A Study of Text Summarization Techniques and Their Applications on Hindi Language

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
Master of Technology
IN
ARTIFICIAL INTELLIGENCE

BY

Rajat Nagpal

Electrical Engineering
Indian Institute of Science
Bangalore – 560 012 (India)

August, 2020

Declaration of Originality

I, **Rajat Nagpal**, with SR No. **04-03-02-10-42-18-1-15533** hereby declare that the material presented in the thesis titled

A Study of Text Summarization Techniques and Their Applications on Indian Languages

represents original work carried out by me in the **Department of Electrical Engineering** at **Indian Institute of Science** during the years **2018-2020**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Advisor Signature

DEDICATED TO

My country. Jai Hind

Acknowledgements

Thank you Mom and Dad for your everlasting support.

Special thanks to my advisor Professor A. G. Ramakrishnan who supported me like a friend and provided me the best environment for a creative research, who always stood by me in my ups and downs, who showed faith in me throughout my research and helped me grow in life.

My heartiest affection to my big brother and mentor Amit Meghanani for his relentless support in making the research journey, a wonderful experience.

Thanks to Devyani Agarwal for helping me through difficult times throughout the research.

Thanks to the open source communities which made my research possible.

Abstract

In recent years, there has been an exponential increase in the amount of textual data from a variety of sources in the growing world of the internet. This volume of text is an important source of information for different fields. There is a great need for shorter summaries of the text which contain salient features of the text to know if the information contained in this volume is what we are looking for. Therefore in this work, we review the main approaches for text summarization, starting from the early techniques, like Topic Words, Word Probability, etc. to the recent ones, like, Text-Rank and Clustering. We apply these techniques on the Hindi language (one of the official languages of India) and compare the results. In addition, we explore various combinations of different sentence representation and extractive summarization techniques on Hindi data, which improves the summary in terms of the rouge score.

We used the Rouge metric which is based on word matching for the evaluation of text summarization for all the experiments. The metric is not satisfactory as far as the problem of summarization is concerned since it does not completely align with the summarization problem. However, this metric is widely accepted and used by the research community. Summaries on Hindi data are generated by using TF-IDF weights, ULF-Fit sentence encoding [1], clustering method [2], text-rank method [3], Maximum Likelihood Estimate (MLE) sentence embeddings [4] , word average sentence embeddings and their combinations. The word vectors used in the MLE sentence embeddings algorithm and word-average sentence embeddings algorithm are trained on the text dataset of the Prime Minister of India, Mr. Narendra Damodardas Modi's speech on "Man ki Baat". The dataset contains 4398 sentences with a word vocabulary size of 12079 words. The ROUGE score is calculated on a test data of 42 sentences taking the average of rouge score between our hypothesis summary and 10 reference summaries. ULM-Fit embedding vectors are trained on the Hindi Wikipedia dataset [5]. We took the pre-trained model and fine-tuned it on our text dataset of Prime Minister's speech on "Man ki Baat" in order to get better sentence representations. We found that ULM-Fit sentence encoding [1] with text-rank algorithm [3] outperforms other techniques in terms of ROUGE Score, giving the ROUGE-1, ROUGE-2 and ROUGE-L scores of 0.53, 0.45 and 0.45, respectively.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is tremendous amount of textual information available in the growing world of internet. The growing availability of the textual content demanded exhaustive research in creating shorter informative summaries of the text. Automatic text summarization can make a huge impact in fetching information in the textual fields.

Automatic summarization is the process of creating fluent summaries which contain the key information about the text and does not change the overall meaning. According to Radef et al. [6] a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that".

Automatic text summarization is a non-trivial and one of the most challenging tasks in natural language processing. Because, when humans summarize a text document, we usually read the whole document and develop our own understanding about it, taking into account our previous knowledge and then we write the summary in our own vocabulary which itself requires a good prior knowledge about the language. Since machines lack human knowledge and linguistic capability, it makes automatic text summarization a very challenging task for the machines.

Automatic text summarization gained attraction as early as the 1950s. An important research was published these days for scientific documents [7]. Luhn et al. [7] used features like word frequency and phase frequency ignoring most common words. Edmundson et al. [8] described a paradigm, in addition to standard frequency of words, used Cue method, Title method and Location method. The sentence score is calculated by searching for the presence of certain cue words in the cue dictionary. The score of a sentence is calculated based on the title or heading words present in the sentence. Location method assumes that sentences in the beginning of the document should be given higher preference than the subsequent ones.

Many advanced techniques for automatic text summarization have been published until 2000s. LexRank [9], a graph based method, is based on the concept of eigenvector centrality in a graph representation of sentences. In this model, adjacency matrix for graph representation is constructed with intra-sentence cosine similarity scores.

There are two different approaches of text summarization, namely, *extractive and abstractive*. As the name suggests, extractive summarization works by extraction of most relevant sentences which can be used as summary of the text. There are some papers that provide overview of extractive summarization techniques [10][11][12]. In contrast, abstractive method creates its own sentences by gathering the information from the given text. Abstractive summarization needs advanced natural language generation techniques to be able to generate its own grammatically correct sentences which convey the salient information of the text. Even though human created summaries are usually abstractive, most of the research until 2017 has focused on extractive summarization. The reason behind this is the more challenging nature of the abstractive summarization problem.

In this survey, we provide an overview of some of the most effective techniques of *extractive and abstractive* summarization. We apply some of the unsupervised extractive summarization techniques on Indian laguage.

Chapter 2

Extractive Summarization

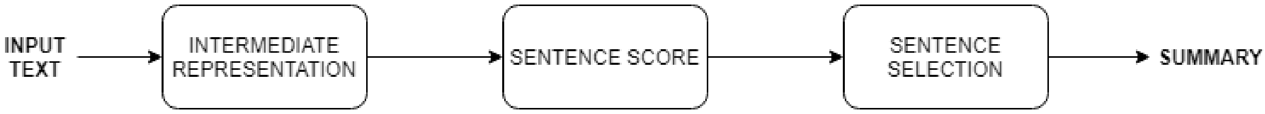


Figure 2.1: Three major steps of every extractive summarization technique

2.1 Intermediate Representation

First step of every summarizer is to create an intermediate representation which captures the salient features of the text. A good intermediate representation can be very helpful to make an informative summary. There are two types of approaches based on the representation: *topic representation and indicator representation*. Topic representation based summarization techniques are divided into frequency-driven approaches, topic word approaches, latent semantic analysis and Bayesian topic models [12]. Indicator representation approaches represents sentences as a combination of features such as sentence length, position in the document, etc.

2.2 Sentence Score

In topic representation approach, sentence score is focused on how well the sentence describes the topic of the document whereas in indicator representation approach, the sentence score is determined by different indicator features of the sentence.

2.3 Sentence Selection

The sentence selection can be converted into an optimization problem where the target is to reduce redundancy and to increase overall importance and coherency. Whereas, some approaches use greedy algorithms to select the sentences.

The summarizer selects top N high ranked sentences to create the summary. The context of the summary can also be used to select the sentences. Another important aspect to consider is the type of documents.

Chapter 3

Evaluation Metrics

The following metrics are used for evaluation of text summarization. Although, the metrics are not satisfactory as far as the problem of summarization is concerned. These metrics are based on word matching and does not completely align with the summarization problem. Till today, these metrics are the best option to evaluate summaries.

3.1 ROUGE Score

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.

Formally, ROUGE-N [13] is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N, where N denotes N gram, is computed as follows:

$$ROUGE-N = \frac{\sum_{S \in R} \sum_{n \in S} C_{match}(n)}{\sum_{S \in R} \sum_{n \in S} C(n)}$$

Where S stands for summary, C(n) denotes total number of n-grams, n stands for the length of the n-gram and $C_{match}(n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, R.

3.2 Recall

It is the ratio of total retrieved correct sentences to the total of the retrieved correct sentences and non-retrieved correct sentences in a document. It can be estimated as follows:

$$Recall = \frac{\sum_{S \in G} \sum_{n \in S} C_{match}(n)}{\sum_{S \in R} \sum_{n \in S} C(n)}$$

Where S stands for summary, C(n) denotes total number of n-grams, n stands for the length of the n-gram and $C_{match}(n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, R. The set G denotes generated summaries.

3.3 Precision

It is the ratio of total retrieved correct sentences to the total of retrieved correct sentences and retrieved incorrect sentences from a document. It can be estimated as follows:

$$Precision = \frac{\sum_{S \in G} \sum_{n \in S} C_{match}(n)}{\sum_{S \in G} \sum_{n \in S} C(n)}$$

Where S stands for summary, C(n) denotes total number of n-grams, n stands for the length of the n-gram and $C_{match}(n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, R. The set G denotes generated summaries.

3.4 F-score

It measures the effectiveness of retrieval with respect to a user, who attaches β times as much importance to the recall as that of precision. The F-score for non-negative real β ($0 \leq \beta < \infty$) is computed as follows:

$$F_\beta = \frac{(1+\beta^2)(P * R)}{(\beta^2 * P + R)}$$

where P denotes precision and R denotes recall.

Chapter 4

Topic Representation Approaches

4.1 Topic Words

It is one of the earliest most commonly used technique which aims to identify words which best represent the topic [7]. An update to Luhn's idea was presented by T. Dunning et al. [14]. They used log-likelihood retio test which proved to be effective and increased the accuracy in multi-document summarization. It is advisable to measure the density of topic signature words instead of the numbers so that it does not depend on the length of the sentence.

4.2 Frequency-driven Approaches

4.2.1 Word Probability

Number of occurence of a word W divided by total number of input words.

$$P(x) = \frac{f(x)}{N}$$

SumBasic approach [15], weight of a sentence S_j is given as follows:

$$W(s_j) = \frac{\sum_{w_i \in s_j} P(w_i)}{|(w_i|w_i \in s_j)|}$$

In the next step, it picks the best scoring sentence that contains the highest probability word. This ensures that the word which represents the topic is included in the summary. Now, Each word weight is updated in the chosen sentence to decrease the chances of the words occurring again. This step takes care of the redundancy.

$$p_{new}(w_i) = p_{old}(w_i)^2$$

SumBasic strategy is a greedy strategy. There are other optimization strategies, like [16][17]

4.2.2 TF-IDF

TF-IDF [18], one of the most loved weighing technique in the field of natural language, automatically takes care of the stop words (that should be omitted from consideration) in the document(s) by giving low weights to words appearing in most documents. The weight of each word w in document d is computed as follows:

$$p(w) = f_d(w) \times log\frac{|D|}{f_D(w)}$$

where $f_d(w)$ is term frequency of word w in the document d, $f_D(w)$ is the number of documents that contain word w and $|D|$ is the number of documents in the collection D.

4.3 Latent Semantic Analysis

Latent semantic analysis [19] is an unsupervised learning method based on semantic representation of the text. It was earlier proposed by Gong and Liu [20] for single document and multi document summarization in the news domain. It assumes that semantically similar words occur in same piece of text. A word-sentence matrix is constructed from the large volume of text, where each row corresponds to a word from the input (n words) and each column corresponds to a sentence (m sentences). Each entry w_{ij} of the matrix is the weight of the word i in sentence j. These weights are computed by TF-IDF weighing scheme and if a word is not present in a particular sentence then that corresponding entry is kept zero.

Singular Value Decomposition (SVD) is used to decrease the number of rows of the matrix while preserving the similarity structure. The SVD of an m×n matrix is given as follows:

$$SVD(A) = U\Sigma V^T$$

Matrix U represents a word-topic matrix. Matrix Σ is a diagonal matrix where each row corresponds to the weight of a topic corresponding to that row. Matrix V^T is the topic-sentence matrix. The matrix D = ΣV^T describes how much a sentence represent a topic, thus, d_{ij} shows the weight of the topic i in sentence j. More advanced variants of LSA technique can be seen here, [21][22][23]

4.4 Bayesian Topic Models

The above given methods have some limitations. These methods do not consider probabilistic approach to topic representation but instead these methods consider sentences as independent of each other and sentence scores are calculated using heuristics instead of having clear probabilistic interpretations. In this way, information is lost by considering the sentences independent of each other.

