Finding the Objective of a Text Document using Summarization Technique and Cosine Similarity

Project Report Submitted in Partial Fulfilment
of the Requirements for the degree of
Master of Computer Application
Of
Jadavpur University
June, 2022

By
Anupal Patra
Master of Computer Application – III
Examination Roll Number: MCA226020
Registration Number: 149883 of 2019 –2020

Under the guidance of
Dr. CHITRITA CHAUDHURI
Associate Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata – 700032, India
June, 2022

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

I hereby forward the project report entitled **"*Finding the Objective of a Text Document using Summarization Technique and Cosine Similarity*"** prepared by **Anupal Patra, Examination Roll no. - MCA226020** and **Registration no. - 149883 of 2019 –2020** under my supervision to be accepted in partial fulfilment for the degree of **Master of Computer Application** in the Faculty of Engineering and Technology of Jadavpur University, Kolkata.

_____

(Dr. Chitrita Chaudhuri)
Associate
Professor
Project
Supervisor
Dept. of Computer Science and
Engineering Jadavpur University
Kolkata – 700032

Countersigned:

_____

Prof. Anupam Sinha
Head, Dept. of Computer Science and
Engineering Jadavpur University
Kolkata – 700032

_____

Prof. Chandan Mazumdar
Dean, Faculty of Engineering and
Technology Jadavpur University
Kolkata – 70032

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL *

The project report entitled "***Finding the Objective of a Text Document using Summarization Technique and Cosine Similarity***" is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessary endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

_____

Final Examination for evaluation
of the project

_____

(Signatures of Examiners)

* Only in case the project report is approved

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this project report contains literature survey and original research work by undersigned candidate, as part of my Master of Computer Application studies.

All information in this document had been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

NAME	**:**	**Anupal Patra**
Examination Roll Number	**:**	**MCA226020**
Registration Number	**:**	**149883 of 2019-2022**
Project Title	**:**	**Finding the Objective of a Text Document using Summarization Technique and Cosine Similarity**
Signature with Date	**:**	

ACKNOWELDGEMENT

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr. Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this project.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Application course.

A special note of thanks goes to Prof. Anupam Sinha, Head, Department of Computer Science and Engineering, Jadavpur University.

I am also thankful to Prof. Chandan Mazumdar, Dean, Faculty of Engineering and Technology, for providing an excellent environment for completion of this project.

I am also indebted to my co-researchers Mr. Anupam Baidya for his seamless co-operation and help in completion of this project. I am thankful to my fellow classmates and my family for constant help and support.

Date:_____ _____

Anupal Patra
Master of Computer Application – III
Examination Roll No. – MCA226020
Registration No:149883 of 2019 – 2020

Contents

List of Tables

List of Figures

Chapter 1: Introduction

The amount of data available online is massively increasing day by day. It is becoming progressively harder for a user to locate the desired information smoothly and efficiently from any article. The task of finding the right document can be made simpler if one can find out the objective of the document in the form of a concise gist. Such a gist usually involves the technique of summarization. Text summarization is the process which can help to extract information from large documents into a shorter version so that one can understand the topic easily and within reasonable time. In this context, one must remember that the process of manual summarization may often be error prone and beyond the ability of the persons concerned.

In research work one often needs to find out which are the most relevant literatures pertinent to the work in hand [1]. So, it is of utmost importance to locate those works. Usually, it is done by paraphrasing some relevant words. The searching techniques involved is bound to produce too many output documents. The task of scanning through the whole set of documents may become too time consuming. An effective brief on the content would be a better choice. So, in this project work the focus is on finding the brief objective of any research paper using automated summarization techniques supported by available deep learning models.

As already mentioned, Automatic Text Summarization (ATS) is a process where some Machine Learning Algorithms are used to extract the gist of the data. It is a very useful and important part of Natural Language Processing (NLP). NLP is an application of Machine Learning which helps the machine to decipher human texts. ATS comes in two forms: extractive summarization and abstractive summarization. An Extractive Summarizer helps to find the most important sentences from the article which can express the internal meaning of the whole document, while the Abstractive Summarizer creates new sentences and phrases that describe the context of the article in short.

The recent trend is to utilize deep learning models in most areas of Machine Learning systems to mimic the human activity accurately. The deep learning model involved in this work is BERT which stands for Bidirectional Encoder Representation from Transformers. It uses networks where every output element is connected to every input element and the weights of the connections are dynamically calculated by the system. BERT has been used here for two purposes: First, it is used to shortlist the probable objective sentences using extractive summarization; Secondly, it also provides the cosine similarity index between the extracted objective and the pre-existing title of a research paper. The average similarity index observed in this work is 81.07% approximately. This score may hopefully be improved by applying the technique on a larger sample than the one used in the present work.

The dissertation is organized as follows:

The 2nd chapter describes some previous state-of-the-art researches which helped to provide the support for the present work. In the 3rd chapter is discussed some of the theories on which the research is based. In the 4th chapter the methodologies and tools are described. The 5th chapter presents the output obtained from the model built with some illustrative visualization. The last chapter helps to conclude with certain provisions of future enhancements.

Chapter 2: Previous Work

Automatic Text Summarization (ATS) as a part of NLP is an important topic for today's world. However, Researchers have been trying to perform automatic text summarization since late 1950s.

- H.P. Luan et al [1] presented a model to find the abstract of technical literature quickly and accurately. They derived the 'significance factor' of a sentence by using word frequency and the relative position of a word in a particular sentence. Sentences with the highest score are treated as abstract.
- H.P. Edmundson et al [2] introduced a new method in automatic extraction. While the previous work have focused only on sentence significance using word frequency, these researchers described three very important additional components – *Cue words* (like some specific words or phrases), *Title or Heading words* and Structural indicators (sentence position) .

Extractive and Abstractive are the two ways by which a text can be summarized.

- R. Khan et al [3] have described a system of Extractive summarization using TF-IDF(Term Frequency and Inverse Document Frequency) to calculate the overall weight of each sentences and K-Means Clustering technique to find the true K value for generating the summery.
- J.L. Neto et al [4] have built a model of ATS with the help of Naïve Bayes and C4.5 algorithm. C4.5 [5] is a machine learning algorithm which is basically used to generate *Decision Tree Classifier* developed by Ross Quinlan.
- A.R. Pal et al [6] represented a ATS model which is based on Lesk Algorithm and Wordnet. Lesk Algorithm solves the word sense disambiguation and was proposed by Micheal E. Lesk[7] .
- R. Barzilay et al [8] used an Sentence Fusion (an abstractive technique) to summarize new articles .
- D. Miller [9] used BERT for text embedding and K-Means for clustering the lectures of M.I.T

Chapter 3: Background Details

3.1 Artificial Intelligence

Artificial Intelligence (AI) is a science and engineering of making intelligent machines, especially intelligent computer programs. It is related to similar task of using computers to understand human intelligent. It is a field in computer science to develop techniques to enable computer system for performing activities that are considered intelligent. [10] [11] [12]

3.2 Machine Learning

Machine Learning, which is an application of AI has ability to learn automatically from some previous knowledge without any specific programming. It focuses on the development of computer programs that can access data and use it to enhance its knowledge base. First of all, we give some data to the computer so that computer can learn the model and then predict the result for a particular input. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. [13]

Machine Learning algorithms can be classified into two categories, Supervised learning and Unsupervised learning. Supervised learning is a learning technique where we give huge amount of data (properly labelled) to the computer. This is called *training data*. Next, we give some data called *test data* to the computer so that it can return the particular label of that data. Unsupervised learning is the process of building a system using data that is neither classified nor labelled and allowing the algorithm to act on that data without guidance. Here the job of the machine is to group the data according to similarities, patterns and differences without any prior knowledge of class value. [14]

Now a days, Machine Learning are widely used in various industries. Some of them are Natural Language Processing, game playing, biometric identification, financial market analysis etc.

3.3 Natural Language Processing (NLP)

Natural Language Processing (NLP) refers to the branch of Computer Science – more specifically, branch of Machine Learning that has ability to understand text and spoken languages, a human can do. Computational linguistics model is combined with statistical, machine learning, and deep learning models in NLP to allow computers processing human language in the form of text or speech data and ‘understand’ its full meaning, including the speaker’s or writer’s intent and sentiment. Now a days, Enterprise solutions use NLP into streamline corporate operations, enhance employee productivity, and simplify mission-critical business processes. [15]

Some of the use cases of NLP –

1. Spam detection of emails and messages
2. Virtual chatbots
3. Text Summarization
4. Social media Sentiment Analysis

3.4 Tokenization

Tokenization is a fundamental NLP task. It is a technique of breaking down a big sentence into smaller chunks word by word using space division or a regular expression, and it is one of the first steps in converting text to numbers.

Example: The quick brown fox jumps over the lazy dog.

