VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

A comprehensive survey on Text Summarization Techniques

under the guidance of
Prof. Saravanakumar Kandasamy
Associate Professor Grade 1
SCOPE
VIT, Vellore
ksaravanakumarvit@gmail.com

Raval Preet Ashishkumar - 20BCE2076
Tejas Ravindra Rote - 20BCE2096
Kartik Tripathi - 20BCE2098
Arush Saxena - 20BCE2106

Vellore Institute Of Technology, Vellore, Tamil Nadu, India

Abstract:

The importance of Text summarization is increasing more and more because of the huge amount of textual content that grows exponentially on the Internet, since the information on the world wide web is growing at an exponential rate, therefore it is necessary to provide the succinct form of the required information without losing its significance. It makes it more time-consuming and difficult for users to obtain relevant information. So, when someone searches for some information, a lot of data is shown which is impossible for a person to read. This brings the importance of text summaries in addressing the problem of how to acquire information and knowledge in a fast, reliable, and efficient way. Researchers have been trying to work and improve text summarization since the 1950s and since then it has improved a lot. Also, text summarization approaches are of three types: extractive, abstractive, or hybrid. Talking about abstractive approaches, it represents the input in an intermediate representation and then generates the summary with the sentence differing from the original document. Extractive approaches choose the most important sentences from the input document and then concentrate it to form a summary. While in hybrid text summaries, it combines both extractive and abstractive approaches. In this work, various ways of text summarization are studied including extractive, abstractive and hybrid approaches. With time, we'll see the importance of text summarization, its different ways and comparison of different datasets used.

1. Introduction

[94]An automatic text summarizing is part of text mining. Automatic Text Summarizer determines the most informative sentences from the entire document and then creates a representative summary. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. The proposed system solves this problem using Glove and Cosine similarity. Glove model is an unsupervised learning algorithm for obtaining vector representation of words. Finally, the sentimental analysis is carried out to determine the attitude or the emotion of the writer using polarity and subjectivity parameters.[95] Extractive based methods consist mainly of three steps: (i) document analysis and representation; (ii) sentence scoring; and (iii) sentence selection. The first step pre-processes and analyzes the documents to build a representation of their content. Based on the latter representation, a score is assigned for each sentence to measure its relevance, and finally, the top-ranked sentences are selected to form the summary. A suitable extractive method must select the relevant sentences that satisfy and optimize coverage and diversity properties and also that minimizes the redundancy between the selected sentences. [96]Word embedding learning algorithms as a method for vector representation of each word or sentence. From a sentiment point of view, their vector representation based on the word embedding algorithm includes the overall sentiment of a sentence. It refers to context-based sentiment analysis, where the word's prior polarity changes concerning a sentence-level sentiment analysis. We need new technology or mechanisms to: (i) tackle the overloading of information; (ii) obtain the information fast and efficiently; (iii) extract the most relevant and vital information; and (iv) sift vast volumes of information. [97]Text summarization technique can be considered as a mechanism to tackle the aforementioned problems. It also helps users to quickly find the required information. Sentiment analysis and text summarization are the essential tasks for Natural Language Processing (NLP) with many applications such as web mining, text mining and data mining. The goal of NLP is to process text using computational linguistics, text analysis, machine learning, statistical and linguistic knowledge to extract significant information. Automatic text summarization (ATS) is defined as the process of extracting important information from one or more text documents in order to produce a shorter version of the original document(s) for the needs of the user (or different users) that is performed by a computer program.[98] As a short version of the input documents, a good summary should convey the most important content in the documents while keeping redundancy to the minimum. Skip-Thought (ST) is an unsupervised model for learning generic and distributed sentences on Recurrent Neural Networks (RNNs)) maps words to a sentence vector, and the decoder predicts the surrounding sentences. Therefore, they have shown that the bi-directional LSTM network with max-pooling trained with the SNLI dataset makes the best sentence encoding method. an innovative method based on an orthogonal basis to combine pre-trained word embeddings into sentence representations.This paper will review each of the papers thoroughly and compare their results/ approaches and express their advantages over each other in further sections.

[1] The problem discussed is to present a simple and efficient extractor architecture which greatly affects the selection of words/phrases from the sentence. Encoder – decoder [6][7][8] mechanism is used in many sequences' generation tasks such as text summarization. This type of framework can not manage particular words in the source text which the model has never seen before which leads to the OOV problem. Many coping methods were introduced to address the OOV problem [10] this was refined [11] with a coverage mechanism that helps the model to prevent repeated method phrases [12]. Another method was proposed which uses encoder – decoder to paraphrase them [9] etc. There are various methods with improvement over previous methods; however, all the standard selector extracts the word that is in the target without its contextual or positional information. Due to which words may be selected which have the wrong context. So, to tackle this problem the authors have proposed [1] a SEGMENT framework that increases the cohesion among phrases, which helps the extractor be aware of the phrases in the source text and target while maintaining a concentration on single words. They have also modified the extractor architecture with the positional context that enriches the area features and a Temporal Convolutional Network (TCN) layer which softly filters the encoding features. [2] Problem is to present a way to keep the same semantics as the source text while abstractive text summarization . In this paper, they have focused on sentence summarization, which is different from standard document summarization as sentences are shorter than document and existing techniques in extractive methods are hard to apply on them like ranking sentences and sentence features[13] etc. Approaches like neural network models have been proposed to focus on designing sophisticated model structures like [13,14] have used selective gate networks to reweigh the source text representation. [15] Have integrated reinforcement learning ,adversarial networks, and recurrent neural networks to improve text generalization representation. [16] applied capsule networks with an adaptive optimizer to enhance the generalization capability from a few data points. [17,18] have used an extractive technique to weight the copy probability and guided the pointer network to copy important words from the source input. To distinguish salient information. On the other hand are various previous approaches for the same problem. extracted and integrated entity information into models, or retrieved summary templates to guide summary generation. For instance, [19] proposed a fact aware neural model, which leveraged open information extraction and dependency parse technologies to extract actual fact descriptions as external entity relation knowledge, to guide summary generation. Authors[2] proposed to use the novel FSum model; it uses the frame semantics guide to generate abstractive summaries. They select text through frame selection which selects the important and relevant frames from the source sentence to guide the summary generation by leveraging summary frames and F-to-F relations. They also designed an interaction between the source sentence representation and frame representation which further helps in learning a better semantic representation. [3] A way/approach to extract highlights from articles without missing annotations and simplify manually annotating new articles. Creating summaries of scientific papers includes dealing with one or more of the following issues: To choose the information that is most likely to appear in the article highlights, (ii) construct a summary of an article, or (iii) discover the keywords that characterize the primary subjects covered by an article. Keywords are single words or phrases (i.e., word combinations) that are often retrieved utilizing keyword extraction algorithms [20]. Abstracts are full-sentence summaries of the most important aspects of a document. Although they are generally accessible for most published articles, extractive summarizing methods may also be used to create them [21]. Here they have addressed the task by using a sentence-based approach, i.e., it identifies a subset of article sentences whose content is worth including in the article highlights. They proposed a method for identifying the top K sentences of a scientific article, where the information found is most likely to be relevant to article highlights. The approach serves two purposes: it aids manual annotation of new articles by providing relevant suggestions to annotators, and it automatically annotates missing highlight information with previous articles. Feature Extraction, which extracts significant features from the complete text of a large group of annotated articles, is the first stage in this method's data analytics process. Model Training produces a regression model on the provided dataset that represents the most relevant correlations between the previously examined data features and the similarity core. Sentence Labeling measures and records the similarity between highlights and article sentences. [4] A way to address problems faced while doing automatic text summarization is that it may not cover all the basic information of the source. Several different approaches have handled the extractive text summarization process. Frequency-based term weighting approaches have been one of the preliminary studies in this area [22]. Subsequently, the latent semantic analysis[23] , hidden Markov models [24], and graph-based unsupervised approaches [25][26][27] have gathered attention. For extractive summarization, a Recurrent Neural Network based Sequence Model (SummaRuNNer) was presented. The SummaRuNNer is a two-layer RNN-based sequence classifier, with the first layer operating at the word level inside each phrase and the second layer running over sentences for classification [28]. [4] a summarization procedure based on ensembled feature space was proposed. Here, the importance of the sentences was determined based on both semantic and syntactic features. To that end, LSTM-NN was proposed in this study to provide a summarization method that handles the joint of semantic and syntactic features. This model is a hierarchical structure where the first layer contains two LSTMs that simultaneously process the semantic and syntactic feature spaces,

