HYBRID TEXT SUMMARIZER FOR BANGLA DOCUMENT

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

Submitted by

Asadullahhil Galib	**15.02.04.022**
Mahimul Islam	**15.02.04.047**
Fariha Nuzhat Majumdar	**15.02.04.054**
Najmul Huda Auvy	**13.02.04.081**

Supervised by

Mohammad Moinul Hoque

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

June 2019

ACKNOWLEDGEMENT

Foremost, we would like to express our sincere gratitude to our supervisor, Mohammad Moinul Hoque, Associate Professor, Department of CSE, AUST, for his continuous support on our thesis research and study and for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. We could not have asked for a better supervisor and mentor for our thesis study. Besides our advisor, we would like to thank the faculties of the department for their support and help. We would also like to thank our mentor Shah Siam Ansary Dhrubo for his wonderful co-operation and support. We would like to thank our families for supporting us spiritually throughout our lives. Last but not the least, we are greatly thankful to the Almighty who granted us patience and motivation to complete our work to our hearts content.

Dhaka
June 2019

Asadullahhil Galib

Mahimul Islam

Fariha Nuzhat Majumdar

Najmul Huda Auvy

ABSTRACT

Automatic text summarization is needed to concisely extract a small subset of text portion from a large text where the isolated text may have sentences which are more significant compared to other sentences in the text. Although there have been a lot of approaches on English summarization, very few works have been done on automatic Bengali text summarization. For the evaluation purpose, a data set was formulated from the scratch with Bangla news documents from a reputed newspaper site with gold standard summary text generated by random people.The current work presents a hybrid approach for dealing with summarization process of Bangla text documents. The hybrid model is introduced with a goal to improve the overall accuracy of the summary text generation. The proposed model is expected to generate a summary text based on sentence scoring, sentiment analysis and interconnection of sentences. The progress of the current work is presented with some empirical verification.

Contents

ACKNOWLEDGEMENT i

ABSTRACT ii

List of Figures v

List of Tables vi

1 Introduction 1
- 1.1 Objective 1
- 1.2 What is Text Summarization 1
- 1.3 Types Of Text Summarization 2
  - 1.3.1 Extractive Summarization 2
  - 1.3.2 Abstractive Summarization 3
- 1.4 Methods of Automatic Text Summarization 3
  - 1.4.1 Generic Summarization 4
  - 1.4.2 Topic Centric Summarization 4
- 1.5 Steps in text summarization 4
- 1.6 Overview 5

2 State Of The Art 6
- 2.1 Documentary Research 6
- 2.2 Research Problems 9
- 2.3 Research Gap 11
- 2.4 Overview 11

3 Proposed Model 12
- 3.1 Introduction 12
- 3.2 Proposed Statement 12
  - 3.2.1 Sentiment Scoring 13
  - 3.2.2 Keyword Ranking 13
  - 3.2.3 Text Ranking 13
- 3.3 Why Hybrid 14

3.4 Detailed Analysis of our proposed method . . . 15
3.4.1 Input Document . . . 15
3.4.2 Preprocessing . . . 16
3.4.3 Document Type Separation . . . 16
3.4.4 Sentence Ranking . . . 16
3.4.5 Generating Output Summary . . . 16
3.5 Overview . . . 16

4 Evaluation Measures of Summary **17**
4.1 Introduction . . . 17
4.2 Evaluation Measures . . . 17
4.2.1 Intrinsic Measure . . . 19
4.3 Overview . . . 20

5 Our Progress Until Now **21**
5.1 Introduction . . . 21
5.2 Progress Report . . . 21
5.2.1 Preparing Datasets for training and Testing . . . 21
5.2.2 Splitting Sentences . . . 22
5.2.3 Keyword Extraction . . . 22
5.2.4 Classification of Documents . . . 23
5.2.5 Score Calculation(IDF) of words . . . 24
5.2.6 Individual Scoring for Each Sentence . . . 25
5.2.7 Sentiment Based Ranking . . . 26
5.2.8 Hybrid Scoring . . . 26
5.3 Experimental Results . . . 27
5.3.1 Experimental Verifications with Evaluation . . . 28
5.4 Sample Generated Output . . . 39
5.5 Limitations . . . 42
5.6 Overview . . . 42

6 Conclusion and Future Works **43**
6.1 Conclusion . . . 43
6.1.1 Brief Survey . . . 43
6.1.2 Difficulties Encountered . . . 43
6.2 Future Work . . . 44
6.3 Overview . . . 45

References **46**

List of Figures

3.1 Proposed Design for Hybrid Summarizer for Bangla Document 14
3.2 A flow chart of the proposed model for Hybrid summarizer for Bengali Document 15

4.1 Classification of summary evaluation measures 18

5.1 Average ROUGE 1 Scores of the system for different methods (1st Setup) . . 35
5.2 Average ROUGE 1 Scores of the system for different methods (2nd Setup) . . 36
5.3 Average ROUGE 2 Scores of the system for different methods (1st Setup) . . 37
5.4 Average ROUGE 2 Scores of the system for different methods (2nd Setup) . . 38

List of Tables

2.1 Scores related to Summarization in several researches 9

5.1 Categories of news and number of documents collected from "Daily Prothom Alo" 22
5.2 Categories and number of documents for training and testing 28
5.3 Average ROUGE 1 Score of the system for Category Accidents (1st Setup). . . 28
5.4 Average ROUGE 2 Score of the system for Category Accidents (1st Setup). . . 29
5.5 Average ROUGE 1 Score of the system for Category Economics (1st Setup). . 29
5.6 Average ROUGE 2 Score of the system for Category Economics (1st Setup). . 29
5.7 Average ROUGE 1 Score of the system for Category Entertainment (1st Setup). 30
5.8 Average ROUGE 2 Score of the system for Category Entertainment (1st Setup). 30
5.9 Average ROUGE 1 Score of the system for Category Politics (1st Setup). . . . 30
5.10 Average ROUGE 2 Score of the system for Category Politics (1st Setup). . . . 31
5.11 Average ROUGE 1 Score of the system for Category Accidents (2nd Setup). . 31
5.12 Average ROUGE 2 Score of the system for Category Accidents (2nd Setup). . 32
5.13 Average ROUGE 1 Score of the system for Category Economics (2nd Setup). 32
5.14 Average ROUGE 2 Score of the system for Category Economics (2nd Setup). 33
5.15 Average ROUGE 1 Score of the system for Category Entertainment (2nd Setup). 33
5.16 Average ROUGE 2 Score of the system for Category Entertainment (2nd Setup). 33
5.17 Average ROUGE 1 Score of the system for Category Politics (2nd Setup). . . . 34
5.18 Average ROUGE 2 Score of the system for Category Politics (2nd Setup). . . . 34

List of Algorithms

1 Unique words detection and frequency calculation 23

2 Classification of unknown document 24

3 Individual Sentence Scoring 25

4 Sentiment wise ranking 26

Chapter 1

Introduction

1.1 Objective

Text summarization requires a short, accurate, and fluent summary of a longer text document. From summary important information can be gained which makes the overall procedure more comfortable and less resources are needed in the long run.

Automatic text summarization ideas are needed to address huge amount of text data available online to discover relevant information faster. Although there have been a lot of approaches on English summarization, very few works have been done on automatic Bengali text summarization.

We have tried to explore a new method for the generation of summary from Bengali documents. We had 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. Our summary generation approach is extractive, i.e., the summaries contain exact sentences present in the document.

1.2 What is Text Summarization

Text summarization is the process of shortening a text document, for creating a summary from the original document by extracting key points. Objective of summarization is to find a subset of data that includes the "information" of the entire set. Document summarization is the process that generates a summary or abstract of the entire document, by finding the most informative sentences.

Automatic summarization is the process of shortening a text document, for creating a summary from the original document by extracting key points with software. Automatic Text Summarization is one of the major applications of Natural Language Processing (NLP). The whole summarization process involves huge time consuming efforts that include reading the entire document, and extracting the important ideas from the raw text takes. Automatic summary software summarize texts of hundreds of words in a split second. Thus, users are able to read less data but still receive the most important information and make impactful conclusions.

Instead of reading full news articles that are full of useless information - the summaries of such web pages can be precise and accurate - but still 20 Percent to 40 Percent the size of the original article.

Machine generated summaries are free from bias. Some software have the unique ability to declare a word whose sentences that include it will automatically appear at the summary. While humans can sometimes miss out an important sentence, but it is impossible for the computers to miss it and thus important sentences will always be mentioned.

With growing digital media and ever growing publishing, who has no time to go through entire articles / documents / books, this technique saves their reading time as they do not have to read huge amount of useless and redundant data.

1.3 Types Of Text Summarization

There are generally two categories of text summarization. They are:

- Extractive Summarization
- Abstractive Summarization

1.3.1 Extractive Summarization

In extraction based summarization, sentences with the most important points that are relevant to the topics are extracted from a piece of text and combined to make a summary. In extractive summarization process, the automatic system pulls out objects from the entire datasets, without modifying the objects. This include key phrase extraction, where the goal is to select individual words or phrases to "tag" a document, and in document summarization, where the objective is to select the sentences without modifying them to create a short summary. Different types of algorithms and methods can be used to gauge the weights of the sentences and then rank them according to their relevance and similarity with one another-and further joining them to generate a summary. In extraction based summarization, scores are provided to sentences using some methods and the sentences that achieve highest scores are included in the summaries. This kind of summary extracts the sentences that contain essential information but it is not always smooth or fluent. Sometimes in the generated summary, there may be lacking of connectivity between adjacent sentences. Our summarizer only generates extractive summaries.

1.3.2 Abstractive Summarization

In abstraction based summarization, advanced deep learning techniques are applied to paraphrase and shorten the original document, just like humans do. Here in the abstraction based summary, there can be sentences or words and phrases that are not present in the original document. Abstractive summarization systems generate new phrases, possibly rephrasing or using words that are not in the original text. Naturally abstractive approaches are harder. In order to generate a perfect abstractive summary, it is important to understand the document first and then express that understanding in short using new sentences and words. Much harder than extractive. Has complex capabilities like generalization, paraphrasing and incorporating real-world knowledge.

Abstraction involves paraphrasing sections of the source document. Abstractive summary is important because as the textual structure present in the summary can vary significantly from the original while some work has been done in abstractive summarization, the majority of summarization systems are extractive.

Structure based approach 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 emphasizes on 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 Methods of Automatic Text Summarization

There are generally two methods of automatic text summarization:

- Generic Summarization

- Topic Centric Summarization

1.4.1 Generic Summarization

Generic summary provides the overall sense of document. It typically contains core information of a document. A document normally consists of sentences that give information to users about a single broad or narrow matter but understanding the matter requires information from other domains as well.

1.4.2 Topic Centric Summarization

The topic centric summaries are generally restricted to one topic. Topic is typically based on the key words that are supplied by human. Summary can be created from single or multiple documents. This kind of summary are going to be impactful in future search engines.

1.5 Steps in text summarization

According to Kamal Sarker [1] text summarization involves preprocessing, stemming, sentence ranking, summary generation. The preprocessing step includes stop-word removal, stemming and breaking the input document in to a collection of sentences.

Using stemming, a word is split into its stem and affix. The design of a stemmer is language specific, and requires some significant linguistic expertise in the language. A typical simple stemmer algorithm involves removing suffixes using a list of frequent suffixes, while a more complex one would use morphological knowledge to derive a stem from the words. Since Bengali is a highly inflectional language, stemming is necessary while computing frequency of a term.

After an input document is formatted and stemmed, the document is broken into a collection of sentences and the sentences are ranked based on two important features: thematic term and position.

The thematic terms are the terms which are related to the main theme of a document. We define the thematic terms are the terms whose TFIDF values are greater than a predefined threshold. The TFIDF value of a term is measured by the product of TF and IDF, where TF (term frequency) is the number of times a word occurs in a document and IDF is Inverse Document Frequency.

The positional score of a sentence is computed in such a way that the first sentence of a document gets the highest score and the last sentence gets the lowest score.

Md. Nizam Uddin and Shakil Akter Khan [2] described extraction based methods for summarizing Bengali documents.

Different methods are used that has been found in the survey to rank the sentences. The methods are as following:
Location method: sentences under headings have higher score. Sentences near beginning or end of document and/or paragraphs have higher score.

Cue method: The Cue method is based on the hypothesis that the probable relevance of a sentence is affected by the presence of pragmatic words

Title: Words in title and in following sentences gives high score.

Term frequency: Words or terms which are frequent in the text are more important than the less frequent. Open class terms are words that change over time.

Numerical data: Sentences containing numerical data is scored higher than the ones without numerical values.

Implementation: Based on the methods mentioned above the Bangla summarizer is implemented. For the implementation the summary size is 40 percent of the actual content.

1.6 Overview

We have gathered a thorough idea about text summarization after discussing different terms related to summarization from this chapter. Text summarization extracts important information from a document and ignores irrelevant information. So there is significant importance of summarization and we have discussed its importance in this chapter. Text summarization is of different types: extractive and abstractive. Different methods can be applied to summarize a document: generic and topic centric. Text or document does through several stages and then a proper summary is generated.