Tokens: [‘The’, ’quick’, ’brown’, ’fox’, ’jumps’, ’over’, ’the’, ‘lazy’, ‘dog’]

3.5 BERT Model

BERT which stands for Bidirectional Encoder Representation from Transformers is a Machine Learning model for NLP. It is highly complex and advanced language model that helps people automate language understanding. It is developed by Google Researchers. BERT is specifically trained on ‘Wikipedia dataset’ and Google’s own ‘Book corpus’ near about 3.3 billion words. It has two versions: $BERT_{base}$ and $BERT_{large}$. BERT is basically an Encoder stack of transformer architecture. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. $BERT_{base}$ has 12 layer of Encoder stack while $BERT_{large}$ has 24 layer of encoder stack. [16]

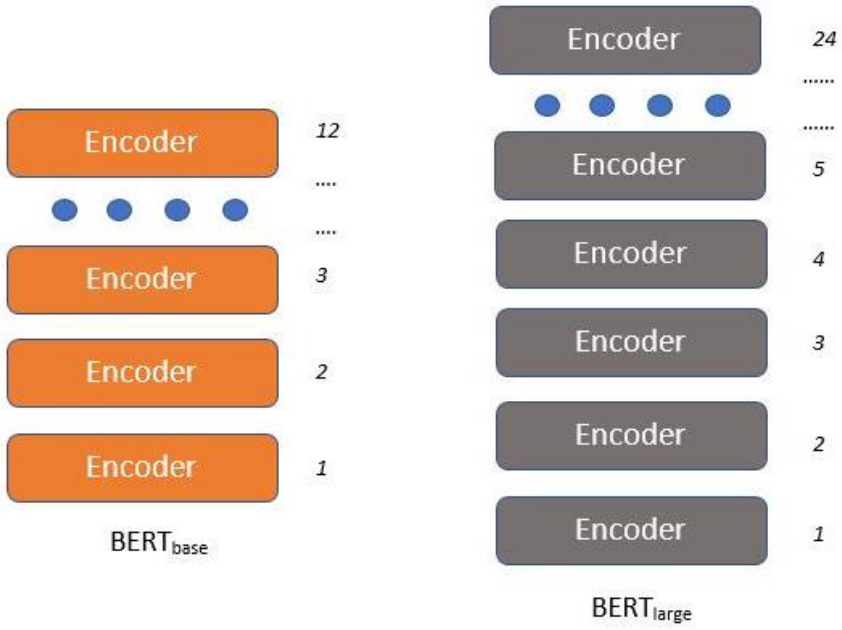

Figure 3.1: BERT Model

3.6 Cosine Similarity

Cosine Similarity is an essential idea in NLP. It assists the cosine of vector angle measurements in computing. The Cosine similarity value usually falls between -1 and +1. The value of +1 indicates that the vectors into consideration are perfectly similar. The vectors under examination are said to be perfectly opposing or different if the value is -1. It is very helpful for tasks like question-answering, document summarization, and semantic text similarity (STS).

Chapter 4: Methodology

In this chapter the detailed description of the model is provided.

4.1 Architecture of the Proposed System

This chapter describes the process utilized in the work. In Figure 4.1 is presented the overview of the proposed System Architecture consisting of the following five modules:

Module 0: Probable phrases with which 'objective' sentences may begin (manually collected).

Module 1: Pre-processing of all text data (Automated NLP procedure).

Module 2: Extraction of possible objective sentences from synonyms of 'objective'.

Module 3: Extraction of possible objective sentences matching with Module 0 collection.

Module 4: Extraction of Summarized objective from output of Module 2 and Module 3.

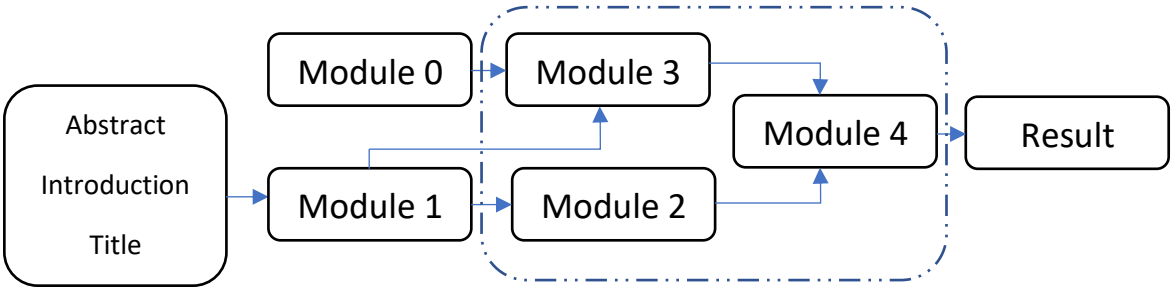

Figure 4.1: Proposed System Architecture

The final result obtained from the system is filtered to select the best quality match with the title. The details of the different modules are explained in the next few sections.

4.2 Module 0

Here, 15 research papers are consulted manually to collect some prospective phrases as objective-statement-beginners. The selection process is incremental. Repetitions may be registered to enhance a importance factor associated with each such phrase. So, when sentences containing those important phrases are encountered in the final round, they would have a highly prioritized chance of appearing in the Objective output. The phrases obtained after screening the papers are listed in Table 4.1 below.

Table 4.1: The List of objective-statement-beginners

Serial No.	Phrase Name	Serial No.	Phrase Name
1	this paper explores	14	This research paper proposes
2	this paper constructs	15	In this paper we focus on
3	This paper aims	16	In this work we propose
4	This paper presents	17	The primary aim
5	This paper describes	18	The aim of this paper
6	This paper gives	19	The objective of
7	This paper focuses	20	We focus on
8	This paper explores	21	We present a model
9	The focus of this paper	22	We deal with
10	This paper reviews and analyses	23	The study evaluated
11	Paper provides	24	The research described
12	This paper is concerned	25	We instantiate the model to utilize
13	The article proposes		

4.3 Module 1

This section describes the whole process of data collection and data pre-processing with some NLP techniques. Figure 4.2 is the represents the outline of Module 1.

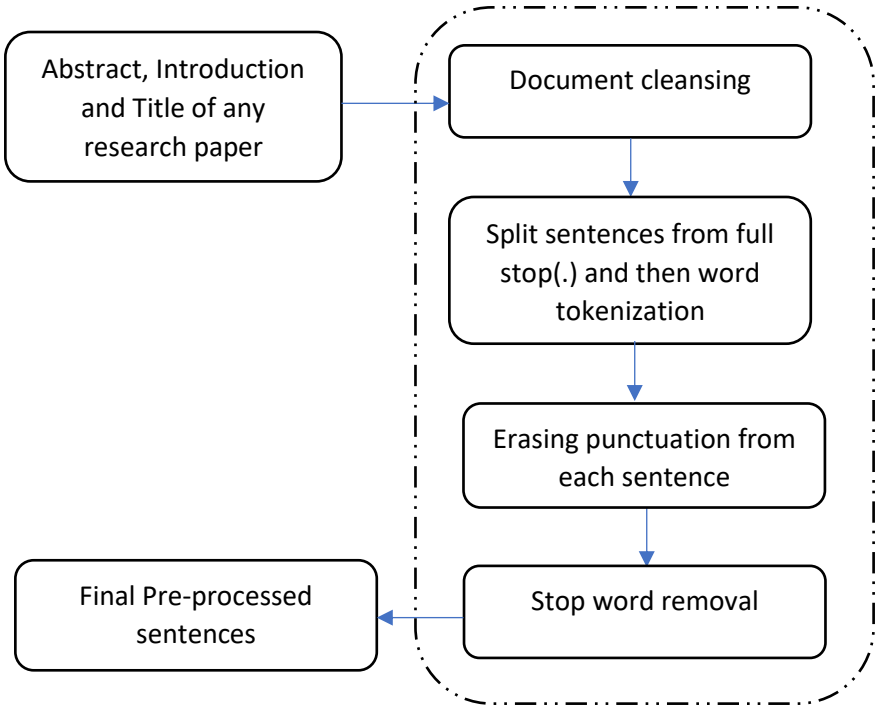

Figure 4.2: System Design of Module 1

4.3.1 Data Collection

Data collection is a method of gathering reliable information from a variety of sources in order to provide any output or insights of any research work. At first, we collect some research paper from online (like Google Scholar). Then we select title, abstract and introduction part of each paper and save them to different files. Next, the filenames are given as input to the model.

4.3.2 Data Pre-processing

On the collected documents, some necessary pre-processing and cleaning task is performed. Several steps are:

- Document Cleansing
- Splitting Sentences and Word Tokenization
- Punctuation removal
- Stop word removal

A. Document Cleansing

Data cleansing is the act of recognising and identifying sections of data that are incomplete, erroneous, inaccurate, or irrelevant, and then replacing, changing, or removing them. In our data we remove the references used in research paper like '[no.]'. We also looked for abbreviated words and attempted to substitute some of the most widely used abbreviated words, such as 'etc','e.g.','i.e.'. We also replace some full stop (.) with comma (,) where name abbreviation is used.