respectively. Then, the outputs of LSTMs were concatenated in a deeper layer to obtain the enhanced feature space. Subsequently, a classification task is performed to determine a degree of importance to the sentences and subsequently select the summary-worthy ones. [5] To offer a solution for automatic text summarization jobs that simultaneously addresses the automatic text summarization problem of generalization. The automatic text summarizing (ATS) work selects the most significant concepts in a text to aid the reader's comprehension of the content. The ATS's role is to synthesize a document by finding (1) the content's key subjects and (2) the relevant ideas within those topics. As a result, current techniques seek to improve their efficacy in locating relevant facts in a text by considering all of the themes present. The most important aspect of the ATS assignment is universality; summarizing a news article is not the same as summarizing financial or medical data, for example. As a result, a number of the solutions on offer have been employed to handle a variety of domain-specific problems. To offer comments for programming language statements, automated summarization techniques were applied. Their technique provides the fundamental principles of a system, which aids in the comprehension of huge, often uncommented code authored by other programmers. The authors observed that individuals prefer to convey personal interests in financial summaries in another use of the ATS assignment. As a result, they created computer-based summaries of earnings reports, simulating the human inclination to exclude generally unimportant or less objective data. Most research studies on EATS were focused on extractive summaries. For instance, they considered key sentences and their positions in the text [31], measured word frequencies [32], or assigned importance levels to the sentences [33]. Supervised and unsupervised approaches are used[34,35,36] to analyze the characteristics retrieved from documents in order to construct models that allow the major components of essential ideas to be discovered. The fundamental disadvantage of supervised methods is that they need a collection of labeled data. Furthermore, the domain of the training samples is frequently too broad to handle fresh multi-domain examples. [5] Hence a method for automated text summarization was developed that uses a vectorial space built by several feature-generation methods. The vectorial space is the foundation of our strategy, which uses a GA to find the optimum grouping of texts. This clustering method organizes a document's phrases according to particular semantic and lexical characteristics. Two approaches were used to get the semantic features: Doc2vec and LDA.

The problem being discussed is the lack of expertise to understand the generated LS and interpret them proficiently, the inability of algorithms to explain the LS that they generate to lay users[72]. Multilingual Verbalization and Summarization for Explainable Link Discovery, the number and size of datasets abiding by the Linked Data paradigm are increasing every day. Discovering links between these datasets is thus central to achieving the vision behind the Data Web[73]. This paper is an extension of previous work (Ahmed et al., 2019) by proposing a generic multilingual approach that allows verbalization of LS in many languages. Ported the LS verbalization framework into German and Spanish, in addition to the English language. The proposed solution is to provide a generic multilingual approach that allows verbalization of LS into understandable natural language. The adequacy and fluency evaluations show that this approach can generate complete and easily understandable natural language descriptions even by lay users[74].
A nature inspired swarm intelligence-based algorithm viz. Firefly algorithm for multi-document text summarization is proposed. It adapted the behavior of firefly swarms to develop an algorithm for optimizing functions with multiple optima. The information on the world wide web is growing at an exponential rate, therefore it is necessary to provide the succinct form of the required information without losing its significance[99]. We needed a solution for multi document extractive text summarization. Multi-document summarization has more challenges as compared to single document summarization. In the past many techniques have been applied like statistical approaches, discourse-approaches, topic-approaches, graph-based, machine learning (Lloret and Palomar, 2012) and meta-heuristic approaches. In the case of the multiple document approach, the genetic algorithm was the first meta-heuristic algorithm used for text summarization. The key features focused on were coverage, non-redundancy, and relevancy to generate a better summary. After observing the results from meta-heuristic approaches, the researchers utilized the performance of multi-document summarization and an algorithm was formed that uses topic relation factor, cohesion factor and readability as fitness functions.
Here the aim is to formulate an approach to create summaries automatically, which in a way would resemble human technique. It proposes a language and domain independent approach for automatic extractive text summarization (EATS) tasks, which is based on a clustering scheme supported by a genetic algorithm (GA), to find an optimal grouping of sentences. The problem here is to create text summarization with automatic keyword extraction. The problem is solved with Belkebir and Guessoum's method, if a sentence belongs to the summary, it is assigned with a label "1", "0" otherwise. The sentences are classified by the AdaBoost algorithm. Soto and García-Hernández invented an automatic summarization system that made use of unsupervised learning. Three text models were used by authors to build a numeric vector. There are many approaches to the problem. The automatic text summarization (ATS) task consists of automatically selecting the key ideas in a text that allow the reader to understand the target document.

