. In machine learning approach, the classifier is trained to identify text spans for summarization. A classifier is trained using a training set of documents and their associative summaries to distinguish between summary and non-summary sentences. After training, such systems operate on unlabeled texts. Such way is usable for scientific approaches but for a large variety of collections, documents are heterogeneous and their summaries depend much more on the content of their texts than on a global class information. When supervised learning is implemented, a training set of documents and associative summaries is needed. In this way we can label the document sentences that are relevant. Labelling a large amount of data is a very lengthy process. So, to improve, semi supervised with self-learning can be explored. In this method, a small amount of data is used for training with large amount of unlabeled document.

Sentence reduction may help to minimize the summary. Six major operations can be used for editing the sentences chosen for this purpose. They include removing extraneous phrases, combining the edited sentences, syntactic transformation, substitution of phrases with paraphrases, substituting sentences with more general description, recording the extracted sentences. As a result of such operation the conciseness of the summary is improved. Syntactic parsing, grammar checking, context information are implemented for this purpose. Such method should do the work in such a way that reduction is done without major loss. Interpretation or topic fusion may also be implemented. During this stage, topics identifies as important are fused, represented in new terms and expressed using new formulations with concepts or words that are not in original texts. This cannot be done without prior knowledge regarding the domain.

1.6 Chapter Overview

In this chapter, we have discussed about the motivation behind automatic text summarization in order to get some initial idea. Text summarization is of different categories and modes. The general approach of summarization goes through some stages. We have explored through all these to gain an insight into the significance of automatic summarization.

Chapter 2

Related Works and Research Predicament

2.1 Literature Review

The field of text summarization is developing day by day. There have been many attempts to construct a summarizer that produces summary that is close to human's work. Most of the works have been done so far are mainly on extractive approach though research on abstractive process is also being done. Different approaches considers different features for improving the results. Sentence features, graph approaches, machine learning are few of the approaches. Some works might even have contradicting motivations. We will be reflecting on some of the works.

Some approaches target sentence features for summarization. Lakshmi, Deepthi and Suresh [3] discussed automatic summarization using font and cue phrase features. Different types of fonts in a document may signify a part of document. Cue phrases also emphasize in sentence significance. Sentence location and length and presence of noun phrase also sometimes may play a part in signifying sentences. The system was tested with 20 documents and it achieved average accuracy of 75%. Srivastava and Gupta [4] proposed a summarization approach that uses key phrases. Presence of cue words, specially prepared lists' words and basic ids put emphasize on sentences. Efat, Ibrahim and Kayesh [5] conducted text summarization on documents of Bengali language. Their method focused on frequency of words, positions of sentences, presence of cue words and documents structure. They used forty five Bengali articles for experiments from various newspapers. Average accuracy of 83.57% was achieved by their approach. Krishna and Reddy [6] experimented with a summarization process of query mode. The sentences of the document that match closest to a given query can be considered as summary sentences. They said that traditional stop words should be used for summarization, rather a list should be prepared specially for summarization purpose. Jing [7] tried summarization through sentence reduction. If sentences are reduced removing extraneous parts, summarization can be done.

In their experiments they achieved average success rate of 81.3% using five-fold validation.

Many have tried to do summarization using machine learning approaches. Neto, Freitas, Kaestner [8] suggested summarization using machine learning techniques. They employed C4.5 and Naïve Bayes classifier on training set and evaluation to testing set. TIPSTER document base was used. They gained compression rate of 10%.Wong, Wi and Li [9] presented summarization using supervised and semi-supervised learning together. They experimented with Probabilistic Support Vector Machine and Naïve Bayes. According to them co-training gives better results than individual approach. DUC 2001 dataset and ROUGE measure was used in their experiments.

Table 2.1: ROUGE evaluation of the approach proposed by Wong, Wu and Li			
Learning Approaches	ROUGE -1	ROUGE -2	ROUGE – L
PSVM	0.358	0.082	0.323
NBC	0.353	0.061	0.317
COT	0.366	0.090	0.329

Graphical approaches have been tried to do summarization in some works. Mihalcea have discussed text summarization in her works of [10] and [11] that is mainly based on graphs. A graph may be constructed using the sentences as nodes and after putting edges, edge weights may be calculated with similarity functions measuring the similarity between two nodes. Ranking of scores gives a score to each sentence and top ranked sentences may be used as a summary.

Sentiment analysis of sentences has been applied to do summarization in some existing approaches. Dabholkar, Patadia and Dsilva [12] presented extractive text summarization with sentiment analysis of sentences. They emphasized on neutral sentences as sentences in summaries should be unbiased and not have any opinionated sentences. Solov'ev, Antonova and Pazel'skaia [13] did summarization of documents using sentiment analysis emphasizing on positive and negative sentences. According to them emotionally charged sentences have more information than others so they should be considered as summary sentences.

2.2 Research Issues

Many research works are being done in the field of text summarization using different approaches. But despite of ongoing experiments, there are still scopes for improvement

and there exist issues that still need to be addressed without which machine summaries cannot be closer to human summaries in quality. We have discussed some issues as in the below:

- **Sentence generation**

The goal of automatic text summarization is to create summary that is similar to human written once. But when humans write summary, they do so in their own expressions that is humans don't normally use direct sentences from document(s) rather they write new and concise sentences that has all the data that is important. For machine it is very difficult as there are many grammatical facts in language and to generate new languages is a very challenging task.

- **Variations of languages**

There are roughly 6500 spoken languages at current time in world. However, about 2000 have fewer than 1000 speakers. Still there are many languages. Automatic summarization in these many languages need a lot of work. Most of the works so far have been done mainly for English language. Before summarization process even begins, machine wound need to detect a language through their patterns and then summarization process would have to work which is very complicated.

- **Multi lingual documents**

The document(s) that are to be summarized can have more than one language in it. To summarize, machine would have to identify the languages of the texts to perform summarization. Hence, summarization process would get very cumbersome.

- **Document formats**

Textual data can be found in various documents. They documents can be in different formats and may need pro processing of many different types to extract the text from document(s).

- **Multiple document summarization**

At a time, summarization may need to be applied to more than one document. In this process to rank or generate sentences in order can be very difficult. Also compression ratio and guarantee of quality might get compromised.

- **Anaphoric links identification**

Anaphora is the use of an expression whose interpretation depends on another expression in context. Different context may give different meaning to one text. Anaphoric link identification is a very difficult task for machines as to get the meaning of something in different context can be quite laborious.

- **Ambiguity perception**

Linguistic components and their meanings can get ambiguous even for humans. Writers sometimes express same thought in contradicting manner. For a machine to summarize such texts is hard.

- **Out of vocabulary problem**

Existing systems often face out of vocabulary problems. Human languages are always changing and new words are being added. Automatic summarizers might not be able to identify words due to lack of proper vocabulary collection.

- **Poor evaluation measures**

Evaluating a summary is a difficult task. Even when humans write summaries, summaries of two people are most likely to differ. People do not normally rank or generate sentences in similar manner as hand written summaries, even though should maintain all necessary data, can be biased of writers perception of the topic. While evaluating a machine summary, to use reference summary and to think of it as the absolute standard may be a bit of hypocrisy. Evaluation technique itself is a ongoing research field.

2.3   Research Problems

In our thesis work, we would try to generate summaries that are good and close to human ones. Hence we want to improve the scores in evaluation matrices. We would explore how works can be done to tackle some hitches and gain better summaries. In our work, we would try to work on the following problems:

- **Feature Selection**

Earlier in this chapter, we have seen that different approaches use various features. Different characteristics of sentences and texts have been used. But to generate a good

summary we cannot use all of these features as there are some features which may even worsen the quality of the generated summary. Hence, it is a task to identify which feature(s) should actually be used for summary generation.

- **Combination of features**

Now, when sentences are written by humans, in their subconscious mind, they don't always use the same way to emphasize on sentences, also different human has their own styles. Hence, it should not be the case that only one feature is used. Rather more than one features should be used but the number should be limited and not all features have the same amount of significance. Also while combining it is important to differentiate among the significance of used features.

- **Improvement in measures**

We hope to improve the generated summaries from the existing techniques. We hope to use eccentric combination of features that will take into consideration the structure of sentences and emphasize of the contained facts and develop the summaries.

Our aim is to find out ways so that summaries can be prepared with features in such a way that the scores such as precision, recall, f-score are improved than the existing systems. Automatic summaries get better in quality as they become similar to human summaries and only then improvements can be achieved.

2.4 Chapter Overview

In this chapter, we have discussed about the state of the art of automatic text summarization. Many different approaches have been considered to generate machine summaries that would be close to human ones but still there lingers scope for improvements. We have also explored some research issues and the research problems that we have tried to resolve in our work.

Chapter 3

Evaluation of Summary

3.1   Introduction

The evaluation of summary is a very subjective process and hence the evaluation is a very difficult task. The goal of automatic summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user or applications needs. It is a subject of debate as what makes a summary a good summary. The summaries can vary to a large degree depending on who is the summarizer. For a summary to be a good summary it has to be comparable to so called good summary. So evaluation of summary is a difficult task.

3.2   Evaluation Measures

There are several measures to evaluate a summary. Josef and Karel [14] discussed about them. The measures are mainly of two kinds. They are:
- Intrinsic measures
- Extrinsic measures

Intrinsic evaluation consists of an isolated system where its results from given tasks are compared to a set of standard answers, predefined by evaluators. It helps to understand the quality of the automatic summary and if it is generated as desired. Extrinsic evaluation is where a system is tested and evaluated in a real world situation where a user has comments on its performance, utility and overall usefulness. Rather than analyzing the sentences of the summaries, extrinsic evaluation try to measure the prospect of using summaries for certain tasks.