As the name suggests, bayesian topic models consider probabilistic approach to topic representation. These ability of representing the documents in details help the summarizer to determine the similarities and differences between the documents used in the summarization. [24]

KL divergence [25] is an amazing measure for scoring sentences because it is based on the fact that summaries produced should tend to be similar to the input documents. The KL is a measure of divergence between two probability distributions P and Q [26]. Given two probability distributions P and Q, the KL divergence for words w can be given as follows:

$$D_{KL}(P||Q) = \sum_{w} P(w) log \frac{P(w)}{Q(w)}$$

It describes how related are the words present in the summary to the words present in the input document, i.e. the KL divergence of a good summary and the input will be low and the KL divergence of a bad summary and the input will be high. So, low KL divergence scores are always preferred for summarization.

4.5 Clustering

Clustering is a simple yet very effective unsupervised learning algorithm in machine learning. It clusters data based on the Euclidean distance between data points. We can use different features to measure sentence similarity to form clusters accordingly. The number of clusters to be formed has to be set beforehand. Clustering can be useful for unsupervised extractive text summarization. In this method, we create sentence encoding vectors for the sentences then we apply clustering. According to the number of clusters, the sentences are grouped. We choose the centroid sentence vector of each cluster. Our assumption is that the centroid sentence is the best sentence that represents the whole cluster. We select the centroid sentence vectors of each cluster as summary sentences. This gives an important sentence from each topic of the input text. Thus, resulting in better summarization.

Chapter 5

The Influence of Context

In order to get better summaries, additional information associated with the context of summarization can be helpful. For example, citations for scientific papers can be useful for generating the summary of scientific articles. Comments on a blog post can be used to find the sentences better representing the context for summarization. These additional information about the text to be summarized can be leveraged to find the importance of different sentences in the text. We define some of the contexts in more detail in the following sections:

5.1 Query-focused Summarization

In query-focused summarization, user has a query related to the input and two factors are considered for this kind of summarization. Firstly, how relevant is the sentence to user asked query and secondly, how well it represents the topic of the input. Above given approaches for topic signature words can be extended to query-focused summarization. For example, the probability of a word appearing in the summary is zero if it is neither present in the query nor it is a topic signature word. A word has probability of 0.5, if it either appears in the query or it is a topic signature word. A word has probability 1 of appearing in the summary if it is present in the query and it is a topic signature word.

Graph based approaches [3] and machine learning approaches have also been implemented to find the patterns for selection of sentences for query-based summarization.

5.2 Web Summarization

Web pages are too complex to be summarized. The textual information they have is often scarce, which makes the summarization, a difficult task. One way to do it is by collecting the sentences from other web pages which are linking to that web page. They might contain useful information about the web page. One of the earlier research in this field is [27], where they fetch the pages having link to that specific web page by using query search engines. Then they analyze the sentences of these web pages and select the best sentences using heuristics. Meishan et al. [28] proposed the summarization by using context from the comments of the blog. More research on web summarization can be found here, [29][30][31].

5.3 Summarization of Scientific articles

Scientific summarization, often called citation based summarization is based on finding other papers that cite the scientific paper to be summarized. Based on the sentences used to describe the citation of the paper, we can find the context of the scientific paper that is cited. Mie et al. [25] propose a language model which gives probability of each word in the sentences which are in context of the citation of our original scientific paper. They get a probability scores of the words and then they use KL divergence [26] to score the importance of the sentences.

Chapter 6

Indicator Representation Approaches

Indicator representation approaches are different from topic representation approaches because they consider different set of features for representation of text rather than representing topic of the text to be summarized. The question arises that how can we get better representation of the text. Graph-based methods and machine learning methods come into picture to get better sentences representations in earlier days. Now-a-days, different deep learning frameworks, like [1] [32], are deployed to get better syntactic and semantic sentence representations. These sentence representations work well in the unsupervised text summarization task where sentence representation plays a major role in determining which sentences will be selected to form the summary.

6.1 Supervised Learning

6.1.1 Machine Learning Techniques

Early works in using machine learning techniques for summarization was done by Kupiec et al. [33]. They used Naive Bayes classifier [34] in which sentences are selected according to a number of features and the classifier is trained on input text data and the extractive summaries corresponding to the text, i.e., labelled data. The classification probabilities are learned statistically using Bayes' rule.

The calculated probability of a sentence to belong to the summary is given as score of the sentence to compete with other sentences for summarization.

Some of the key features used in that paper [33] are title words in the sentence, position of the sentence in the input text, presence of upper-case letters, sentence length, etc. Other widely used machine learning techniques in summarization, [35][36].

Naive Bayes [34], decision trees [37], Hidden Markov models [38], support vector machines

[39] and Conditional Random Fields [40] are widely used machine learning techniques for summarization. Most of the machine learning classifier assume that the sentences are independent that leads to loss of information whereas Hidden Markov Models [38] and Conditional Random Fields [40] often outperforms other techniques because of considering dependency between sentences.

The major disadvantage of using machine learning techniques is that they need labeled training data to train the classifier which may not be the case, always.

6.1.1.1 Semi-supervised learning approach

Wong et al. [41] proposed a method of semi supervised learning for extractive text summarization. Semi-supervised learning approaches are good choice when we have limited labelled data. In semi-supervised learning, we make use of unlabelled data with small amount of labelled data. The classifier is first trained with the labelled data then it is tested on unlabelled data and the samples with top scores are considered for training along with the labelled data in the next iteration. In this process, we get more labelled data with each iteration. The reliability of the classifier trained on unlabelled data(which is labelled by the classifier itself) depends on a number of other factors.

6.2 Unsupervised learning

6.2.1 Graph-based Methods

Graph-based methods, such as TextRank [3], represent the text document as a connected graph, where each node of the graph represents the sentences and each edge represents the similarity between the two accompanying nodes. The question arises that how do we represent the sentences as the nodes and how do we define the weights of the edges connecting the nodes. Most common technique that is used for sentence representation is TF-IDF weights of words present in the sentence and the weights of the edges are measured by applying cosine similarity. A common technique to check the connection between two sentences is by deciding a threshold from your model, if the sentence similarity is greater than certain threshold then the two nodes are connected otherwise they are not. This helps in making sub-graphs of the entire document graph that leads to classification of the documents into different topics. The sentence with more number of connections is likely to be center of the the text document and it is more likely to be included in the summary. One key thing that needs to be taken into account is that how well the sentence representation represents the syntactic and semantic relations of the sentence because this is the base on which the above classification is done.

Most of the recent researches have applied TF-IDF weights on different graph-based techniques. We apply Universal language Model Fine-tuning, ULM-Fit [1] sentence encoding for sentence representation which has both, syntactic and semantic information of the sentence and it has proved to be better than TF-IDF weights which consider only frequency of the words in the sentence.