Recently ats have been using unsupervised approaches, because they do not require a set of pre-labeled items for training a classifier. Most EATS research studies have been focused on extractive summaries.
The focus is on how to score the salience of candidate text summary units like sentences, clauses, etc. to which the key is the Extractive single-document summarization task. With the amount of data being generated in the Web age, ATS plays an increasingly important role in addressing the problem of how to acquire information and knowledge in a fast, reliable, and efficient way. Moreover, there is no guarantee that the summaries generated by ATS are grammatically correct and have the same meaning as the original documents have. To make the results more faithful, motivated by discourse structure theories, incorporation of discourse structures into extractive summarization generation. ATS has a remarkable effect in the modern Web age. The usage of attention mechanism into text summarization was first brought to prominence by Rush et al. An unsupervised Chinese-oriented rhetorical parsing method is first proposed in the paper as it leverages the idea of translation and embeds the Chinese and English texts in the same latent space.
A novel hybrid approach for generating abstractive text summaries by combining fuzzy logic rules with bidirectional long short-term memory which further produces abstract summaries. Abstractive text summarization using attentional recurrent neural network (sequence-to-sequence) models have proven to be very effective. The proposed approach utilizes fuzzy measures and inference to extract textual information from the document to find the most relevant sentences. This paper is based on the English language and has both extractive and abstractive summarization. The purpose of text summarization is to find out the major concept and produce a text of shorter length compared to the source document. The main problem in the area of nlp was to locate the significant parts present in the input document. As the amount of data online is growing tremendously and with much wider information from different sources, it makes it more time consuming and difficult for users to obtain relevant information. So, when someone searches for some information, a lot of data is shown which is impossible for a person to read. This brings the importance of generating automatic text summaries. Most work in the field of extraction-based text summarization has been done where a summary is generated from the sentences already existing in the source text.

The problem of obtaining specific information about a determined topic as quickly as possible is answered in [56]. Due to the increasing size of digital information, retrieving specific information quickly has become a hassle along with the study of people's opinion about different parameters being an important aspect when retrieving information. Hence, the authors have tried to solve these issues and develop an efficient text summarization technique. In this era, there is an urgent need for short and precise sentiment-oriented summaries so as to get quick responses. There are generic/ query-focused, abstractive/ extractive, single-document/ multi-document approaches for text summarization. The proposed approach is Query-focused Sentiment-Oriented Multi-Objective Crow Search Algorithm (QSO-MOCSA) which is a combination of many desirable aspects of summarization techniques. The problem has been approached before using various methods like CCNU [60], IIITSum system [61], ITALICA system, NUS system [62], PloyU system [63], IITSummarizers model [64], QMOS method [65], most of which are based on sentiment analysis. The Query-focused Sentiment-Oriented Multi-Objective Crow Search Algorithm (QSO-MOCSA) was designed, implemented and tested for solving the problem.

In [57], the aim is to systematically measure the energy consumption characteristics of many different summarization algorithms on mobile devices for sustainable computing and find out which algorithms are more energy-hungry and reduce energy consumption. There is a need for the text summarization algorithms to be energy-efficient as users find it easier to read summary of contents rather than the information overload on hand-held devices. Hence, an energy-efficient algorithm is developed. There are a lot of active and increasing hand-held device users who need summarized results for saving a great amount of time and energy. Hence, the chosen problem focuses on research on green computing for mobile devices which is very much needed. There have been lots of different approaches like, graph-based, semantic-based, optimization-based, fuzzy-logic-based among others. Additionally, there are several application-level summarization techniques, like microblog summarization. In recent years, neural network-based summarization algorithms have been developed. Also, with the advent of transformer-based language models, there have been researches on how pre-trained models like BERT can be used for text summarization. For solving the problem at hand, energy consumption, summary quality and execution time of the selected summarization algorithms are measured and energy-efficient hybrid summarization algorithms are developed by combining previously available algorithms while monitoring their energy consumption as well as the quality of summary obtained.

[58] deals with the problem of redundancy in extractive text summarization and hence aims to propose a text summarization technique which includes only those sentences in summary, which represent the maximum of the topics embedded in the given text document. The primary shortcoming associated with extractive text summarization is redundancy, where more than one sentence representing a similar type of information is incorporated in summary. Hence, including only those sentences in summary, which represent the maximum of the topics embedded in the given text document is a big issue that needs to be tackled. The problem is chosen because it is difficult for human beings to manually summarize the large text documents in an efficient manner. The main objective is to address the redundancy problem associated with summarization methods and include only those sentences in summary, which represent the maximum of the topics embedded in the given text document. The concerned problem has been confronted before by Ferreira et al. (2013), Lloret and Palomar (2009), Mani and Bloedorn (1998), Abdi, Shamsuddin, Hasan, and Piran (2018), Mutlu et al. (2019), but finding the best set of features for the text summarization techniques is still a challenging task. So, the proposed solution is based on the similarity of sentences with the topic word embedded in the input text. In the proposed technique, the sentences which are closer (or similar) to the topic words of the given document are included in the summary. It consists of four steps namely; Preprocessing, Vector Generation, Relevance finding, Ranking and summary generation; in that order.

[59] solves the problem of finding a way to present a summary of a text document in a way that allows multi-dimensional integration with the user. Because of the amount of different types of data for analysis, the extraction of specific information from them is more and more complicated, there is a need for better text summarizers. The authors have tried to solve this problem by finding a way to present a summary of a text document in a way that allows multi-dimensional integration with the user. The problem is chosen so as to get a shorter version of the source text, ensuring the meaning and main components of the original. In the past the business data had been used mostly for operational data processing. There are already some existing approaches to multidimensional text data like DocCube[66], XML-OLAP[67], Document Cube[68], Topic Cube[69], etc. The proposed solution is multidimensional text summarization since the user can create different dimensions. In the proposed method, an external, internet knowledge database is being used for analyzing the document and creating a semantic relationship between the concepts in the document which is an entirely new solution to the concerned problem. The solution consists of four steps, i.e., document creation, document summarization, creating semantic relations and Data analysis using OLAP.

1.1 Important Definitions: -

R1: - ROUGE-1 refers to the overlap of ***unigram*** *(each word)* between the system and reference summaries.

R2: - ROUGE-2 refers to the overlap of ***bigrams*** between the system and reference summaries.

RL: - Longest Common Subsequence (LCS) based statistics. Longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.