3.3   Study of Intrinsic measures

Intrinsic measures are more crucial to determine the effectiveness and efficiency of generated summary than extrinsic measures. It reflects on the text quality and its materials

and through numerical scores help to compare with the provided summary of any dataset. As intrinsic measures are of many emplacement, we give them more importance. We have discussed some intrinsic measures in below:

- ➢ **Text quality evaluation:** There are several aspects for quality measures. They are as below:
  - **Grammatically:** The summary should not contain non-textual items or punctuation errors or incorrect words.
  - **Non redundancy:** Summary should not contain redundant information.
  - **Reference quality:** The nouns and pronouns should be clearly referred to in the summary.
  - **Coherence and structure:** The summary should have good structure and the sentences should be coherent.
- ➢ **Content Evaluation:** The content evaluation measures are the assessments of how well generated summaries is close to gold summaries.
  - **Co selection:**
    - o **Precision:** Precision (P) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary.
    - o **Recall:** Recall (R) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the ideal summary.
    - o **F-score:** F-score is a composite measure that combines precision and recall. It is the harmonic average of precision and recall.
    - o **Mean reciprocal rank:** The mean reciprocal rank evaluates possible responses to a samples, ordered by probability of correctness.
    - o **Mean average precision:** Mean average precision is the mean of the average precision scores for each try.
  - **Content based:**
    - o **Cosine similarity:** Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.
    - o **Unit overlap:** Unit Overlap between sentences of summaries may be calculated.
    - o **Longest common sub sequence:** The third content-based measure is called Longest Common Subsequence (LCS).
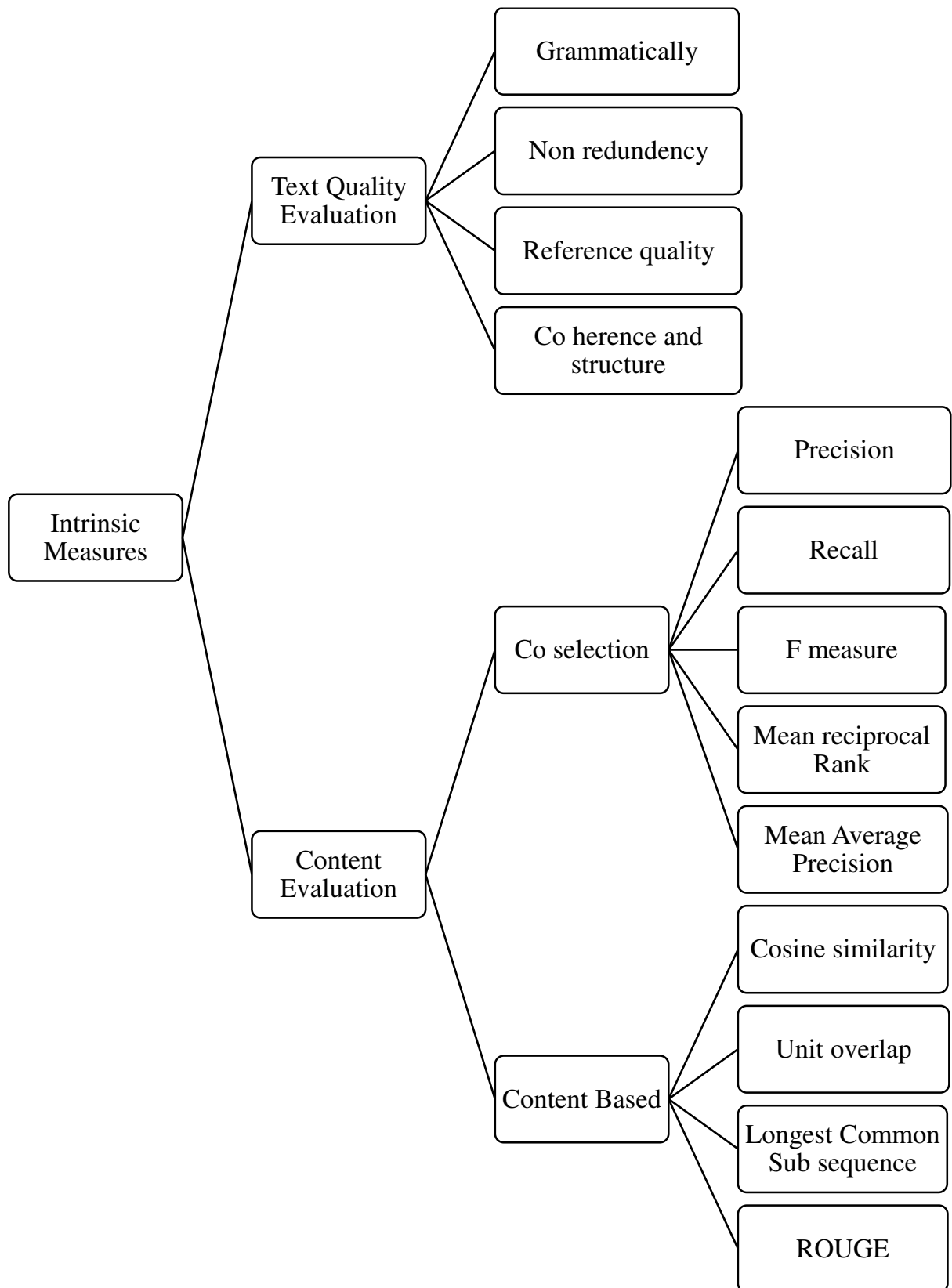
Figure 3.1: Different Intrinsic measures for summary evaluation

- **N-gram Co-occurrence Statistics – ROUGE:** The ROUGE family of measures, which are based on the similarity of n-grams, was firstly introduced in 2003 in DUC conference. Suppose a number of annotators created reference summaries – reference summary set (RSS). Simply, ROUGE-n is overlap of n-grams between the system and reference summaries. ROUGE-1 refers to the overlap of 1-gram (each word) between the system and reference summaries. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

3.4   Chapter Overview

In this chapter, we have discussed about the importance of evaluation of summary. Evaluation conveys us how good is the machine generated summary in quality. We have also explored different evaluation measure and their significance.

Chapter 4

Proposed Method

4.1 Introduction

We have done text summarization using the extractive approach. In this approach we select the sentences that best represent the document. For this, all the sentences are scored based on some criteria. Then they are sorted and using the top most sentences, summary is formed.

4.2 Selected Criteria

We have selected three criteria for scoring the sentences. They are as below:
- Interconnection among sentences
- Prioritization of neutral sentences
- Primacy of sentences with key phrases

4.2.1 Inter connection of sentences

The idea is that we find out how a sentence is connected to other ones that is how similar of dissimilar they are or the data in the sentences overlay. The sentence that has the more score from interconnection with others has more significance, hence should be in the summary. What we do for this is that, we construct a graph using the sentences. In the graph, each sentence acts like a node. Edges are then formed between the nodes. Now, weights are assigned to each edge. The weights are calculated as the cosine similarity between sentences.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Cosine similarity between x and y can be calculated as in eqn 4.1.

$$cosine(x, y) = \frac{\sum_{i=0}^{n-1} x_i y_i}{\sqrt{\sum_{i=0}^{n-1} x_i{}^2} \sqrt{\sum_{i=0}^{n-1} y_i{}^2}} \qquad (4.1)$$

After weight are assigned, the graph is complete. Now, how sentences are interconnected is to be found out. For this PageRank algorithm can be used. PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. It was originally designed as an algorithm to rank web pages.

Algorithm 4.1 : Steps for scoring sentences based on interconnection	
Step 1	Start
Step 2	Input Document
Step 3	Split the document into sentences
Step 4	Construct a text graph where each sentence is a node
Step 5	The weight of the edge between two nodes is the corresponding cosine similarity value of sentences
Step 6	Calculate PageRank of each node
Step 7	End

$W_1 =$ *cosine similarity (S_1, S_2)*
$W_2 =$ *cosine similarity (S_1, S_3)*
$W_3 =$ *cosine similarity (S_2, S_3)*
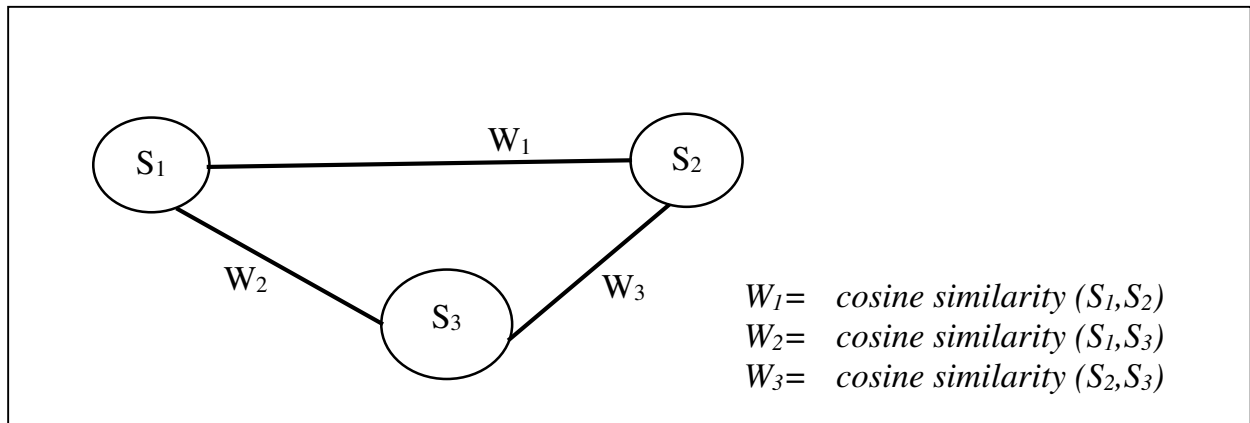
Figure 4.1: A sample text graph

Considering Figure 4.1, suppose there is a document with only three sentences. Then, each sentence is considered as a node and an edge is put between two nodes. The weights of an edge is the cosine similarity of the sentences of the corresponding nodes. Then, PageRank algorithm is applied.

4.2.2   Prioritization of neutral sentences

In the summary, the sentences with the most data should be included. Neutral sentences don't have any positive or negative opinions or biased information. Rather they hold facts about a matter. Hence, identifying neutral sentences, they should be included in the summary.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. Sentiment analysis

is done to select the portions of the text that conveys the most message in any deed. Through sentiment analysis it is possible to find out the sentence polarity and identify the neutral sentences.

Sentiment polarity can be determined with Naïve Bayes classifier. Naive Bayes is a kind of classifier which uses the Bayes Theorem. Bayes theorem works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, the probability of an event can be calculated using its prior knowledge.