Chapter 2

State Of The Art

2.1 Documentary Research

In today's fast emerging world text summarization is one of the most required tools for understanding significant information of a document. Sometimes due to large amount of data and various amount of sources it becomes really tough to understand the gist of documents. Selecting information from very large data is difficult for human beings. To manually summarize information available on the internet is really challenging, complicated and difficult task. There have been some researches on automatic Bengali text summarization. We have studied some papers regarding this topic.
Kamal Sarkar [1] [3] has discussed text summarization for single document of Bengali language signifying the impact of thematic term feature and position feature of sentences. In linguistics, thematic feature means to relate to the theme of a writing. In the experiment, LEAD baseline has been used for comparison of produced summary. LEAD has been defined in DUC 2001 and DUC 2002 conferences. The discussed work mainly is of three phases. The phases are preprocessing, sentence ranking and summary generation. Stop words removal, stemming and tokenization has been done as preprocessing. Sentence ranking has been done with thematic terms and sentence position. Thematic terms have been defined as the terms with higher TFIDF value than predefined values. As for position of sentences, first to last sentences have been scored as higher to lower. Too short and too long sentences has not been preferred as summary sentences. After combination of scores, sentences have been ranked and order of sentences has been maintained as input document. It has been ensured in the work that the length of produced summary and reference summary remain same. Also in terms of recall, precision and f-score, the proposed system as given better results. It has been argued that the proposed model has outperformed the compared systems. The average unigram based Recall score: 0.4122. The score for LEAD baseline: 0.3991.

Md. Nizam Uddin and Shakil Akter Khan [2] experimented with Bengali text summarization taking into consideration several criteria. They have put significance on sentence location, cue phrase presence, title word presence, term frequency and numerical data. They have argued that sentences which appear in the first or last of a passage, are of more importance. They have said that presence of cue phrases, words from title, words with high frequency and numerical data also put importance on a text. They have tested their approach using documents from the newspaper the "Daily Prothom Alo". They have considered two facts for evaluating a summary and that are the information in a summary and its size. They have concluded that the summary should be of forty percent of the main document. Average accuracy is 71.3 Percent.

Amitava Das and Sivaji Bandyopadhyay [4] has summarized Bengali documents using sentiment information. They have tried to identify the sentiment information in a document and they aggregate that for generating summary. They have used a classifier based on support vector machine. Three kinds to feature have been considered which are lexico-syntactic, syntactic and of discourse level. Parts of speech, SentiWordNet, frequency, stemming, chunk label, dependency parsing depth, title of document, first paragraph, term distribution, collocation have been used as features in the work. After theme detection, theme clustering has been done and with theme relational graph using PageRank algorithm,
summary sentences have been identified. For corpus, documents from Bengali newspaper "AnandabazarPatrika" have been used. Annotation of sentence level subjectivity and discourse level theme words has been done manually from portion of the corpus. It has been said that the summarization system has achieved precision: 72.15 percent, recall: 67.32 percent, f-score: 69.65 percent

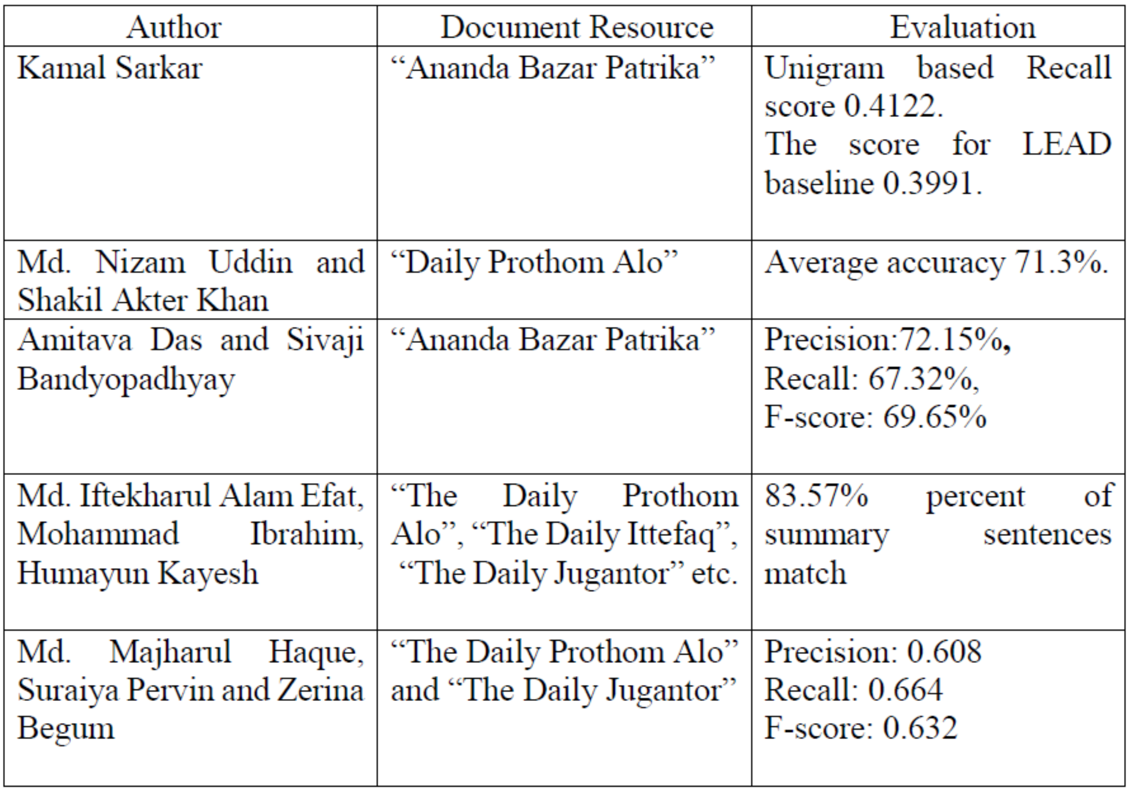
Md. Iftekharul Alam Efat, Mohammad Ibrahim, Humayun Kayesh [5] have discussed Bengali summarization taking into consideration of several attributes. They have calculated sentences' scores based on aspects such as frequency, sentence position, cue phrases' presence and skeleton of document. Before scoring of sentences, pre-processing such as tokenization, stop words removal, stemming have been done. After calculating scores based on various aspects, final sentences' scores have been calculated as weighted summation of the scores of individual features. For experiments, dataset has been made using articles from various newspapers like "The Daily Prothom Alo", "The Daily Ittefaq", "The Daily Jugantor" etc. They have argued that 83.57 percent percent of summary sentences match to human summaries.

Haque, Pervin and Begum [6] have done text summarization with Bengali documents using sentence ranking and clustering. The summarization has been done through several phases. In the preprocessing phase, stop words removal and stemming have been done. Then sentences have been ranked with term frequency calculation for each sentence and sentence frequency. Here, sentence frequency has been introduced for removing redundant sentences. If overlap ratio of two sentences have been shown over or equal to sixty percent, then the smaller sentence falls out of consideration and importance of larger sentence increases. Sentences have been clustered using cosine similarity to group similar sentences. Then summary has been generated by selecting sentences from clusters based on volumes of clusters. To build the corpus for experiment, Bengali newspapers "The Daily Prothom Alo" and "The Daily Jugantor" have been used. After evaluation, precision, recall and f-score values have been calculated as 0.608, 0.664 and 0.632 respectively.

Haque, Pervin and Begum [7] have discussed Bangla summarization using key phrases. They have argued that key phrases with three or four terms are better for summarization tasks. They have discussed that from first to last, sentences' scores should decrease from more to less and sentences with numerical figures should be given importance. After combining the scores, sentences have been ranked. Dataset has been made with four hundred newspaper documents that are of wide varieties. Using ROUGE-1 and ROUGE-2, they said that summaries' quality has improved.

Akter et al. [8] proposed an extractive summarization using K-means clustering algorithm. They have used clustering to tackle both single and multiple document summarization issues. At first some pre-processing such as noise words and stop words removal, tokenization and stemming are done. Later, using TFIDF word scores have been calculated. Afterword, sentences have been scored with word scores, position and cue or skeleton words presence. Later, the ranked sentences have been saved in a file. For multi document summarization, all sentences from different inputs have been saved in a file and after merging, sort has been done in descending order. Then K-means clustering has been done taking the score of greatest and lowest score of sentences. After that summary sentences have been picked as the top thirty percent of the input document(s).
Paul et al. [9] have discussed about summarization process with sentence ranking based on clustering. For this, at first noise words, stop words have been removed and stemming has been done. Then, clustering of sentences has been done. Term frequency matrix has been used for summarization. Sentences have been given scores based on four different aspects. Based on term frequency, term frequency and clustering, TFIDF, TFIDF and clustering the four methods have been implemented. For experiments, corpus has been made from newspaper articles. Using ROUGE-N, it has been argued that better results has been gained with term frequencies.

Table 2.1: Scores related to Summarization in several researches



Author	Document Resource	Evaluation
Kamal Sarkar	"Ananda Bazar Patrika"	Unigram based Recall score 0.4122. The score for LEAD baseline 0.3991.
Md. Nizam Uddin and Shakil Akter Khan	"Daily Prothom Alo"	Average accuracy 71.3%.
Amitava Das and Sivaji Bandyopadhyay	"Ananda Bazar Patrika"	Precision:72.15%, Recall: 67.32%, F-score: 69.65%
Md. Iftekharul Alam Efat, Mohammad Ibrahim, Humayun Kayesh	"The Daily Prothom Alo", "The Daily Ittefaq", "The Daily Jugantor" etc.	83.57% percent of summary sentences match
Md. Majharul Haque, Suraiya Pervin and Zerina Begum	"The Daily Prothom Alo" and "The Daily Jugantor"	Precision: 0.608 Recall: 0.664 F-score: 0.632

2.2 Research Problems

According to Kamal Sarkar [1] [3] the performance of the proposed system depends on preprocessing, stemming, sentence ranking. It may be improved by improving stemming process, exploring more features and applying learning algorithm for effective feature combination. From the table shown in the previous section, we clearly see that there is an opportunity to improve the unigram based recall score and LEAD baseline score in the summarizer they have proposed.
According to Md. Nizam Uddin and Shakil Akter Khan [2] the main limitation of the Bangla summarizer is it just extracts some sentences from the given text which is much more different from the human generated summary. Another limitation is sometime the sentences which come early in the text have higher possibility to appear in the summary.

Amitava Das and Sivaji Bandyopadhyay [4] stated that, the evaluation result of their present summarization system for Bengali documents is reasonably good but still not outstanding. The main limitation of their system occurs for subjectivity identifier. The recall value of the classifier is higher than it which order to present them so that the whole text makes sense for the user. They prefer the original order of sentences as they occurred in original document.

Our aim is to generate summaries from given Bengali document(s) that are good and close to human created ones. To gain better scores in evaluation matrices we would explore in which way we should proceed to make better summaries.

The Problems with extractive text summarization [10] [11] [12] are:

- Sentences selected for summary generally longer, so unnecessary parts of the sentences for summary also get included they consume space.
- If summary size is not long enough, the important information scattered in various statements cannot captured using extractive summarization.
- Information which is clashing may not be presented accurately.
- Sentences frequently contain pronouns. They lose their referents when used out of context. If irrelevant sentences are clubbed together, may lead to confusing understanding of anaphors which will result in erroneous representation of original information.
- The same problem is with multi document summarization, because extraction of text is performed on different sources. Post processing can be used to deal with these troubles, for example, replacing pronouns with their background, replacing relative temporal expression with actual dates etc.

Our goal is to work on the above problems. Besides, feature selection and combination is also one important factor. All of the features cannot be used in a good summary as some features may worsen the quality of the summary. Again combining two or more features according to hierarchy of significance results in good summarization. So, we focused on this factor.

We have proposed a model to generate summaries so that the features are selected in a way that the scores such as precision, recall, f-measure are improved and the generated summaries match with the gold summaries which are human created.

2.3 Research Gap

Research gap refers to research problems which have not been answered appropriately or at all in a given field of study. Research gap makes others' research publishable. Because it shows that existing research is not being duplicated and it shows that research has been conducted in order to fulfill that gap in the literature.

We have discussed about various types of research problems in the previous section. In our perspective, combining different methods can reduce the problems. So, we are working on applying different methods of scoring in our summarizer so that it can overcome the problems and generate better summaries.

2.4 Overview

We have discussed about the state of the art related to text summarization for Bengali document. In this chapter we came to know different literature reviews and the problems and issues that have been faced regarding text summarization. More focus is given to extractive approaches of text summarization.

Chapter 3

Proposed Model

3.1 Introduction

We are devising a summarizer that generates extractive summary of Bengali documents. We give score to the sentences based on some criteria to select the best ones that represent the gist of a given document.

3.2 Proposed Statement

We are combining 3 approaches to determine the ranking of the sentences of the given document:

- Sentiment Ranking
- Keyword Ranking
- Text Ranking

We will add weight to each of them and the final ranking will be based on the combination of the three.
Say,
SS*W1 + KR*W2 + TR*W3 = Final Score of a sentence
W1 = A percentage of the total sentiment score
W2 = A percentage of the total keyword score
W3 = A percentage of the total text ranked score

3.2.1 Sentiment Scoring

"PolyGlot" is a natural language pipeline that supports massive multilingual applications. It supports Sentiment Analysis for 136 Languages. Unique Words Polarity can be calculated, and then sentences can have a scoring based on the neutral words (Words with polarity score 0) present in each sentence. More neutral words mean more score.