MRR Value: - The Mean Reciprocal Rank (MRR) evaluates the responses retrieved, in correspondence to a query, given their probability of correctness. This evaluation metric is typically used in information retrieval tasks quite often.

F-Measure: - F-Measure provides a single score that balances both the concerns of precision and recall in one number

LDA: - Latent Dirichlet Allocation (LDA) algorithm is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories.

Doc2Vec: - Doc2vec is a generalization of the word2vec approach and an NLP tool for modeling documents as vectors. It is recommended that you first learn about the word2vec strategy before learning about doc2vec.

TF–IDF: - TF*IDF is an information retrieval strategy that considers the frequency of a phrase (TF) as well as the inverse document frequency (IDF) (IDF). Each word or term in the text has a TF and IDF score associated with it. The TF*IDF weight of a phrase is equal to the product of its TF and IDF ratings.

OHE: - Preprocessing categorical features for machine learning models using one hot encoding is a frequent practice. This form of encoding produces a new binary feature for each potential category and assigns a value of 1 to each sample's feature that matches to its original category.

The Eagle algorithm:- It is a supervised machine-learning algorithm able to learn LS using genetic programming.

The Wombat algorithm:- It implements a positive-only learning algorithm for automatic LS finding based on generalization via an upward refinement operator.

Raven algorithm:- It is an active learning approach that treats the discovery of specifications as a classification problem.

Adequacy:- The state or quality of being adequate.

Micro-planner:- Micro-planner planning is responsible for lexical selection and ellipsis, a form of abstraction of filtered data.

Meta-heuristic approaches:- Metaheuristic is an approach method based on a heuristic method that does not rely on the type of the problem.

Fitness function:- The fitness function is constant whose values can be changed to fit the user's needs.

JS-divergence:- The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions.

Cosine similarity:- Cosine similarity is one of the metrics to measure the text-similarity between two documents.

Cohesion Factor (CF):- Cohesion Factor (CF) determines whether sentences in the summary are talking about the same topic or not.

Readability Factor (RF): - Readability is a measure of how easy a piece of text is to read.

Latent Dirichlet allocation (LDA):- latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

lexical chain -a sequence of related words in writing, spanning short (adjacent words or sentences) or long distances (entire text).

2. Related works

[94]In this section, we concentrate on extractive text summarization methods focusing on different features such as statistical, linguistic, and semantic. They have discussed various methods using statistical fea[1]tures like term frequency, cue phrases, title similarity, and sentence position.[95] In 1950, Luhn (1958) coined the idea of ATS and extended it by Edmundson and Wyllys (1961) by incorporating statistical features such as word frequency, frequency of title word in a sentence, cue phrases, and sentence position in the input document. Namita Mittal et al proposed a text summarization approach based on the removal of redundant sentences. The summarization takes place in two stages where the input of a stage is the output of the previous stage and after each stage, the output of the summary is less redundant than the previous one. In this approach, more than 60% of the generated sentences match with the original input text.[96] Traditional sentence embedding methods are based on weighting and averaging word vectors of their constituents to construct sentences' vectors. Recently, pre-trained sentence embedding models have emerged as important methods for learning contextual representations. Most of these models are pre-trained using language modeling tasks on large text corpora.[97] The existing sentence embedding methods can be classified according to the learning paradigm into two categories: i) parameterized methods and ii) non-parameterized methods. Extractive summarization approaches attempt to generate the final summary by selecting a set of salient sentences from source document(s) that are most informative and relevant.[98] To tackle this challenging task, a variety of techniques and methods have been applied in extractive summarizers so far. One of the first attempts at the automatic text summarization field was done in 1985 which extracts important sentences based on word frequency. Sentiment analysis — with the huge amount of user-generated texts, extraction of significant information from numerous documents has gained much attention from the community of NLP. Sentiment analysis is an active research area of NLP that aims to identify subjective information and determine the sentiment orientation (e.g., positive or negative) of a given text.

[37] When encoder- decoder is used for sequence generation tasks including text summarization the framework is not able to manage words which are not in the source input that the model has never seen before which leads to OOV problem.[12] Another paper have worked on the previous Pointer networks and further refined it by a coverage mechanism that helps prevent repeated phrases by informing their history attention. [28] Other approaches like to select text and summarize the two important texts from the source a joint extractive-abstractive model was suggested which helps compress the information/ important sentences from the source text then use those texts to train the abstractive system. [9] Following the previous paper to select the important phrases first and to paraphrase them later a model was suggested which selected important words first and then used an encoder- decoder model to paraphrase them later. [38] The previous system that selected words first and then trained the abstractive model on the selected words the methods proposed here guides the abstractive system with compressed representation of keyword extracts to constrain copying words. [39] To deal with multi-label text categorization as existing approaches to multi-label text categorization fall short to extract local semantic information and hence, they model label correlation between them . [40] They have used neural Attention model for abstractive Sentence summarization or proposed a full data – driven approach to abstractive sentence summarization. [41] a framework/model or architecture was proposed to use standard neural sequence to sequence models for abstractive text summarization without reproducing the factual details inaccurately and not to repeat the details. [42] Another approach was to use a sequence-to-sequence model for abstractive text summarization as it has degenerated attention distribution so propose a method to produce attention distribution considering both quality and diversity without breaking end to end architecture. [43] Proposed a way to extract highlights of scientific articles without missing annotations. [44] Following on from the previous paper they have proposed an approach to extract key phrases from research papers by leveraging citation networks. [45] Introduced a new dataset for summarisation of computer science publications by exploiting a large resource of author-provided summaries and showing straightforward ways of extending it further. [46] Also on a similar concept of sequence-to-sequence model for abstractive summarization like [42], they proposed a way/approach to summarize documents using sequence to sequence models and tackle the problem of abstractive summarization of scientific articles. They generated sentence-level summaries of each paragraph. The intermediate results are exploited to automatically generate title and abstract using different types of Neural Network models. [47] Proposed a way to use a limited amount of multi-sentence training data or to use a data driven approach to sequence to sequence modeling to short text summarization of new articles. This work focuses on extractive summarization rather than on an abstractive approach. [28] To present an approach for extractive text summarization which can also categorize the information in the text presented as novel, content etc. Also a way to train extractive models on human generated reference summaries without the need for sentence-level extractive labels. [48] To propose a method or an approach to summarize single documents . Also to advise an approach to use illegal documentation for summarization which

can help law – enforcement in identifying.[31] This paper has proposed a technique or an approach to extract the most informative sentences of a given text by extracting a number of features from each sentences and then evaluating the importance of the sentence of the source text. [32] Provided a solution to problems faced by the word frequency-based methods for extractive summarization which are easy to implement and yield reasonable results but face limitations like ignoring the role of context; they offer uneven coverage of topics in a document, and sometimes are disjointed and hard to read. [33] The task of constructing a shorter version of a document while keeping its main information content is known as a single document summary. In this research, they conceive extractive summarization as a sentence ranking job and present a unique training method that uses a reinforcement learning goal to globally maximize the ROUGE evaluation metric. [49] This paper's core idea is to rank Maximal Frequent Sequences (MFS) in order to find the most significant information in a text. In the term selection stage, MFS are treated as nodes in a graph, and then ranked using a graph-based method in the term weighting step. [50] A method for generating document extracts is proposed in this work. This excerpt is made up of the most important sentences from the current document. These sentences were chosen based on their resemblance to a virtual paragraph with the same name (VP). The VP is made up of the keywords that were acquired using the four n-gram approaches used in this study.