Naive Bayes classifier assumes that all the features are unrelated to each other. Presence or absence of a feature does not influence the presence or absence of any other feature. Training with pre-classified data helps to do sentiment analysis. When a parser parses the text, the subjectivity and the polarity of every adjective is clustered; after that, a simple probability distribution table is given out. A probability distribution chart is confined between 0.0 and 1.0.

4.2.3   Primacy of sentences with key phrases

A key phrase is made up of multiple keywords or is a combination of keywords. Key phrases provide a compact representation of a document's content. Ideally they represent in condensed form the essential content of a document. There are specific key phrases for different topics. These key phrases being in a sentence, signify them and give a boost. Hence, if a sentence of a certain topic has key phrases of that topic, it has much more significance than others and help to create summaries of documents.

4.3   Devised approach

For summarization, we have devised an approach. The process consists of several steps. The main phases for formation of summary are preparation of a phrase list and summarization process execution. The steps are illustrated in the flowchart of the figure 4.2 and the steps are listed in figure 4.3.

4.4   Detailed analysis of proposed model

For text summarization, we score the sentences based on interconnectivity, neutral contents and presence of key phrases. Before going straight to summarization process, at first we prepare five key phrase lists for the topics. After summarization, the evaluation of the generated summaries are done. Then, comparisons are done with three different other summarization approaches.

4.4.1　Key phrase Extraction

Key phrases are independent of corpus that is if a key phrase list is prepared from a document of certain type, the key phrase list is also applicable for other documents of same
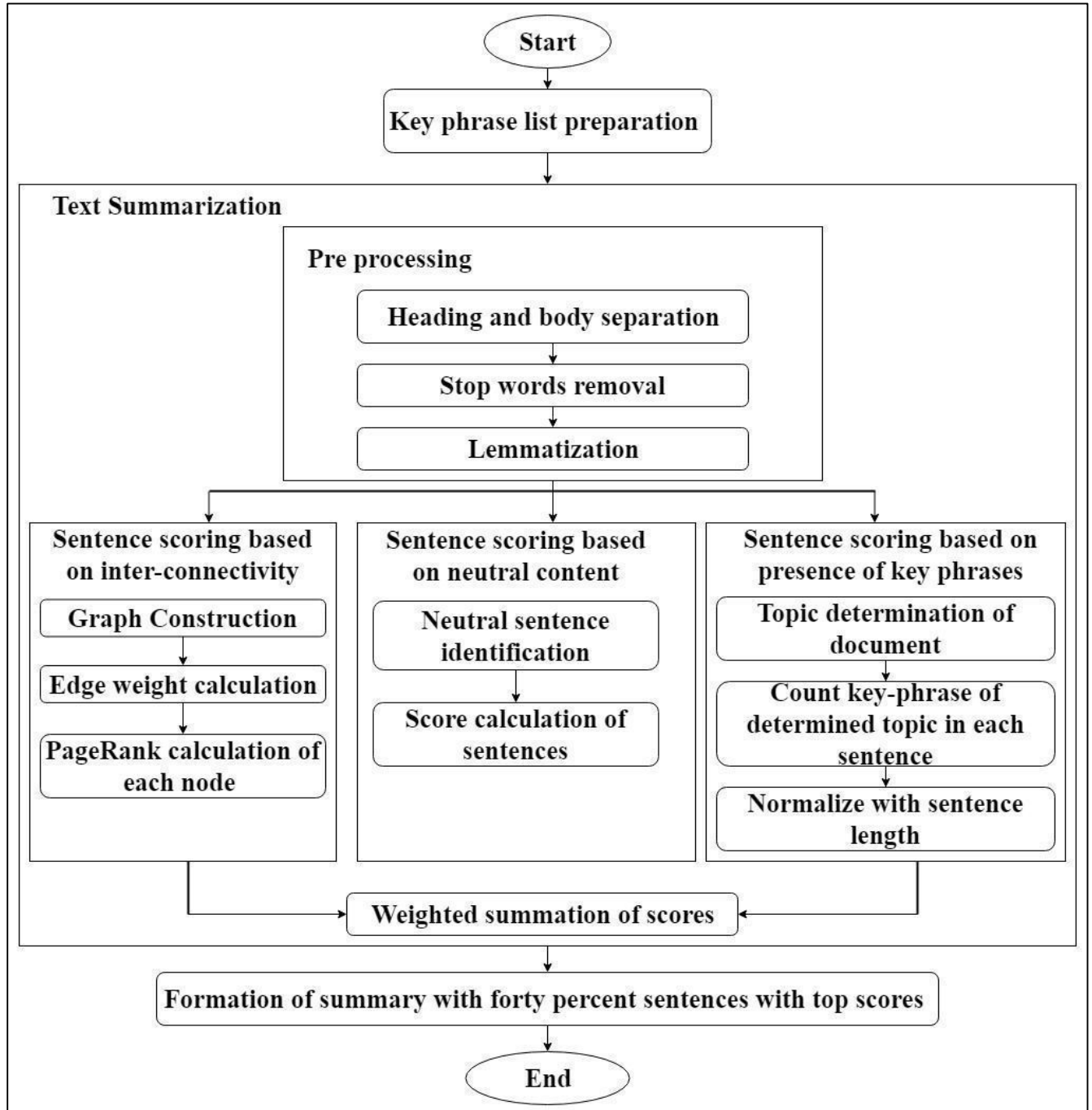
Figure 4.2: Flowchart of steps of the devised approach

- Key phrase list formation
- Text summarization
  - Pre-processing
    - Heading and body separation
    - Stop words removal
    - Lemmatization
  - Topic determination of document
  - Scoring of sentences
    - Scoring based on sentence interconnectivity
      - Graph construction
      - Edge weight calculation
      - PageRank of each node calculation
    - Scoring in accordance with neutral content
      - Neutral sentence detection
      - Value calculation for differentiated sentences
    - Scoring on presence of key phrases
      - Document topic determination
      - Count number of key-phrases present of determined topic in each sentence
      - Normalize the count with sentence length
  - Forming summary with forty percent sentences of topmost scores

Figure 4.3: List of steps of devised approach

type. Stuart Rose, Dave Engel, Nick Cramer and Wendy Cowley [15] discussed about key phrase extraction in their research work. They proposed Rapid Automatic Keyword Extraction which is unsupervised, domain independent and language independent and extracts keyword from individual documents. It is very simple yet has computational efficiency. The steps of key phrase extraction are shown in figure 4.4.

To find out the rank of each candidate key phrase, at first the scores of each word of a key phrase is calculated and those scores are summed. The frequency of a word indicates how many times it exists in the candidate key phrases in total. The degree of a word is the summation of counts of words that make up the candidate key-phrases that the word belongs to. Score of a word is then calculated using eqn 4.2.

$$score(word) = \frac{degree(word)}{frequency(word)} \qquad (4.2)$$

We generate the key phrases for each topic individually and then save the top half cause phrase of lower score are not of much significance.

| Algorithm 4.2: Steps for key phrase extraction	
Step 1	Start
Step 2	Input document
Step 3	Split the document into an array of words
Step 4	Break array of words at word delimiters
Step 5	Split the words into sequences of contiguous words
Step 6	Break each sequence at a stop-word. Consider each sequence as a candidate keyword
Step 7	Calculate the "score" of each individual word in the list of candidate keywords using the metric $degree(word)/frequency(word)$
Step 8	For each candidate keyword, add the word scores of its constituent words to find the candidate keyword score.
Step 9	End
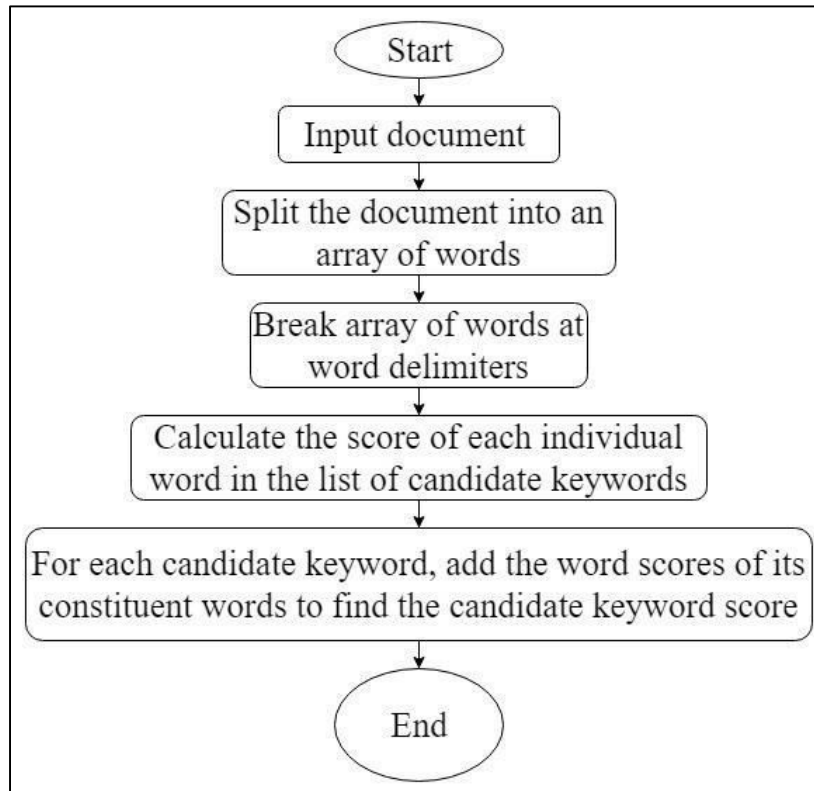
Figure 4.4: Flowchart of Steps of key-phrase extraction

4.4.2   Text Summarization execution

Text summarization is done after the key phrase lists are prepared. This stage is of several processes that are to be discussed in the following sections.

4.4.2.1 Pre processing

After each document is input, at first we identify the heading of the news article. We work with the body of the news article. Document is split to sentences. From each sentence, stop words are removed then lemmatization is done. Lemmatization is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form. Lemmatized form is effective to calculate cosine similarity and identify neutral sentences.