B. Splitting Sentences and Word Tokenization

The whole document has been chopped from full stop (.) and then is stored in an array. Tokenization is a process of breaking larger text into smaller one. Here we tokenize whole document word by word. This process is also known as Word segmentation.

C. Punctuation Removal

In this section we try to remove the several punctuations like '[', ']', '(',')', '.', '\n', '', '-' from the sentence array so that we can identify correct words needed for this work.

D. Stop Word Removal

Stop words are terms that are filtered out of natural language data at the time of processing since they usually refer to the most prevalent words in a language. We do not need to prepare stop word list because NLTK has already a pre-defined list. By using the list we remove those words from our document.

When pre-processing steps are completed, we collect the tokens and match with Synonymous words of 'Objective' by the help of 'Synonym Tree'. In the next section we have discussed about it.

After these following steps the data is prepared and pre-processed. Now we can use the data as input to our system.

4.4 Module 2

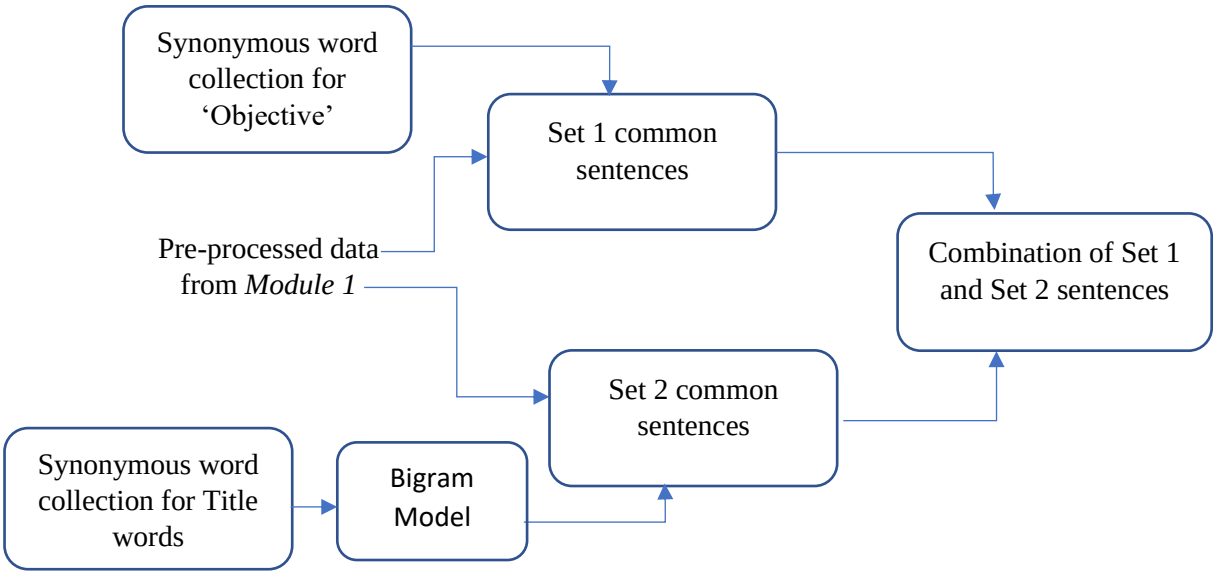

Figure 4.3: System Design of Module 2

4.4.1 Synonymous Word Collection Process for any Root Word

The data structure used in this technique is a hierarchical one which can represent relationships between different nodes. This is designated as a ***synonym tree*** which can store synonymous words with a level value indicative of its proximity to the original word placed at the root of the tree. At each hierarchically consequent level are placed child nodes carrying words with meaning related to the parent node word. The new words are obtained by an exhaustive search mechanism which peruses through a Standard English dictionary. Unless the word picked up is already placed elsewhere in the tree, it is assigned to the next child node position.

The process continues till the search terminates with no new words being found, or the user determines to terminate the search prematurely on reaching a prefixed limit value on the number of words in the bags.

A. Synonymous 'Objective' Word Collection

We have collected the synonyms of 'Objective' up to depth five and store them into an array. Now we match the words with tokens which we have taken earlier. If any match is found during the search process, we collect those sentences from abstract as well as introduction part within the module ***Set 1 common sentences.***

B. Synonymous 'Title Word' Collection

First of all, the title words are pre-processed and turn into tokens. After that, each token has been passed through the Synonym tree up to depth 2 and collected the words.

A schematic representation of the tree follows: -

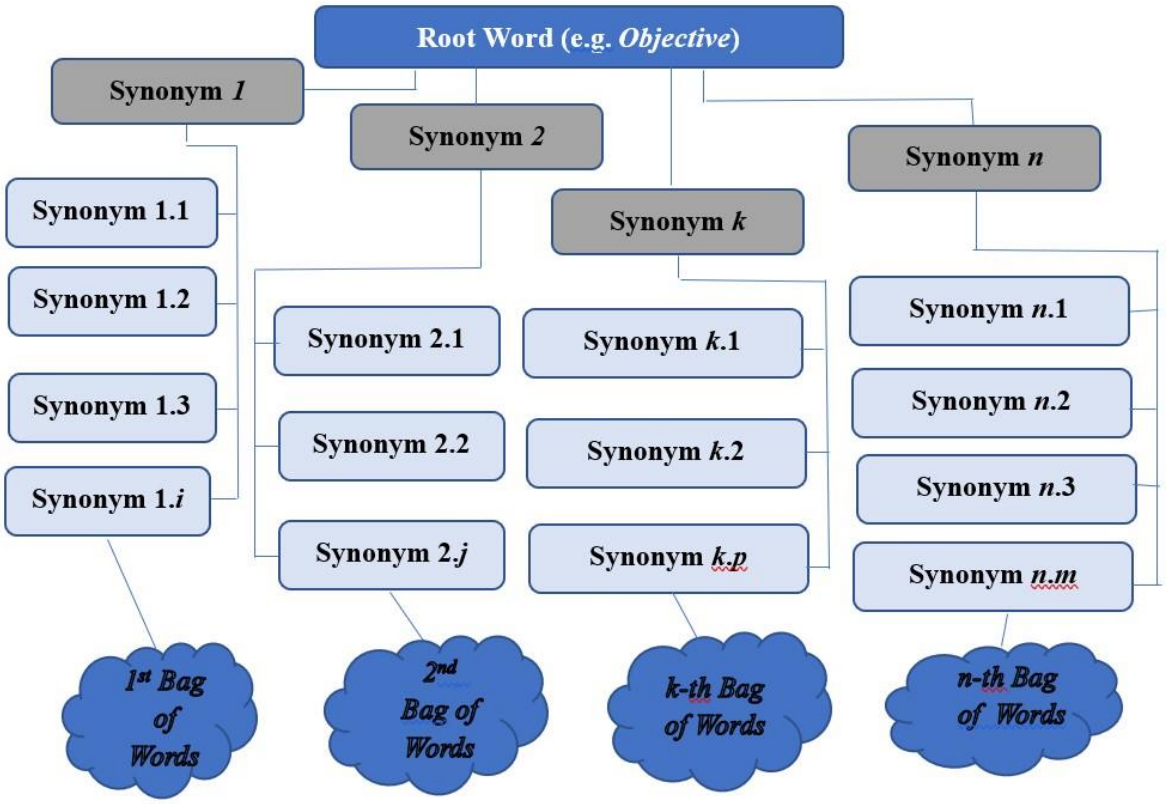

Figure 4.4: Design of Synonymous Tree

C. Algorithm of 'Synonymous Tree'

Step 1: Input the word (e.g. objective) and depth value

Step 2: If the depth = 0 then return the word

Step3: If depth ≠ 0 then insert it to an array

Step 4: Take two variables first_pos = 0 and last_pos = 1

Step 5: Find the Synonym of the words between first_pos and last_pos. Insert the unique words only into the array

Step 6: Now first_pos and last_pos are pointing to the previous and next size of array respectively.

Step 7: Set depth = depth+1 and go to Step 5 until depth reaches the specified value

Step 8: Print the array

Step 9: End

4.4.2 'Bigram' Model

Language Models are one of the most important part NLP. A bigram is a two-element sequence derived from a string of tokens, which are often letters, syllables, or words. The bigram is a n-gram model where n=2. The bigram model uses only the conditional probability of one preceding word to estimate the likelihood of a word given all previous words.

For example: Given a sentence 'This is my research paper'. If we extract the bigram from the sentence then

Sentence	Bigram
This is my pen	('This', 'is'), ('is', 'my'), ('my', 'pen')

- **'Bigram' generation for Title words**

We apply this Bigram Language model into the synonyms of title words. After that we try to match if the bigrams are existed in the abstract and introduction sentences. If any match is found then we collect those sentences. Now we calculate the weight of the sentences based on bigram frequency. Then the sentences are collected and stored in the module ***set 2 common sentences*.**

4.4.3 Model 2 Output : Combination of Set 1 and Set 2 Sentences

The process output from set 1 and set 2 common sentences modules are next combined, keeping in view the following norms:

- Priority settings should be checked. Higher priority gets preference.
- Repetitive sentences should be given higher priority. Only one copy maintained.
- Sentences with priority less than a pre-set threshold value should be excluded.