1.1) Declarative Link Discovery frameworks, which rely on complex LS to express the conditions necessary for linking resources within RDF datasets[75]. 1.2) The verbalization of semantic data involved within such approaches. For example, it expands on an approach for converting RDF triples into Polish[76]. 1.3) The summarization of LS, which is related to work in the area of text summarization with a focus on sentence scoring techniques. The work surveys many sentence scoring techniques[77]. 1.4) newer work, the Wombat algorithm implements a positive-only learning algorithm for automatic LS finding based on generalization via an upward refinement operator[78]. 1.5) a generic multilingual approach that allows verbalization of LS in many languages, i.e., converts LS into understandable natural language text.

2.1) A cat swarm optimization algorithm for multi-document summarization was proposed in Rautray and Balabantaray. The model was compared with a harmonic search algorithm-based summarizer and particle swarm-based summarizer[79]. 2.2) An ATS (Automatic Text Summarization) model for single and multi-document, proposed in Ali and Malallah. Two main factors were relevance and redundancy[80]. 2.3) A shark smell optimization method based on multi-document for summarization was proposed in Verma and Om (2019). The main features were coverage, non-redundancy and relevancy[81]. 2.4) A novel text clustering technique called ensemble clustering method for text summarization was proposed in Lee et al[82]. 2.5) bio-inspired algorithms such as Fruit Fly Optimization (FOA), a new method for finding global optimization based on the food finding behavior of the fruit fly[83].

3.1) Ledeneva's method: the sequences of n-grams are extracted from the text by using a model of maximal frequent sequences[84]. 3.2) In Belkebir and Guessoum's method each sentence in a document is labeled ''1'' if it belongs to a summary, and the remaining sentences are labeled "0". Then, the authors generate a variety of features[85]. 3.3) Fattah and Ren proposed a method that is like that of Belkebir and Guessoum in that a summarizer that can be trained by using a variety of extracted features is applied[86]. 3.4) Unsupervised machine learning approaches have been utilized by applying clustering algorithms to group sentences based on the structure and frequency of the words[87]. 3.5) extractive summaries are generated by ranking sentences in the source document. The author's proposal is based on three general components: a sentence and document encoder and a sentence extractor[88].

4.1) Zheng explored how to enable humans to use big knowledge correctly and effectively in the biomedical domain .4.2) knowledge-based text summarization methods 4.3) Marcu's earlier works on RST parsing and applications on text summarization. 4.4) Attention mechanism into text summarization was first brought to prominence by Rush et al. This attentional encoder–decoder abstractive model was trained on a large-scale Gigaword dataset. 4.5) PacSum: is a typical unsupervised, directed text graph-based extractive model, in which BERT was employed as a sentence encoder to compute sentence similarity for better measuring sentence centrality.

5.1) Suanmali et al proposed a fuzzy logic method to extract the important sentences on the basis of sentence score. Sentence score was calculated using nine features. 5.2) Witte and Bergler presented a cluster graph algorithm based on a fuzzy set for the analysis of documents. It shows how extraction of some common and distinctive topic of a document is more informative than keyword extraction. 5.3) Dixit and Apte present an extractive summarization in which fuzzy techniques were used for sentence selection. The input text of 30 documents was taken where the sentence score is calculated for each. 5.4) Chopade et al. proposed a hybrid model consisting of a neural network and fuzzy rules for extractive text summarization. RBM (Restricted Boltzmann machine) consists of one input, two hidden and one output layer for training the system. 5.5) Sahba et al proposed a novel model for text summarization based on fuzzy features and attention-based sequence-to-sequence model. This model benefits the state-of the-art approach to both summarization techniques, extractive and abstractive [89][90][91][92][93].

In the CCNU summarization system which was introduced in [60], syntactic-based anaphora resolution and sentence compression algorithms are used for summarization tasks. After that, significance of terms is obtained by frequency-related topic significance and query-related significance. In [70], Update Summarization is discussed. The basic summarization of IIITSUM is enhanced with support vector regression so as to better estimate the combined effects in ranking. In [62], several new approaches concerning update and opinion summarization tasks. The previous timestamped graph approaches are improved upon by incorporating information about temporal ordering of events. At TAC 2008, PolyU [63] participated in three of the TAC 2008 tasks, including the update text summarization track, the opinion text summarization track and the query answering track. Three independent systems for these tracks were submitted. At TAC 2008, IIT Kharagpur [64] proposed a statistical model for opinion extraction and the subsequent summarization of the extracted opinions for the opinion summarization task.