3.2.2 Keyword Ranking

Keywords are extracted based on each category. Then according to their frequency, score can be given. More appearances in each category means higher score. After analyzing at the datasets it can be determined that key phrases don't put extra value on top of keywords. As there are minimal number of key phrases in our datasets.

3.2.3 Text Ranking

TextRank, is a graph-based ranking model for text processing which can be used in order to find the most relevant sentences in text and also to find keyword. Sentences are like nodes connected with each other based on their matching words. For example in cosine similarity, two sentences are plotted in a graph . The angle between them depends on their cosine similarity. The more cosine similarity will mean the lesser angle between them.

We plan to use "word to vec" as it seems to be really important in finding correlated sentences. The angle that determines the correlation perhaps gets closer than any other method that is used in the case ,e.g: cosine similarity.

3.3 Why Hybrid

Our aim is to try to combine three approaches to generate output summary. Our proposed design is a combination of three approaches: Sentiment Scoring, Phrase Ranking and Text Ranking. We can consider our summarizer for Bengali document a hybrid one.

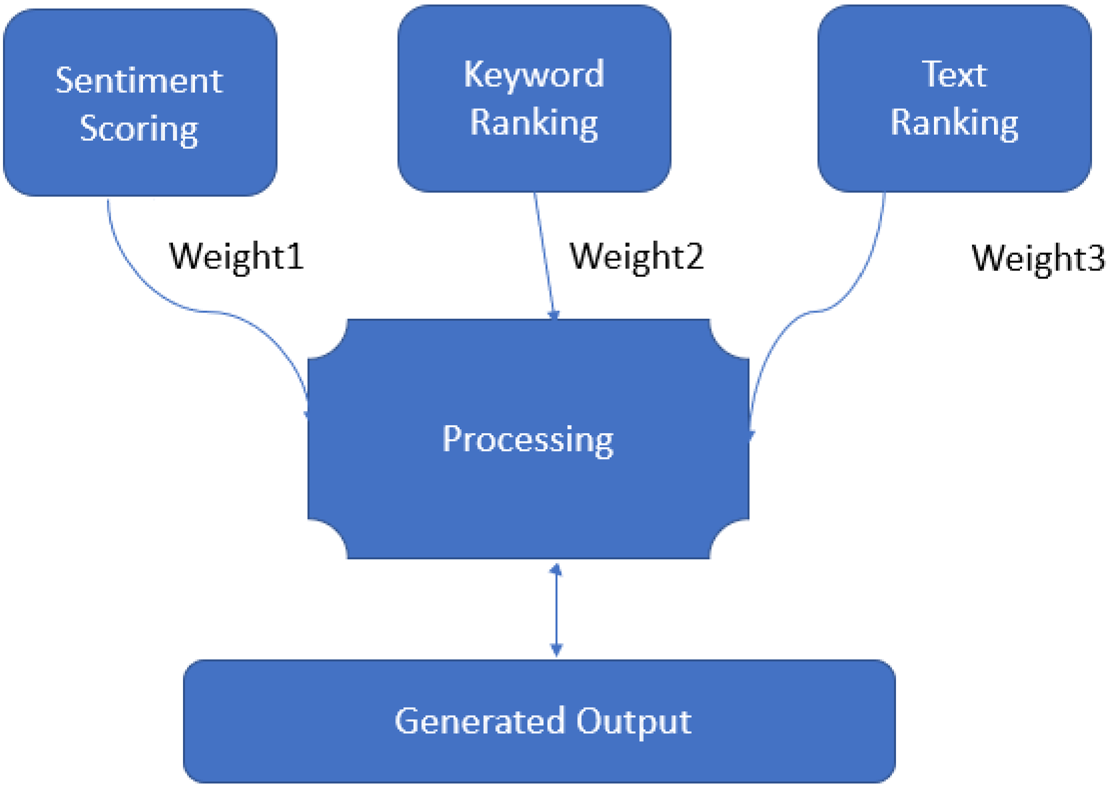


Figure 3.1: Proposed Design for Hybrid Summarizer for Bangla Document

We can consider our summarizer for Bengali document a hybrid one.

3.4 Detailed Analysis of our proposed method

We will have a thorough idea about the steps we will be following in our devised summarizer.

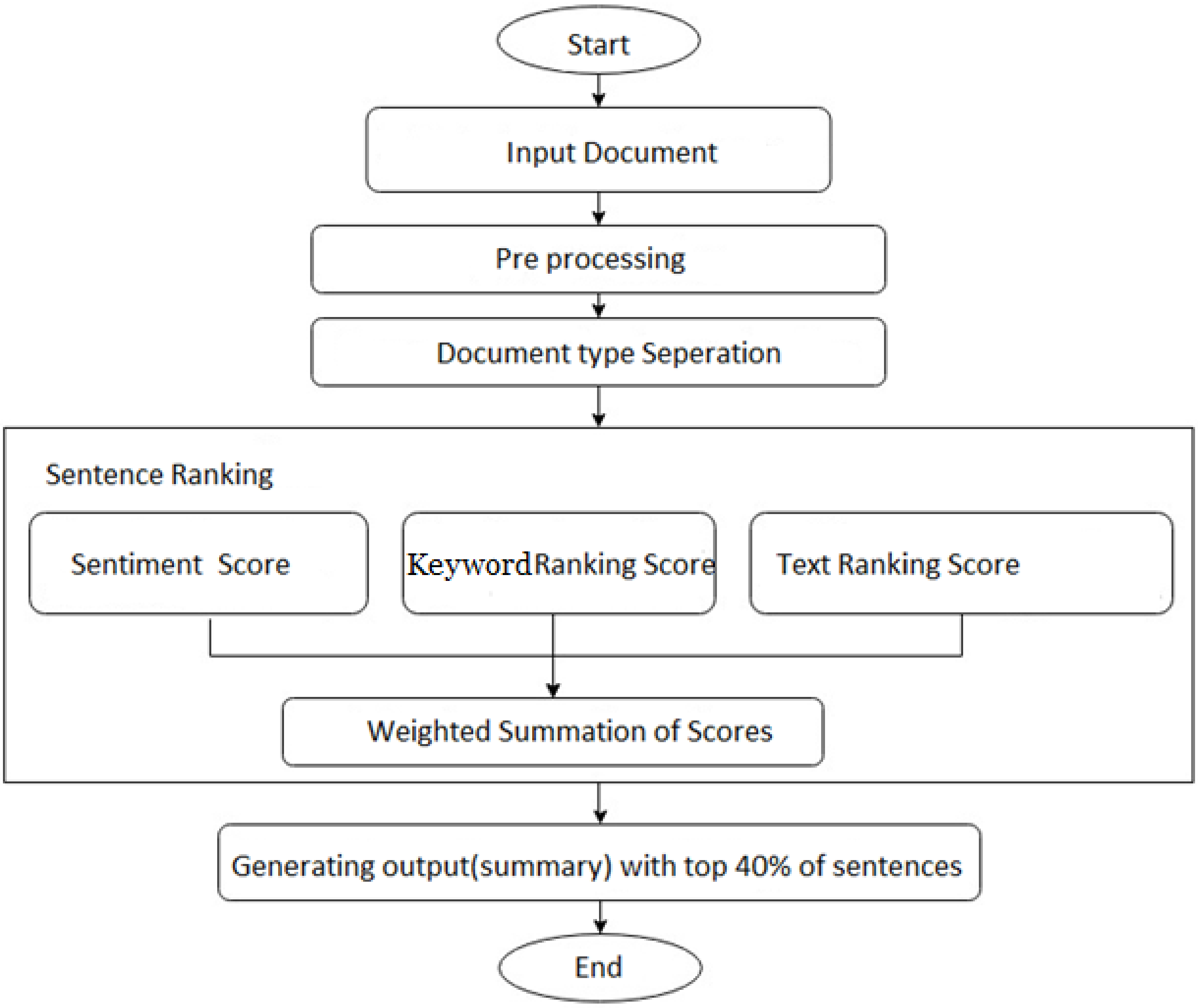


Figure 3.2: A flow chart of the proposed model for Hybrid summarizer for Bengali Document

3.4.1 Input Document

We can use any Bengali news document as input of the summarizer. But the document has to be of a specific category. For starters, it can be any news from our datasets, that are reserved for testing or training purpose.

3.4.2 Preprocessing

Preprocessing involves any type of processing performed on raw data to prepare it for another processing procedure. We are performing preprocessing by splitting the input documents into words. We have used stopword removal to remove irrelevant sentences that won't be added to the generated summaries.

3.4.3 Document Type Separation

Type separation involves separating documents in different categories. Before processing, the input document will be categorized as a specific type. We are working on summarizing four (4) categories of news: Politics, Economics, Entertainment and Accidents. So, the input document will be categorized in any of the stated document types.

3.4.4 Sentence Ranking

Sentence ranking involves ranking the sentences of a document based on different scoring techniques. We have stated that we are working on three (3) methods: sentiment scoring, phrase ranking and text ranking. To get the highest ranked sentences, we are mainly trying to combine these three approaches. We add weights to the different approaches and by adding the weighted scores we get a weighted summation of scores.

3.4.5 Generating Output Summary

After sentences are ranked, they are sorted in descending order. Highest ranked sentences get the highest probability of being selected. Top 40 Percent of the highest ranked sentences are selected to generate summary.

3.5 Overview

In this chapter, we have discussed about the proposed methods we have intended to use in our summarizer. We have also given solid reasons why we named it a Hybrid summarizer. We have mentioned the steps we are following to generate summaries. In the next chapter we will have a thorough idea about our progress until now.

Chapter 4

Evaluation Measures of Summary

4.1 Introduction

Automatic summarization takes an information source, extract content from it, and users can get the gist content of a document. As automatic summarization is out of bias and not human created, it may lack the importance and expression of human mind and sometimes it cannot express the main idea of the document properly. So, it is important to understand what makes a summary a good summary. To identify good summaries, it is important to evaluate summaries using several measures. So, evaluation of summary is an important task.

4.2 Evaluation Measures

There have been many researches on summary evaluation and many measures have been invented. Josef Steinberger, Karel JeËĞzek [13] stated that, intrinsic content evaluation is the main approach for determining summary quality that is often done by comparison with an ideal or reference or gold summary. For sentence extractions, it is measured by co-selection. It detects the number of ideal sentences in the automatic summary. Content-based measures compare the words that are present in a sentence. Their advantage is that they can compare both human and automatic summaries and can extract with human abstracts that contain newly written sentences. Another major group are extrinsic methods. They are also known as task-based methods. They measure the performance of using the summaries for a certain task. So basically, there are two types of measures:

1.Intrinsic Measures
2.Extrinsic Measures

Saiyed Saziyabegum, Dr. Priti S. Sajja [10] have defined intrinsic and extrinsic measures. Intrinsic measures check the summarization system itself. It determines the quality of summary by comparing the automatically generated summary and the human made summary. Summary is evaluated regarding two aspects: Quality or informativeness. The informativeness of a summary is evaluated by comparing it with a human-made summary.
Extrinsic measures check the summarization on how it can influence the completion of some other task such as text classification, information retrieval, answering of question etc. It evaluates the quality of summarization using different tasks like reading comprehension, relevance assessment etc.
Our focus is on mainly Intrinsic Measure.

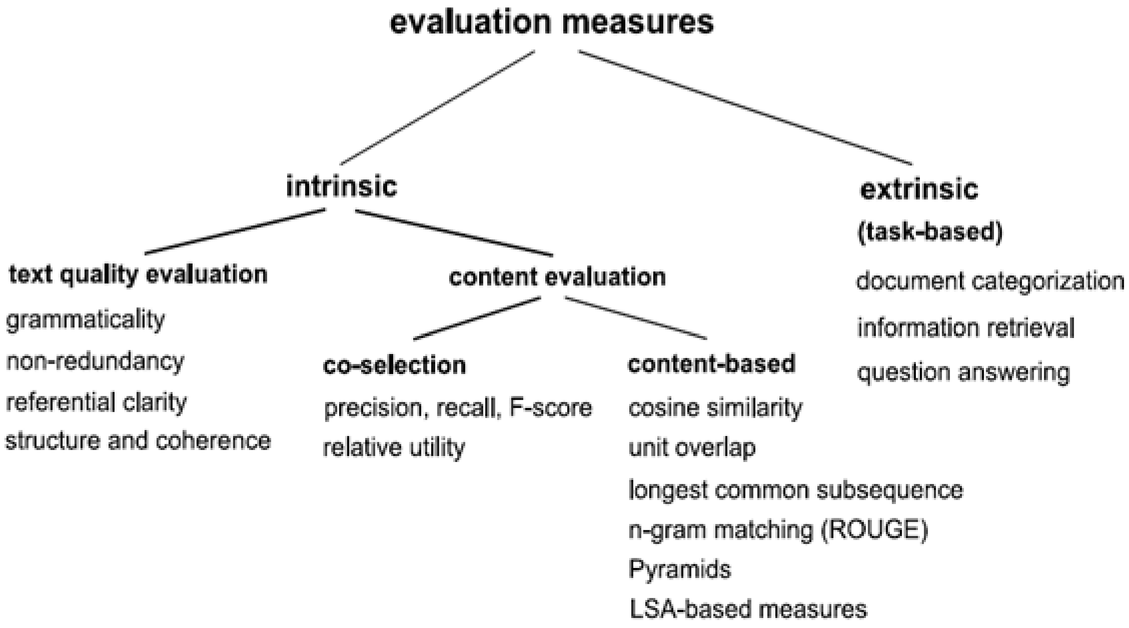


Figure 4.1: Classification of summary evaluation measures

4.2.1 Intrinsic Measure

From the figure 4.1, we see that intrinsic measures can be classified into two another classes: 1. Text Quality Evaluation 2. Content Evaluation. We will have a detailed idea about different types of intrinsic measures from this section.