Chapter 7

Application on Hindi Language

7.1 TF-IDF Method

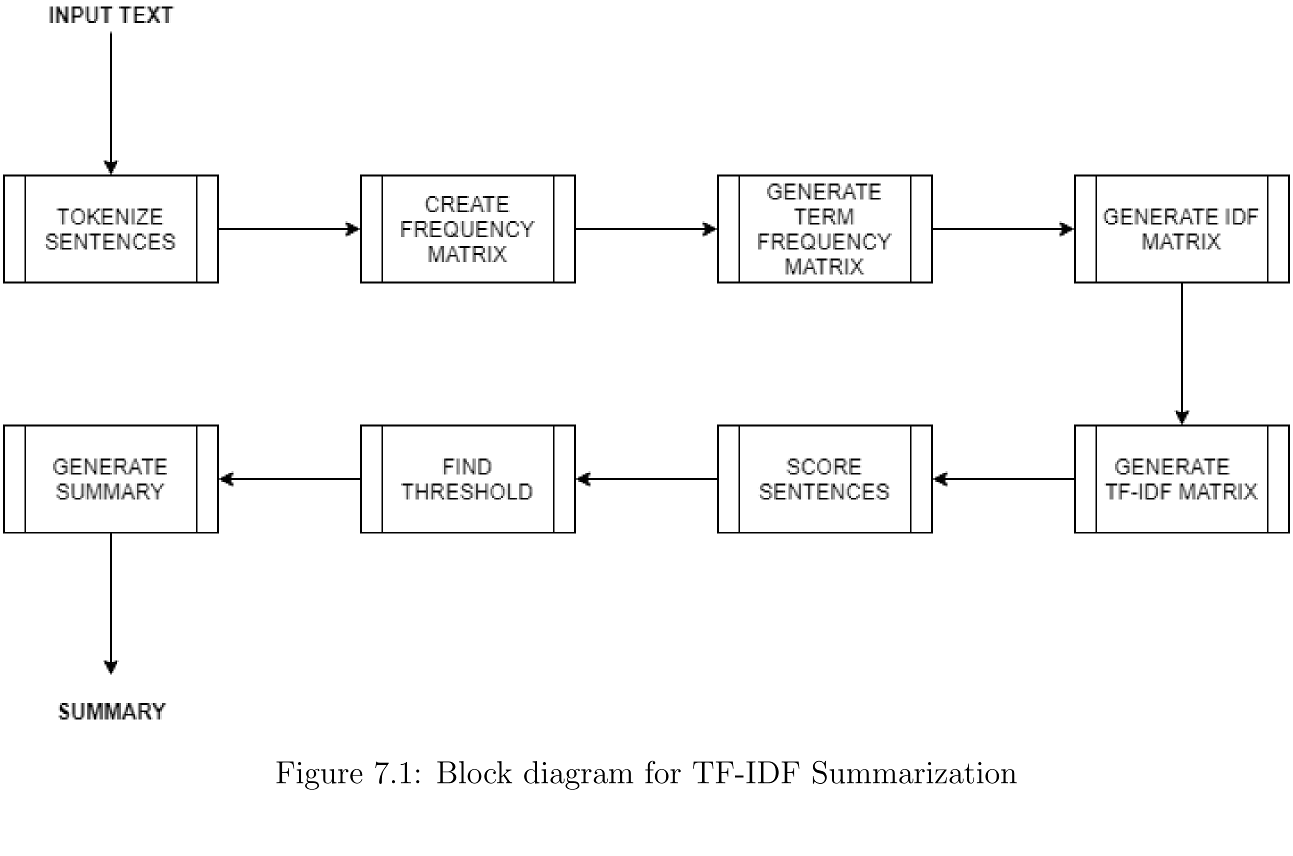


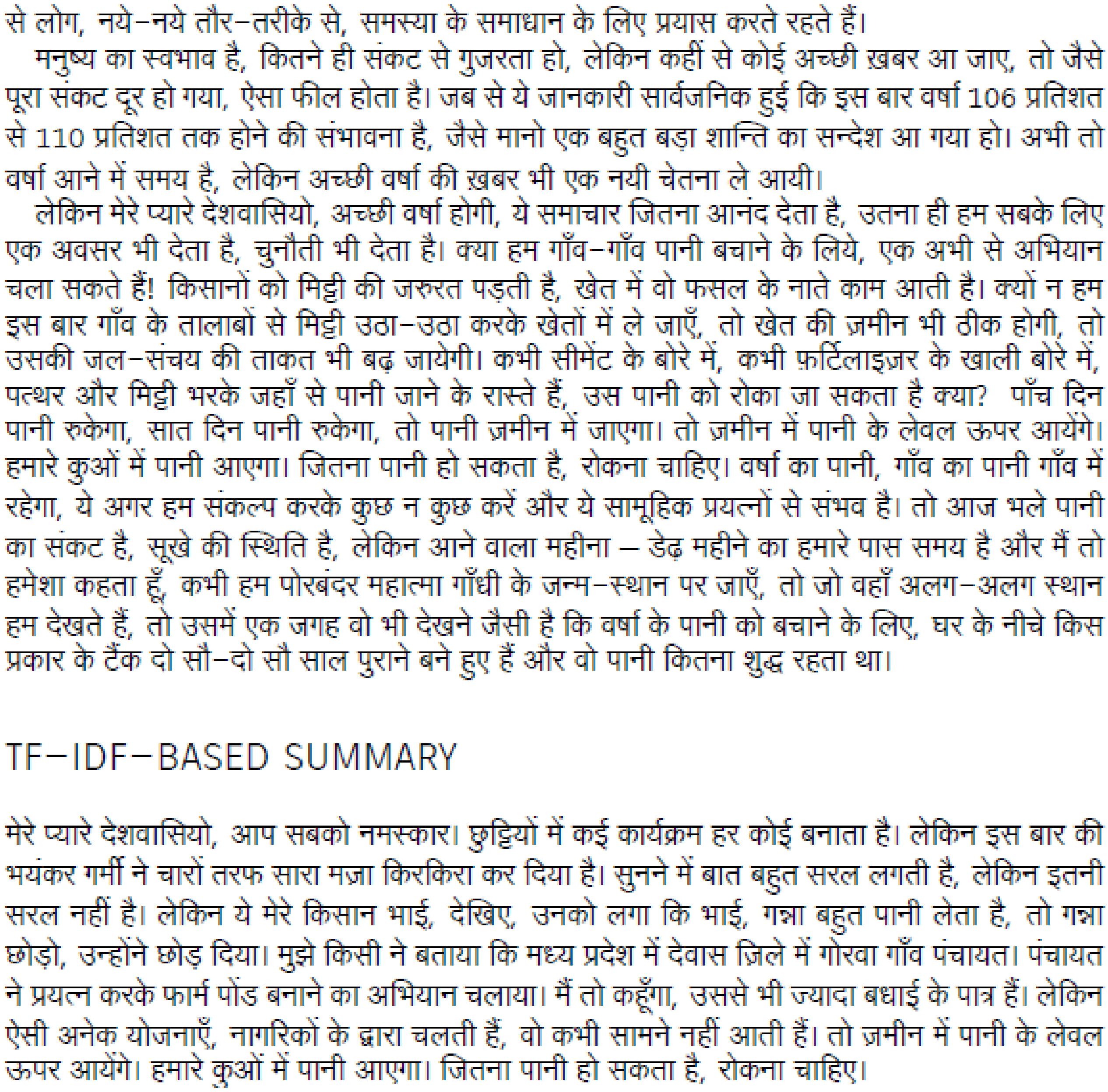
Figure 7.1: Block diagram for TF-IDF Summarization

We used TF-IDF features for text summarization on Hindi text. The text has been taken from the speech of our beloved Prime Minister, Mr. Narendra Modi on “Man ki baat”. The threshold for sentence selection has been chosen as the average sentence score and the sentence selection for generating the summary is done using sentences with sentence score above the threshold value. The text and the summary are as follows:

TEXT

मेरे प्यारे देशवासियो, आप सबको नमस्कार। छुट्टियों में कई कार्यक्रम हर कोई बनाता है। और छुट्टियों में आम का सीजन होता है, तो ये भी मन करता है कि आम का मज़ा लें और कभी ये भी मन करता है कि कुछ पल दोपहर को सोने का मौका मिल जाए, तो अच्छा होगा। लेकिन इस बार की भयंकर गर्मी ने चारों तरफ सारा मज़ा किरकिरा कर दिया है। देश में चिंता होना बहुत स्वाभाविक है और उसमें भी, जब लगातार सूखा पड़ता है, तो पानी–संग्रह के जो स्थान होते हैं, वो भी कम पड़ जाते हैं। कभी–कभार एन्क्रोचमेंट के कारण, सिल्टिंग के कारण, पानी आने के जो प्रवाह हैं, उसमें रुकावटों के कारण, जलाशय भी अपनी क्षमता से काफी कम पानी संग्रहीत करते हैं और सालों के क्रम के कारण उसकी संग्रह–क्षमता भी कम हो जाती है। सूखे से निपटने के लिए पानी के संकट से राहत के लिए सरकारें अपना प्रयास करें, वो तो है, लेकिन मैंने देखा है कि नागरिक भी बहुत ही अच्छे प्रयास करते हैं। कई गाँवों में जागरूकता देखी जाती है और पानी का मूल्य क्या है, वो तो वही जानते हैं, जिन्होनें पानी की तकलीफ़ झेली है। और इसलिए ऐसी जगह पर, पानी के संबंध में एक संवेदनशीलता भी होती है और कुछ–न–कुछ करने की सक्रियता भी होती है। मुझे कुछ दिन पहले कोई बता रहा था कि महाराष्ट्र के अहमदनगर ज़िले के हिवरे बाज़ार ग्राम पंचायत और वहाँ के गाँव वालों ने पानी को गाँव के एक बहुत बड़े संवेदनशील इशू के रूप में एड्रेस किया। जल संचय करने की इच्छा करने वाले तो कई गाँव मिल जाते हैं, लेकिन इन्होंने तो किसानों के साथ बातचीत करके पूरी क्रॉपिंग पैटर्न बदल दी। ऐसी फसल, जो सबसे ज्यादा पानी उपयोग करती थी, चाहे गन्ना हो, केला हो, ऐसी फसलों को छोड़ने का निर्णय कर लिया। सुनने में बात बहुत सरल लगती है, लेकिन इतनी सरल नहीं है। सबने मिल करके कितना बड़ा संकल्प किया होगा?   किसी कारख़ाना वाला पानी का उपयोग करता हो, कहोगे, तुम कारख़ाना बंद करो, क्योंकि पानी ज्यादा लेते हो, तो क्या परिणाम आएगा, आप जानते हैं। लेकिन ये मेरे किसान भाई, देखिए, उनको लगा कि भाई, गन्ना बहुत पानी लेता है, तो गन्ना छोड़ो, उन्होंने छोड़ दिया। और पूरा उन्होंने फ्रूट और वेजिटेबल, जिसमें कम–से–कम पानी की ज़रूरत पड़ती है, ऐसी फसलों पर चले गए। उन्होंने स्प्रिंकलर , ड्रिप इरीगेशन , टपक सिंचाई, वाटर हार्वेस्टिंग , वाटर रिचार्जिंग – इतने सारे इनिशिएटिव लिये कि आज गाँव पानी के संकट के सामने जूझने के लिए अपनी ताकत पर खड़ा हो गया। ठीक है, मैं एक छोटे से गाँव हिवरे बाज़ार की चर्चा भले करता हूँ, लेकिन ऐसे कई गाँव होंगे। मैं ऐसे सभी गाँववासियों को भी बहुत–बहुत बधाई देता हूँ आपके इस उत्तम काम के लिए।

मुझे किसी ने बताया कि मध्य प्रदेश में देवास ज़िले में गोरवा गाँव पंचायत। पंचायत ने प्रयत्न करके फार्म पोंड बनाने का अभियान चलाया। करीब 27 फार्म पोंडस बनाए और उसके कारण ग्राउंड वाटर लेवल में बढ़ोत्तरी हुई, पानी ऊपर आया। जब भी पानी की ज़रूरत पड़ी फ़सल को, पानी मिला और वो मोटा–मोटा हिसाब बताते थे, करीब उनकी कृषि उत्पादन में 20 प्रतिशत वृद्धि हुई। तो पानी तो बचा ही बचा और जब पानी का वाटर टेबल ऊपर आता है, तो पानी की क्वालिटी में भी बहुत सुधार होता है। और दुनिया में ऐसा कहते हैं, शुद्ध पीने का पानी जीडीपी ग्रोथ का कारण बन जाता है, स्वास्थ्य का तो बनता ही बनता है। कभी–कभार तो लगता है कि जब भारत सरकार रेलवे से पानी लातूर पहुँचाती है, तो दुनिया के लिए वो एक ख़बर बन जाती है। ये बात सही है कि जिस तेज़ी से रेलवे ने काम किया, वो बधाई की पात्र तो है, लेकिन वो गाँव वाले भी उतने ही बधाई के पात्र हैं। मैं तो कहूँगा, उससे भी ज्यादा बधाई के पात्र हैं। लेकिन ऐसी अनेक योजनाएँ, नागरिकों के द्वारा चलती हैं, वो कभी सामने नहीं आती हैं। सरकार की अच्छी बात तो कभी–कभी सामने आ भी जाती है, लेकिन कभी हम अपने अगल–बगल में देखेंगे, तो ध्यान में आएगा कि सूखे के खिलाफ़ किस–किस प्रकार



से लोग, नये–नये तौर–तरीके से, समस्या के समाधान के लिए प्रयास करते रहते हैं।

मनुष्य का स्वभाव है, कितने ही संकट से गुजरता हो, लेकिन कहीं से कोई अच्छी ख़बर आ जाए, तो जैसे पूरा संकट दूर हो गया, ऐसा फील होता है। जब से ये जानकारी सार्वजनिक हुई कि इस बार वर्षा 106 प्रतिशत से 110 प्रतिशत तक होने की संभावना है, जैसे मानो एक बहुत बड़ा शान्ति का सन्देश आ गया हो। अभी तो वर्षा आने में समय है, लेकिन अच्छी वर्षा की ख़बर भी एक नयी चेतना ले आयी।

लेकिन मेरे प्यारे देशवासियो, अच्छी वर्षा होगी, ये समाचार जितना आनंद देता है, उतना ही हम सबके लिए एक अवसर भी देता है, चुनौती भी देता है। क्या हम गाँव–गाँव पानी बचाने के लिये, एक अभी से अभियान चला सकते हैं! किसानों को मिट्टी की जरुरत पड़ती है, खेत में वो फसल के नाते काम आती है। क्यों न हम इस बार गाँव के तालाबों से मिट्टी उठा–उठा करके खेतों में ले जाएँ, तो खेत की ज़मीन भी ठीक होगी, तो उसकी जल–संचय की ताकत भी बढ़ जायेगी। कभी सीमेंट के बोरे में, कभी फ़र्टिलाइज़र के खाली बोरे में, पत्थर और मिट्टी भरके जहाँ से पानी जाने के रास्ते हैं, उस पानी को रोका जा सकता है क्या? पाँच दिन पानी रुकेगा, सात दिन पानी रुकेगा, तो पानी ज़मीन में जाएगा। तो ज़मीन में पानी के लेवल ऊपर आयेंगे। हमारे कुओं में पानी आएगा। जितना पानी हो सकता है, रोकना चाहिए। वर्षा का पानी, गाँव का पानी गाँव में रहेगा, ये अगर हम संकल्प करके कुछ न कुछ करें और ये सामूहिक प्रयत्नों से संभव है। तो आज भले पानी का संकट है, सूखे की स्थिति है, लेकिन आने वाला महीना – डेढ़ महीने का हमारे पास समय है और मैं तो हमेशा कहता हूँ, कभी हम पोरबंदर महात्मा गाँधी के जन्म–स्थान पर जाएँ, तो जो वहाँ अलग–अलग स्थान हम देखते हैं, तो उसमें एक जगह वो भी देखने जैसी है कि वर्षा के पानी को बचाने के लिए, घर के नीचे किस प्रकार के टैंक दो सौ–दो सौ साल पुराने बने हुए हैं और वो पानी कितना शुद्ध रहता था।

TF–IDF–BASED SUMMARY

मेरे प्यारे देशवासियो, आप सबको नमस्कार। छुट्टियों में कई कार्यक्रम हर कोई बनाता है। लेकिन इस बार की भयंकर गर्मी ने चारों तरफ सारा मज़ा किरकिरा कर दिया है। सुनने में बात बहुत सरल लगती है, लेकिन इतनी सरल नहीं है। लेकिन ये मेरे किसान भाई, देखिए, उनको लगा कि भाई, गन्ना बहुत पानी लेता है, तो गन्ना छोड़ो, उन्होंने छोड़ दिया। मुझे किसी ने बताया कि मध्य प्रदेश में देवास ज़िले में गोरवा गाँव पंचायत। पंचायत ने प्रयत्न करके फार्म पोंड बनाने का अभियान चलाया। मैं तो कहूँगा, उससे भी ज्यादा बधाई के पात्र हैं। लेकिन ऐसी अनेक योजनाएँ, नागरिकों के द्वारा चलती हैं, वो कभी सामने नहीं आती हैं। तो ज़मीन में पानी के लेवल ऊपर आयेंगे। हमारे कुओं में पानी आएगा। जितना पानी हो सकता है, रोकना चाहिए।

Figure 7.2: A sample input and its summary using TF-IDF weights. There are 42 sentences in the input. ROUGE-1, ROUGE-2, ROUGE-l scores are 0.37, 0.2 and 0.3, respectively. The same input text is used with all the implemented techniques to illustrate their outputs, which are given in subsequent figures.