The collected objective sentences are finally stored in a data repository called ***Objective 1***.

4.5 Module 3

In this module we collect the pre-processed sentences from Module 1 and compare them with Module 0 phrases to generate all matched sentences. The collection is preserved in a data repository named ***Objective 2***.

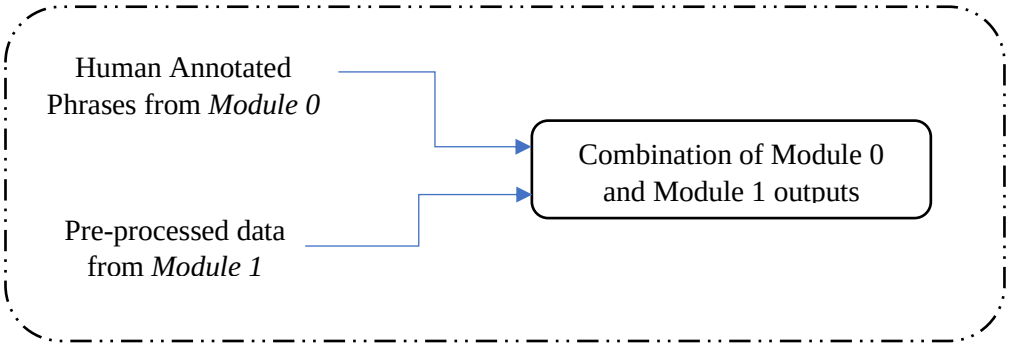

Figure 4.5: Schematic Diagram for Module 3

4.6 Module 4

The outputs Objective 1 and Objective 2 from **Module 2 and Module 3** respectively are processed in this Module. The combined Objective 3 is produced after checking and excluding any repeated sentence. The BERT model is next applied on this collective Objective 3 sentences to generate an extractive summary. The procedure is described in the following sections.

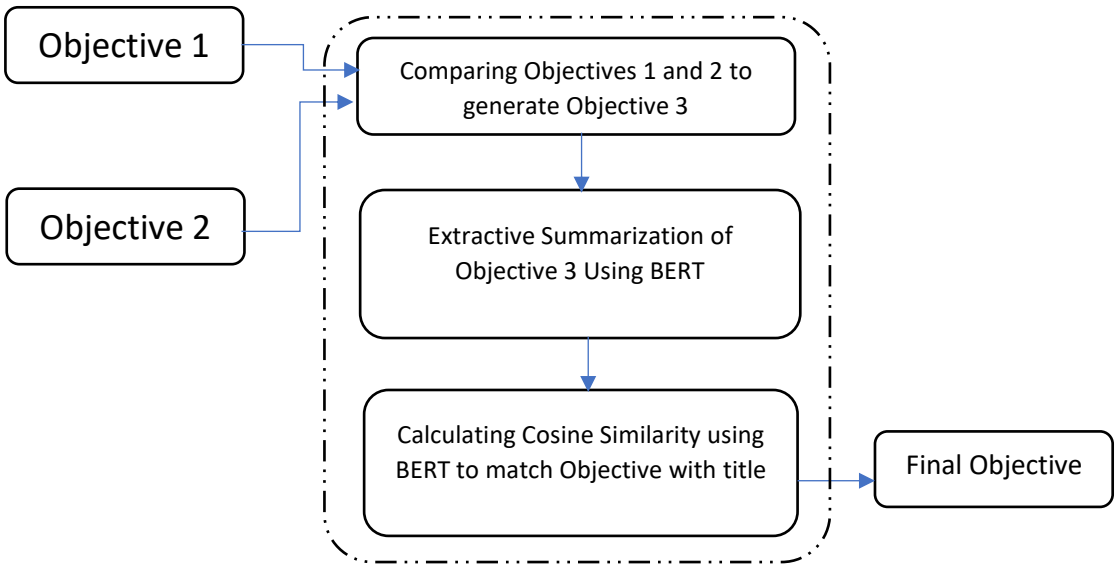

Figure 4.6: Schematic Diagram for Module 4

4.6.1 Extractive Summarization using BERT

Extractive summarization is a method of ATS. Here we use BERT's inbuilt Extractive summarizer to extract the candidate sentences from objective 3. We filtered a maximum of three sentences and generated all combinations of those sentences to be tried out for best match in the next part.

4.6.2 Calculation of Cosine Similarity using BERT

This section aims to establish a link between the title and each sentence combination obtained from the previous part. Here we use BERT model to find the cosine similarity index. The combination with the best result is selected as the prospective Objective of the document.

Besides the machine generated objectives, some human annotator generated objective (Paper no. 1 to 7) are also matched with the Title to find the Cosine Similarity Index and compared with similar values for machine objective.

In the next Chapter are presented the Results obtained from the proposed system along with some analysis on these results.

Chapter 5: Results and Analysis

In this Chapter the Results obtained from the experiments involving 15 journal papers are presented with detailed visualization techniques. The following few sections discusses these results in some detail.

5.1 Collective Objective-statement-beginners

A collection of objective phrases have already been shown in Table 4.1 in the last Chapter. These indicate the prospective beginnings of Objective sentences a human observer determined from an available paper corpus. The process is a manual one.

5.2 Synonym tree outputs as bag of '*Objective'* words

The following figure 5.1 depicts the generation of the first synonym tree.

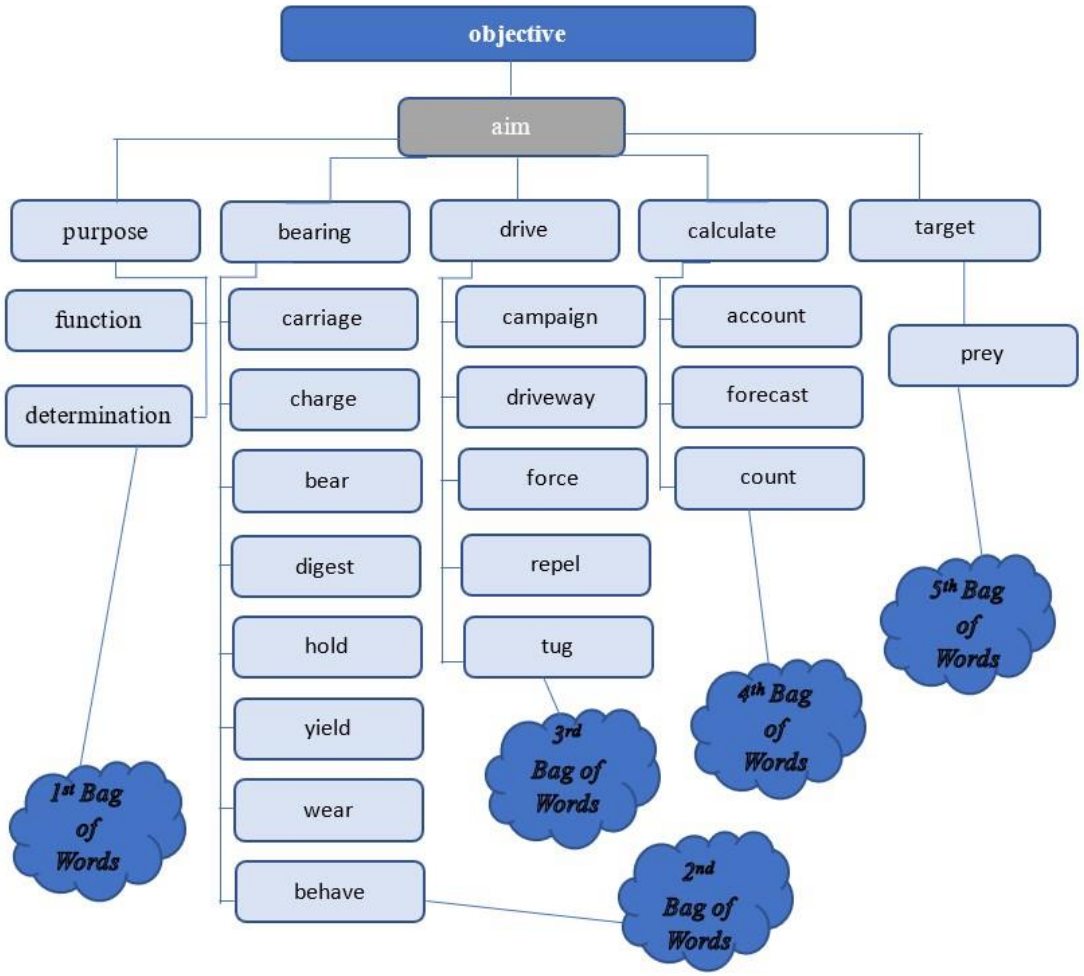

Figure 5.1: Generation of Synonyms of the word *Objective*

Table 5.1: Final Bag of 'Objective' words

Serial No.	Bag of Words
1st	affair , routine, serve, officiate, decision
2nd	passenger_car, baby_buggy, care, mission, cathexis, bang, commission, accusation, tear, appoint, commit, consign, agitate, load, blame, compilation, clasp, appreciation, delay, detention, handle, cargo_area, keep, have, deem, harbor, restrain, retain, accommodate, prevail, contain, reserve, defend, oblige, defy, apply, control, halt, carry, declare, agree, output, return, give_way, render, concede, move_over, give, succumb, clothing, break, tire, act
3rd	political_campaign, crusade, military_unit, violence, power, effect, force_out, coerce, impel, push, wedge, pull, storm, rebuff, disgust, tugboat, lug
4th	history, report, explanation, score, bill, prognosis, bode, consider, reckon
5th	Raven

The tree produces bags of words containing all possible words related to the word 'Objective' using Tree based Synonymous word Collection technique . In NLTK library of python, the inbuilt 'Wordnet' dictionary is used here to find the words. In this case, the synonym tree is developed upto depth 5.