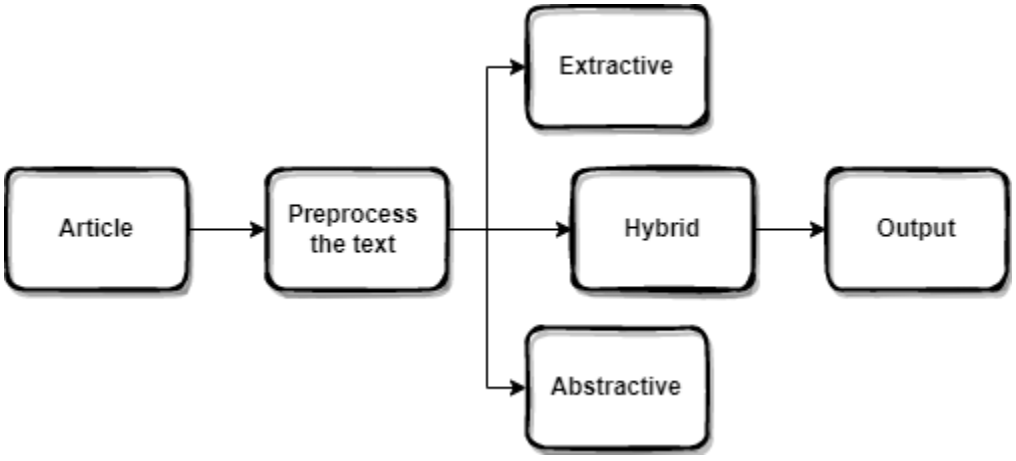
[71] deals with the creation of automatic summaries/abstracts. Statistical information derived from word frequency and distribution is used by the machine to compute a relative measure of significance, first for individual words and then for sentences. Sentences scoring highest in significance are extracted and printed out to become the "auto-abstract." The paper "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization" introduces a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. LexRank, a new approach for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences, is considered. The paper "Graph-Based Text Summarization Using Modified TextRank" deals with managing vast information on the internet through summarization. A graph-based text summarization method has been described which captures the aboutness of a text document. The method has been developed using modified TextRank computed based on the concept of PageRank defined for each page in the Web pages. The paper "Using latent semantic analysis in text summarization and summary evaluation" focuses on generic text summarization. The authors propose two new evaluation methods based on LSA, which measure a content similarity between an original document and its summary. In the paper "COSUM: Text summarization based on clustering and optimization", the authors have discussed the problem of providing wide topic coverage and diversity in a summary. Then they have proposed a two-stage sentences selection model based on clustering and optimization techniques, called COSUM. At the first stage, to discover all topics in a text, the sentence set is clustered by using the k-means method. At the second stage, for selection of salient sentences from clusters, an optimization model is proposed. The paper "Assessing sentence scoring techniques for extractive text summarization" describes and performs a quantitative and qualitative assessment of 15 algorithms for sentence scoring available in the literature. NATSUM approach in the paper "NATSUM: Narrative abstractive summarization through cross-document timeline generation" is centered on generating a narrative chronologically ordered summary about a target entity from several news documents related to the same topic. To achieve this, the system creates a cross-document timeline where a time point contains all the event mentions that refer to the same event.

The paper "Machine learning of generic and user-focused summarization" focuses on finding a salience function which determines what information in the source should be included in the summary. This method describes the use of machine learning on a training corpus of documents and their abstracts to discover salience functions which describe what combination of features is optimal for a given summarization task. The method addresses both "generic" and user-focused summaries. The paper "QMOS: Query-based multi-documents opinion-oriented summarization" presents the QMOS method, which employs a combination of sentiment analysis and summarization approaches. It is a lexicon-based method for query-based multi-documents summarization of opinions expressed in reviews. QMOS combines multiple sentiment dictionaries to improve word coverage limits of the individual lexicon. The paper "Multi-document extractive text summarization: A comparative assessment on features" focuses on extracting informative summaries from multiple documents using commonly used hand-crafted features from the literature. It recommends the use of fuzzy systems based on a feature vector and a fuzzy rule set for extractive text summarization. DocCube - DocCube treats several document facts as dimensions. Multidimensional visualization provides the user with the opportunity to learn the connections between documents. URL links give us direct access to the exploration of the content of the text. At any time, the user can have direct access via a link to documents associated with the selected dimension values. XML-OLAP - All documents represent facts data and dimensional data. In XML-OLAP, the query result returns a text cube. Text cube contains words, paragraphs or clusters. Document Cube - Keywords as author surname, publication date or title are uses like multidimensional data. Document cube proposes to link every document with keywords and other similar materials. This way is possible to hierarchically navigate. The query result is a text cube where cells consist of keywords to the relevant text. Topic Cube - OLAP must support roll-up and drill-down. The main idea is to use a hierarchical topic tree as a hierarchy for the dimension of the text. This structure allows the

user to drill down along this tree and discover the content of text documents to display different levels of topics. The first level in the tree contains the details of the issues, the second level is more general types, and the last level includes the aggregation of all topics. The authors proposed a thematic scope that calculates the probability that the document contains a topic. These measures allow users to know which topic dominates in the document collection. R-cube - Users provide a list of keywords, and then documents and facts related to the selected context are downloaded. Each paper describes the facts chosen according to their occurrence frequency.