The topic of the text given above is “decreasing water level in the fields and what techniques can be implemented to raise the ground water level.”

Since, there is no perfect evaluation metric for text summarization, the accuracy of the summarizer is debatable.

In our point of view, the first sentence in the summary is redundant. It has least relevance to the topic of concern. TF-IDF features are not able to capture the topic of the text and does not give satisfactory results. Thus, we need better representation of the text and better algorithms to summarize the text. In the following chapters, we will be discussing about the graph based methods for summarization. We have implemented the graph based methods and clustering methods on hindi text example considered above with better representation vectors for the sentence and the results are impressive.

7.1.1 Evaluation

Since, there are no reference summaries for the given text, we created 10 reference summaries from 10 different people who know Hindi literature well. We calculated the average evaluation scores from these summaries as shown below:

ROUGE-N	**F-Score**	**Precision**	**Recall**
ROUGE-1	0.37	0.61	0.27
ROUGE-2	0.2	0.34	0.15
ROUGE-L	0.3	0.42	0.23

Table 7.1: ROUGE-N evaluation for Hindi text summarization using TF-IDF features using threshold for sentence selection as the average sentence score. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

In ROUGE-L, “L” stands for longest common subsequence. An advantage of using longest common subsequence is that, it require in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, we don’t need to redefine n-gram length.

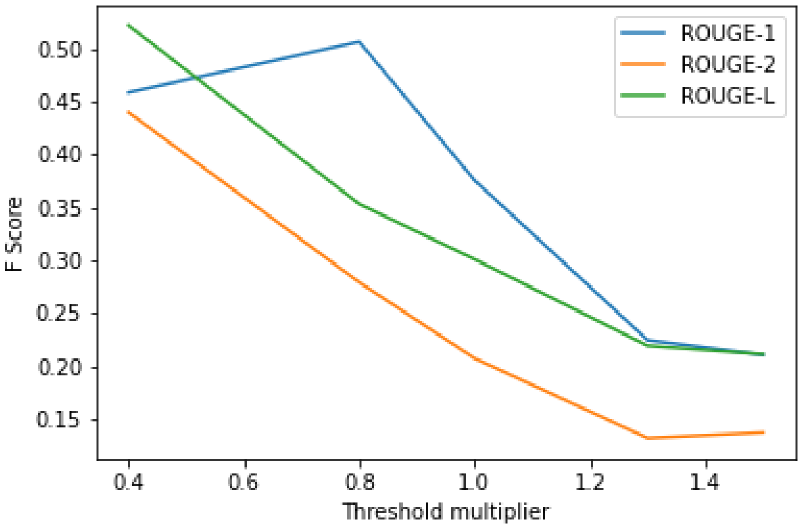


Figure 7.3: [Best viewed in color] F1 Score using TF-IDF features using different threshold multipliers. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

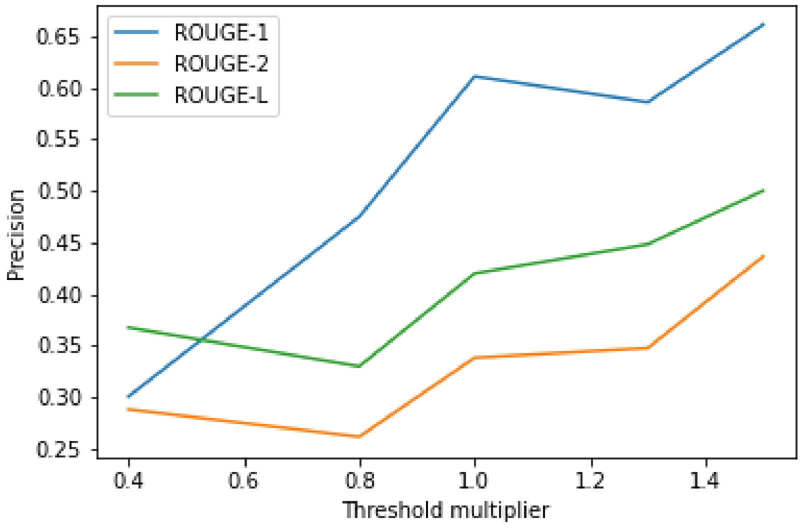


Figure 7.4: [Best viewed in color] Precision using TF-IDF features using different threshold multipliers. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

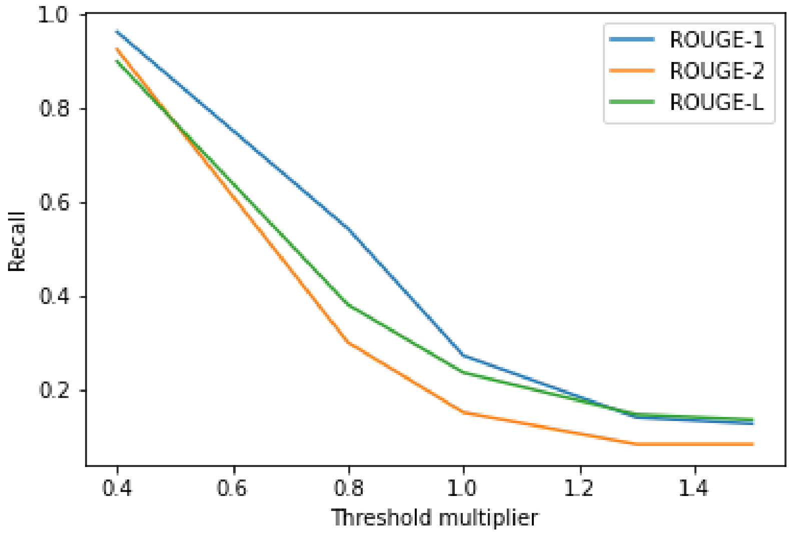


Figure 7.5: [Best viewed in color] Recall using TF-IDF features using different threshold multipliers. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

7.2 Clustering Method

In this method, we use ULM-Fit [1] sentence embedding vectors for Hindi data-set rather than TF-IDF features that are used in the previous work. These embedding vectors are trained on Hindi Wikipedia dataset [5]. We took the pre-trained model and fine-tuned it on our dataset of Prime Minister Narendra Damodardas Modi’s speech on “Man ki baat” in order to get better sentence representations.

We followed the approach given in [2]. Below is the block diagram of our model:

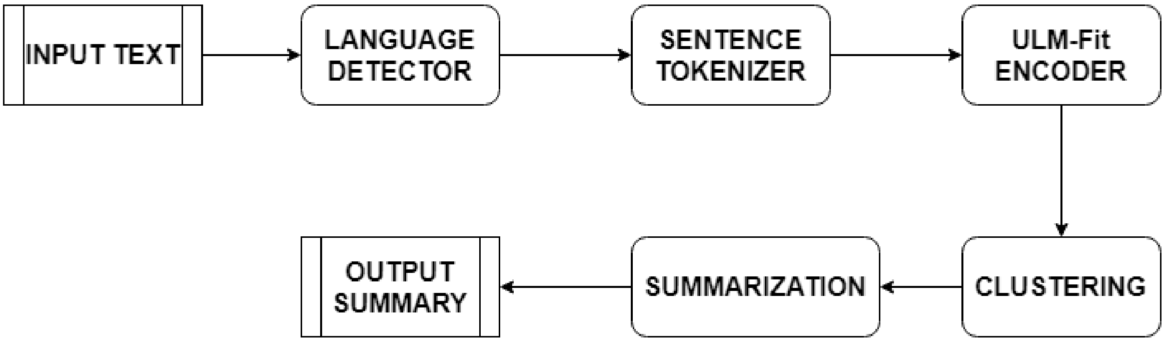
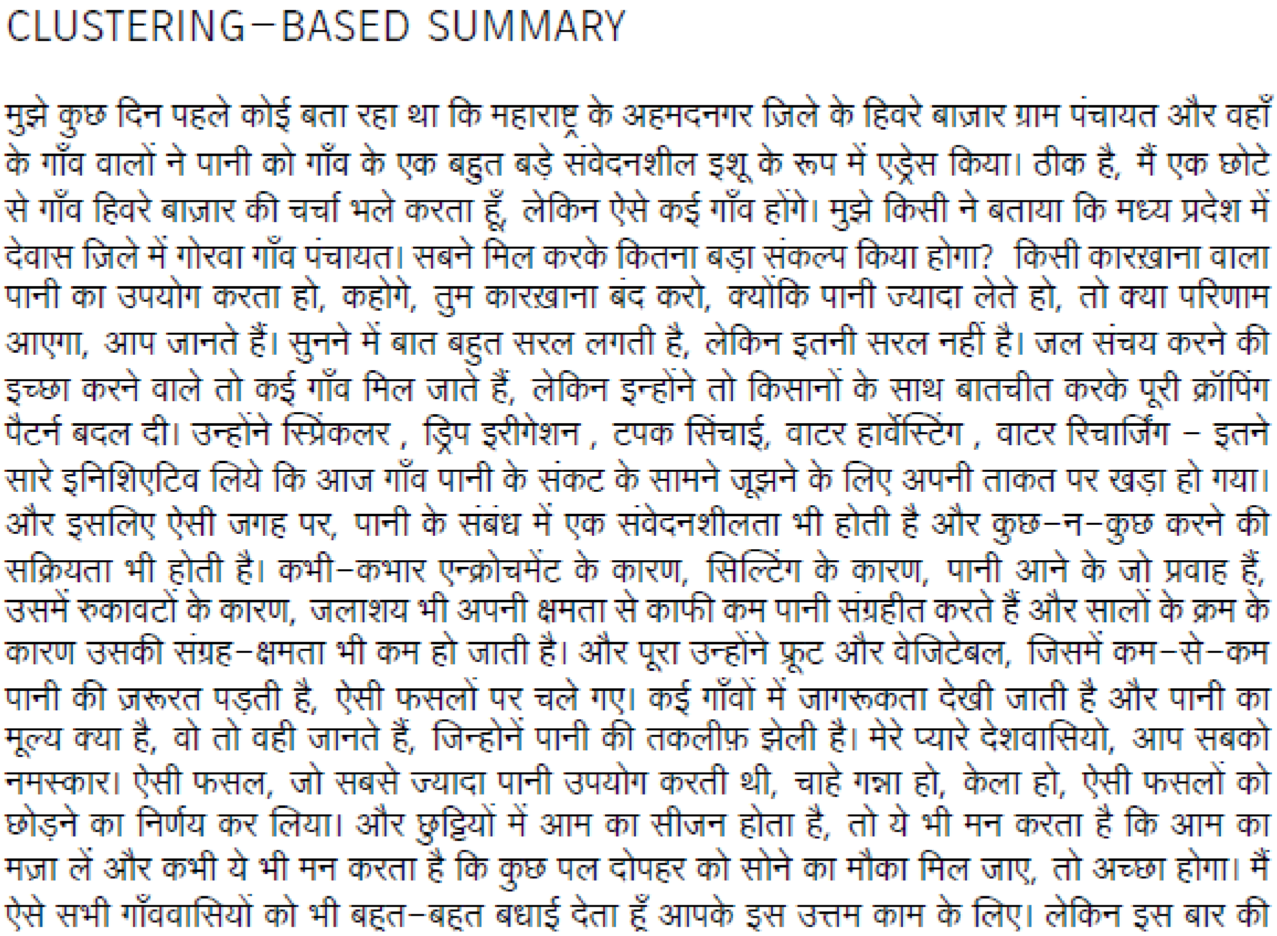