5.3 Synonym tree output for bag of Title Words

Next is developed the synonym tree collecting bag of words related to the title of each paper belonging to the dataset one by one. Here the output is displayed for the first one in the series.

Title -1 : Depression detection using machine learning

Root words are depression, detection, using, machine, learning

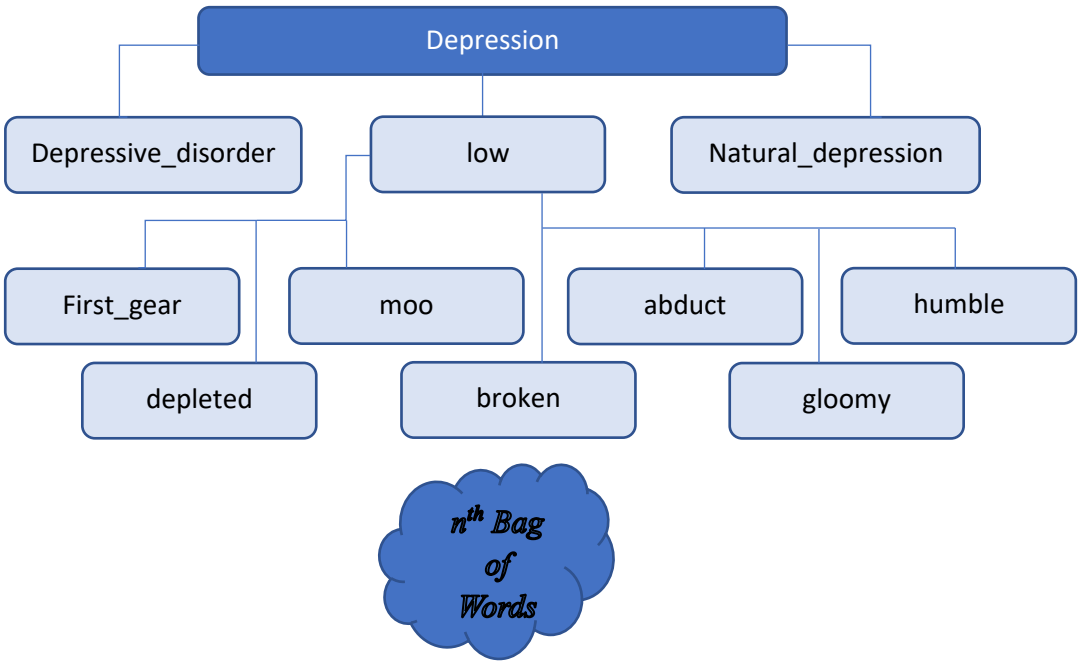

Figure 5.2: Generation of Synonyms of the word *Depression*

Table 5.2: Synonym tree for Paper Title 1 ["depression detection using machine learning"]

Serial no.	Title word (Root)	Related Bag of words
1	Depression	Depressive_disorder, low, … crack, fracture, collapse
2	Detection	signal_detection
3	Using	Practice, Exploration, … , give, entrust, invest
4	Machine	Car, cable_car
5	Learning	Learn, determine, teach, …, specify, fall, finalize

5.4 List of Summarization Output

This section discusses the output obtained after summarizing all the collected sentences gathered by considering the words and phrases accumulated from the previous three sections. The output of the same is reproduced in Table 5.3 below.

Table 5.3: Selection Before and After Extractive Summarization

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	Sentence 2	Sentence 3
1	hence, we come forward to provide an effective method to detect depression using machine learning . the primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions according to the results.	hence, we come forward to provide an effective method to detect depression using machine learning.	the primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions according to the results.	
2	this paper gives an overview of deep learning and then provides a comprehensive survey of its current applications in sentiment analysis. this paper first gives an overview of deep learning and then provides a comprehensive survey of the sentiment analysis research based on deep learning . learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results . along with the success of deep learning in many application domains, deep learning is also used in sentiment analysis in recent years . since about a decade ago, deep learning has emerged as a powerful machine learning technique (goodfellow, bengio, & courville, 2016) and produced state-of-the-art results in many application domains, ranging from computer vision and speech recognition to nlp . applying deep learning to sentiment analysis has also become very popular recently.	this paper gives an overview of deep learning and then provides a comprehensive survey of its current applications in sentiment analysis.	learning has emerged as a powerful machine learning technique that learns multiple layers of representations or features of the data and produces state-of-the-art prediction results.	since about a decade ago, deep learning has emerged as a powerful machine learning technique (goodfellow, bengio, & courville, 2016) and produced state-of-the-art results in many application domains, ranging from computer vision and speech recognition to nlp.

Table 5.3: Continued..

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	**Sentence 2**	**Sentence 3**
3	this article proposes a sentiment analysis model of youtube video comments, using a deep neural network. they claimed to be ranked number 2 among 11 teams on the twitter sentiment analysis campaign organised by semeval-2015 [4].	this article proposes a sentiment analysis model of youtube video comments, using a deep neural network.	they claimed to be ranked number 2 among 11 teams on the twitter sentiment analysis campaign organised by semeval-2015 [4].	
4	the semantic component of our model learns word vectors via an unsupervised probabilistic model of documents. they encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space.	the semantic component of our model learns word vectors via an unsupervised probabilistic model of documents.	they encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space.	
5	most of existing solutions to twitter sentiment analysis basically only consider textual information of twitter messages, and struggle to perform well when facing short and ambiguous twitter messages. recent studies show that sentiment diffusion patterns on twitter have close relationships with sentiment polarities of twitter messages. then we consider the inter-relationships between textual information of twitter messages and sentiment diffusion patterns, and propose an iterative algorithm called sentidiff to predict sentiment polarities expressed in twitter messages. although sentiment diffusion patterns have close relationships with sentiment polarities of twitter messages, existing work on twitter sentiment analysis basically only considers the textual information of twitter messages, but ignores sentiment diffusion information. considering the shortcomings of existing solutions to twitter sentiment analysis that only consider textual information and the close relationships between sentiment diffusion patterns and sentiment polarities of twitter messages, we argue that the best strategy is to fuse textual information of twitter messages and sentiment diffusion information in a supervised learning framework. the objective of sentiment analysis on twitter data is to classify the sentiment polarity of a twitter message as positive, neutral or negative.	most of existing solutions to twitter sentiment analysis basically only consider textual information of twitter messages, and struggle to perform well when facing short and ambiguous twitter messages.	considering the shortcomings of existing solutions to twitter sentiment analysis that only consider textual information and the close relationships between sentiment diffusion patterns and sentiment polarities of twitter messages, we argue that the best strategy is to fuse textual information of twitter messages and sentiment diffusion information in a supervised learning framework.	the objective of sentiment analysis on twitter data is to classify the sentiment polarity of a twitter message as positive, neutral or negative.

Table 5.3: Continued..

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	**Sentence 2**	**Sentence 3**
6	this paper constructs a depression detection model based on the features of depressed users derived from psychological observations. firstly, a sentiment analysis method is proposed utilizing vocabulary and man-made rules to calculate the depression inclination of each micro-blog . this paper applies data mining to psychology area for detecting depressed users in social network services. this paper applies data mining techniques to psychology, specifically the field of depression, to detect depressed users in social network services (sns).	this paper constructs a depression detection model based on the features of depressed users derived from psychological observations.	firstly, a sentiment analysis method is proposed utilizing vocabulary and man-made rules to calculate the depression inclination of each micro-blog.	this paper applies data mining to psychology area for detecting depressed users in social network services.
7	this paper focuses on the extractive based summarization using k-means clustering with tfidf (term frequency-inverse document frequency) for summarization . numerous applications are using text summarization that's why the practitioners required a means for generating the summaries and provide the privilege for a decision that excluding demand of reading the whole document and also parallel reliable enough in detailing central ideas.	this paper focuses on the extractive based summarization using k-means clustering with tfidf (term frequency-inverse document frequency) for summarization .	numerous applications are using text summarization that's why the practitioners required a means for generating the summaries and provide the privilege for a decision that excluding demand of reading the whole document and also parallel reliable enough in detailing central ideas.	
8	several research efforts have been done on keyword extraction . existing methods on keyword extraction have been done mainly by using a predefined controlled-vocabulary, which cannot process the unknown words/phrases.	several research efforts have been done on keyword extraction.	existing methods on keyword extraction have been done mainly by using a predefined controlled-vocabulary, which cannot process the unknown words/phrases.	
9	we will present a summarization procedure based on the application of trainable machine learning algorithms which employs a set of features extracted directly from the original text . text processing is a research field that is currently extremely active.	we will present a summarization procedure based on the application of trainable machine learning algorithms which employs a set of features extracted directly from the original text.	text processing is a research field that is currently extremely active.	