4.2.1.1 Text Quality Evaluation

Under this class we see text quality can be evaluated in different ways. Such as,
Grammaticality : The text contains textual items and punctuation errors or incorrect words should not be contained in the text.
Non-redundancy :The text should not contain unnecessary information.
Reference clarity : The nouns and pronouns of sentences should be clearly referred to in the summary e.g., the pronoun 'he/ she' has to mean somebody in the context of the summary.
Coherence and structure : The summary should have good structure and the sentences need to be coherent. This cannot be done automatically. Marks are assigned to each summary.

4.2.1.2 Content Evaluation

Under this class we see that contents can be evaluated by co-selection and based on the contents. **Co-selection** involves precision, recall, f-score and relative utility.

Precision :Precision is defined as the number of sentences present in both system generated and gold summaries divided by the number of sentences in the system generated summary.
Recall : Recall is defined as the number of sentences occurring in both system generated and gold summaries divided by the number of sentences in the gold summary.
F-score : F-score is a measure that combines precision and recall. The basic way to compute the F-score is to calculate a harmonic average of precision and recall. The formula is given below:
F = 2 * P * R / (P+R) *Content based evaluation* involves cosine similarity, ROUGE etc. measures. **Cosine Similarity**: Cosine similarity is a commonly used approach to match similar documents. It refers to counting the maximum number of common words between the documents. It is basically a metric that is used to measure how similar the documents are in respect to their size. In terms of mathematics, this method measures the cosine of the angle between two vectors that are projected in a multi-dimensional axis.

ROUGE : It is the abbreviation of Recall-Oriented Understudy for Gisting Evaluation. It refers to a set of metrics and a software package. It is mainly used for evaluating summarization. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. It measures the quality of summary by counting overlapping units such as the n-gram, sequences of words i.e., which word is appearing after which and word pairs between the generated and reference summary.
ROUGE-N: It is defined as the overlapping of N-grams between the system's generated and reference summaries.
* ROUGE:1 determines the overlap of 1-gram (each word) between the system generated and reference summaries. In this measure, summary is evaluated comparing the precision, recall and f-measure scores for 1-gram.
* ROUGE:2 determines the overlap of bigrams between the system generated and reference summaries. In this measure, summary is evaluated comparing the precision, recall and f-measure scores for bigrams.

4.3 Overview

We discussed about the different measures to evaluate whether a summary is good or bad. We broadly explained the main classifications evaluation measure: intrinsic and extrinsic and their different types.

Chapter 5

Our Progress Until Now

5.1 Introduction

We are discussing our progress to inform our supervisor and associates about the progress we've made on our summarizer up until now in this chapter. We are reassuring our recipients that we are making progress, and we hope that it will be complete by the expected date. We will provide our recipients with a brief look at the findings and work of the thesis.

5.2 Progress Report

In this section we will try to give a thorough idea about what tasks we have done so far. Relevant discussions, calculations and algorithms will be discussed here.

5.2.1 Preparing Datasets for training and Testing

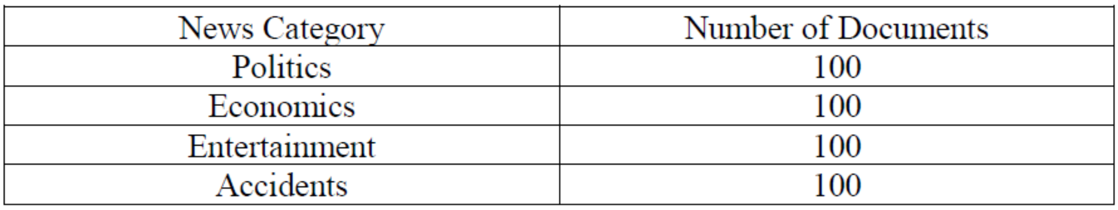
There is No Dataset Available for our Topic. We researched on different domains and created our own datasets from "Daily Prothom Alo" newspaper. We have collected about 400 news documents from 4 different categories: Accident, Entertainment, Economics and Politics. From each category we collected 100 news. Extractive summaries have been generated by random people and are considered as gold summaries. Each summary is around 40 percent of the original document. We have used 2 setups for our research.

In the first setup, Random 50 percent of the actual news are used for training the summarizer. Because training summarizer with 100 percent of the document results in overfitting. The rest 50 percent are used for testing purpose, meaning generating the summaries.

In the second setup, Random 30 percent of the actual news are used for training the sum-

marizer. The rest 70 percent are used for testing purpose.
The summaries generated by the summarizer are then compared with the gold summaries and generating rouge scores.

Table 5.1: Categories of news and number of documents collected from "Daily Prothom Alo"



News Category	Number of Documents
Politics	100
Economics	100
Entertainment	100
Accidents	100

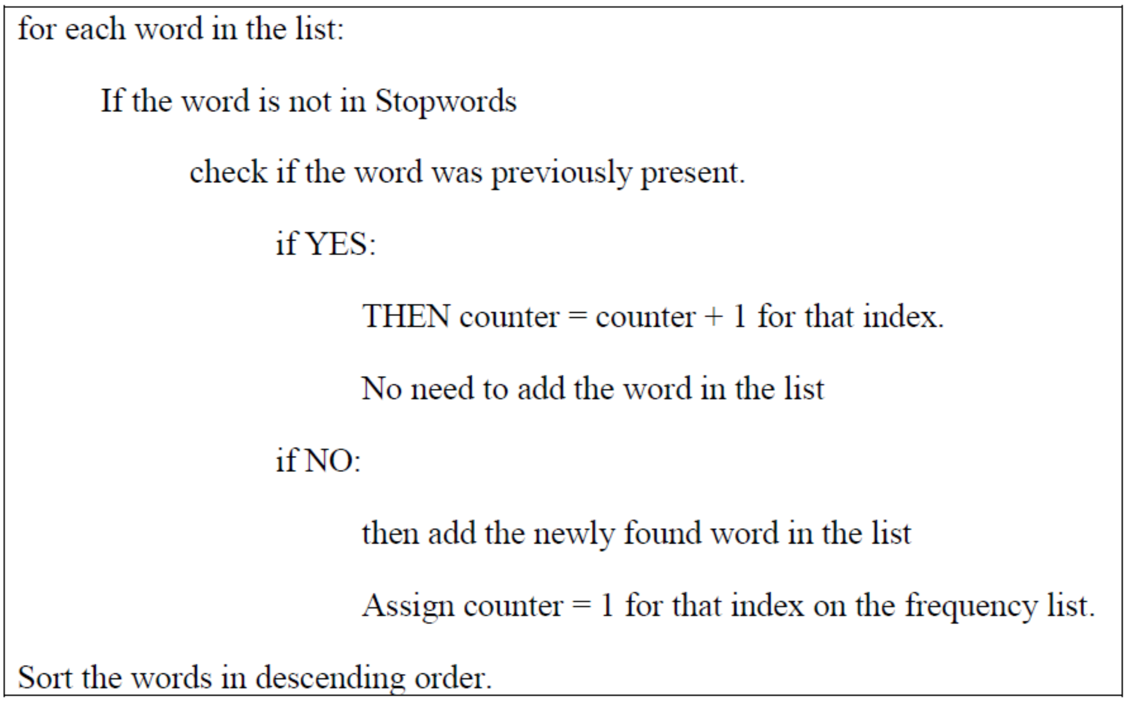
5.2.2 Splitting Sentences

We need to separate the sentences. We separated all the sentences from all the documents of each class and prepared text document for each class by using sentence splitter of Polyglot where all the appropriate sentences from the news of each class are present. The Polyglot splitter automatically splits each sentence finding ('|' or '.') from the whole document.
We used a Stop-word list to remove the unnecessary and irrelevant sentences. We used the Stop-word list of Bengali-Sentiment-Analysis. [14] We extracted sentences from 50 documents and added them in a text file for each category e.g., accidentsentences.txt.

5.2.3 Keyword Extraction

We have found out that there are very few key phrases in Bengali documents. So, we are working on keywords instead of key phrases. There are four categories of documents in our system and we have already mentioned them. So, we have extracted the keywords from all the documents in each category and calculated their frequencies. Tokenization in polyglot relies on the Unicode Text Segmentation algorithm. We created a list for storing all the unique words and a list for storing their frequencies. We stored the keywords with their respective scores in a document. The algorithm is given below:

Algorithm 1: Unique words detection and frequency calculation



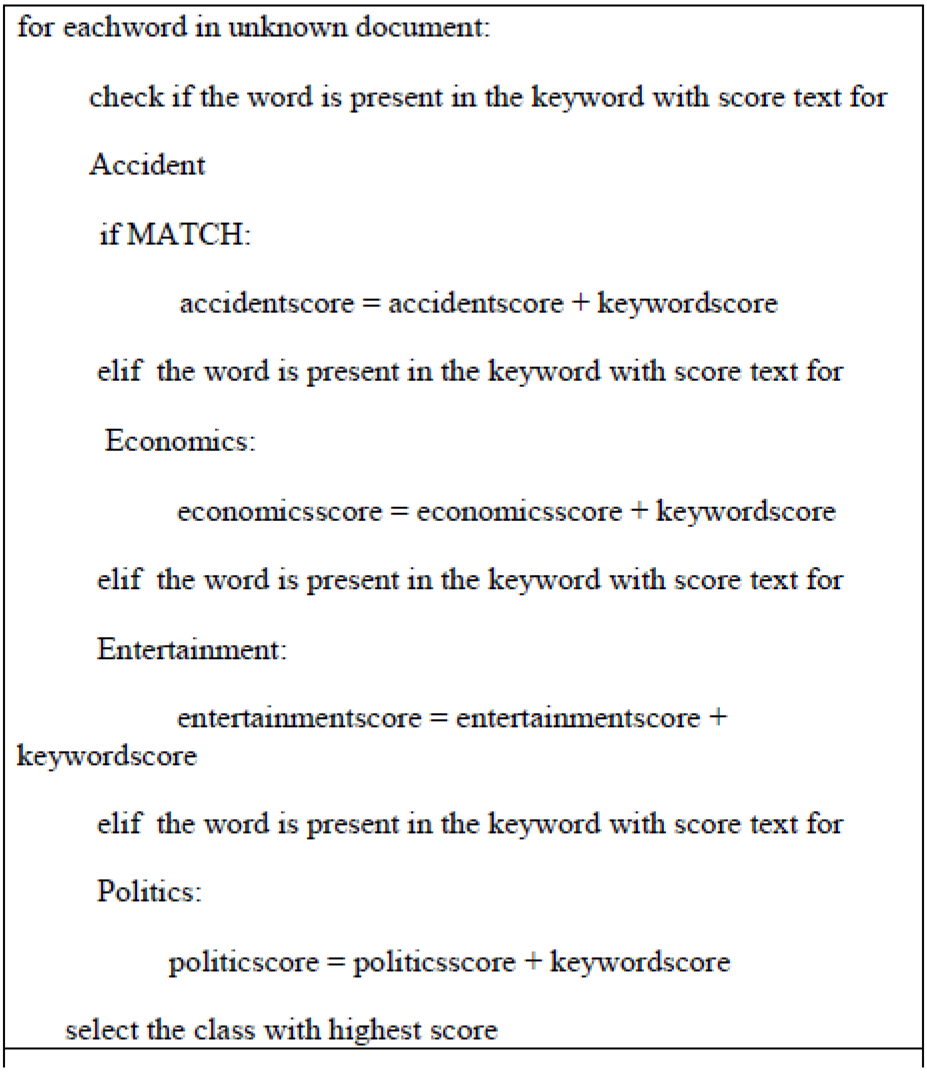
```
for each word in the list:
    If the word is not in Stopwords
        check if the word was previously present.
            if YES:
                THEN counter = counter + 1 for that index.
                No need to add the word in the list
            if NO:
                then add the newly found word in the list
                Assign counter = 1 for that index on the frequency list.
Sort the words in descending order.
```

After calculating frequency score, the words are sorted in descending order (High frequently used words to low frequently used words).