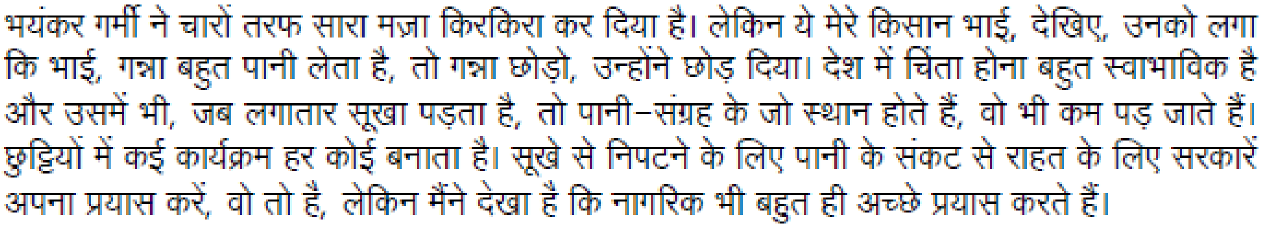
Figure 7.6: Block diagram for Cluster-based Summarization using ULM-Fit [1] sentence encoding for vector representation

In figure 7.6, Language detector is a toolkit used to determine the language of the text. Sentence tokenizer splits the text into constituent sentences. ULM-Fit encoder [1] encodes the sentences into 400 dimensional vectors. The encoded sentences of the text are clustered together based on sentence similarity. The sentences corresponding to the sentence embedding closest to the cluster centers are chosen as the summary sentences. We used ULM-Fit [1] sentence encoding for text summarization on Hindi text. The text has been taken from the speech of our beloved Prime Minister, Mr. Narendra Modi on ”Man ki baat”. The summary length has been chosen as 50%. The summary for the above given text sample in 7.2 is as follows:



CLUSTERING−BASED SUMMARY

मुझे कुछ दिन पहले कोई बता रहा था कि महाराष्ट्र के अहमदनगर ज़िले के हिवरे बाज़ार ग्राम पंचायत और वहाँ के गाँव वालों ने पानी को गाँव के एक बहुत बड़े संवेदनशील इशू के रूप में एड्रेस किया। ठीक है, मैं एक छोटे से गाँव हिवरे बाज़ार की चर्चा भले करता हूँ, लेकिन ऐसे कई गाँव होंगे। मुझे किसी ने बताया कि मध्य प्रदेश में देवास ज़िले में गोरवा गाँव पंचायत। सबने मिल करके कितना बड़ा संकल्प किया होगा? किसी कारख़ाना वाला पानी का उपयोग करता हो, कहोगे, तुम कारख़ाना बंद करो, क्योंकि पानी ज्यादा लेते हो, तो क्या परिणाम आएगा, आप जानते हैं। सुनने में बात बहुत सरल लगती है, लेकिन इतनी सरल नहीं है। जल संचय करने की इच्छा करने वाले तो कई गाँव मिल जाते हैं, लेकिन इन्होंने तो किसानों के साथ बातचीत करके पूरी क्रॉपिंग पैटर्न बदल दी। उन्होंने स्प्रिंकलर , ड्रिप इरीगेशन , टपक सिंचाई, वाटर हार्वेस्टिंग , वाटर रिचार्जिंग – इतने सारे इनिशिएटिव लिये कि आज गाँव पानी के संकट के सामने जूझने के लिए अपनी ताकत पर खड़ा हो गया। और इसलिए ऐसी जगह पर, पानी के संबंध में एक संवेदनशीलता भी होती है और कुछ–न–कुछ करने की सक्रियता भी होती है। कभी–कभार एन्क्रोचमेंट के कारण, सिल्टिंग के कारण, पानी आने के जो प्रवाह हैं, उसमें रुकावटों के कारण, जलाशय भी अपनी क्षमता से काफी कम पानी संग्रहीत करते हैं और सालों के क्रम के कारण उसकी संग्रह–क्षमता भी कम हो जाती है। और पूरा उन्होंने फ्रूट और वेजिटेबल, जिसमें कम–से–कम पानी की ज़रूरत पड़ती है, ऐसी फसलों पर चले गए। कई गाँवों में जागरूकता देखी जाती है और पानी का मूल्य क्या है, वो तो वही जानते हैं, जिन्होंने पानी की तकलीफ़ झेली है। मेरे प्यारे देशवासियो, आप सबको नमस्कार। ऐसी फसल, जो सबसे ज्यादा पानी उपयोग करती थी, चाहे गन्ना हो, केला हो, ऐसी फसलों को छोड़ने का निर्णय कर लिया। और छुट्टियों में आम का सीजन होता है, तो ये भी मन करता है कि आम का मज़ा लें और कभी ये भी मन करता है कि कुछ पल दोपहर को सोने का मौका मिल जाए, तो अच्छा होगा। मैं ऐसे सभी गाँववासियों को भी बहुत–बहुत बधाई देता हूँ आपके इस उत्तम काम के लिए। लेकिन इस बार की



भयंकर गर्मी ने चारों तरफ सारा मज़ा किरकिरा कर दिया है। लेकिन ये मेरे किसान भाई, देखिए, उनको लगा कि भाई, गन्ना बहुत पानी लेता है, तो गन्ना छोड़ो, उन्होंने छोड़ दिया। देश में चिंता होना बहुत स्वाभाविक है और उसमें भी, जब लगातार सूखा पड़ता है, तो पानी–संग्रह के जो स्थान होते हैं, वो भी कम पड़ जाते हैं। छुट्टियों में कई कार्यक्रम हर कोई बनाता है। सूखे से निपटने के लिए पानी के संकट से राहत के लिए सरकारें अपना प्रयास करें, वो तो है, लेकिन मैंने देखा है कि नागरिक भी बहुत ही अच्छे प्रयास करते हैं।

Figure 7.7: A sample input 7.2 and its summary using Clustering Algorithm [2] and ULM-Fit encoding [1]. There are 42 sentences in the input. The summary length is 50% of the input text. ROUGE-1, ROUGE-2, ROUGE-l scores are 0.5, 0.32 and 0.28, respectively.

7.2.1 Evaluation

We calculated the average evaluation scores from the 10 reference summaries as shown below:

ROUGE-N	**F-Score**	**Precision**	**Recall**
ROUGE-1	0.5	0.41	0.65
ROUGE-2	0.32	0.26	0.42
ROUGE-L	0.28	0.24	0.35

Table 7.2: ROUGE-N evaluation for Hindi text summarization using ULM-Fit [1] features using clustering-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

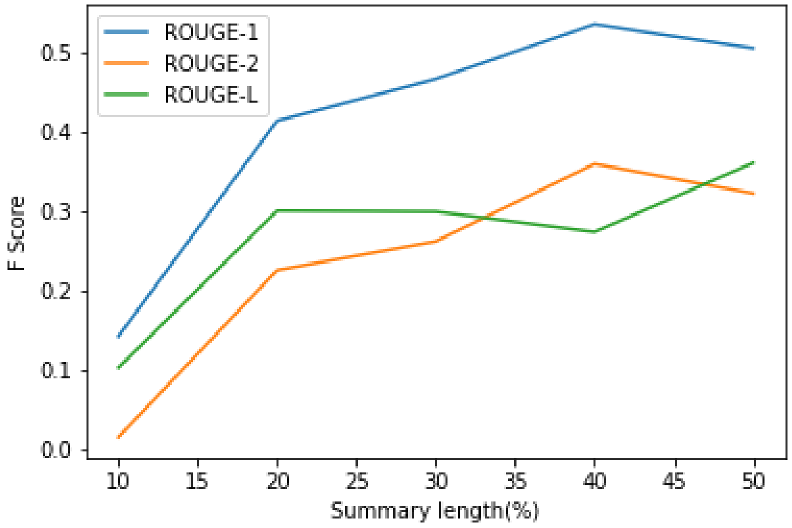


Figure 7.8: [Best viewed in color] F1 Score using ULM-Fit sentence encoding [1] and clustering method for different summary lengths. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

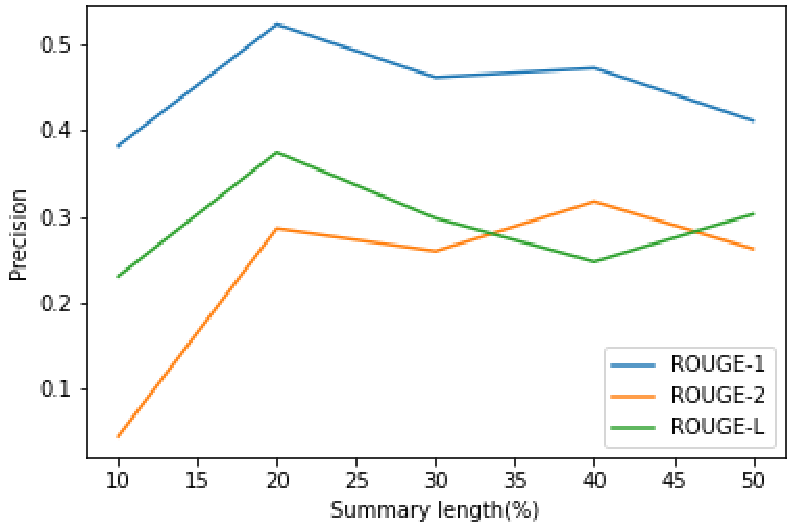


Figure 7.9: [Best viewed in color] Precision using ULM-Fit sentence encoding [1] and clustering method for different summary lengths. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Modi on “Man ki Baat”. The input text contains 42 sentences.

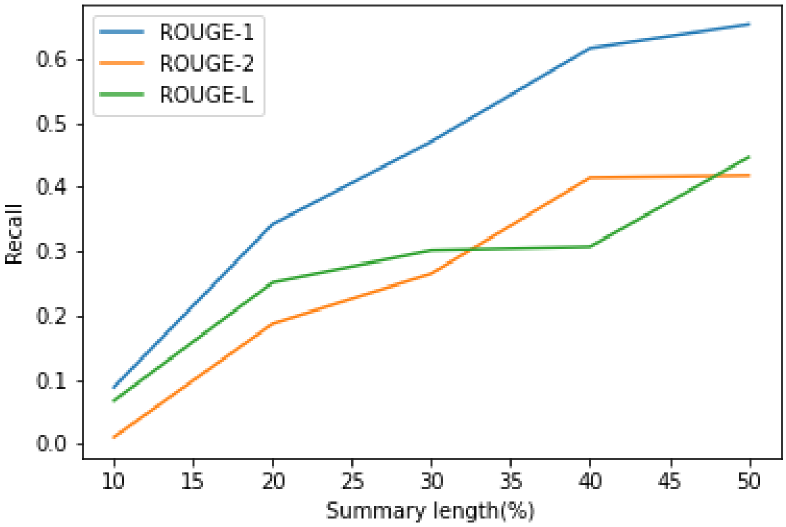


Figure 7.10: [Best viewed in color] Recall using ULM-Fit sentence encoding [1] and clustering method for different summary lengths. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

7.3 TextRank Method (Graph-based)

We used ULM-Fit [1] sentence encoding for text summarization on Hindi text. The text has been taken from the speech of our beloved Prime Minister, Mr. Narendra Modi on "Man ki baat". The algorithm is inspired by the Page-Rank algorithm used by google for ranking pages. In Text-Rank algorithm [3], we treat sentence vectors as nodes of the graph and the similarity between those sentences as weights of the edges. The ranking is based on the score of a particular sentence according to the connectivity of the sentence vector in the graph. The most connected sentence is ranked as the first sentence and the least connected sentence is given the last rank. In the previous work [3], they have used TF-IDF weight vectors for sentence representation whereas we use ULM-Fit sentence encoding which keeps track of semantic as well as syntactic relationship of the sentences.