Table 5.3: Continued..

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	**Sentence 2**	**Sentence 3**
10	the materials failure forecasting method for volcanic eruptions (ffm) analyses the rate of precursory phenomena . the materials failure law in general describes terminal failure of metals, rocks, soils, concrete, or polymers . the technique is not to be used indiscriminately, but only for situations in which specific types of eruption precursors display an accelerating pattern, this paper aims to provide practical techniques and guidelines for applying the ffm approach to volcano eruptions . voight (1988) proposed a general materials failure law governing accelerating creep to characterize precursory phenomena . it may be linked to other established equations of accelerating creep (voight, 1989; cornelius and scott, 1993), and indeed the relation was recognized in surface displacements preceding failure of large-scale slope models (fukuzono, 1985) _ voight (1988) demonstrated that the materials failure forecasting method (ffm) based on this failure law was applicable to some examples of medium to large andesitic and dacitic eruptions.	the materials failure forecasting method for volcanic eruptions (ffm) analyses the rate of precursory phenomena.	the materials failure law in general describes terminal failure of metals, rocks, soils, concrete, or polymers.	it may be linked to other established equations of accelerating creep (voight, 1989; cornelius and scott, 1993), and indeed the relation was recognized in surface displacements preceding failure of large-scale slope models (fukuzono, 1985) _ voight (1988) demonstrated that the materials failure forecasting method (ffm) based on this failure law was applicable to some examples of medium to large andesitic and dacitic eruptions.
11	to support users wisely in choosing tourism activities, a recommendation or suggestion system using natural language processing (nlp), which is a subdomain of artificial intelligence (ai), is proposed . this paper explores and compares the nlp techniques that are currently applied to the existing recommendation systems. this paper aims to compare the nlp techniques for the recommender system to apply the techniques that are most useful for our future recommender systems.	to support users wisely in choosing tourism activities, a recommendation or suggestion system using natural language processing (nlp), which is a subdomain of artificial intelligence (ai), is proposed.	this paper explores and compares the nlp techniques that are currently applied to the existing recommendation systems.	this paper aims to compare the nlp techniques for the recommender system to apply the techniques that are most useful for our future recommender systems.

Table 5.3: Continued..

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	**Sentence 2**	**Sentence 3**
12	we propose a two-stage method that leverages the concept of a dynamically reconfigurable metasurface antenna (dma) in a narrow frequency band in which the effects of the wall are minimum to perform twi . two-dimensional (2d) dynamically reconfigurable metasurface aperture is presented to perform frequency selective through wall imaging (twi) with an unknown structure of the wall . based on these characteristics, a narrow band frequency selective window is identified . second, a dma consisting of an array of tunable metamaterial elements is used for twi in the identified frequency selective window . wall imaging (twi) has become a promising way to detect and recognize various objects in a plethora of applications, such as rescue operations, surveillance, and reconnaissance [1]–[4] . using dynamically modulated apertures, the radiation of spatio-temporally varying modes is achieved by electronically tuning the radiating metamaterial elements on the aperture over a narrow operating bandwidth or even at a single frequency [31]–[38].	we propose a two-stage method that leverages the concept of a dynamically reconfigurable metasurface antenna (dma) in a narrow frequency band in which the effects of the wall are minimum to perform twi.	two-dimensional (2d) dynamically reconfigurable metasurface aperture is presented to perform frequency selective through wall imaging (twi) with an unknown structure of the wall.	second, a dma consisting of an array of tunable metamaterial elements is used for twi in the identified frequency selective window.
13	this review paper provides a comprehensive overview, discusses challenges and opportunities, and indicates future directions for: (a) enabling technologies needed to make body area sensing and stimulation a reality, and (b) emerging bioelectromagnetics applications that may readily benefit from such technologies . this can be enabled through next-generation implants and wearables that, in turn, enable emerging bioelectromagnetics applications such as neurosensing, neurostimulation, and innovative imaging modalities [2].	this review paper provides a comprehensive overview, discusses challenges and opportunities, and indicates future directions for: (a) enabling technologies needed to make body area sensing and stimulation a reality, and (b) emerging bioelectromagnetics applications that may readily benefit from such technologies.	this can be enabled through next-generation implants and wearables that, in turn, enable emerging bioelectromagnetics applications such as neurosensing, neurostimulation, and innovative imaging modalities [2].	

Table 5.3: Continued..

Paper no.	Before Extraction	After Extraction Using BERT		
		Sentence 1	**Sentence 2**	**Sentence 3**
14	the recent emergence of machine learning approaches for enhancing wireless communications and empowering them with much-desired intelligence holds immense potential for redefining wireless communication for 6g . the evolving communication systems will be bottlenecked in terms of latency, throughput, and reliability by the underlying signal processing at the physical layer . in this position letter, we motivate the need to redesign iterative signal processing algorithms by leveraging deep unfolding techniques to fulfill the physical layer requirements for 6g networks . specifically, deep unfolded signal processing is presented by sketching the interplay between domain knowledge and dl . applying traditional signal processing and deep learning approaches independently entails significant computational and memory constraints . to this end, we present a general deep unfolding methodology that can be applied to iterative signal processing algorithms . deep learning (dl) [8]—a subset of ml—has been applied to wireless communication problems, such as signal recognition [9]–[11], detection, characterization, channel estimation, optimal network resource allocation [12], error correction coding schemes, and other physical layer applications [13]–[15] . although from a 6g latency and data speed requirements perspective, the upper layer enhancements would be constrained by the physical layer signal processing capability.	the recent emergence of machine learning approaches for enhancing wireless communications and empowering them with much-desired intelligence holds immense potential for redefining wireless communication for 6g.	in this position letter, we motivate the need to redesign iterative signal processing algorithms by leveraging deep unfolding techniques to fulfill the physical layer requirements for 6g networks.	applying traditional signal processing and deep learning approaches independently entails significant computational and memory constraints.
15	the main objective of this paper is to give an excellent outcome of mri brain cancer classification using support vector machine . the reduced features are submitted to a support vector machine for training and testing.	the main objective of this paper is to give an excellent outcome of mri brain cancer classification using support vector machine.	the reduced features are submitted to a support vector machine for training and testing.	

Here, BERT has been used for extractive summarization technique. An upper limit of three sentences has been considered. All combination of these sentences are compared with the title of the paper as indicated in the next section.

5.5 Using BERT for Cosine Similarity

All combination of the summarized sentences are compared here with the title of the paper.The most similar combination is selected as the final objective of the paper.The result of the experiment is presented in the Table 5.4 shown below. The best one is highlighted for each paper.

Table 5.4: Cosine Similarity between Title and Objective sentences

Paper no.	Sentence 1	Sentence 2	Sentence 3	Sentence1 + Sentence2	Sentence1 + Sentence3	Sentence2 + Sentence3	Sentence1 + Sentence2 + Sentence3
1	0.859	0.528		0.735			
2	0.88	0.658	0.501	0.781	0.598	0.6	0.65
3	0.958	0.446		0.697			
4	0.615	0.505		0.6			
5	0.551	0.659	0.754	0.6	0.643	0.683	0.627
6	0.694	0.793	0.764	0.821	0.768	0.82	0.84
7	0.815	0.552		0.698			
8	0.596	0.508		0.54			
9	0.889	0.516		0.798			
10	0.793	0.628	0.742	0.804	0.771	0.764	0.794
11	0.796	0.536	0.53	0.823	0.784	0.552	0.79
12	0.605	0.799	0.692	0.711	0.672	0.836	0.723
13	0.806	0.847		0.856			
14	0.884	0.739	0.503	0.877	0.849	0.707	0.841
15	0.76	0.45		0.8			

Inference from the table is that, 12 out of 15 cases the Cosine Similarity is greater than 80%. Summarization for papers 4 and 8 have the least resemblance with their respective titles, less than 65%. Similarity for paper no. 5 is around 75%. On an average the Cosine Similarity value is found to be 81.02%.