5.2.4 Classification of Documents

Classification involves determining class for unknown documents. For classifying an unknown document, we have matched each and every word of the document with four documents (each document for each category) prepared earlier that stores the keywords with scores. If a match is found, the corresponding score in that category (same word can have different score in different category) is added. Thus, we have chosen the class with the highest score among those four scores. We can understand the concept by the following algorithm:

Algorithm 2: Classification of unknown document



```
for eachword in unknown document:
        check if the word is present in the keyword with score text for
        Accident
         if MATCH:
                  accidentscore = accidentscore + keywordscore
         elif  the word is present in the keyword with score text for
          Economics:
                  economicsscore = economicsscore + keywordscore
         elif  the word is present in the keyword with score text for
         Entertainment:
                  entertainmentscore = entertainmentscore +
keywordscore
         elif  the word is present in the keyword with score text for
         Politics:
                  politicscore = politicsscore + keywordscore
     select the class with highest score
```

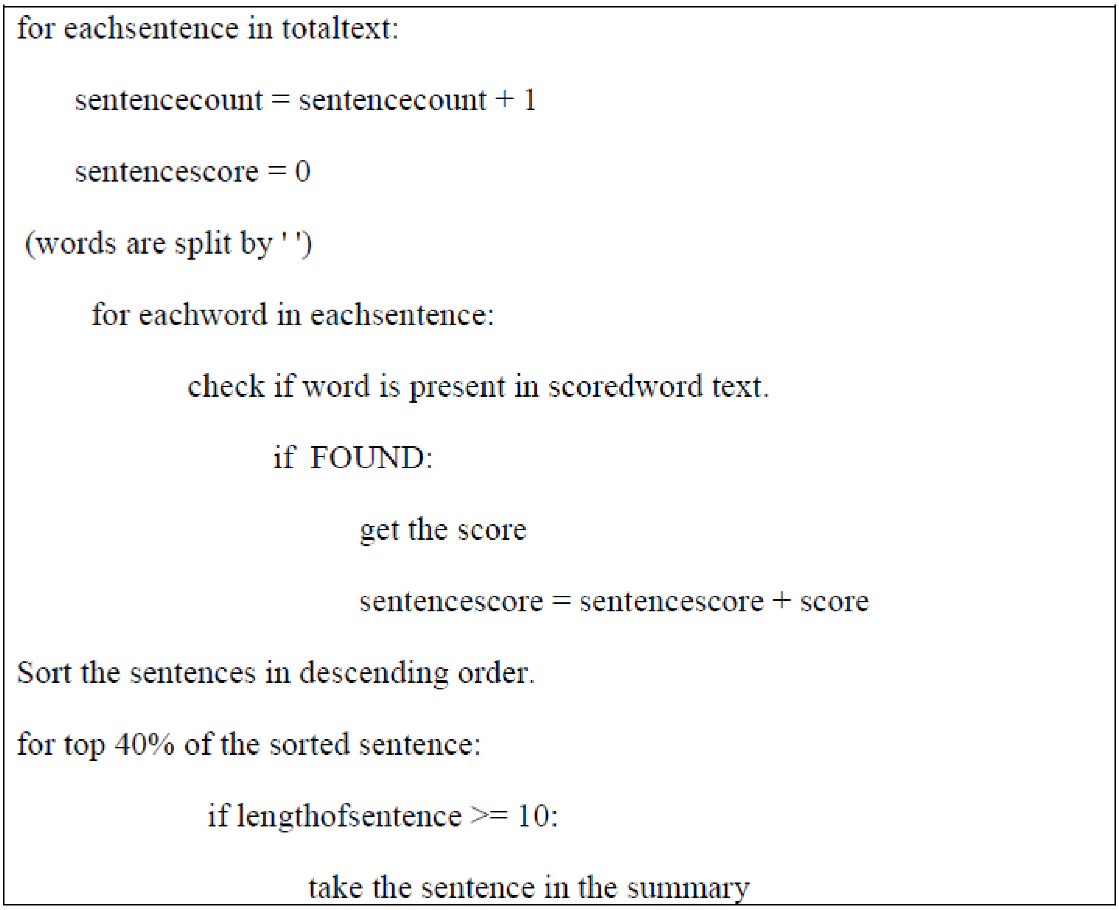
5.2.5 Score Calculation(IDF) of words

We calculated the score of each word by using the following formula:
Word score = Word Frequency / No of words present in all documents in each category.

5.2.6 Individual Scoring for Each Sentence

Using text function of polyglot, we stored the previously scored words in a variable (textual). For each document from each category, every sentence is read and stored in a variable (totaltext). The sentences are split by polyglot splitter. Every word in each sentence is matched with the words previously stored with frequency score and the score is taken. For each word in a sentence, the scores are added and the sentence score is stored in a new list.

Algorithm 3: Individual Sentence Scoring



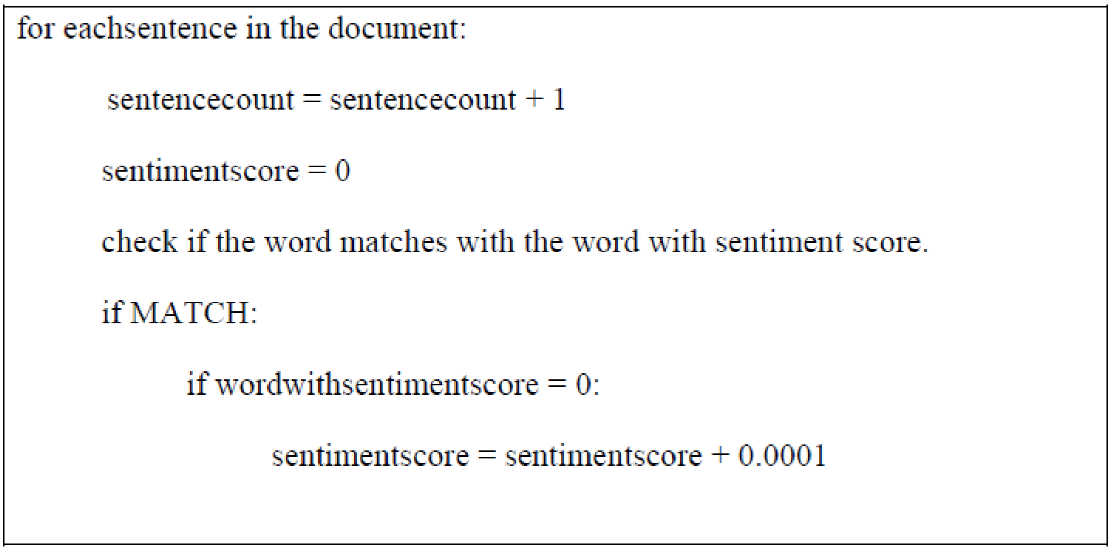
```
for eachsentence in totaltext:
    sentencecount = sentencecount + 1
    sentencescore = 0
 (words are split by ' ')
     for eachword in eachsentence:
            check if word is present in scoredword text.
                  if  FOUND:
                        get the score
                        sentencescore = sentencescore + score
Sort the sentences in descending order.
for top 40% of the sorted sentence:
               if lengthofsentence >= 10:
                      take the sentence in the summary
```

After individual scoring of sentences, they are sorted in descending order (Higher scored sentences to lower scored sentences). Then top 40 percent sentences are selected for summarization.

5.2.7 Sentiment Based Ranking

For sentiment scoring, we have used polyglot. Here polarity of each unique word has been calculated. Words have been scored based on their polarity. Using a function of polyglot we calculated the polarity of the words (+1 for positive words, -1 for negative words and 0 for neutral words). For sentiment wise ranking, we used the following algorithm:

Algorithm 4: Sentiment wise ranking



```
for eachsentence in the document:
    sentencecount = sentencecount + 1
    sentimentscore = 0
    check if the word matches with the word with sentiment score.
    if MATCH:
        if wordwithsentimentscore = 0:
            sentimentscore = sentimentscore + 0.0001
```

After the words are ranked sentiment wise, for each sentence in a document the sentiment scores of the words are added to score the sentences sentiment wise. Then sentences are sorted in descending order (Sentences with higher sentiment score to sentences with lower sentiment score). After that top 40 percent sentences are selected for summarization.

5.2.8 Hybrid Scoring

We have combined the two methods: keyword ranking and sentiment score ranking and calculated a combined score. We have added a weight to keyword ranking of a sentence and the weight is 40 percent of the sentence's total key word ranking. Similarly, a weight of 60 percent of the total sentiment score is added to the sentence's sentiment score and both weighted scores are summed up to calculate the hybrid score.
Let, Keyword Ranking * W1 + Sentiment scoring * W2 = Combined Score of a sentence
W1 = 40 percent of the total keyword ranking score
W2 = 60 percent of the total sentiment score

5.3 Experimental Results

We have stated earlier that we have generated summaries based on keyword and sentiment scoring and also combining both approaches. We have evaluated our generated summaries based on ROUGE score. 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.

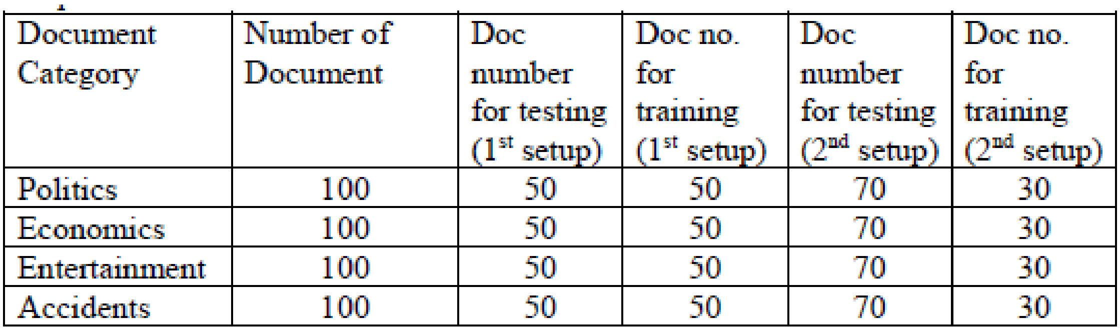
We have evaluated our generated summaries taking 50 percent, 30 percent of the total documents respectively for training purpose and 50 percent, 70 percent of the total documents for testing in keyword extraction and hybrid method.

That means, we have done evaluation in two setups. In 1st setup, the system is trained with 50 percent documents and has been tested for the rest 50 percent documents.

In 2nd setup, the system is trained with 30 percent documents and has been tested for the rest 70 percent documents.
We have evaluated our generated summaries taking 50 percent , 30 percent of the total documents respectively for training purpose and 50 percent , 70 percent of the total documents for testing in keyword extraction and hybrid method.
That means, we have done evaluation in two setups. In 1st setup, the system is trained with 50 percent documents and has been tested for the rest 50 percent documents.
In 2nd setup, the system is trained with 30 percent documents and has been tested for the rest 70 percent documents.For summary based on sentiment we have used 100 percent of the documents because we don't have to train the dataset as we have used polyglot polarity scores for calculating positive, negative and neutral sentences.

We calculated ROUGE 1 and 2 precision, recall and f-measures for both experiments. We have used the following formulas to calculate ROUGE precision, recall and f-measure.
ROUGE precision = No. of overlapping words / No. of words in the gold summary
ROUGE recall = No. of overlapping words / No. of words in the generated summary
F-measure = 2 * Precision * Recall / (Precision + Recall).
We can get a clear overview if we see the tables and charts regarding ROUGE evaluation.
For our first two setups i.e., keyword ranking and sentiment scoring we will now see a table that mentions the categories and number of documents used for the setups.

Table 5.2: Categories and number of documents for training and testing



Document Category	Number of Document	Doc number for testing (1st setup)	Doc no. for training (1st setup)	Doc number for testing (2nd setup)	Doc no. for training (2nd setup)
Politics	100	50	50	70	30
Economics	100	50	50	70	30
Entertainment	100	50	50	70	30
Accidents	100	50	50	70	30

5.3.1 Experimental Verifications with Evaluation

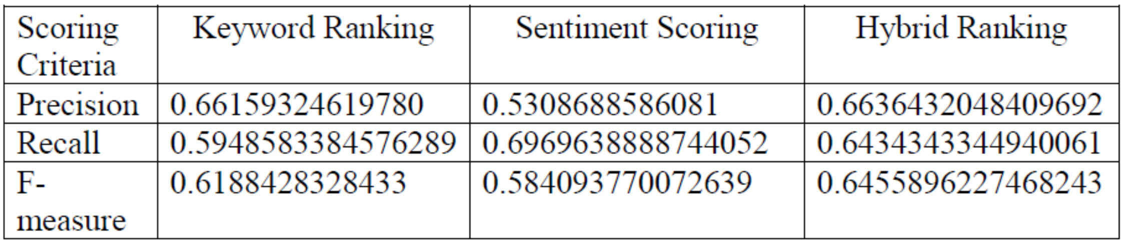
We have evaluated our generated summaries using ROUGE 1 and ROUGE 2 measures. We have already stated that, there are two types of setups we have experimented (training 50 percent , 30 percent and testing 50 percent , 70 percent respectively). In this section we will see the ROUGE scores in both setups along with their bar charts.