4.4.2.2 Topic determination of the document

Now we find out the key phrases of the particular document. Then we calculate the number how many key phrase belong to each class. The document is determined to the topic whose key phrases it has most. We use the key phrase list generated before. Identification of the topic of the input document is done as below:

$$\forall j \neq i; \; document \in topic_i \; ;$$

$$when \; \max\_count(\; document \; key \; phrase) \in topic_i \; keyphrase$$

4.4.2.3 Scoring of sentences

Now we assign score to each sentence on the basis of three criteria.

$$score(sentence_j) = \sum_{i=1}^{3} criterion_i score(sentence_j) \times weight_i \qquad (4.3)$$

$$criterion_1 score(sentence) = interconnectivity \; score(sentence) \qquad (4.4)$$

$$criterion_2 score(sentence) = \frac{0.5}{sentence \; length} \; if \; sentence \; is \; neutral \qquad (4.5)$$

$$criterion_2 score(sentence) = 0 \; if \; sentence \; isn't \; neutral \qquad (4.6)$$

$$criterion_3 score(sentence) = \frac{no \; of \; document \; topic \; keyphrases}{sentece \; length} \qquad (4.7)$$

For the first criterion, we construct a graph from the sentences where each sentence is a node and edges between sentences are put. The weights of the edges are calculated with eqn 4.1. Cosine similarity is assigned as the edge weights. Then PageRank of each node is calculated and that is the score for this criterion.

For second criterion, if find the sentiment polarity of score of each sentence. If the sentence is neutral, it is thought to be free of opinion and should contain facts. Hence, for neutral sentences score is calculated, otherwise it is zero.

We have already devised how to determine the topic of the document and we have a key phrase list for each topic. Now, we take each sentence and find out how many key phrase are there of the document topic in the sentence and then divide that count with number of tokens the sentence has. This is the score of third criterion. Now, the division is important cause, sentence length also matters.

Suppose we have two sentences, $sentence_m$ and $sentence_n$. Now, $sentence_m$ has n_1 number of key phrases and $sentence_n$ has n_2 number of key phrases of the document type, where $n_1 > n_2$. Now, it may seem like $sentence_m$ has more significance. But, if number of tokens in $sentence_m$ are far more than $sentence_n$, then only counting the key phrase does not do justice, rather number of tokens has to come into consideration.

After for each criterion scores are calculated, we calculate the weighted sum. Heuristically, first criterion should have more weight as interconnection between two sentences is of great significance. Then criterion two has importance as key-phrases are very crucial to emphasize on a sentence. Neutral sentences are important but just because a sentence is not neutral, it should not be excluded as sometimes sentences with emotional charge can hold valuable information about a particular topic in a document. In our implementation, we have taken the values of $weight_1$, $weight_2$ and $weight_3$ as 0.55, 0.15 and 0.30.

4.4.2.4　Formation of summary

In our experiment, the summaries are created with forty percent of the input documents. Hence we identify the top forty percent of sentences by sorting them in descending order of sentence score. Then summaries are formed with the identified sentences maintaining the sentence order of original document.

4.5　Chapter Overview

In this chapter, we have proposed an approach for automatic summarization using combination of three criteria and discussed about how it can be implemented.

Chapter 5

Experimental Verification

5.1 Experimental Results

For our implementation, we have used the **BBC News Summary** dataset from kaggle [16]. This is a dataset of extractive text summarization. This dataset is from the original **BBC Datasets** that are on Insight Resources [17]. This dataset is provided for use as benchmarks for machine learning research. D. Greene and P. Cunningham [18] used this dataset for their research. The dataset contains total of 2,225 documents. The documents are of five types which are business, entertainment, politics, sport and tech. Table 5.1 shows the number of documents for each topic.

Table 5.1: Document topics and no of documents of each topic in BBC Dataset	
Document topics	Number of documents
Business	510
Entertainment	386
Politics	417
Sport	511
Tech	401

BBC News Summary dataset has extractive reference summary for each news articles. The news articles consists of headlines and bodies.

We have used a portion of data only for preparing key phrase list of each topic. The rest of the data is used for summarization. Same data is not used for both to avoid the over-fitting issue. Also, the portion that is used for key phrase extraction is less or equal to the other portion, this portion should never be more. In the experiment, we did the work for three times. In first set up, 33.33% data was used for key phrase; then 2nd time, 40% and in the 3rd set up, 50% of data was used for key phrase extraction. Table 5.2 shows the number of documents used to prepare key phrase lists for each type. The documents used for key phrase extraction are not used for summarization process.

For evaluation, we have used **ROUGE** measure. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the

candidate summary and reference summary. Chin-Yew Lin [19] proposed this **ROUGE** measure that stands for Recall-Oriented Understudy for Gisting Evaluation. Recall in the context of ROUGE simply means how much of the reference summary is the system summary. If just individual words are being considered, it can be computed as the ratio of number of overlapping words and number of words in the reference summary. However, a generated summary can be extremely long, capturing all words in the reference summary. But, much of the words in the system summary may be useless, making the summary unnecessarily verbose. This is where precision comes into play. In terms of precision, we find out how much of the system summary was in fact relevant or needed. Precision is measured as the ratio of number of overlapping words and number of words in the generated summary.

Table 5.2: Total documents and number of documents used for key phrase extraction and summarization in different set ups							
Document Type	No of docs	1st Set up		2nd Set up		3rd Set up	
		No of Doc for key phrase	No of Doc for summary	No of Doc for key phrase	No of Doc for summary	No of Doc for key phrase	No of Doc for summary
Business	510	170	340	204	306	255	255
Entertainment	386	128	258	154	232	193	193
Politics	417	139	278	167	250	208	209
Sport	511	170	341	204	307	255	256
Tech	401	133	268	160	241	200	201

$$ROUGE\ Recall = \frac{number\ of\ overlapping\ words}{number\ of\ words\ in\ the\ refernce\ summary} \qquad (5.1)$$

$$ROUGE\ Precision = \frac{number\ of\ overlapping\ words}{number\ of\ words\ in\ the\ generated\ summary} \qquad (5.2)$$

After document type determination, scoring of sentences are done and then summary sentences are selected using weighted summation of scores. **ROUGE 1** refers to overlap of unigrams between the system summary and reference summary. The **ROUGE-1** measures of our experiment is shown in the table 5.3 and then plotted in figures 5.1. For different set ups, **ROUGE 1** measures of each document type is in tables 5.4, 5.5, 5.6 and the information is illustrated in the figures 5.2, 5.3, 5.4.

Table 5.3: **ROUGE 1** measures of the implemented approach in different set ups			
	1st set up	2nd set up	3rd set up
F score	0.6453	0.6435	0.6458
Precision	0.7135	0.7115	0.7129
Recall	0.5939	0.5921	0.5953

Table 5.4: **ROUGE 1** measures of different document types in first set up					
	Document types				
	Business	Entertainment	Politics	Sports	Tech
F score	0.6408	0.6378	0.6362	0.6584	0.6473
Precision	0.7259	0.6986	0.6956	0.7171	0.7196
Recall	0.5790	0.5921	0.5899	0.6135	0.5922

Table 5.5: **ROUGE 1** measures of different document types in 2nd set up					
	Document types				
	Business	Entertainment	Politics	Sports	Tech
F score	0.6419	0.6356	0.6347	0.6515	0.6539
Precision	0.7276	0.6986	0.6961	0.7094	0.7249
Recall	0.5796	0.5883	0.5869	0.6075	0.5990

Table 5.6: **ROUGE 1** measures of different document types in third set up					
	Document types				
	Business	Entertainment	Politics	Sports	Tech
F score	0.6399	0.6411	0.6276	0.6552	0.6551
Precision	0.7274	0.7036	0.6865	0.7098	0.7284
Recall	0.5770	0.5943	0.5815	0.6131	0.5986
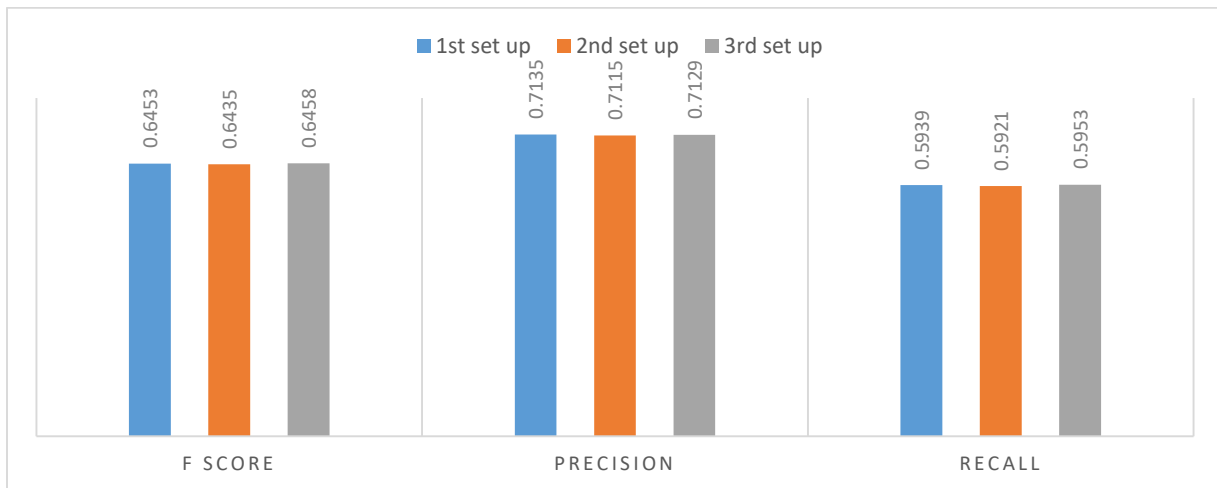
Figure 5.1: **ROUGE 1** measures of implemented approach in different set ups