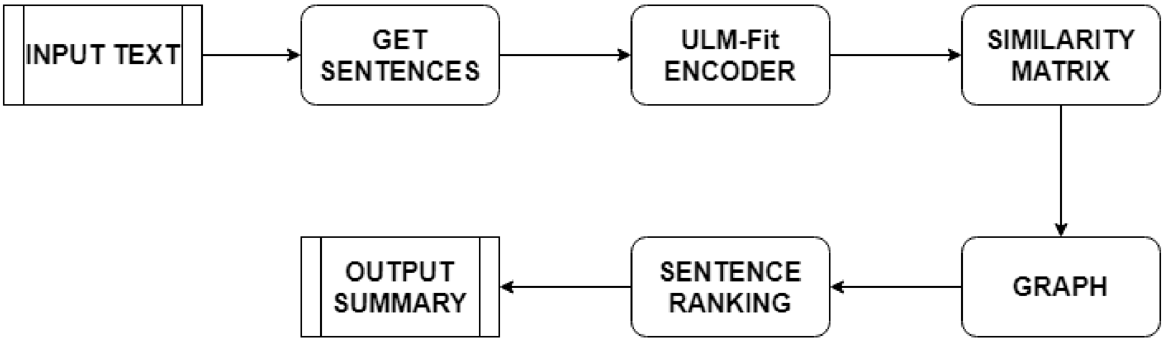
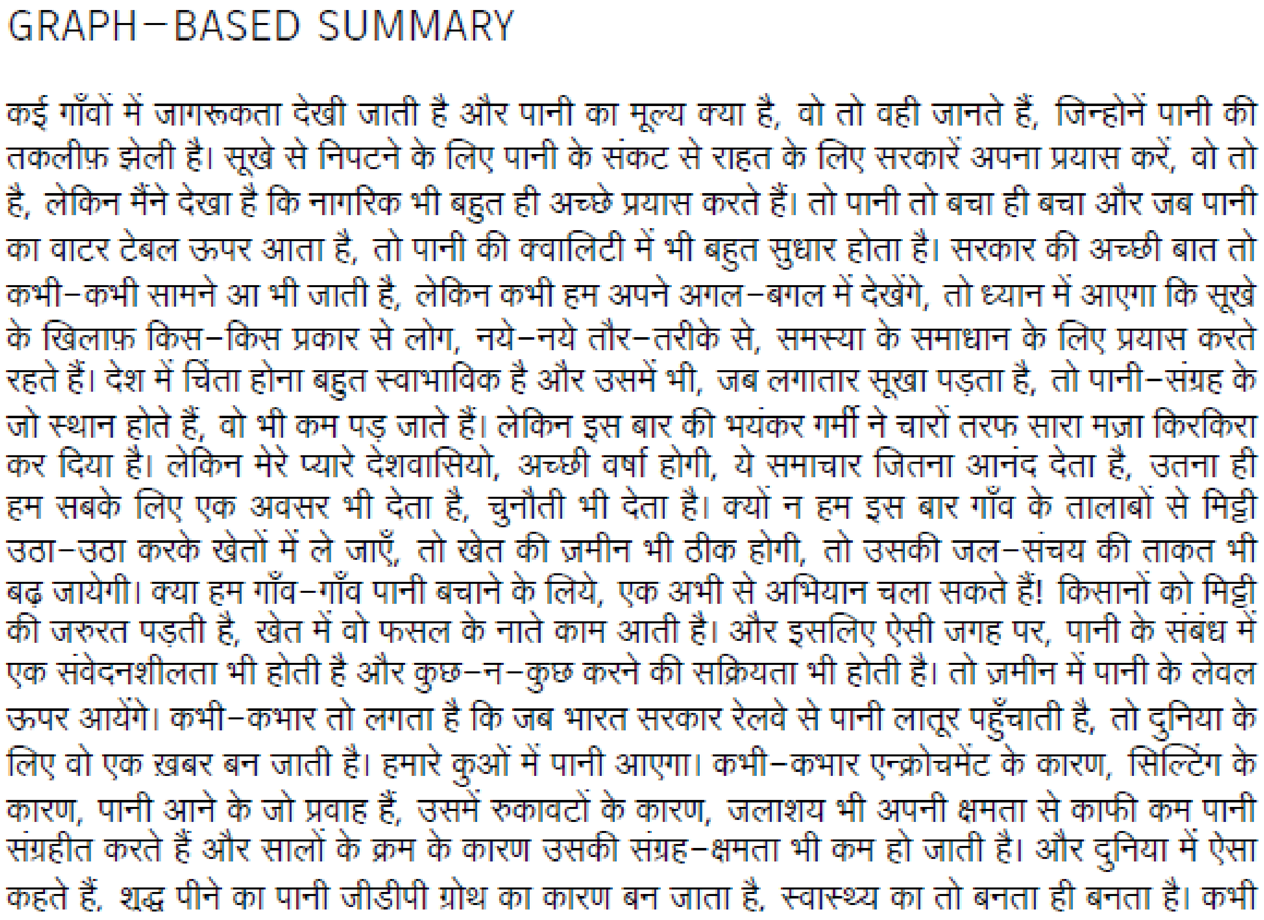


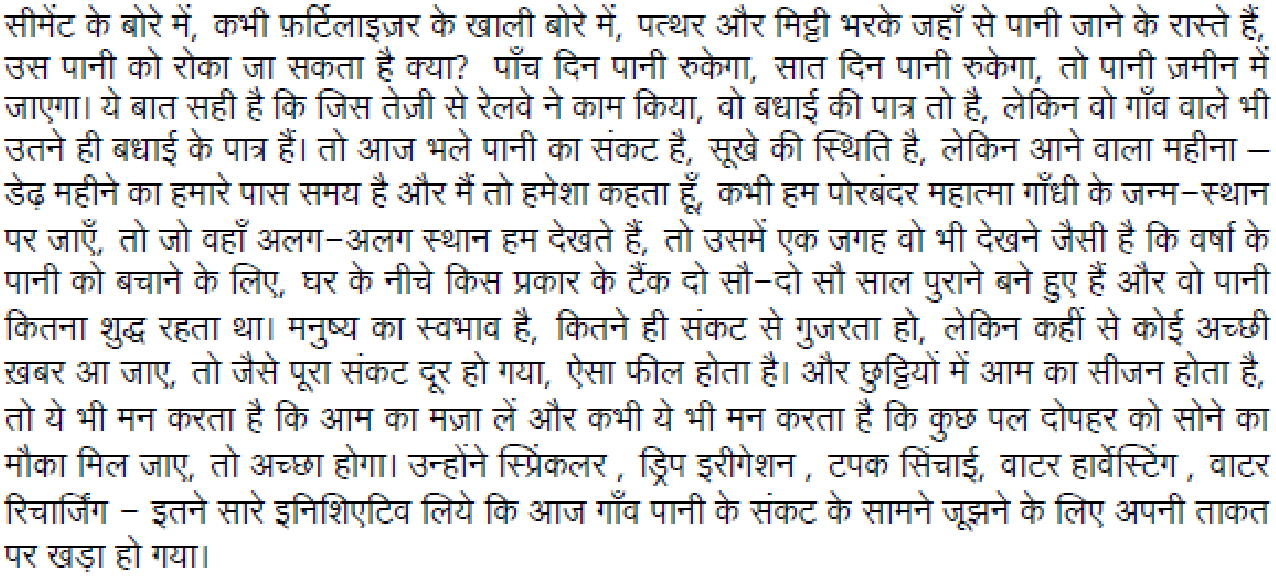
Figure 7.11: Block diagram for Graph-based Summarization using ULM-Fit [1] sentence encoding for vector representation

The summary length has been chosen as 50%. The summary for the above given text sample in 7.2 is as follows:



GRAPH−BASED SUMMARY

कई गाँवों में जागरूकता देखी जाती है और पानी का मूल्य क्या है, वो तो वही जानते हैं, जिन्होनें पानी की तकलीफ़ झेली है। सूखे से निपटने के लिए पानी के संकट से राहत के लिए सरकारें अपना प्रयास करें, वो तो है, लेकिन मैंने देखा है कि नागरिक भी बहुत ही अच्छे प्रयास करते हैं। तो पानी तो बचा ही बचा और जब पानी का वाटर टेबल ऊपर आता है, तो पानी की क्वालिटी में भी बहुत सुधार होता है। सरकार की अच्छी बात तो कभी−कभी सामने आ भी जाती है, लेकिन कभी हम अपने अगल−बगल में देखेंगे, तो ध्यान में आएगा कि सूखे के खिलाफ़ किस−किस प्रकार से लोग, नये−नये तौर−तरीके से, समस्या के समाधान के लिए प्रयास करते रहते हैं। देश में चिंता होना बहुत स्वाभाविक है और उसमें भी, जब लगातार सूखा पड़ता है, तो पानी−संग्रह के जो स्थान होते हैं, वो भी कम पड़ जाते हैं। लेकिन इस बार की भयंकर गर्मी ने चारों तरफ सारा मज़ा किरकिरा कर दिया है। लेकिन मेरे प्यारे देशवासियो, अच्छी वर्षा होगी, ये समाचार जितना आनंद देता है, उतना ही हम सबके लिए एक अवसर भी देता है, चुनौती भी देता है। क्यों न हम इस बार गाँव के तालाबों से मिट्टी उठा−उठा करके खेतों में ले जाएँ, तो खेत की ज़मीन भी ठीक होगी, तो उसकी जल−संचय की ताकत भी बढ़ जायेगी। क्या हम गाँव−गाँव पानी बचाने के लिये, एक अभी से अभियान चला सकते हैं! किसानों को मिट्टी की जरुरत पड़ती है, खेत में वो फसल के नाते काम आती है। और इसलिए ऐसी जगह पर, पानी के संबंध में एक संवेदनशीलता भी होती है और कुछ−न−कुछ करने की सक्रियता भी होती है। तो ज़मीन में पानी के लेवल ऊपर आयेंगे। कभी−कभार तो लगता है कि जब भारत सरकार रेलवे से पानी लातूर पहुँचाती है, तो दुनिया के लिए वो एक ख़बर बन जाती है। हमारे कुओं में पानी आएगा। कभी−कभार एन्क्रोचमेंट के कारण, सिल्टिंग के कारण, पानी आने के जो प्रवाह हैं, उसमें रुकावटों के कारण, जलाशय भी अपनी क्षमता से काफी कम पानी संग्रहीत करते हैं और सालों के क्रम के कारण उसकी संग्रह−क्षमता भी कम हो जाती है। और दुनिया में ऐसा कहते हैं, शुद्ध पीने का पानी जीडीपी ग्रोथ का कारण बन जाता है, स्वास्थ्य का तो बनता ही बनता है। कभी



सीमेंट के बोरे में, कभी फ़र्टिलाइज़र के खाली बोरे में, पत्थर और मिट्टी भरके जहाँ से पानी जाने के रास्ते हैं, उस पानी को रोका जा सकता है क्या? पाँच दिन पानी रुकेगा, सात दिन पानी रुकेगा, तो पानी ज़मीन में जाएगा। ये बात सही है कि जिस तेज़ी से रेलवे ने काम किया, वो बधाई की पात्र तो है, लेकिन वो गाँव वाले भी उतने ही बधाई के पात्र हैं। तो आज भले पानी का संकट है, सूखे की स्थिति है, लेकिन आने वाला महीना – डेढ़ महीने का हमारे पास समय है और मैं तो हमेशा कहता हूँ, कभी हम पोरबंदर महात्मा गाँधी के जन्म–स्थान पर जाएँ, तो जो वहाँ अलग–अलग स्थान हम देखते हैं, तो उसमें एक जगह वो भी देखने जैसी है कि वर्षा के पानी को बचाने के लिए, घर के नीचे किस प्रकार के टैंक दो सौ–दो सौ साल पुराने बने हुए हैं और वो पानी कितना शुद्ध रहता था। मनुष्य का स्वभाव है, कितने ही संकट से गुजरता हो, लेकिन कहीं से कोई अच्छी ख़बर आ जाए, तो जैसे पूरा संकट दूर हो गया, ऐसा फील होता है। और छुट्टियों में आम का सीजन होता है, तो ये भी मन करता है कि आम का मज़ा लें और कभी ये भी मन करता है कि कुछ पल दोपहर को सोने का मौका मिल जाए, तो अच्छा होगा। उन्होंने स्प्रिंकलर , ड्रिप इरीगेशन , टपक सिंचाई, वाटर हार्वेस्टिंग , वाटर रिचार्जिंग – इतने सारे इनिशिएटिव लिये कि आज गाँव पानी के संकट के सामने जूझने के लिए अपनी ताकत पर खड़ा हो गया।

Figure 7.12: A sample input 7.2 and its summary using TextRank Algorithm [3] and ULM-Fit encoding [1]. There are 42 sentences in the input. The summary length is 50% of the input text. ROUGE-1, ROUGE-2, ROUGE-l scores are 0.53, 0.45 and 0.45, respectively.