The next Table 5.5 compares how each paper fares against the average and puts the categorical description accordingly. The result is summarized in a visual bar-chart model displayed in Figure 5.3 next, which clearly indicates that 80% of the samples (12 out of 15) show near or above average similarity. The next Table 5.6 demonstrates the statistics collectively, and the following pie-chart in Figure 5.4 represents the same visually.

Table 5.7 gives the paper-wise descriptive textual objective along with the title and the cosine-similarity value between the two. As already mentioned, Papers 4 and 8 show the lowest similarity values. The probable reason may be the objective missing some vital title words marked in red in the table.

Table 5.5: Comparison between Maximum and Average Cosine Similarity

Paper no.	Cosine Similarity Index		
	Max	**Above / Below Average**	**Category**
1	0.859	0.044571429	Above 80%
2	0.88	0.065571429	Above 80%
3	0.958	0.143571429	Above 80%
4	0.615	-0.199428571	Below 65%
5	0.754	-0.060428571	65% - 80%
6	0.84	0.018571429	Above 80%
7	0.815	0.000571429	Above 80%
8	0.596	-0.218428571	Below 65%
9	0.889	0.074571429	Above 80%
10	0.804	-0.010428571	Above 80%
11	0.823	0.008571429	Above 80%
12	0.836	0.021571429	Above 80%
13	0.856	0.041571429	Above 80%
14	0.884	0.069571429	Above 80%
15	0.8	-0.010214286	Above 80%
Average	0.810714286	**11 above , 4 below**	

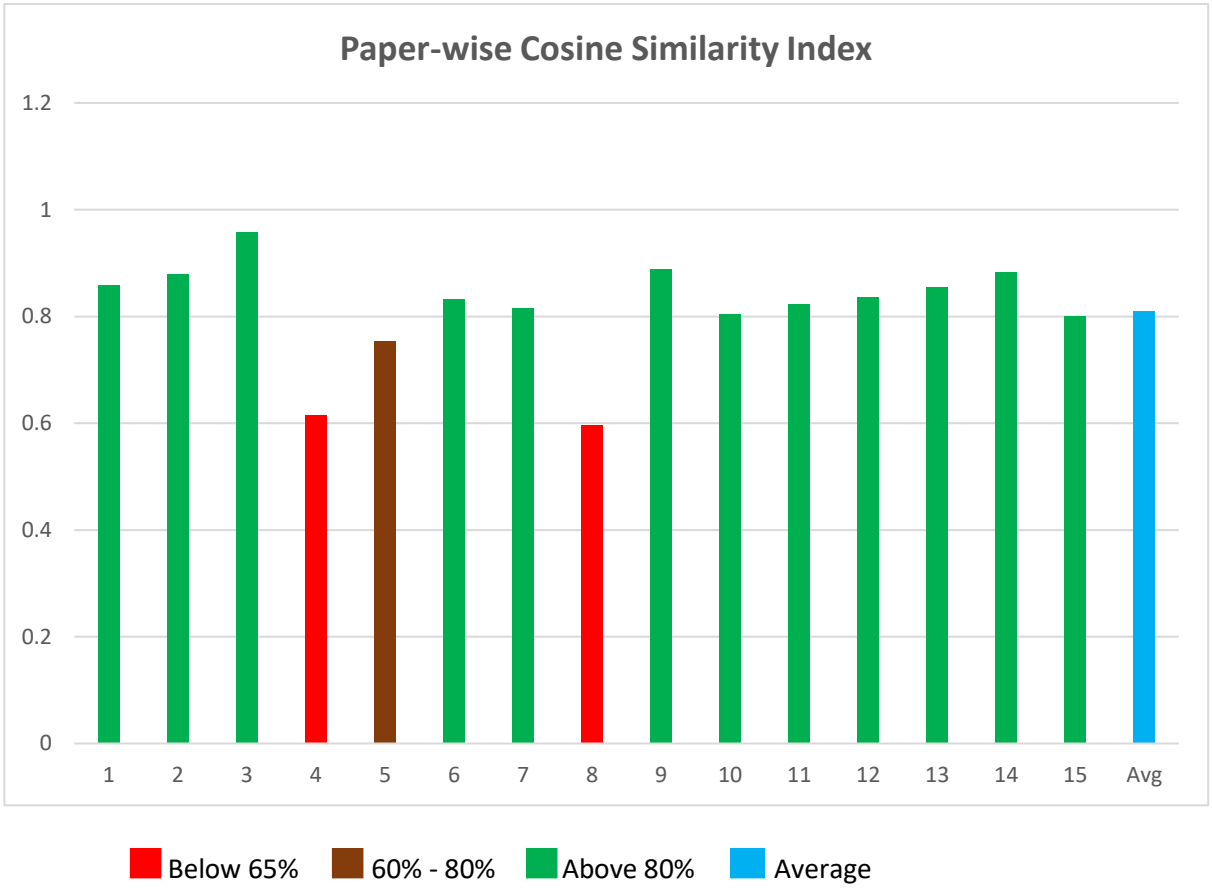

Figure 5.3: Graphical Representation: Cosine Similarity - Title vs. Objective

Table 5.6 : Category-wise Paper Count

Category	Paper Count
Above 80%	12
65%-80%	1
Below 65%	2

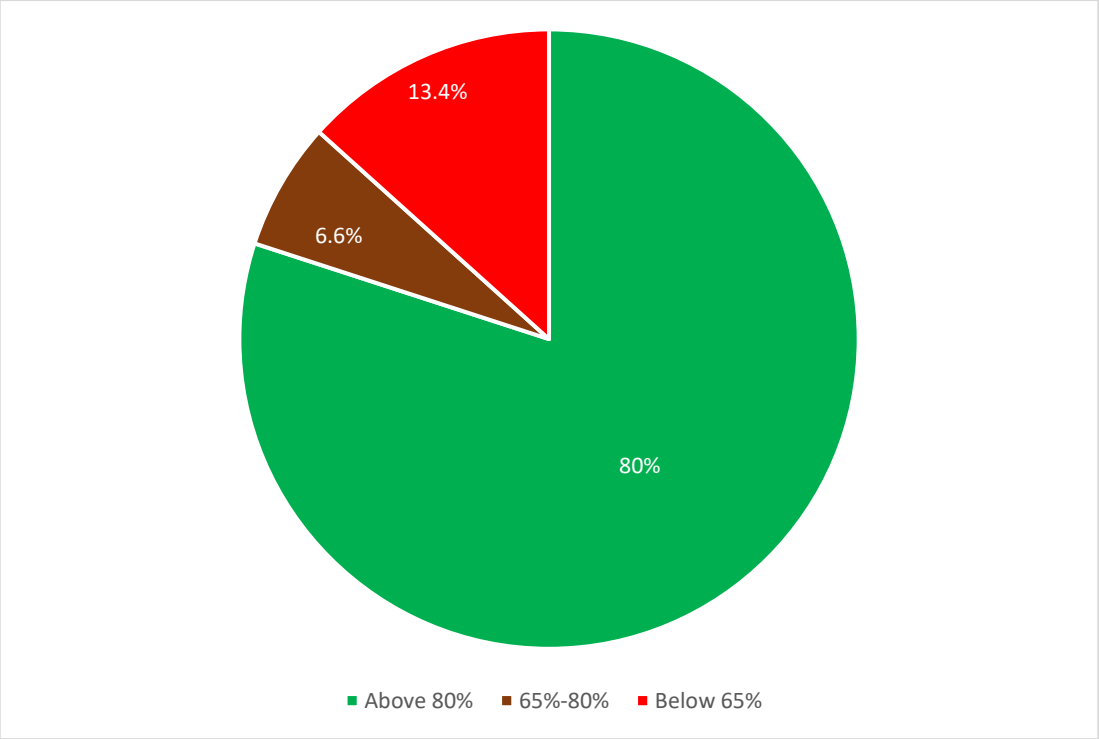

Figure 5.4: Ratio amongst the Three Categories

Table 5.7: System Generated Objective List for Papers

Paper no.	Title	Objective	Cosine Similarity
1	depression detection using machine learning	hence, we come forward to provide an effective method to detect depression using machine learning.	0.859
2	deep learning for sentiment analysis : a survey	this paper gives an overview of deep learning and then provides a comprehensive survey of its current applications in sentiment analysis.	0.88
3	sentiment analysis of youtube video comments using deep neural networks	this article proposes a sentiment analysis model of youtube video comments, using a deep neural network.	0.958
4	learning word vectors for sentiment analysis	the semantic component of our model learns word vectors via an unsupervised probabilistic model of documents.	0.615
5	sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis	the objective of sentiment analysis on twitter data is to classify the sentiment polarity of a twitter message as positive, neutral or negative.	0.754

Table 5.7: Continued…..