ROUGE being a comparatively new measure, works really well as it works with n grams. F score, precision, recall of normal methods other than **ROUGE** work with whole sentence, as a result for comparing a generated summary to gold summary is not a good approach cause without being the same exact sentence, two sentences can give same meaning. From our experimental results, it can be said that the results from third set up are relatively better
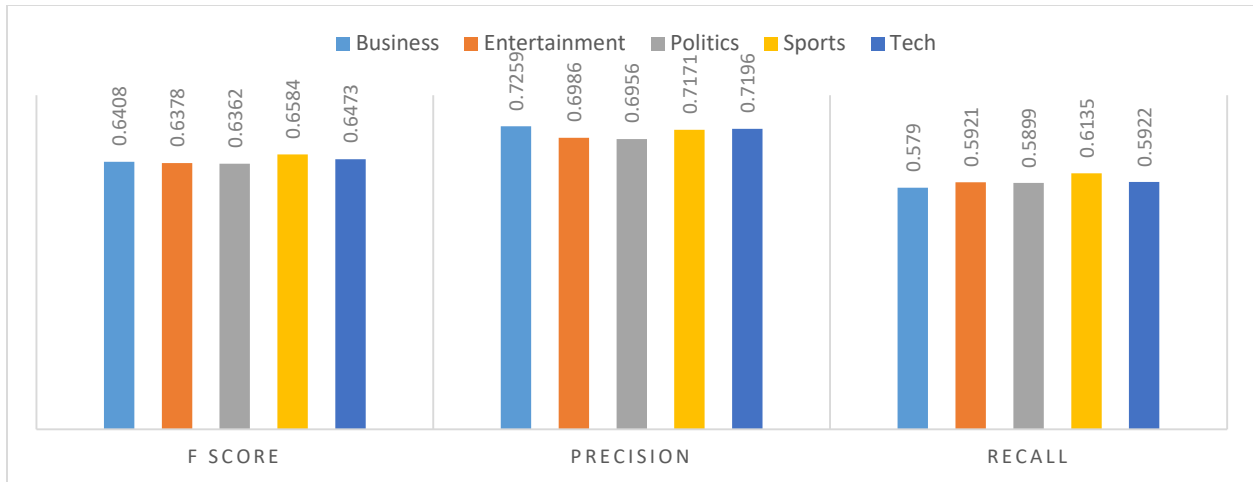
Figure 5.2: **ROUGE 1** measures of different document types in first set up
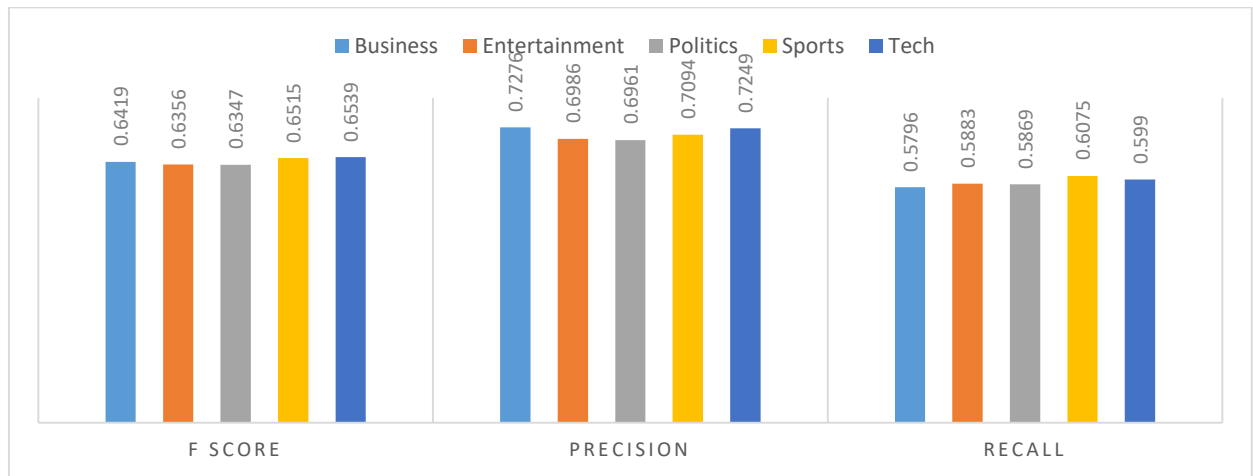
Figure 5.3: **ROUGE 1** measures of different document types in second set up
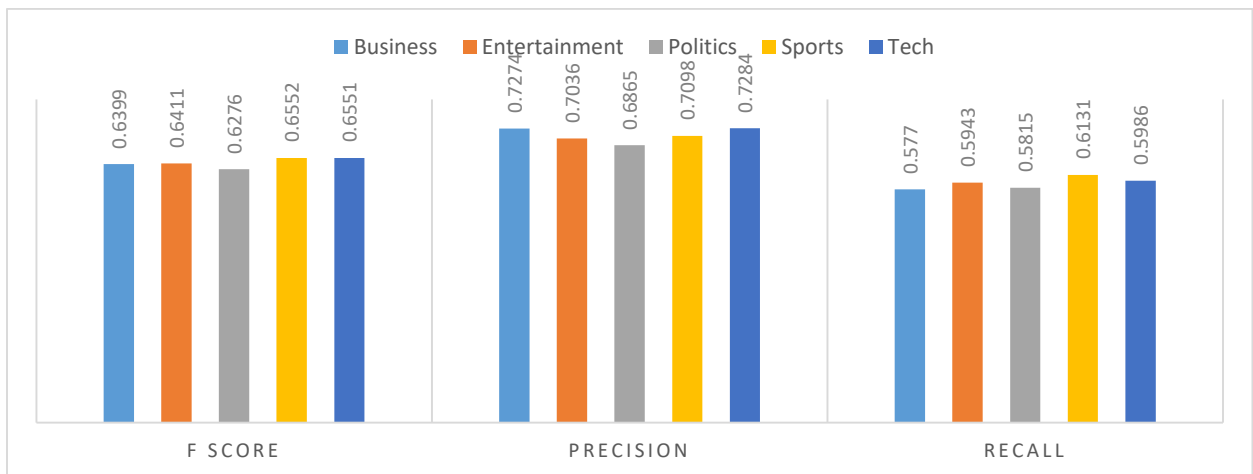
Figure 5.4: **ROUGE 1** measures of different document types in third set up

than others. Also, as we have worked with five different types of documents, it can be noticed that Sports type of documents have given better results than other types of documents.

5.2   Comparison of summary

To know how is our system among the existing ones, we have tried to compare the summaries generated by our implementation with three others. They generate extractive summaries using own approaches. They are as below:

- **Gensim**

Gensim is a robust open source vector space modelling and topic modelling toolkit implemented in Python. Its license is The GNU Lesser General Public License. Its original author is Radim Řehůřek. It is developed by RaRe Technologies. The gensim implementation uses BM25 Ranking function that is based on the probabilistic retrieval framework. For long documents it does not work well. It can be found on the gensim site [20].

- **summarizer 0.0.7**

summarizaer 0.0.7 is a package in the Python Package Index Repository[21] . It is of The MIT License. Its author is Eric Bower. The package works with sentence features. It uses title feature, sentence length, sentence position, keyword features. The feature weights are fine tuned.

- **TextSummarization**

TextSummarization[22] is an online summarization tool. The professional API of it is on Mashape[23] . It uses advanced Natural Language Processing and Machine Learning technologies.

The evaluation results of Gensim, summarizer 0.0.7 and TextSummarization are shown in the table 5.7. For Gensim and summarizer 0.0.7 all 2,225 documents were used for creating summaries but for TextSummarization, three hundred documents were used where each topic has sixty documents. The comparisons of different approaches are plotted in the figures 5.5, 5.6, 5.7.

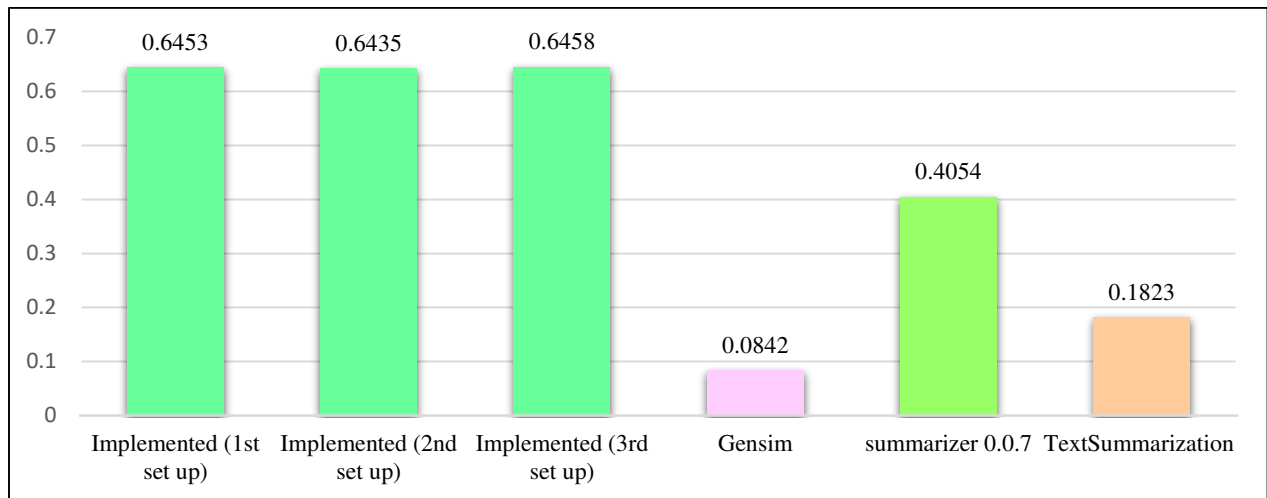| Table 5.7: **ROUGE 1** measures of summaries of Gensim, summarizer 0.0.7 and TextSummarization			
	Gensim	summarizer 0.0.7	TextSummarization
F score	0.0842	0.4054	0.1823
Precision	0.2522	0.6756	0.2864
Recall	0.0525	0.3006	0.1513