7.3.1 Evaluation

We calculated the average evaluation scores from the 10 reference summaries as shown below:

ROUGE-N	F-Score	Precision	Recall
ROUGE-1	0.53	0.39	0.84
ROUGE-2	0.45	0.33	0.7
ROUGE-L	0.45	0.35	0.61

Table 7.3: ROUGE-N evaluation for Hindi text summarization using ULM-Fit [1] features using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

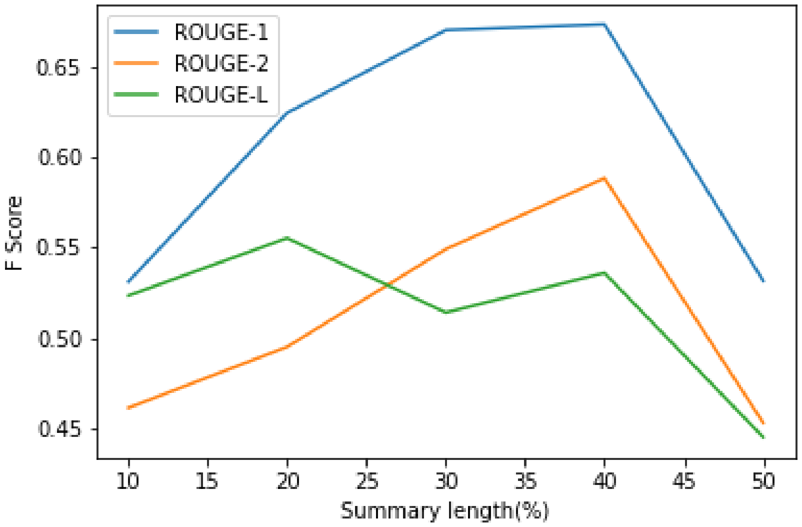


Figure 7.13: [Best viewed in color] F1 Score using ULM-Fit sentence encoding [1] and graph-based method for different summary lengths. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

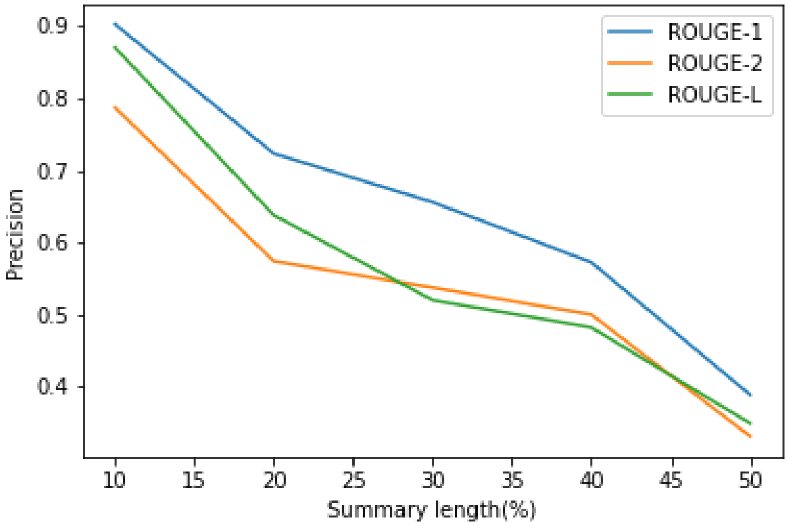


Figure 7.14: [Best viewed in color] Precision using ULM-Fit sentence encoding [1] and graph-based method for different summary lengths. The same text is taken from the speech of the Prime Minister of India, Mr. Narendra Modi on "Man ki Baat". The input text contains 42 sentences.

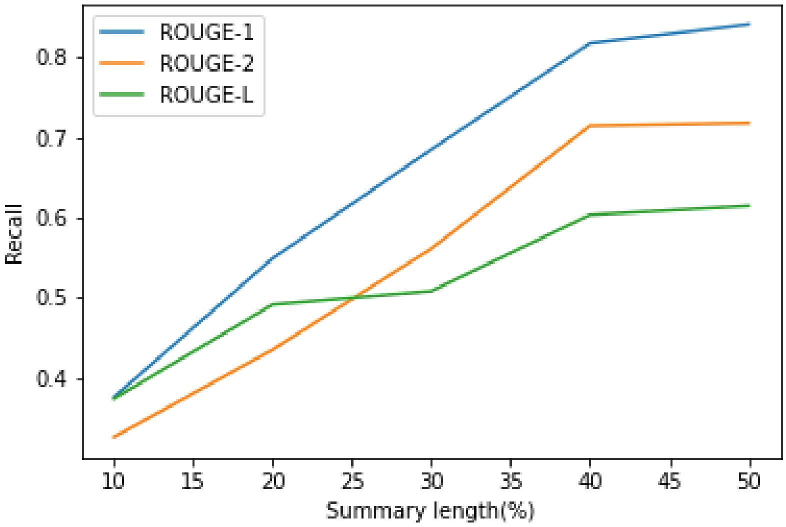


Figure 7.15: [Best viewed in color] Recall using ULM-Fit sentence encoding [1] and graph-based method for different summary lengths. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

7.4 A Combination of Maximum likelihood Estimate Sentence Embeddings with TextRank Method(Graph-based) [1] [4]

A very simple but effective method to get the sentence representation vectors which beats most of the naive deep learning frameworks was given by Arora et. al. [4]. The word vectors used in the algorithm are trained on the whole speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The dataset contains 4398 sentences with a word vocabulary size of 12079 words. The word vectors are then used to construct the sentence embeddings and producing summaries as given in the algorithm 1.

```
Algorithm 1: Text Summarization using MLE sentence embedding and Text-Rank
  Input: Word embeddings $\{v_w : w \in \mathcal{V}\}$ a set of sentences S, parameter a and
         estimated probabilities $\{P_w : w \in \mathcal{V}\}$
  Output: Summary
  for all sentences s in $\mathcal{S}$ do
      $v_s \leftarrow \frac{1}{|S|} \sum_{w \in S} \frac{a}{a+p(w)}$ ;
      endfor ;
  Form a matrix $\mathcal{X}$ whose columns are $\{v_s : s \in \mathcal{S}\}$, and let u be its first singular vector ;
  for all sentences s in S do
      $v_s \leftarrow v_s - uu^T v_s$ ;
      endfor ;
  SIM-MAT ⟵ COSINE-SIMILARITY($v_s$);
  SCORE ⟵ PAGE-RANK(SIM-MAT);
  RANKED SENTENCES ⟵ SORT($\mathcal{S}$);
  return RANKED SENTENCES ;
```

Algorithm 1 summarizes our setup. In Algorithm 1, $\mathcal{S}$ denotes the set of sentences, P_w denotes the word occurrence probability and $\mathcal{V}$ denotes the word vocabulary.

This sentence embedding method performs better than the normal word average method. The weighing term $a/(a + p(w))$ is known as *smooth inverse frequency.* These simple yet useful embedding perform better than some of the deep learning frameworks in a number of tasks. The first singular vector of the matrix has been removed from all the vectors for common component removal.

7.4.1 Evaluation

We calculated the average evaluation scores from the 10 reference summaries as shown below:

ROUGE-N	F-Score	Precision	Recall
ROUGE-1	0.495	0.39	0.67
ROUGE-2	0.25	0.20	0.35
ROUGE-L	0.276	0.235	0.335

Table 7.4: ROUGE-N evaluation for Hindi text summarization using MLE sentence embeddings [4] using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

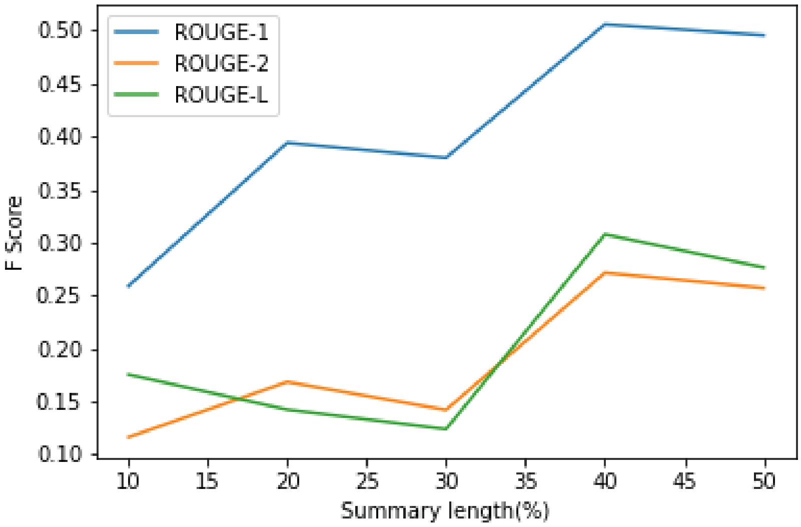


Figure 7.16: [Best viewed in color] F1 Score using MLE sentence embeddings [4] using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

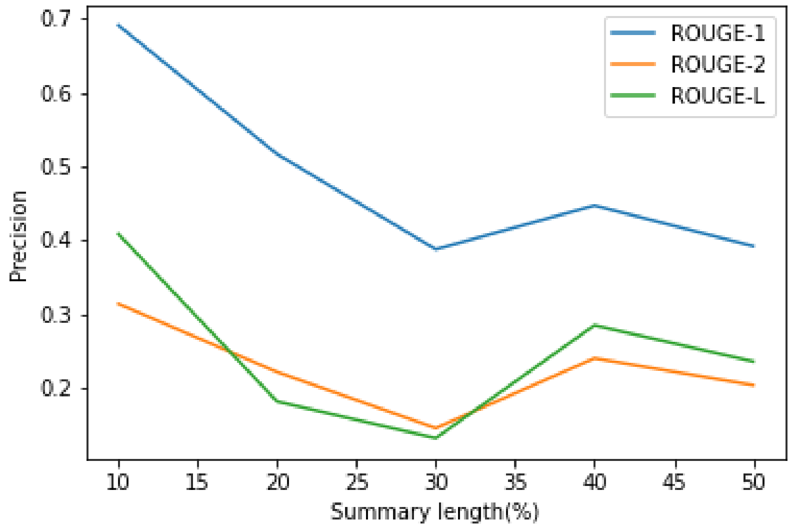


Figure 7.17: [Best viewed in color] Precision using MLE sentence embeddings [4] using graph-based summarization for 50% summary length. The same text is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

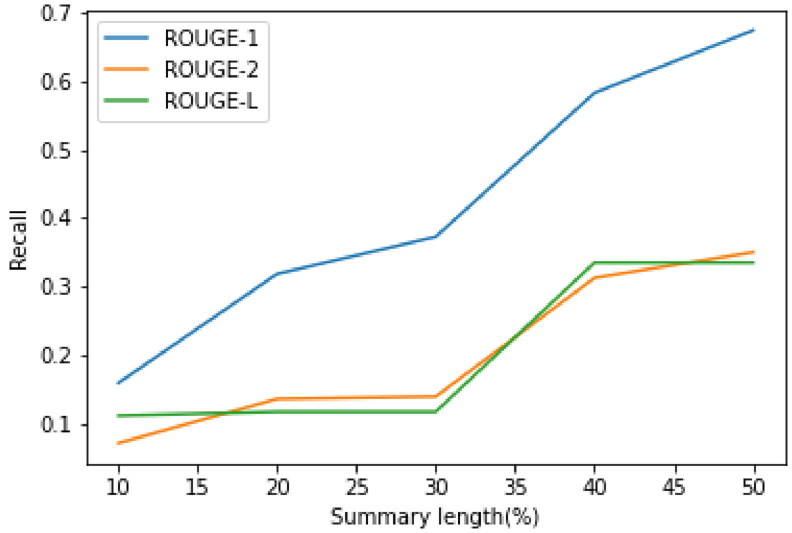


Figure 7.18: [Best viewed in color] Recall using MLE sentence embeddings [4] using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

7.5 A Combination of Word average with Text-Rank Method(Graph-based)

This is the naive approach for producing sentence embedding. In this approach, we make use of word vectors. The word vectors used in the algorithm are trained on the whole speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The dataset contains 4398 sentences with a word vocabulary size of 12079 words. The word vectors are obtained according to Word2Vec paper [42]. The word vectors of all the words of a sentence are calculated using Word2Vec model trained before. All the word vectors of a particular sentence are added and divided by the length of the sentence. This is called word-average sentence embeddings. We applied text-rank algorithm on these sentence embeddings and performed summarization.