5.3.1.1 ROUGE Scores of Different Categories for 1st Setup

For Category Accidents:

Rouge 1 precision, recall and f-measure scores are shown in the table below:

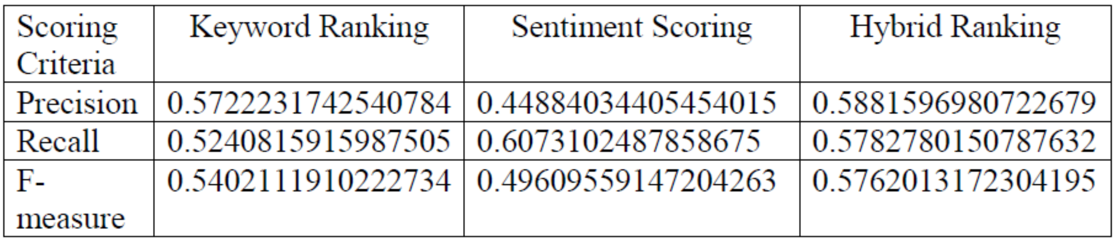
Table 5.3: Average ROUGE 1 Score of the system for Category Accidents (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.66159324619780	0.5308688586081	0.6636432048409692
Recall	0.5948583384576289	0.6969638888744052	0.6434343344940061
F-measure	0.6188428328433	0.584093770072639	0.6455896227468243

Rouge 2 precision, recall and f-measure scores are shown in the table below:

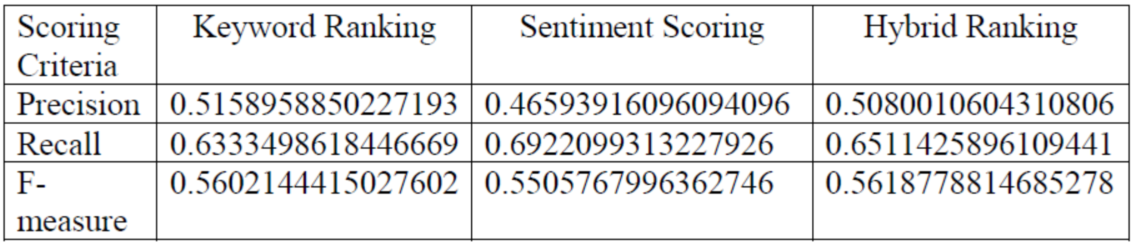
Table 5.4: Average ROUGE 2 Score of the system for Category Accidents (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.5722231742540784	0.44884034405454015	0.5881596980722679
Recall	0.5240815915987505	0.6073102487858675	0.5782780150787632
F-measure	0.5402111910222734	0.49609559147204263	0.5762013172304195

For Category Economics: Rouge 1 precision, recall and f-measure scores are shown in the table below:

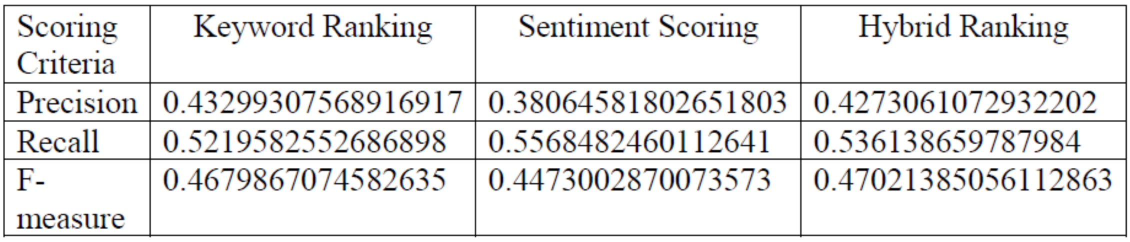
Table 5.5: Average ROUGE 1 Score of the system for Category Economics (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.5158958850227193	0.46593916096094096	0.5080010604310806
Recall	0.6333498618446669	0.6922099313227926	0.6511425896109441
F-measure	0.5602144415027602	0.5505767996362746	0.5618778814685278

Rouge 2 precision, recall and f-measure scores are shown in the table below:

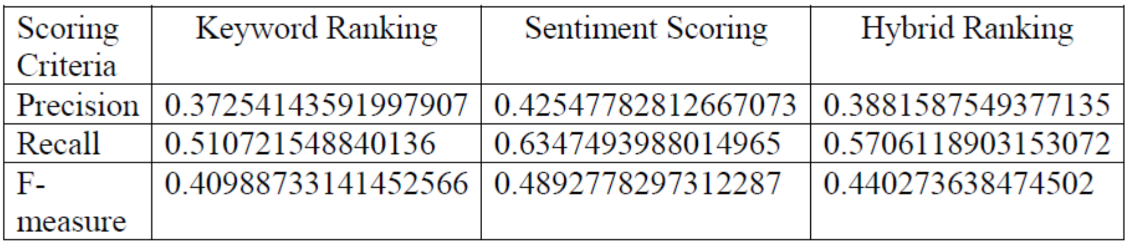
Table 5.6: Average ROUGE 2 Score of the system for Category Economics (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.43299307568916917	0.38064581802651803	0.4273061072932202
Recall	0.5219582552686898	0.5568482460112641	0.536138659787984
F-measure	0.4679867074582635	0.4473002870073573	0.47021385056112863

For Category Entertainment: Rouge 1 precision, recall and f-measure scores are shown in the table below:

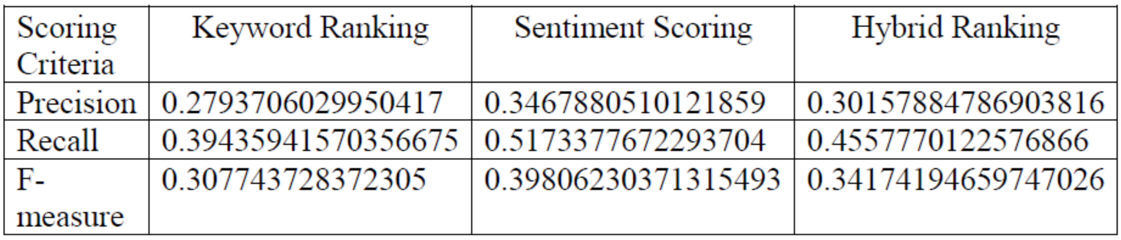
Table 5.7: Average ROUGE 1 Score of the system for Category Entertainment (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.37254143591997907	0.42547782812667073	0.3881587549377135
Recall	0.510721548840136	0.6347493988014965	0.5706118903153072
F-measure	0.40988733141452566	0.4892778297312287	0.440273638474502

Rouge 2 precision, recall and f-measure scores are shown in the table below:

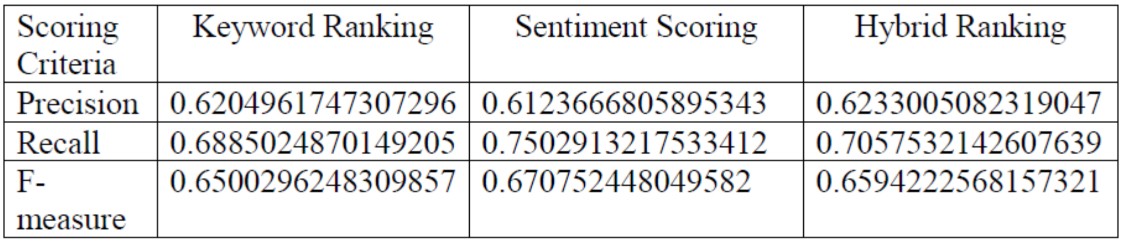
Table 5.8: Average ROUGE 2 Score of the system for Category Entertainment (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.2793706029950417	0.3467880510121859	0.30157884786903816
Recall	0.39435941570356675	0.5173377672293704	0.4557770122576866
F-measure	0.307743728372305	0.39806230371315493	0.34174194659747026

For Category Politics: Rouge 1 precision, recall and f-measure scores are shown in the table below:

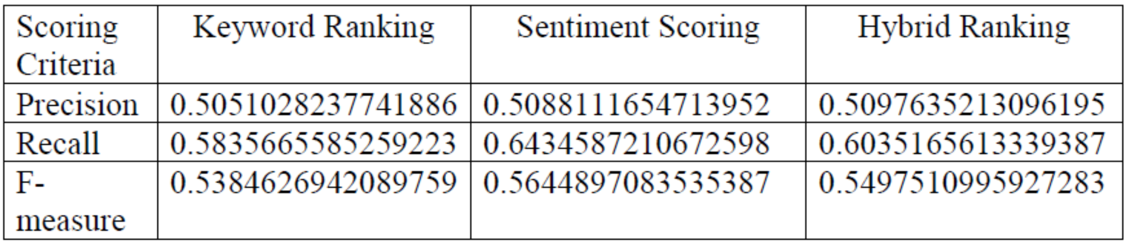
Table 5.9: Average ROUGE 1 Score of the system for Category Politics (1st Setup).



Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.6204961747307296	0.6123666805895343	0.6233005082319047
Recall	0.6885024870149205	0.7502913217533412	0.7057532142607639
F-measure	0.6500296248309857	0.670752448049582	0.6594222568157321

Rouge 2 precision, recall and f-measure scores are shown in the table below:

Table 5.10: Average ROUGE 2 Score of the system for Category Politics (1st Setup).

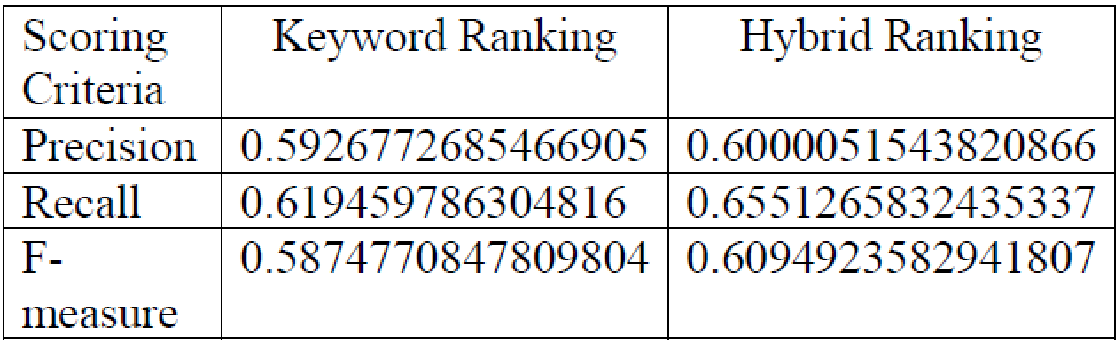


Scoring Criteria	Keyword Ranking	Sentiment Scoring	Hybrid Ranking
Precision	0.5051028237741886	0.5088111654713952	0.5097635213096195
Recall	0.5835665585259223	0.6434587210672598	0.6035165613339387
F-measure	0.5384626942089759	0.5644897083535387	0.5497510995927283

5.3.1.2 ROUGE Scores of Different Categories for 2nd Setup

We haven't calculated ROUGE scores for sentiment-based ranking in our second setup. **For Category Accidents:** Rouge 1 precision, recall and f-measure scores are shown in the table below:

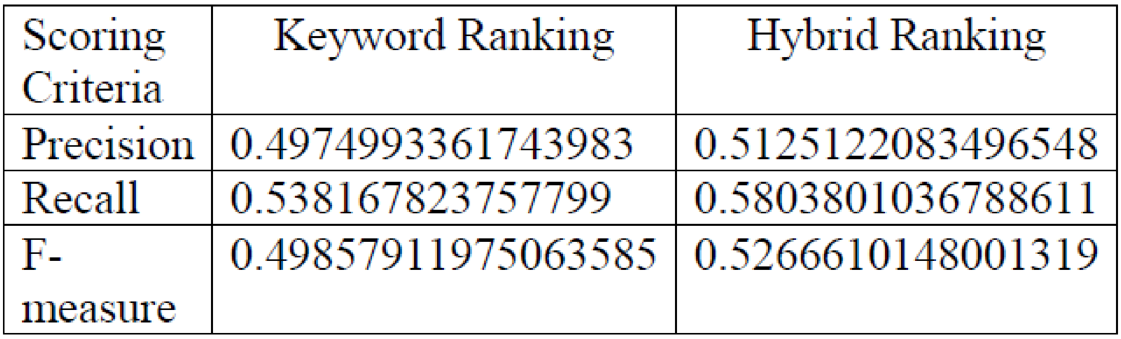
Table 5.11: Average ROUGE 1 Score of the system for Category Accidents (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.5926772685466905	0.6000051543820866
Recall	0.619459786304816	0.6551265832435337
F-measure	0.5874770847809804	0.6094923582941807

Rouge 2 precision, recall and f-measure scores are shown in the table below:

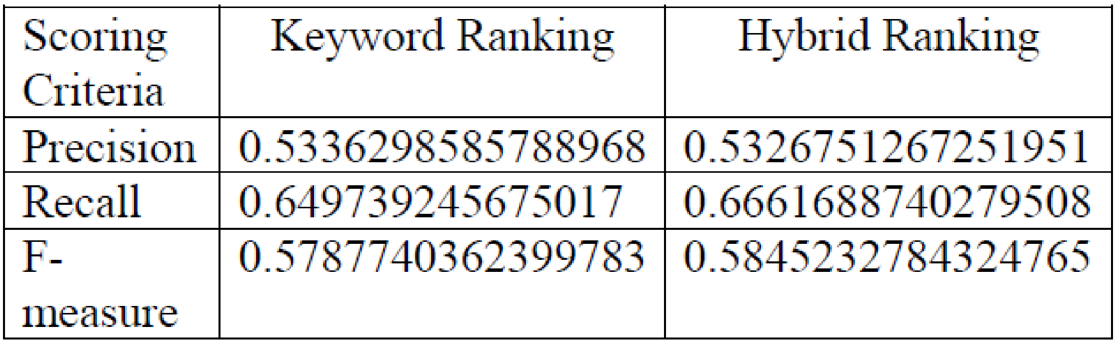
Table 5.12: Average ROUGE 2 Score of the system for Category Accidents (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.4974993361743983	0.5125122083496548
Recall	0.538167823757799	0.5803801036788611
F-measure	0.49857911975063585	0.5266610148001319