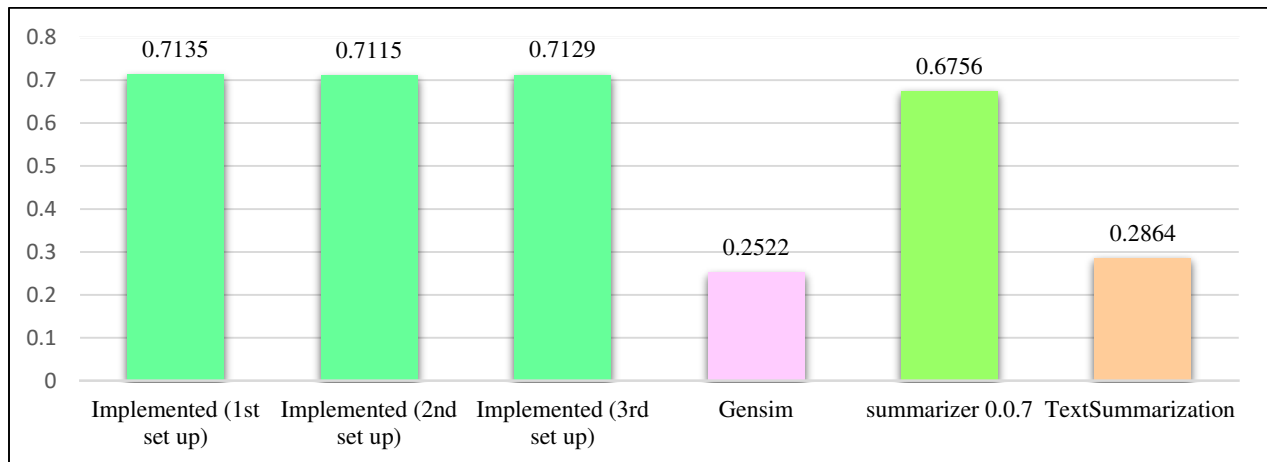Figure 5.5: Comparison of **ROUGE 1** F Score of different approaches

Figure 5.6: Comparison of **ROUGE 1** Precision of different approaches
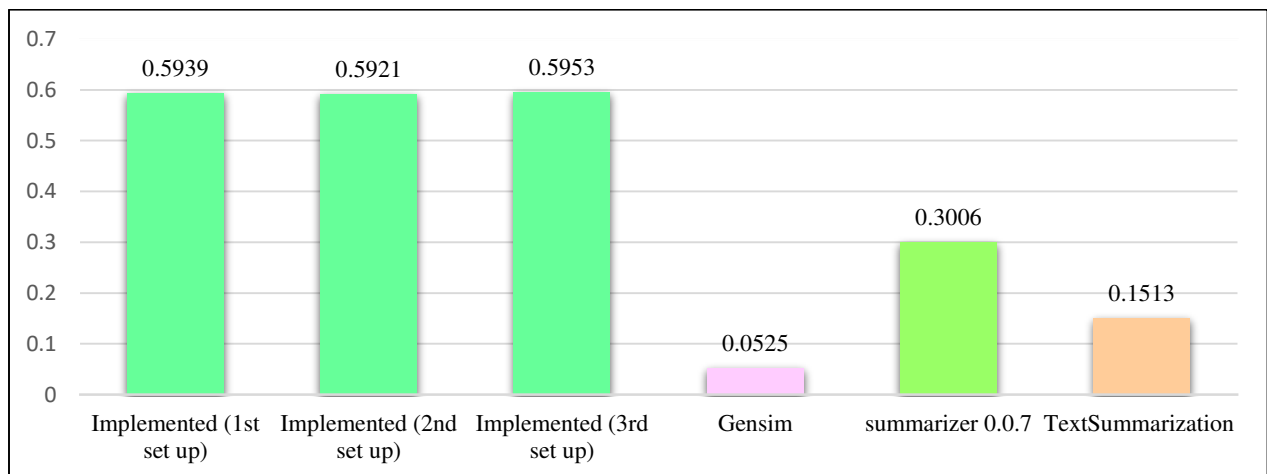

Figure 5.7: Comparison of **ROUGE 1** Recall of different approaches

It can be said that our system works quite good among the compared to ones. Gensim has given the lowest scores. TextSummarization online summarizer has given higher scores than Gensim and summarizer 0.0.7 package has given even higher scores but its scores are not higher than ours.

5.3   Sample of generated summarization outcomes

Table 5.8 : Sample of text summarization results
Document Type: Business **No:** 495

Input:
BBC poll indicates economic gloom

Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening.

Most respondents also said their national economy was getting worse. But when asked about their own family's financial outlook, a majority in 14 countries said they were positive about the future. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster. The poll found that a majority or plurality of people in 13 countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. Those surveyed in three countries were split. In percentage terms, an average of 44% of respondents in each country said the world economy was getting worse, compared to 34% who said it was improving. Similarly, 48% were pessimistic about their national economy, while 41% were optimistic. And 47% saw their family's economic conditions improving, as against 36% who said they were getting worse.

The poll of 22,953 people was conducted by the international polling firm GlobeScan, together with the Program on International Policy Attitudes (Pipa) at the University of Maryland. "While the world economy has picked up from difficult times just a few years ago, people seem to not have fully absorbed this development, though they are personally experiencing its effects," said Pipa director Steven Kull. "People around the world are saying: 'I'm OK, but the world isn't'." There may be a perception that war, terrorism and religious and political divisions are making the world a worse place, even though that has not so far been reflected in global economic performance, says the BBC's Elizabeth Blunt.

The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia. China has seen two decades of blistering economic growth, which has led to wealth creation on a huge scale, says the BBC's Louisa Lim in Beijing. But the results

also may reflect the untrammelled confidence of people who are subject to endless government propaganda about their country's rosy economic future, our correspondent says. South Korea was the most pessimistic, while respondents in Italy and Mexico were also quite gloomy. The BBC's David Willey in Rome says one reason for that result is the changeover from the lira to the euro in 2001, which is widely viewed as the biggest reason why their wages and salaries are worth less than they used to be. The Philippines was among the most upbeat countries on prospects for respondents' families, but one of the most pessimistic about the world economy. Pipa conducted the poll from 15 November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews. The interviews took place between 15 November 2004 and 5 January 2005. The margin of error is between 2.5 and 4 points, depending on the country. In eight of the countries, the sample was limited to major metropolitan areas.

Gold Summary:
Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. The poll found that a majority or plurality of people in 13 countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. The poll of 22,953 people was conducted by the international polling firm GlobeScan, together with the Program on International Policy Attitudes (Pipa) at the University of Maryland. "While the world economy has picked up from difficult times just a few years ago, people seem to not have fully absorbed this development, though they are personally experiencing its effects," said Pipa director Steven Kull. " There may be a perception that war, terrorism and religious and political divisions are making the world a worse place, even though that has not so far been reflected in global economic performance, " says the BBC's Elizabeth Blunt. The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia. China has seen two decades of blistering economic growth, which has led to wealth creation on a huge scale, says the BBC's Louisa Lim in Beijing. The BBC's David Willey in Rome says one reason for that result is the changeover from the lira to the euro in 2001, which is widely viewed as the biggest reason why their wages and salaries are worth less than they used to be. The Philippines was among the most upbeat countries on prospects for respondents' families, but one of the most pessimistic about the world economy. Pipa conducted the poll from 15 November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews.

Experimental Set up 01 outcome:
Number of extracted key phrases in input: 14

Number of extracted key phrases identified
as business type: 5, *as entertainment type*: 4, *as politics type*: 2,
as tech type   : 2, *as sport type*   : 1.
Document identified as: Business type.

Generated Summary:
In percentage terms, an average of 44% of respondents in each country said the world economy was getting worse, compared to 34% who said it was improving. The poll found that a majority or plurality of people in 13 countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. Most respondents also said their national economy was getting worse. The Philippines was among the most upbeat countries on prospects for respondents' families, but one of the most pessimistic about the world economy. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster. The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia. Pipa conducted the poll from 15 November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews. "People around the world are saying: 'I'm OK, but the world isn't'." Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. "While the world economy has picked up from difficult times just a few years ago, people seem to not have fully absorbed this development, though they are personally experiencing its effects," said Pipa director Steven Kull.

Evaluation result:
ROUGE 1 F-score: 0.6789, ROUGE 1 Precision: 0.7708, ROUGE 1 Recall: 0.6066

Experimental Set up 02 outcome:
Number of extracted key phrases in input: 16

Number of extracted key phrases identified
as business type: 6,   *as entertainment type*: 4,   *as politics type*: 3,
as tech type   : 2,   *as sport type*   : 1.
Document identified as : Business type

Generated Summary:
In percentage terms, an average of 44% of respondents in each country said the world economy was getting worse, compared to 34% who said it was improving. The poll found that a majority or plurality of people in 13 countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. The Philippines was among the most upbeat countries on prospects for respondents' families, but one of the most pessimistic about the world economy. Most respondents also said their national economy was getting worse. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster. The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia. Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. Pipa conducted the poll from 15

November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews. "While the world economy has picked up from difficult times just a few years ago, people seem to not have fully absorbed this development, though they are personally experiencing its effects," said Pipa director Steven Kull. "People around the world are saying: 'I'm OK, but the world isn't'."

Evaluation result:
ROUGE 1 F-score: 0.6789, ROUGE 1 Precision: 0.7708, ROUGE 1 Recall: 0.6066

Experimental Set up 03 outcome:
Number of extracted key phrases in input: 18

Number of extracted key phrases identified
as business type: 7,   *as entertainment type*: 5,   *as politics type*: 3,
as tech type   : 2,   *as sport type*   : 1.
Document identified as : Business type

Generated Summary:
In percentage terms, an average of 44% of respondents in each country said the world economy was getting worse, compared to 34% who said it was improving. The poll found that a majority or plurality of people in 13 countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. The Philippines was among the most upbeat countries on prospects for respondents' families, but one of the most pessimistic about the world economy. Most respondents also said their national economy was getting worse. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster. Those surveyed in three countries were split. The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia. Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. Pipa conducted the poll from 15 November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews. "While the world economy has picked up from difficult times just a few years ago, people seem to not have fully absorbed this development, though they are personally experiencing its effects," said Pipa director Steven Kull.

Evaluation result:
ROUGE 1 F-score: 0.6789, ROUGE 1 Precision: 0.7708, ROUGE 1 Recall: 0.6066

Gensim approach:
Generated Summary:
Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. The poll found that a majority or plurality of people in 13

countries believed the economy was going downhill, compared with respondents in nine countries who believed it was improving. In percentage terms, an average of 44% of respondents in each country said the world economy was getting worse, compared to 34% who said it was improving. The countries where people were most optimistic, both for the world and for their own families, were two fast-growing developing economies, China and India, followed by Indonesia.