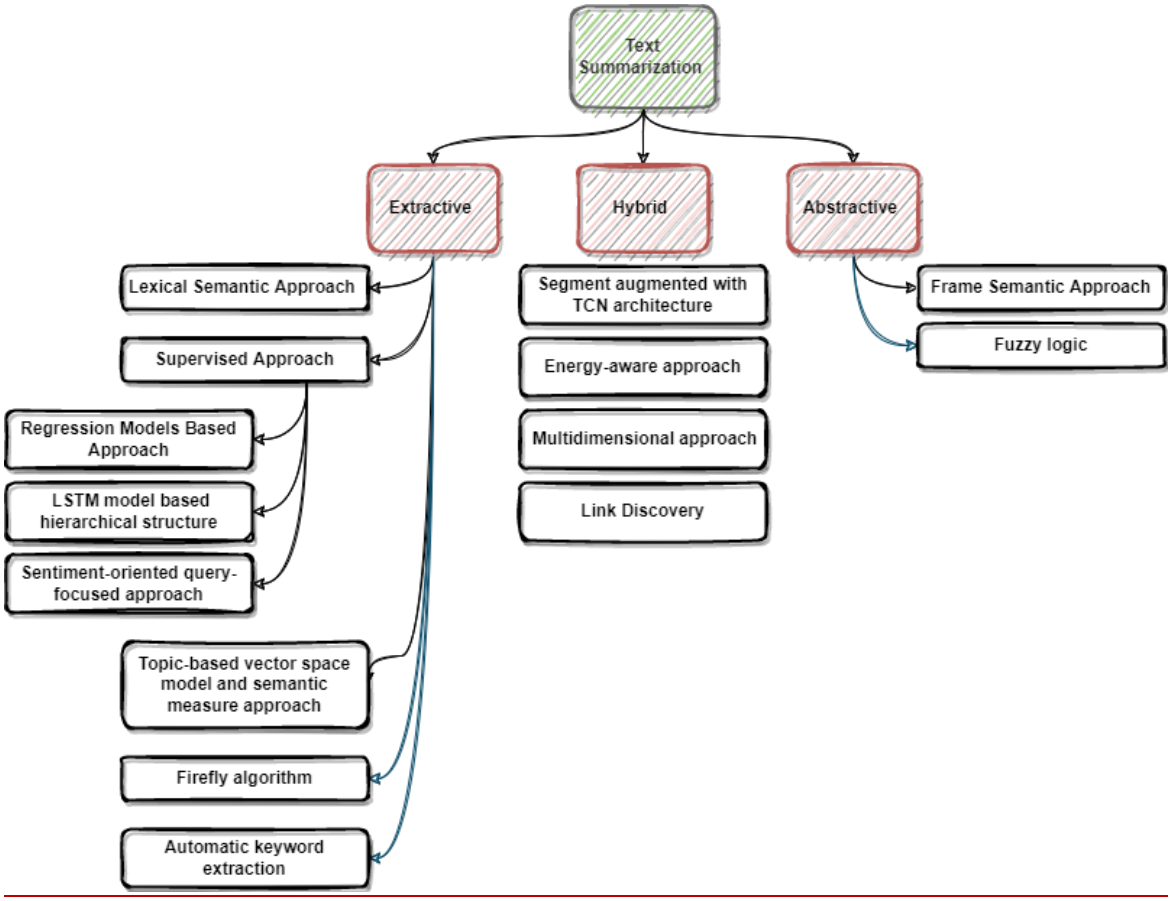
Algorithm/ Approaches discussed in proposed study		Algorithm/ Approaches discussed in Survey Paper	Description	Survey Paper
	1	Latent semantic analysis, Sentence based extraction,Statistical based, Complex network approach, Non - negative matrix factorization , MR, GA, FFNN, GMM and PNN based models	This document gives a summary of recent text summarization extraction strategies developed over the last decade. Their prerequisites are outlined, and their benefits and drawbacks are weighed. A few strategies for abstracting and summarizing multilingual texts are also described. Another problematic aspect of this type of research is summary evaluation. As a consequence, both internal and extrinsic summary assessment methodologies, as well as text summarizing evaluation conferences and workshops, are well addressed.	[100]
Temporal convolutional Network (TCN), Segment, Encoder- decoder approach ,	2	Intermediate representation, sentence score, topic representation, Frequency - driven , Bayesian topic models, Latent semantic analysis	They have reviewed different processes for summarization and described the effectiveness and shortcomings of different methods.	[101]
Frame semantic guided network, LSARank, PageRank algorithm, Partitional Clustering algorithms , LDA algorithm ,	3	Latent semantic analysis, Sentence Clustering , topic representation, Frequency - driven , Bayesian topic models	They have pointed out some of the peculiarities of the task of summarization which have posed challenges to machine learning approaches for the problem, and some of the suggested solutions.	[102]
	4	Supervised , unsupervised , keyphrase extraction, Graph - based ranking , topic based clustering , Simultaneous Learning , Language Modeling	Has presented a survey of state of the art in automatic key phrase extraction and examining the major sources of error made by existing systems and discussing the challenges ahead.	[103]

3. Architecture



The general methods/approaches used in the given paper below are some architecture of some of these methods below sections that cover detailed explanations of above given approaches. [1] Architecture which uses encoder-decoder concept includes, Extractor which has positional context, segment and TCN and an Abstractor which has gated attention. [2] To our knowledge, the author provided a novel FSum model, which is the first attempt to use Frame semantics to drive the development of abstractive summarization. 2. Author created a new Frame Selection module that uses summary Frames and F-to-F relationships to choose key and relevant Frames from the source sentence to guide the summary production. 3. The interaction method between the source sentence representation and the Frame representation was also devised by the author, which makes learning a better semantic representation more easier. 4. Experiments on the Gigaword dataset and the DUC 2004 dataset revealed that their suggested FSum model outperforms existing state-of-the-art approaches. [3] Paper proposed a way to extract highlights from the articles, (1)Feature extraction: it extracts salient information from the full text of a large set of annotated articles. (2)Sentence labeling: it measures and stores the similarity between article sentences and highlights. (3)Model training: it generates a regression model on the prepared dataset that describes the most significant correlations between the analyzed data features and the similarity score. (4)Model application: it applies the model on test data in order to predict sentence-level scores and to rank sentences by decreasing similarity score. [4] The proposed SIGIR 2018 corpus was handled by an extractive text summarization task. The problem was handled as sentence ranking and classification by basing the semantic features, syntactic features, and ensembled features. The initial layer of this model has a hierarchical structure, with two LSTMs analyzing the semantic and syntactic feature spaces at the same time. The outputs of LSTMs were concatenated in a deeper layer, and the enhanced feature space was then sent to a fully connected 2-layer neural network for classification. The produced summaries were assessed using ROUGE, and two types of Lead techniques, TextRank, and two state-of-the-art deep learning models (SummaRuNNer and BanditSum) were empirically compared. [5] Following this flowchart, the first step in the proposed summarizing process is to convert the texts to numerical vectors by using various ways to create the features. As a result, each phrase in a document is represented by a numerical vector. The quality of the clustering is assessed using the Silhouette index, which, in conjunction with a GA, aids in identifying the optimal approximate number of clusters . The above stages result in a clustering representation, in which key phrases are chosen from each cluster as follows: The word distribution in the document to be summarized is obtained using an LDA model. This distribution makes a connection between the term and its likelihood of appearing in the document.

4. Classification and Methods of Text Summarization



4.1 Extractive Text summarization: -

Summarized text is obtained by computing cosine similarity and extracting higher-ranked Then the summarized text is also subjected to speech conversion using Google out to determine the attitude or the emotion of the writer using polarity and subjectivity. The proposed system solves this problem using Glove and Cosine similarity. Glove model is an unsupervised learning algorithm for obtaining vector representation of words. It takes into account the order of words in sentences, unlike Word2Vec and fastText models. The cosine similarity determines the similarity between the vectors of words. Then the summarized text is obtained by computing cosine similarity and extracting higher ranked sentences. Then the summarized text is also subjected to speech conversion using the Google Text-to-Speech engine. Finally, the sentimental analysis is carried out to determine the attitude or the emotion of the writer using polarity and subjectivity parameters. Polarity identifies positive and negative statements. Subjective sentences generally refer to personal opinion, emotion or judgment. We are facing an inevitable and challenging problem of information overload, which has highlighted the need to develop relevant and specific tools to alleviate this problem by allowing users to save time and resources, as well as to find the most suitable information for their needs. Automatic Text Summarization (ATS), by condensing texts while preserving their important aspects can help to process this ever-growing text collection efficiently. We propose an unsupervised method for generic extractive multi-document summarization based on the centroid approach and sentence embedding representations. For this reason, we will briefly present the recent sentence embedding models as well as some previous unsupervised extractive methods to make the paper self-contained for reading. For the reader who is interested in a detailed overview of automatic text summarization approaches and methods, he may refer to the recent surveys in the field.