For Category Economics: Rouge 1 precision, recall and f-measure scores are shown in the table below:

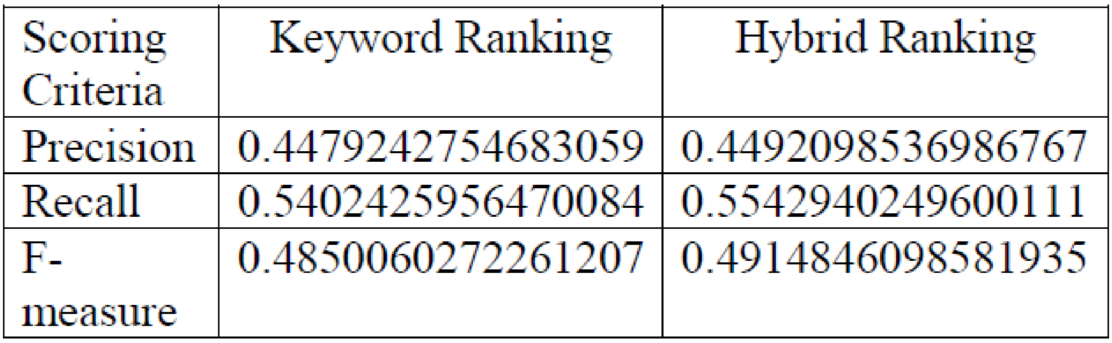
Table 5.13: Average ROUGE 1 Score of the system for Category Economics (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.5336298585788968	0.5326751267251951
Recall	0.649739245675017	0.6661688740279508
F-measure	0.5787740362399783	0.5845232784324765

Rouge 2 precision, recall and f-measure scores are shown in the table below:

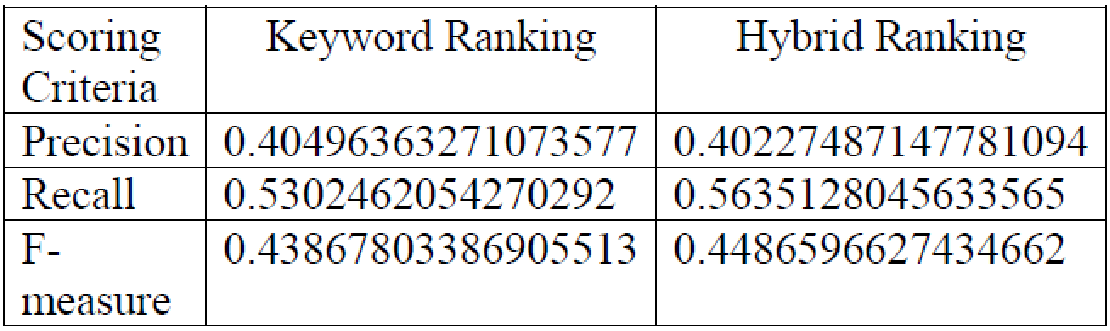
Table 5.14: Average ROUGE 2 Score of the system for Category Economics (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.4479242754683059	0.4492098536986767
Recall	0.5402425956470084	0.5542940249600111
F-measure	0.4850060272261207	0.4914846098581935

For Category Entertainment: Rouge 1 precision, recall and f-measure scores are shown in the table below:

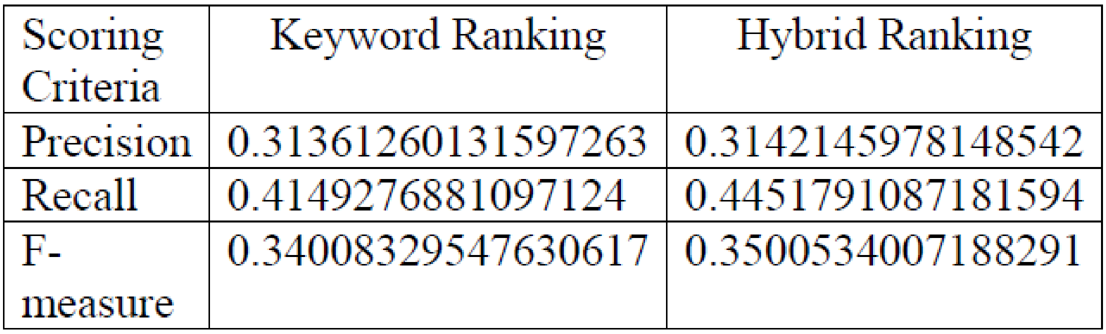
Table 5.15: Average ROUGE 1 Score of the system for Category Entertainment (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.40496363271073577	0.40227487147781094
Recall	0.5302462054270292	0.5635128045633565
F-measure	0.43867803386905513	0.4486596627434662

Rouge 2 precision, recall and f-measure scores are shown in the table below:

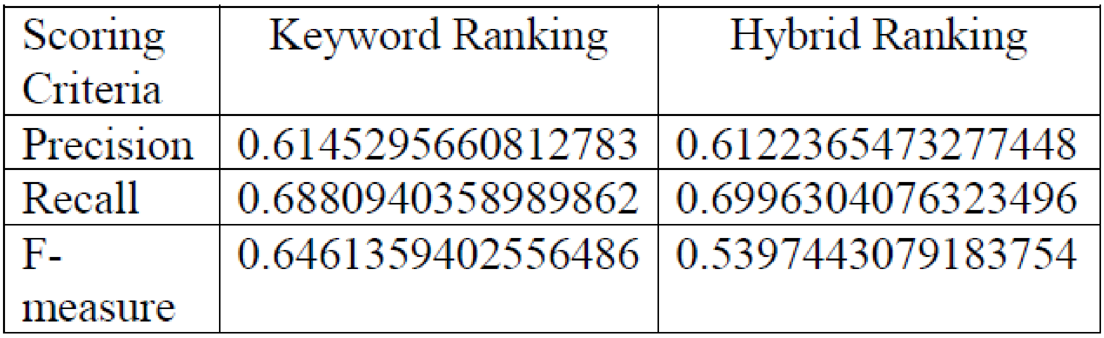
Table 5.16: Average ROUGE 2 Score of the system for Category Entertainment (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.31361260131597263	0.3142145978148542
Recall	0.4149276881097124	0.4451791087181594
F-measure	0.34008329547630617	0.3500534007188291

For Category Politics: Rouge 1 precision, recall and f-measure scores are shown in the table below:

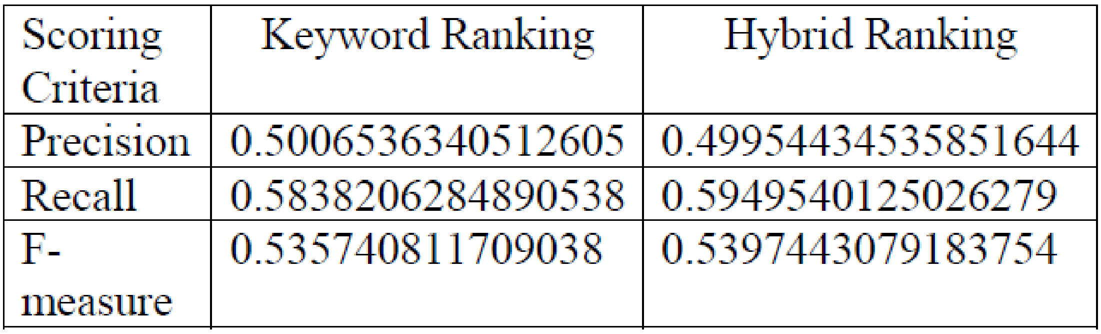
Table 5.17: Average ROUGE 1 Score of the system for Category Politics (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.6145295660812783	0.6122365473277448
Recall	0.6880940358989862	0.6996304076323496
F-measure	0.6461359402556486	0.5397443079183754

Rouge 2 precision, recall and f-measure scores are shown in the table below:

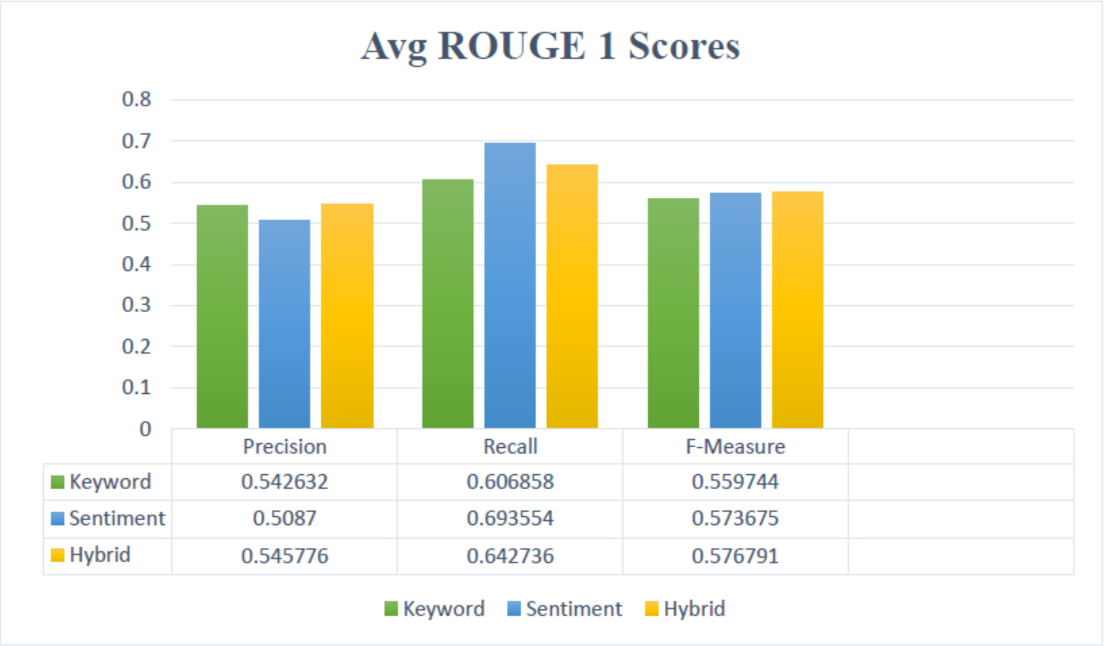
Table 5.18: Average ROUGE 2 Score of the system for Category Politics (2nd Setup).



Scoring Criteria	Keyword Ranking	Hybrid Ranking
Precision	0.5006536340512605	0.49954434535851644
Recall	0.5838206284890538	0.5949540125026279
F-measure	0.535740811709038	0.5397443079183754

5.3.1.3 Average ROUGE 1 Score Calculation for Different Methods

We will see tables and bar diagrams of average ROUGE 1 Scores (both setups) for different methods in this section.
Bar charts along with the table showing average Rouge 1 scores based on Keyword, Sentiment and Hybrid Scoring in 1st Setup is shown below:

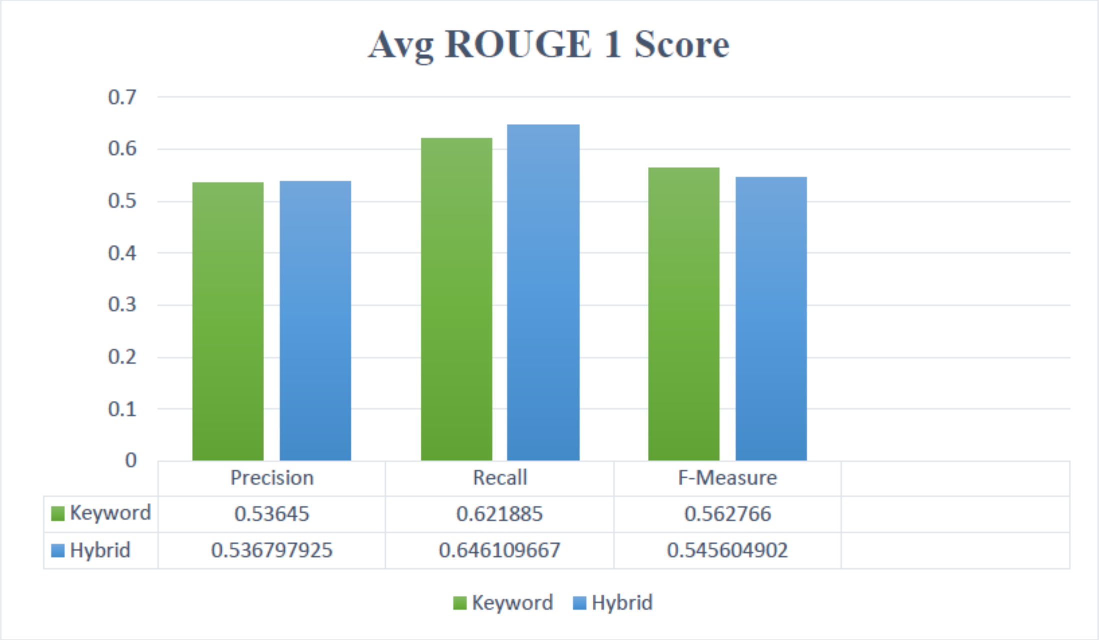


	Precision	Recall	F-Measure
🟩 Keyword	0.542632	0.606858	0.559744
🟦 Sentiment	0.5087	0.693554	0.573675
🟨 Hybrid	0.545776	0.642736	0.576791

Figure 5.1: Average ROUGE 1 Scores of the system for different methods (1st Setup)

From the chart we can say in terms of Precision, Hybrid ranking gets slightly better Scores than Keyword ranking. In terms of Recall, Sentiment scoring is the highest, but in terms of F-Measure Hybrid ranking again beats them all. Keyword based ranking seems not too good.