Evaluation result:
ROUGE 1 F-score: 0.1599, ROUGE 1 Precision: 0.9412, ROUGE 1 Recall: 0.0874

summarizer 0.0.7 approach:
Generated Summary:
Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster. Pipa conducted the poll from 15 November 2004 to 3 January 2005 across 22 countries in face-to-face or telephone interviews.

Evaluation result:
ROUGE 1 F-score: 0.3030, ROUGE 1 Precision: 0.7292, ROUGE 1 Recall: 0.1913

TextSummarization-Online Summarizer:
Generated Summary:
Most respondents also said their national economy was getting worse. Citizens in a majority of nations surveyed in a BBC World Service poll believe the world economy is worsening. But the results also may reflect the untrammelled confidence of people who are subject to endless government propaganda about their country's rosy economic future, our correspondent says. The interviews took place between 15 November 2004 and 5 January 2005. And 47% saw their family's economic conditions improving, as against 36% who said they were getting worse. Almost 23,000 people in 22 countries were questioned for the poll, which was mostly conducted before the Asian tsunami disaster.

Evaluation result:
ROUGE 1 F-score: 0.3197, ROUGE 1 Precision: 0.5, ROUGE 1 Recall: 0.2349

5.4   Chapter Overview

In this chapter, we have discussed about the experimental results of our work and later comparing against some existing systems, explored the quality of our method against others that gives us an insight about the quality of our generated summaries.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

Our summarization work gives good results. After comparing it to other systems, we have also gained satisfactory scores against them. But the implementation of our approach have several difficulties and lacking.

6.1.1 Analysis

We have done text summarization in extractive approach using sentence ranking. We have used some interesting features and have tried to combine them properly. Graph approach with sentiment analysis and key phrase extraction have been used. This approach has given satisfactory results as we have seen it has given better results against other existing systems.

6.1.2 Difficulties

In our system, there are some hitches in creating summaries. Some of them are as below:

- **Similarity function**

In our summarization system, when graph has been constructed, to calculate weight of the edge between two nodes, we had used cosine similarity function for calculating similarity of the sentences of the nodes. But cosine similarity works with frequency of words. It does not take context or meaning into consideration, which may lead to bad summaries as words in different context have different meanings.

- **Limitations of workable input formats**

Our system mainly works with text files that has a heading and a body. Our used dataset is a new summary dataset. So, inputs with other patterns may not give appropriate summary

as data may need to be processed differently.

- **Dependency on other research problems**

In our work, we have used key phrase extraction and sentiment analysis. These two are ongoing research problems themselves. Still works are being done on both of them. Though our idea is commendable but as we depend on them, performance of ours may vary with theirs.

- **Unexplored research problems**

As we have discussed before, there lingers many research issues and problems in the field of text summarization. Issues like multi document, multi lingual, multi format documents need to be addressed as they are of huge significance.

6.2 Future Works

Text summarization is not of bad quality but still there are scopes for work. The future scopes are as below:

- **Multi document summarization**

Our future target is to continue to try to do summarization of multiple documents. Single document summarization may lessen data for analysis work but several single documents might need to be summarized. So, if summarization can be done on various documents at the same time that would be more convenient. Also we will try to tackle the problems that come from various document format types.

- **Testing on more datasets**

We will experiment with more datasets. That will give opportunity to understand the scope of our approach and do improvements.

6.3 Chapter Overview

In this chapter, we have drawn conclusion to our work. We have explored the eventual situation of our work, its difficulties and lacking along with scopes for future improvements.

Appendix A

Examples of Key phrases of different topics

Table Appx A 01 : Samples of key phrases of different topics	
Topic	Samples of key phrases
Business	market activity, market value, global insight, news agency, growth rate, subsidiary brands, public finances, economic growth, mobile phone network, annual gross national income, rising oil prices, break even position , food insecurity, oil supply deal, consumer spending grew, gross domestic product figure, main global growth engine, mobile phone market, average revenues per passenger, international energy agency, sports utility vehicles, local newspapers , long term growth, big positive step, tax rates, news international unit, benchmark interest rate, chief executive, new technology , dominant market position, real estate firm, wall street journal, geographical area, nationwide building society, high oil prices, online service free, future economic growth, labor department, foreign operators, national interest grounds, economic growth rate, income tax cuts, account deficit, flight test system, overseas companies, house price, skilled workers, product prices, average induction time, fixed exchange rate, mortgage loans, company share price, private property rights, new growth opportunity, national security grounds, strong supportive role, internet service providers, mutual fund shares, competitive products, early retirement, extraordinary circumstances, national union, private firm, global network, economic crisis, economic discipline, consumer confidence, energy costs, direct investment, speed broadband, supply management, raw materials, highest rate, international markets, global information, car market, new facilities, financial advisors, stock markets, online advertising, labour market, construction sector, capital goods, good management, domestic economy, profits margins, monthly basis, energy investments, international roaming, early trade, development costs, criminal charges, restructuring plans, transition government, profit growth, total earnings, share value, unemployment statistics, market expectations, strong competition, management control, public transport, economic expansion, overall rate, profitable operations, domestic consumption, exploration portfolio, positive territory, lower profits, economic issues, disease treatment, annual increase, food assistance, exchange commission, remittance charges, growing market, economic

	assessment, public finances, pay rise, market concerns, training centre, transitional year, important year, calendar year, economic policies, criminal penalties, key indicators
Entertainment	record industry, grand jury prize, music industry sector, current concern, super bowl, controversial film shows, cinema ticket, box office glory, academy awards ceremony, million pound industry, novel vanity fair, beach resort, chart top spot, special projects manager, best international album, golden globes ceremony, phonographic industry, worldwide tour, memorable character, touring attraction, new track, best adapted writer, animated comedy, big festive hits, short acoustic set, virgin radio, first novel prize, golden envelopes, film event, sexual acts, public performance, unfortunate events, text messages, multiple nominations, film star, office hit, judge panel, music campaign, social life, motion picture, music fans, new record, young stars, national treasure, production company, new releases, independent film, crime writer, production feature, famous venue, literary talent, pop chart, digital chart, musical act, new chart, several awards, hit film, singles chart, boxing drama, international film, record sales, download sales, record label, wedding date, former head, annual survey, comedy meet, known albums, every week, chart success, many critics, biggest stars, new life, every song, morning show, winning films, special effects, international icon, national theatre, new research, festival ballet, school play, heart surgery, christian community, dance music, film career, good morning, first choice, famous work, latest film, final nominees, top names, film editing, classic film, huge success, award glory, winning performance, commercial success, music piracy, supporting roles, real target, last time, worth showing, famous sale, dance school, music veterans, music scene, surreal film, winning role, official website, strong performance, toughest critics, biography category, biggest opening, recent examples, recent biopic, coming artists, respective categories, original decision, final stages, final decision, great thriller, historical significance, rap culture, outstanding newcomer, worldwide fame, past week, public performance, young age, best parties, good thing, creative achievement, film adaptation, commercial gain, title role, live coverage, wardrobe malfunction, video sales, universal rule, ugly duckling, stolen items, sound mixing, sexual abuse, serious trouble, science, received well, producer, physical sales, perfectly adequate, official source, melancholia looks, medieval history, legal system, hard work, growing list, government grant, government funding, freelance photographer, creative community, catholic community, film biography, personal life, original version, ever seen, later taken, opening night, adult viewers, national body, career achievement, glittering ceremony, state prison, released songs, greatest contributors, developments lead, death threats,

	ever known, surprise winners, made exclusively, mainstream, police praise, pharmaceutical industry, critical success, cottage industry, first appeared, redemption song, promising playwright
Politics	public sector, street policy unit, global warming, human rights act, secondary education system , controversial labour plans, animal welfare groups, union safety effect, free personal care policy, human rights groups, efficient asylum system, law commission chairman, human rights records, national audit office, human rights implications, new poaching laws, human rights grounds, new immigration plans, public sector unions, innocent people, civil service jobs, constitutional crisis, commercial services union, road safety bill, good discussion, public service unions, control orders, level playing field, wrong direction, new race strategy, weekly press conference, id card plans, public order act, general election, government scheme, agricultural policy, criminal offence, law commission report, bad law, racial equality, draft bill, home insulation, economic stability, labour manifesto, common law, young people, political culture, efficient public transport, visa processing system, animal welfare, shadow cabinet, hugely disappointing, particular community , democratic accountability, immediate effect, maternity leave, secretary general, general secretary, electoral commission, audit commission, foreign suspects, immigration controls, trial scheme, local police, last month, education policies, voluntary work, international plans, new bill, public sector, public donations, current scheme, race riots, pension scheme, primary concern, terrorist suspect, immigration problems, health service, environmental groups, terror plans, small minority, immigration status, pension age, high incidence, genuine refugees, diplomatic service, economic migration, opposition members, court action, national statistics, national borders, poll suggests, illegal immigration, current ban, muslim country, proper control, safety whistleblowers, progress check, take responsibility, strong panel, major factor, public meeting, publish figures, political appointees, work force, death penalty, annual target, new voice, genuine concerns, visa application, similar system, new arrangements, public confidence, current rate, election manifesto, commons inquiry, changing legislation, state funding, election stance, party politics, terror attacks, safety inspections, trading day, powerful tool, future policy, wide platform, health conditions, health agenda, multiple parties, democratic society, immediate election, major advance, many areas, equal access, leading expert, major changes, major step, low incomes, retroviral treatment, represented groups, research team, big change, strategic priorities, security agencies, national institutions, moral responsibility, modern government, personal responsibility, weekly payment, annual limits, preceding years, public spotlight, public