Paper no.	Title	Objective	Cosine Similarity
6	a depression detection model based on sentiment analysis in micro-blog social network	this paper constructs a depression detection model based on the features of depressed users derived from psychological observations . firstly, a sentiment analysis method is proposed utilizing vocabulary and man-made rules to calculate the depression inclination of each micro-blog . this paper applies data mining to psychology area for detecting depressed users in social network services.	0.84
7	extractive based text summarization using k-means and tf-idf	this paper focuses on the extractive based summarization using k-means clustering with tfidf (term frequency-inverse document frequency) for summarization.	0.815
8	keyword extraction using support vector machine	several research efforts have been done on keyword extraction.	0.596
9	automatic text summarization using a machine learning approach	we will present a summarization procedure based on the application of trainable machine learning algorithms which employs a set of features extracted directly from the original text.	0.889
10	graphical and pc-software analysis of volcano eruption precursors according to the materials failure forecast method (ffm)	the materials failure forecasting method for volcanic eruptions (ffm) analyses the rate of precursory phenomena . the materials failure law in general describes terminal failure of metals, rocks, soils, concrete, or polymers.	0.804
11	comparative study on natural language processing for tourism suggestion system	to support users wisely in choosing tourism activities, a recommendation or suggestion system using natural language processing (nlp), which is a subdomain of artificial intelligence (ai), is proposed . this paper explores and compares the nlp techniques that are currently applied to the existing recommendation systems.	0.823
12	frequency selective computational through wall imaging using a dynamically reconfigurable metasurface aperture	two-dimensional (2d) dynamically reconfigurable metasurface aperture is presented to perform frequency selective through wall imaging (twi) with an unknown structure of the wall . second, a dma consisting of an array of tunable metamaterial elements is used for twi in the identified frequency selective window.	0.836

Table 5.7: Continued…..

Paper no.	Title	Objective	Cosine Similarity
13	next-generation healthcare: enabling technologies for emerging bioelectromagnetics applications	this review paper provides a comprehensive overview, discusses challenges and opportunities, and indicates future directions for: (a) enabling technologies needed to make body area sensing and stimulation a reality, and (b) emerging bioelectromagnetics applications that may readily benefit from such technologies . this can be enabled through next-generation implants and wearables that, in turn, enable emerging bioelectromagnetics applications such as neurosensing, neurostimulation, and innovative imaging modalities.	0.856
14	redefining wireless communication for 6g: signal processing meets deep learning with deep unfolding	the recent emergence of machine learning approaches for enhancing wireless communications and empowering them with much-desired intelligence holds immense potential for redefining wireless communication for 6g.	0.884
15	mri brain cancer classification using support vector machine	the main objective of this paper is to give an excellent outcome of mri brain cancer classification using support vector machine . the reduced features are submitted to a support vector machine for training and testing.	0.8

The following Table 5.8 displays the Human Annotator generated Objectives for the first 7 Papers. The Similarity Index with the corresponding titles are also provided in the Table, and the Machine generated similarities for those Papers are also included for comparison. It is found that in most cases (6 out of 7) the machine gives the better Similarity Index. This indicates that human bias and errors may actually produce detrimental results under ordinary circumstances. The single instance where the human objective beats the machine generated one is for paper 4. The slight improvement may have been caused due to the inclusion of the green highlighted word which was missed by the machine generated objective, as already pointed out.

Table 5.8: Sample Human Annotator Generated Objective List

Paper no.	Title	Human Annotator Phrase	Title Matching	
			Human	Machine
1	depression detection using machine learning	In this paper, the authors detect fluctuations in the depression level of patients by monitoring their mood changes through text processing using Machine Learning algorithms. The graphical representations of the findings are more accurate and reduce the work of psychologists by half.	0.690	0.859
2	deep learning for sentiment analysis : a survey	The survey gives a overview about deep learning and its applications. It also provides a comprehensive study of its efficacy in sentiment analysis.	0.795	0.88
3	sentiment analysis of youtube video comments using deep neural networks	The objective of the paper is to propose a sentiment analysis model which classifies comments on youtube videos as negative, positive and neutral classes using deep neural networks. These are then compared with classifications by human annotators.	0.791	0.958
4	learning word vectors for sentiment analysis	The author built a model to capture both semantic and sentiment similarities among words. The model uses the vector representation of words to predict the sentiment annotations on contexts in which the words appear.	0.787	0.615
5	sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis	The author proposes a model which demonstrates how probabilities of correct classification of tweet messages increase when sentiment polarities of tweet texts and their retweets are consistent with sentiment reversals predicted by sentiment classifiers based on sentiment diffusion patterns.	0.644	0.754
6	a depression detection model based on sentiment analysis in micro-blog social network	The author constructs a depression detection model based on sentiment features extracted from texts written by Micro-blog users. The model uses 3 classifiers and annotated observations of some psychologists in the process.	0.838	0.84
7	extractive based text summarization using k-means and tf-idf	Here, k-means clustering algorithm is applied to find abstractive summary of a document, using Term Frequency-Inverse Document Frequency. The true k value is found using Elbow method and Silhouette method.	0.754	0.815

Chapter 6: Conclusion and Future Scope

This project work proposes a model to extract the objective of a research paper, by employing some NLP techniques and deep learning methods using BERT. The experiments yield a success rate of 81.07% when the system generated objectives are compared with their corresponding paper titles. In an additional experiment some of the papers are reviewed by Human Annotators and the objectives assessed by them are similarly compared with the paper title: in most of the observed cases the similarity index is higher for the machine generated objectives. This indicates that the proposed model helps to overcome the shortcomings of human endeavours to a large extent.

Besides this, the advantages of the proposed model are manifold. The researchers now face a tremendous task searching for the correct corpus to select – they have to read through the bulk of the papers before rejecting the same. With this type of an automated system, they can achieve the goal in a very short time and with little effort. They just need to go through the short objectives produced by the proposed system. While collecting other document corpuses too the system can help one to select precise contextual matters. For example, Journalists can concentrate on focused articles of importance, medical and law practitioners can pick and choose documented evidence of interest, and historians can dig up the strategic foundations already presented in the past in support of their theories. These are only a few instances where the system can be usefully applied.

The shortcomings of the system obviously lie first and foremost in the comparatively low success rate (81.07%), although this is better than the cumulative rate of the corresponding manual system (75.7% for the first seven papers). Another disadvantage is its inability to assess Abstractive Summary so far. Here a larger corpus, stronger hardware supports and extensive use of further deep learning techniques may help to alleviate the situation. Another improvement may be achieved by extending the corpus of objective-statement-beginners automatically and incrementally using case-based-reasoning concepts. For this purpose one may think of preserving within a case base short-listed top-ranking objective sentences obtained from new documents as and when they are being processed by the system. So, the knowledge base gets automatically enriched and can supply a better Module 0 output without further costs or human intervention.

In the application area, such a system may gain confidence enough to categorise unknown authors using advanced authorship attribution techniques, and easily pin down the identity of the miscreants of unsolved crimes, by studying past signature characteristics of enlisted persons. In short, with a little effort the system can be upgraded to a versatile and competent agent fairly mimicking and even excelling the original human counterparts in various tasks.

References

1. H. P. Luhn, “The Automatic Creation of Literature Abstracts”, Presented at IRE National Convention, New York, 159-165, 1958
2. Edmundson, H. P. New methods in automatic extracting. Journal of the Association for Computing Machinery 16 (2) (1969) 264-285
3. R. Khan, “Extractive based Text Summarization Using K-Means and TF-ID” , I.J. Information Engineering and Electronic Business, 2019, 3, 33-44
4. Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach”, Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.
5. J. Ross Quinlan. "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc., 1993
6. A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet", Advance Computing Conference (IACC) 2014 IEEE International, pp. 1169-1173, 2014,
7. M. Lesk,"Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC, 1986.
8. Barzilay, R., & McKeown, K. R. “Sentence Fusion for Multidocument News Summarization”. Computational Linguistics, 31(3), 297–328, 2005
9. D. Miller, "Leveraging BERT for Extractive Text Summarization on Lectures," arXiv, 2019.
10. D. T. Pham and P.T.N. Pham, “Artificial Intelligence in engineering”, International Journal of Machine Tools and Manufacture, Volume 39 Issue 6, June 1999, Pages 937-949
11. Mahind, R., Patil, A.: A review paper on general concepts of artificial intelligence and machine learning. Int. Adv. Res. J. Sci. Eng. Technol. (IARJSET) 4(4), 79–82 (2017)
12. A. Pannu, "Artificial intelligence and its application in different areas", *Artif. Intell.*, vol. 4, no. 10, pp. 79-84, 2015.
13. Tom M Mitchell, “Machine learning”, McGraw-Hill Education; 1st edition (March 1, 1997)
14. J. Han, M. Kamber, J.Pei, “Data mining : concepts and techniques”, Waltham, MA : Morgan Kaufmann/Elsevier, 3rd Edition ,2012.
15. J. Eisenstein, “Introduction to Natural Language Processing ”, MIT Press, 2019.
16. S. Ravichandiran, “Getting started with google bert;build and train state-of-the-art natural language processing models using bert”, packt publishing, 2021.