Bar charts along with the table showing average Rouge 1 scores based on Keyword, Sentiment and Hybrid Scoring in 2nd Setup is shown below:



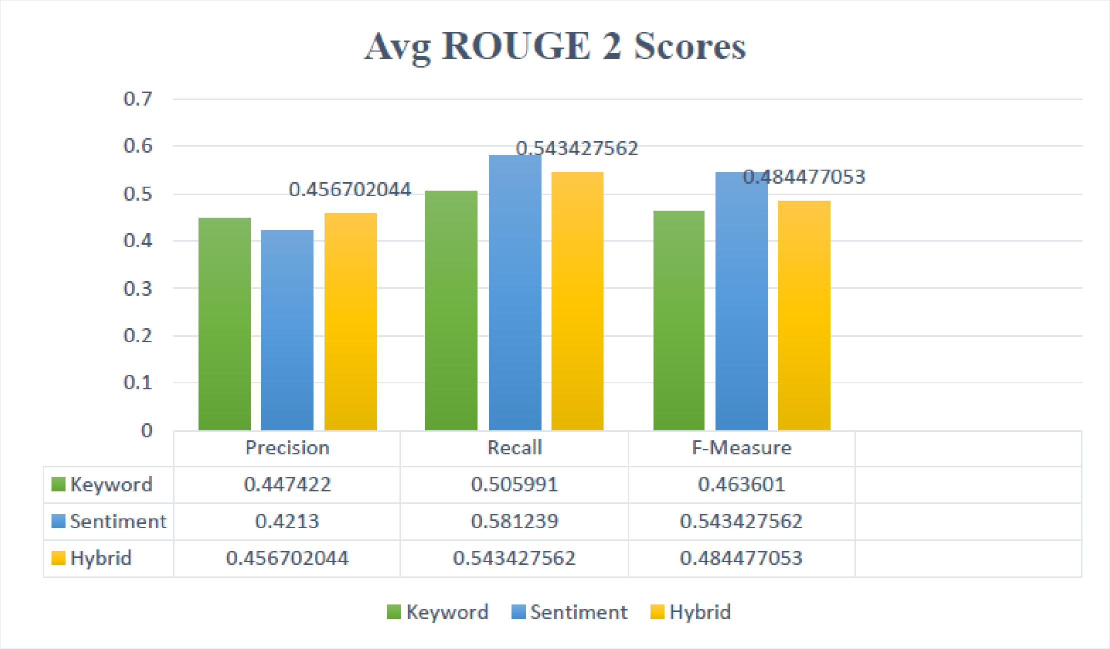
	Precision	Recall	F-Measure
🟩 Keyword	0.53645	0.621885	0.562766
🟦 Hybrid	0.536797925	0.646109667	0.545604902

Figure 5.2: Average ROUGE 1 Scores of the system for different methods (2nd Setup)

From the chart we can say in terms of Precision, Hybrid ranking gets slightly better Scores than Keyword ranking. In terms of Recall, Sentiment scoring is the highest, also in terms of F-Measure Sentiment scoring again beats them all. Keyword based ranking seems okay in this case.

5.3.1.4 Average ROUGE 2 Score Calculation for Different Methods

We will see tables and bar diagrams of average ROUGE 2 Scores (both setups) for different methods in this section.
Bar charts along with the table showing average Rouge 2 scores based on Keyword, Sentiment and Hybrid Scoring in 1st Setup is shown below:



	Precision	Recall	F-Measure
🟩 Keyword	0.447422	0.505991	0.463601
🟦 Sentiment	0.4213	0.581239	0.543427562
🟨 Hybrid	0.456702044	0.543427562	0.484477053

Figure 5.3: Average ROUGE 2 Scores of the system for different methods (1st Setup)

From the chart we can say in terms of Precision, Hybrid ranking gets slightly better Scores than Keyword ranking. In terms of Recall, Sentiment scoring is the highest, but in terms of F-Measure Hybrid ranking again beats them all. Keyword based ranking seems not too good.

Tables and bar chart for average ROUGE 2 scores based on Keyword, Sentiment and Hybrid Scoring in 2nd Setup is shown below:

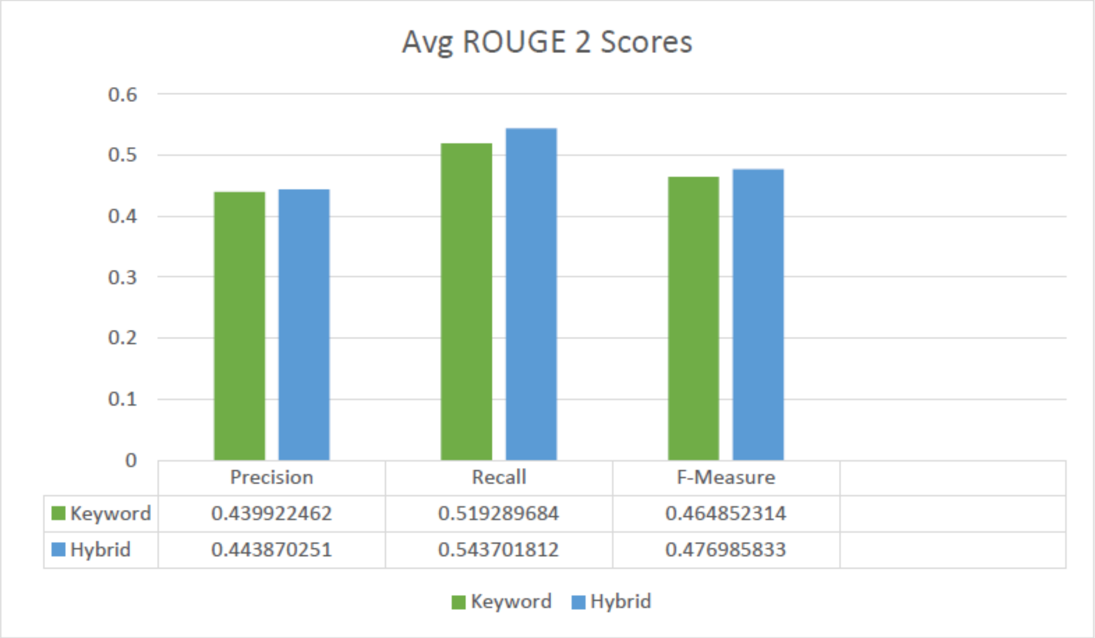


Figure 5.4: Average ROUGE 2 Scores of the system for different methods (2nd Setup)

From the chart we can say in terms of Precision, Hybrid ranking gets slightly better Scores than Keyword ranking. In terms of Recall, Sentiment scoring is the highest, also in terms of F-Measure Sentiment scoring again beats them all. Keyword based ranking seems okay in this case.

5.4 Sample Generated Output

.

.

.

5.5 Limitations

There are some limitations in our system. We have found out some problems in the second method i.e., sentiment scoring.

*We have stated before that, in sentiment scoring method, we have given the neutral words in sentences higher scores in order to select the sentence for summary generation i.e., more neutral words mean more score. We have given higher priority to the sentences that have neutral words. If a sentence has only neutral words that should be selected for the summary. But a sentence that has only neutral words gets less score than a sentence that has positive or negative words along with more neutral words than the previous sentence. So, the sentence with only neutral words doesn't get selected which should not be the case.

*We have used polyglot in our system to determine the polarity of the words. But the sentiment of polyglot cannot always determine the polarity correctly. So, it is a limitation of our system.

5.6 Overview

In this chapter we gained knowledge and a clear overview of the system we are working on for generating automatic summary. We have given a brief idea about how much of the work is complete, what part of the work is currently in progress and how the work is going in general. From this chapter, the recipients will get a chance to evaluate our work and to suggest if any changes need to be done or not.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

We have discussed about the overview of our proposed system and how much progress there have been so far. We also discussed about the evaluation measures, experimental verifications and how much accurate result our system provides. We have encountered some difficulties in our works.

6.1.1 Brief Survey

We are working on generating automatic extractive summaries of Bengali documents. We have proposed a system that combines three methods of generating automatic summary. We have prepared our datasets and have been able to generate summaries based on key phrase ranking, sentiment scoring and combining both of the methods. We haven't done the text ranking part yet. We hope that we will be able to finish our work in time.

6.1.2 Difficulties Encountered

We have encountered some difficulties in our work. Some of them are mentioned below:
*We have mentioned that we have created our own summaries and considered them as gold summaries. But we cannot assure that the summaries are 100 percent correct.

*Bengali is a sophisticated language. But very few works have been done for text summarization for Bengali documents. As a result, there are very few resources as well as predefined libraries are available for Bengali documents.

*We have used polyglot library for determining the polarity of words. Polyglot is a predefined, open source library that covers 136 languages including Bengali. So naturally it is less trained library than the library which works for only one language. Polyglot sometimes lacks crucial words that must be included in the summary.

*Another vital point is that, polyglot most of the times splits sentences with full stops (.). Polyglot will split the sentence into two sentences: one before the full stop has occurred and other one after the full stop to "|".
It has a huge impact on the generated summary. As a result, we get less ROUGE score. This is a great issue.

6.2 Future Work

We have proposed a hybrid model and only half of our work is complete. So, in future we will try to complete the following tasks:

*Text Ranking - We haven't completed this part yet. This is a graph-based ranking. Graph-based ranking algorithms are a way of deciding the importance of a vertex within a graph. We will give highest weight to text ranking method. Because this is an imperial method as the score is based on the connectivity of related sentences.

*FastText comparison with polyglot - We have used polyglot for now but in future we may use another library fastText for sentiment analysis depending on the result. FastText is a Library that uses Word to Vector method for 157 languages. Pre-trained word vectors are distributed over 157 languages using fastText. We will compare the result and see the outcome.

*We will also try to provide more importance on sentences that based on headlines and numerical values.

*We haven't calculated the accuracy of classifying documents, in future our plan is to calculate it.

6.3 Overview

In this chapter, we have discussed about the whole idea of our work shortly. We have given a clear overview of what difficulties we have faced and we will be doing in future. We hope we can make a better summarizer for Bengali document.

References

[1] K. Sarkar, "Bengali text summarization by sentence extraction," *arXiv preprint arXiv:1201.2240*, 2012.

[2] M. N. Uddin and S. A. Khan, "A study on text summarization techniques and implement few of them for bangla language," in *2007 10th international conference on computer and information technology*, pp. 1–4, IEEE, 2007.

[3] K. Sarkar, "An approach to summarizing bengali news documents," in *proceedings of the International Conference on Advances in Computing, Communications and Informatics*, pp. 857–862, ACM, 2012.

[4] A. Das and S. Bandyopadhyay, "Topic-based bengali opinion summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 232–240, Association for Computational Linguistics, 2010.

[5] M. I. A. Efat, M. Ibrahim, and H. Kayesh, "Automated bangla text summarization by sentence scoring and ranking," in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 1–5, IEEE, 2013.

[6] M. M. Haque, S. Pervin, and Z. Begum, "Automatic bengali news documents summarization by introducing sentence frequency and clustering," in *2015 18th International Conference on Computer and Information Technology (ICCIT)*, pp. 156–160, IEEE, 2015.

[7] M. M. Haque, S. Pervin, and Z. Begum, "Enhancement of keyphrase-based approach of automatic bangla text summarization," in *2016 IEEE Region 10 Conference (TENCON)*, pp. 42–46, IEEE, 2016.

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

[9] A. Paul, M. T. Imtiaz, A. H. Latif, M. Ahmed, F. A. Adnan, R. Khan, I. Kadery, and R. M. Rahman, “Bangla news summarization,” in *International Conference on Computational Collective Intelligence*, pp. 479–488, Springer, 2017.

[10] S. Saziyabegum and P. S. Sajja, “Literature review on extractive text summarization approaches,” *International Journal of Computer Applications*, vol. 156, no. 12, 2016.

[11] J. Lin, “Summarization,” *Encyclopedia of database systems*, pp. 2884–2889, 2009.

[12] J. C. Cheung, “Comparing abstractive and extractive summarization of evaluative text: controversiality and content selection,” *B. Sc.(Hons.) Thesis in the Department of Computer Science of the Faculty of Science, University of British Columbia*, 2008.

[13] J. Steinberger and K. Ježek, “Evaluation measures for text summarization,” *Computing and Informatics*, vol. 28, no. 2, pp. 251–275, 2012.

[14] Imran, “Bengali-sentiment-analysis.” `https://github.com/Imran-cse/Bengali-Sentiment-Analysis`, 2019.