	confirmation, specific investigation, minimum access, scheme known, quite clear, busy lives, terrorism powers, personal life, high quality, real risk, powerful figure, good opportunity, policy agenda, think strategically, tax revenues, right time, commons clashes, long term, effective use, special intelligence, transport improvements, recent weeks, huge campaign, first constituency, large crowds, highland council, economic output, economic cycle, whole system, efficient system, next stage, ethnic minorities, important act, legal team, written notices, big turnout, educational needs, annual limit, long way, separate inquiry, industrial history, private meeting, forthcoming referendum, flawed research, intelligence sources, serious concerns, upgrading security, takes decisions, take steps, take advantage, slight increases, significant factor, self defence, rich agenda, quite significant, quite dangerous, prompted censure, prompted accusations
Sport	domestic sporting authorities, former club, top form possible transfer fee, bad knee injury, easier scientific route, consecutive losses, bit nervous, official warning, strong physical display, famous winner, new coach, weight problems, disciplinary committee, illegal approach, title race, deserved winners, unbeaten record, outdoor season, finished third, cup games, veteran keeper, epic battle, new club, best striker, striker played, tough league, final jump, last summer, big stars, new deal, home tie, never scored, football match, club football, national trials, sport performance, impressive year, top three, final sprint, banned substances, disciplinary decision, great year, team honours, new motivation, trial race, young players, winning return, seriously damage, match official, several injury, tremendous strike, drugs charges, qualifying time, decorated competitors, last goal, new evidence, next weekend, drugs test, round victory, winning mentality, difficult match, games history, got lucky, top sprinters, serious injury, next competition, new start, star attractions, difficult side, leading mark, fastest time, teenage stardom, technical committee, put pressure, impressive start, penalty area, home game, strong team, successful return, foot injury, winning form, real struggle, real prospect, difficult year, current decision, young lad, sufficient punishment, every season, loan move, injury problems, ankle injury, premiership game, personal level, first appearance, go head, vital part, better athlete, international friendly, long term, national players, next tournament, first choice, easy decision, tremendous form, hard game, press charges, international perspective, wildly contradictory, substantially reduced, stumbled upon, prohibited substances, overall discipline, opening ceremony, main rival, legal implications, legal advise, last experience, investment package, intensive treatment, huge determination, huge asset, head wide, halted illegally, detailed talk, anybody else, amateur athletics, different set, almost certain, similar

	injury, doping offences, home turf, several chances, difficult time high level, top scorer, everyone thinking, something special time penalty play across, still competitive, recent history, leg injury, game beyond, first leg, winning performances,
Tech	profitable services, internet browsing, high speed downlink, digital video broadcasting, home theatre, deep computing, mobile phone technology, web tracking, voice recognition software, instant messenger, search engine, media ventures, pirated films, video clips , online journals , alternative domains, games developers, radio frequency, software patents, domain names, infected users, open source developers , hidden data, music video show, smart tags feature, desktop search market, personal video recorder, mobile music services, digital screen network, virtual world, video games, fast internet access, computer security firms, portable media player, mobile phone operators, next generation games, wireless computing rules, windows operating system, online security issues, fast net connection, faster connections, law enforcement, broadcast compressed signals, exciting new services, companion robot, real world benefits, intensive games, technical websites report, physical memory cards, great product innovation, customer demand, mobile analyst, video messaging, security firm, subscription fee, download speeds, internet explorer, mobile technology, portable players, fundamental problem, available alternative, radio signals, windows messenger, worldwide trend, wireless technology, broadband speeds, multimedia messaging, spam messages, portable technologies, music software, cultural differences, accessible browser, mobile companies, operating systems, data network, wireless net, alternative browsers, information systems, intelligent card, illegal content, game developers, audio technology, portable computers, internet connections, music industry, hard drive, financial affairs, new technology, security companies, data theft, latest news, cell processor, video streaming, traditional entertainment, internal networks, alternative operator, storage technologies, digital lifestyle, digital imaging, home customers, wireless broadband, web link, telecoms infrastructure, principal analyst, current technology, mobile marketing, commercial advantage, movie player, software developers, consumer groups, technology development, video signal, instant access, spread information, legal services, smart chip, games consoles, surveillance cameras, modest speeds, computer servers, virtual desktop, different format, net browser, storage media, computer programs, software industry, wireless gaming, operating system, browsing population, home entertainment, internet protocol, cultural influence, consumer technologies, virus software, chip technology, mainstream games, report reveals, high quality, mainstream radio, sensitive information,

	next evolution, virtual human, transmission capacities, global community, mobile category, broadband media, information devices, strong competition, internet skills, product management, commercial information, sharing systems, data networks, multiplayer gaming, data transmission, online accounts, data rates, top spot, wireless connectivity, test software, special offers, recent activity, legitimate companies, game release, software vulnerabilities, market share, extra navigation, web pages, peer software, legal action, spam domain, fake domain, research team, key customers, computer experts, mainstream media, complex calculations, surveillance system, powerful servers, online education, personal websites, conventional chips, virtual camera, mobile communications, first campaign, rate files, miles away, large numbers, individual voice, web address, gaming device, important gadget, common use, technology correspondent, spam campaign, server sites, leaked information, theme park, space weather, polling station, political propaganda little confidence, holding page, head office, global operation, fatal force, earth simulator, picture quality, sharing networks, better chips, malicious code, low cost, small businesses, customer service, new version, tech gadgets, discarded phones, address book, single device, security breach, poor security, web experience, big memory, software tools, test versions, simple tool, web surfers, web station, web page, proved controversial, ongoing communication, games grew, freely available, similar model, system upgrade, highly developed, criminal attack, current connection, improved graphics, default settings, current issues, special deals, battery life, future versions, wireless cybercrime, wireless client, greater risks, research director, digital hole, latest tool, huge changes

Appendix B

Source code of Implementation

The source code of our work is has been done available on GitHub [24]. It is in the following link as https://git.io/fp9B0 .

BIBLIOGRAPHY

1. Narendra Andhale, L.A.Bewoor, "An Overview of Text Summarization Techniques", in proceedings of 2016 International Conference on Computing Communication Control and automation (ICCUBEA), pp. 1-7, August, 2016, Pune, India.

2. Eduard Hovy, "Text Summarization", in Ruslan Mitov, editor, The Oxford Handbook of Computational Linguistics (1 ed.), Chapter 32, pp. 583 - 598, 2003, Oxford University Press, Oxford, United Kingdom.

3. S.V.S.S Lakshmi, K.S. Deepthi, C. Suresh, "Text Summarization Basing on Font and Cue-Phrase Feature for a Single Document", in proceedings of the 49th Annual Convention of the Computer Society of India CSI, Volume 2, pp. 537-542, December, 2014, Hyderabad, Telangana, India.

4. Noopur Srivastava and Bineet Kumar Gupta, "An Algorithm for Summarization of Paragraph Upto One Third with the Help of Cue Words Comparison", International Journal of Computer Science and Applications, Volume 5, No. 5, pp. 167-171, 2014, United Kingdom.

5. Md. Iftekharul Alam Efat, Mohammad Ibrahim and Humayun Kayesh , "Automated Bangla Text Summarization by Sentence Scoring and Ranking", in proceedings of 2013 International Conference on Informatics, Electronics and Vision (ICIEV), pp. 1-5, May, 2013, Dhaka, Bangladesh.

6. R.V.V. Murali Krishna and Ch. Satyananda Reddy, "A Sentence Scoring Method for Extractive Text Summarization based on Natural Language Queries", International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, pp. 259-262, May, 2012.

7. Hongyan Jing, "Sentence Reduction for Automatic Text Summarization", in proceedings of the Sixth conference on Applied natural language processing, pp. 310- 315, May, 2000, Seattle, Washington, USA.

8. Joel Larocca Neto, Alex A. Freitas, and Celso A. A. Kaestner, "Automatic Text Summarization Using a Machine Learning Approach", in proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence, pp. 205-215, November, 2002, Porto de Galinhas, Recife, Brazil.

9. Kam – Fai Wong, Mingli Wu and Wenjie Li, "Extractive Summarization Using Supervised and Semi-Supervised Learning", in the proceedings of the

22nd International Conference on Computational Linguistics, pp. 985 – 992, August, 2008, Manchester, United Kingdom.

10. Rada Mihalcea and Paul Tarau, "TextRank: Bringing Order into Texts", in the proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), July, 2004, Barcelona, Spain.

11. Rada Mihalcea, "Graph based ranking algorithm for sentence extraction applied on text summarization", in the proceedings of the ACL 2004 on Interactive poster and demonstration sessions, July, 2004, Stroudsburg, PA, USA.

12. Salil Dabholkar, Yuvraj Patadia and Prajyoti Dsilva, "Automatic Document Summarization using Sentiment Analysis", in Proceedings of 2016 International Conference on Informatics and Analytics (ICIA), August, 2016, Pondicherry, India.

13. Solov'ev A.N., Antonova A.Ju. and Pazel'skaia A.G., "Using Sentiment-Analysis for Text Information Extraction", in proceedings of the 2012 Computational Linguistics and Intelligent Technology, pp. 616-627, August, 2012, Moscow, Russia.

14. Josef Steinberger, Karel Jezek, "Evaluation Measures For Text Summarization", Computing and Informatics, Vol. 28, pp. 1001–1026, March, 2009, Bratislava, Slovakia.

15. Stuart Rose et al., "Automatic Keyword Extraction from Individual Documents", in Jacob Kogan, Editor, Text Mining: Applications and Theory, February, 2010, John Wiley & Sons Ltd., New York, USA.

16. "Kaggle", https://www.kaggle.com, Last accessed date: 8 October 2018.

17. "Insight Resources", http://mlg.ucd.ie , Last accessed date: 8 October 2018.

18. D. Greene and P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", in the proceedings of the 23rd International Conference on Machine Learning, pp. 377-384, June, 2006, Pittsburgh, Pennsylvania, USA.

19. Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of summaries", in proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), January, 2004, Barcelona, Spain.

20. "genism: Topic Modelling for Humans", https://radimrehurek.com/gensim/, Last accessed date: 8 October, 2018.

21. "The Python Package Index", https://pypi.org/, Last accessed date: 8 October, 2018.

22. "TextSummarization", https://textsummarization.net/, Last accessed date: 8 October, 2018.

23. "Mashape- Free API Management Platform & Marketplace", https://market.mashape.com/ , Last accessed date: 8 October, 2018.

24. "GitHub Inc.", "https://github.com/", Last accessed date: December 10, 2018.