7.5.1 Evaluation

We calculated the average evaluation scores from the 10 reference summaries as shown below:

ROUGE-N	**F-Score**	**Precision**	**Recall**
ROUGE-1	0.5	0.38	0.73
ROUGE-2	0.36	0.27	0.53
ROUGE-L	0.27	0.22	0.36

Table 7.5: ROUGE-N evaluation for Hindi text summarization using word-avg sentence embeddings using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

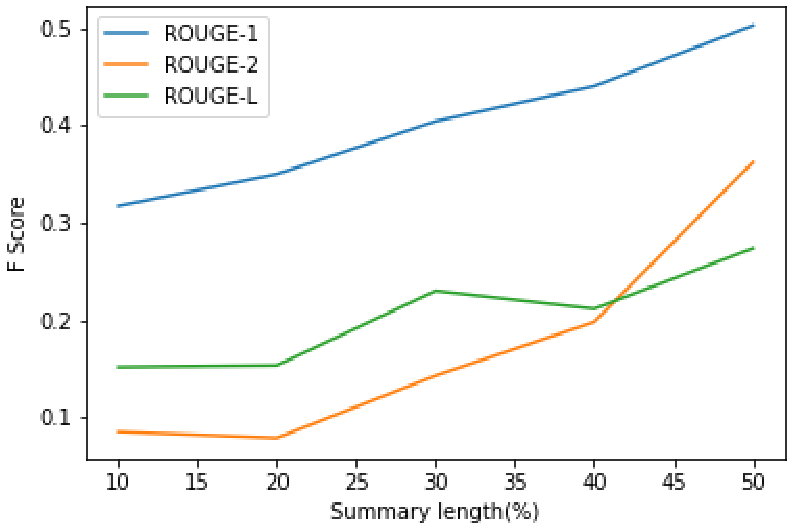


Figure 7.19: [Best viewed in color] F1 Score using word-avg sentence embeddings using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

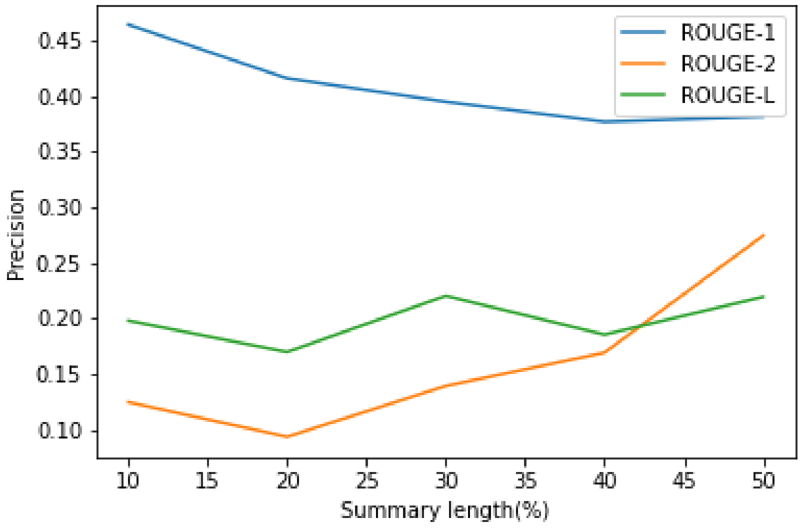


Figure 7.20: [Best viewed in color] Precision using word-avg sentence embeddings using graph-based summarization for 50% summary length. The same text is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on "Man ki Baat". The input text contains 42 sentences.

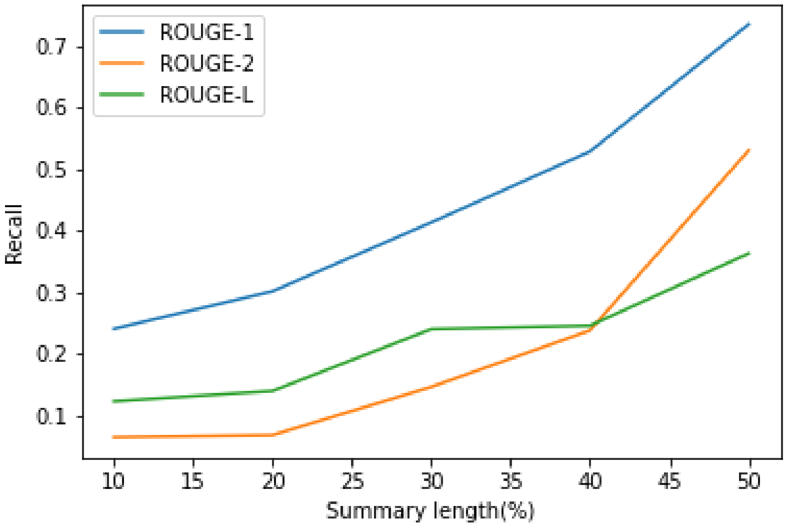


Figure 7.21: [Best viewed in color] Recall using word-avg sentence embeddings using graph-based summarization for 50% summary length. The same text has been used which is taken from the speech of the Prime Minister of India, Mr. Narendra Damodardas Modi on “Man ki Baat”. The input text contains 42 sentences.

Chapter 8

Conclusion and Future Work

In this work, we applied different combinations of the known extractive text summarization techniques and sentence embedding techniques on Hindi. We found that the sentence embeddings with more semantic and syntactic information give better results than other techniques. We also observed that simple statistical techniques are good enough to compete with deep learning techniques. A major drawback in this research was that we did not get supervised dataset for text summarization of Indian languages. We feel that there is a strong need for the same. One use-case of extractive text summarization is to divide the text into different topics. This can be used for a speech recognition model with huge vocabulary size. The problem of huge vocabulary can be narrowed down to some extent if we know the topic of the speech to be spoken by the user at the recognition time.

References

[1] Jeremy Howard and Sebastian Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[2] Aishwarya Padmakumar and Akanksha Saran, "Unsupervised text summarization using sentence embeddings," Tech. Rep., Technical Report, University of Texas at Austin, 2016.

[3] Rada Mihalcea and Paul Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma, "A simple but tough-to-beat baseline for sentence embeddings," 2016.

[5] Gaurav: Goru001, "Hindi wikipedia dataset," https://www.kaggle.com/disisbig/hindi-wikipedia-articles-172k, May 2018.

[6] Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown, "Introduction to the special issue on summarization," *Comput. Linguist.*, vol. 28, no. 4, pp. 399–408, Dec. 2002.

[7] Hans Peter Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[8] Harold P Edmundson, "New methods in automatic extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 1969.

[9] Günes Erkan and Dragomir R Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[10] Elena Lloret and Manuel Palomar, "Text summarisation in progress: a literature review," *Artificial Intelligence Review*, vol. 37, no. 1, pp. 1–41, 2012.

[11] Horacio Saggion and Thierry Poibeau, “Automatic text summarization: Past, present and future,” in *Multi-source, multilingual information extraction and summarization*, pp. 3–21. Springer, 2013.

[12] Ani Nenkova and Kathleen McKeown, “A survey of text summarization techniques,” in *Mining text data*, pp. 43–76. Springer, 2012.

[13] Chin-Yew Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.

[14] Ted Dunning, “Accurate methods for the statistics of surprise and coincidence,” *Computational linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[15] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova, “Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion,” *Information Processing & Management*, vol. 43, no. 6, pp. 1606–1618, 2007.

[16] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki, “Multi-document summarization by maximizing informative content-words.,” in *IJCAI*, 2007, vol. 7, pp. 1776–1782.

[17] Rasim M Alguliev, Ramiz M Aliguliyev, Makrufa S Hajirahimova, and Chingiz A Mehdiyev, “Mcmr: Maximum coverage and minimum redundant text summarization model,” *Expert Systems with Applications*, vol. 38, no. 12, pp. 14514–14522, 2011.

[18] Gerard Salton and Christopher Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[19] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[20] Yihong Gong and Xin Liu, “Generic text summarization using relevance measure and latent semantic analysis,” in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 19–25.

[21] Makbule Gulcin Ozsoy, Ilyas Cicekli, and Ferda Nur Alpaslan, “Text summarization of turkish texts using latent semantic analysis,” in *Proceedings of the 23rd international*

conference on computational linguistics. Association for Computational Linguistics, 2010, pp. 869–876.

[22] Ben Hachey, Gabriel Murray, and David Reitter, "Dimensionality reduction aids term co-occurrence based multi-document summarization," in *Proceedings of the workshop on task-focused summarization and question answering*, 2006, pp. 1–7.

[23] Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek, "Two uses of anaphora resolution in summarization," *Information Processing & Management*, vol. 43, no. 6, pp. 1663–1680, 2007.

[24] Inderjeet Mani and Eric Bloedorn, "Summarizing similarities and differences among related documents," *Information Retrieval*, vol. 1, no. 1-2, pp. 35–67, 1999.

[25] Qiaozhu Mei and ChengXiang Zhai, "Generating impact-based summaries for scientific literature," in *Proceedings of ACL-08: HLT*, 2008, pp. 816–824.

[26] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[27] Einat Amitay and Cécile Paris, "Automatically summarising web sites: is there a way around it?," in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 173–179.

[28] Meishan Hu, Aixin Sun, and Ee-Peng Lim, "Comments-oriented blog summarization by sentence extraction," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 901–904.

[29] Meishan Hu, Aixin Sun, and Ee-Peng Lim, "Comments-oriented document summarization: understanding documents with readers' feedback," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 291–298.

[30] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita, "Summarizing microblogs automatically," in *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics.* Association for Computational Linguistics, 2010, pp. 685–688.

[31] Beaux P Sharifi, David I Inouye, and Jugal K Kalita, "Summarization of twitter microblogs," *The computer journal*, vol. 57, no. 3, pp. 378–402, 2014.

[32] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011.

[33] Julian Kupiec, Jan Pedersen, and Francine Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 1995, pp. 68–73.

[34] Irina Rish et al., "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, pp. 41–46.

[35] Liang Zhou and Eduard Hovy, "A web-trained extraction summarization system," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 205–211.

[36] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu, "Applying regression models to query-focused multi-document summarization," *Information Processing & Management*, vol. 47, no. 2, pp. 227–237, 2011.

[37] J. Ross Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[38] Lawrence Rabiner and B Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[39] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[40] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, ICML '01, pp. 282–289, Morgan Kaufmann Publishers Inc.

[41] Kam-Fai Wong, Mingli Wu, and Wenjie Li, "Extractive summarization using supervised and semi-supervised learning," in *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, 2008, pp. 985–992.

[42] Yoav Goldberg and Omer Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.