4.1.1 Lexical Semantic Approach: -

[5] The automated text summarization (ATS) job chooses the most important concepts in a text to help the reader understand the target material. In general, the ATS job is to synthesize a document by identifying (1) the content's primary subjects and (2) the significant concepts within those topics. As a result, existing approaches strive to enhance their effectiveness in finding significant data in a text by taking into account all of the themes present. The ATS task's key issue is generality; summarizing a news item is not the same as summarizing financial or medical information, for example. As a result, several of the offered solutions have been used to solve various domain-specific challenges. A method for automated text summarization was developed that uses a vectorial space built by several feature-generation methods. The vectorial space is the foundation of our strategy, which uses a GA to find the optimum grouping of texts. This clustering method organizes a document's phrases according to particular semantic and lexical characteristics. Two approaches were used to get the semantic features: Doc2vec and LDA.

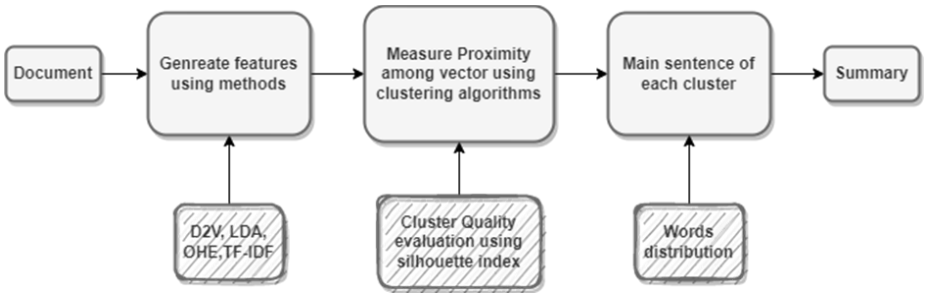


Figure 4.1.1.1 This flowchart explains the method proposed by the Author in [5].

Following this flowchart, the first step in the proposed summarizing process is to convert the texts to numerical vectors by using various ways to create the features. As a result, each phrase in a document is represented by a numerical vector. The quality of the clustering is assessed using the Silhouette index, which, in conjunction with a GA, aids in identifying the optimal approximate number of clusters . The above stages result in a clustering representation, in which key phrases are chosen from each cluster as follows: The word distribution in the document to be summarized is obtained using an LDA model. This distribution makes a connection between the term and its likelihood of appearing in the document.

4.1.2 Supervised Approach: -

4.1.2.1 Regression Models Based Approach : -

[3] When the summary is produced by some of the established graph – based methods/ models they tend to be less concise and are mostly focused on sentences that are produced by regression approaches. Hence propose a model/approach or a system to capture the most relevant sentence level information and its correlation to the content of the highlight.This work proposes a method for identifying the top K sentences of a scientific article, where the information found is most likely to be relevant to article highlights. The approach serves two purposes: it aids manual annotation of new articles by providing relevant suggestions to annotators, and it automatically annotates missing highlight information with previous articles. Feature Extraction, which extracts significant features from the complete text of a large group of annotated articles, is the first stage in this method's data analytics process. Model Training produces a regression model on the provided dataset that represents the most relevant correlations between the previously examined data features and the similarity core. Sentence Labeling measures and records the similarity between highlights and article sentences.

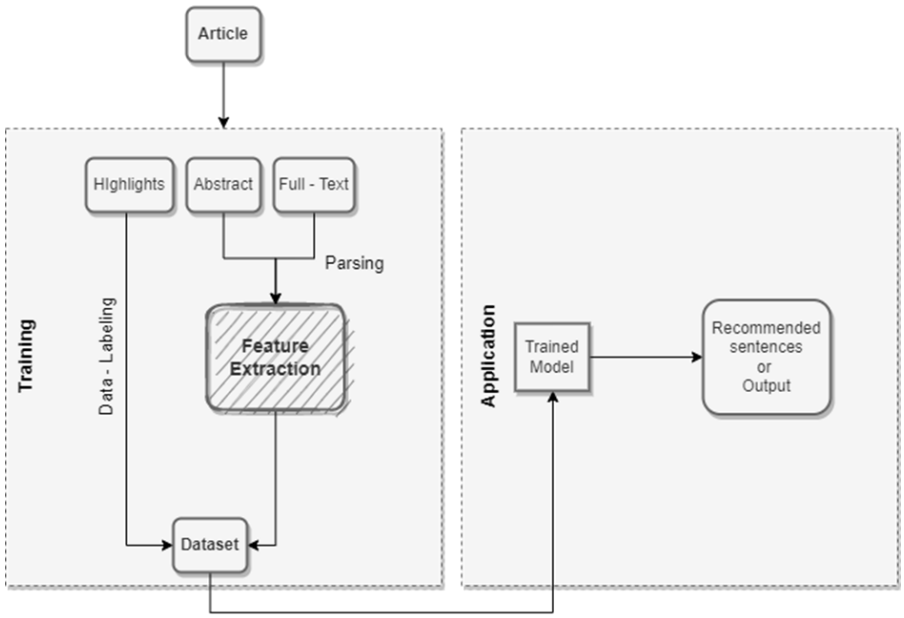


Figure 4.1.2.1.2 Architecture proposed for FSum Model for text summarization[3].

4.1.3 LSTM model based hierarchical structure for processing semantic and syntactic features of text for summarization.

[4] The process of creating a condensed version of a single document or a group of documents is known as text summarizing. Automatic summarizing of text documents is a difficult topic since it is critical that the generated summaries contain as much fundamental information as possible from the original document(s). This subject has been researched in the literature using two main strategies: abstraction and extraction. The summary process of humans has been emulated in abstractive text summarization. People summarize papers by collecting key information and rearranging it in their own unique words. Imitating such a process allows for the creation of more natural and artifact-like summaries.the proposed SIGIR 2018 corpus was handled by an extractive text summarization task. The problem was handled as sentence ranking and classification by basing the semantic features, syntactic features, and ensembled features. The initial layer of this model has a hierarchical structure, with two LSTMs analyzing the semantic and syntactic feature spaces at the same time. The outputs of LSTMs were concatenated in a deeper layer, and the enhanced feature space was then sent to a fully connected 2-layer neural network for classification. The produced summaries were assessed using ROUGE, and two types of Lead techniques, TextRank, and two state-of-the-art deep learning models (SummaRuNNer and BanditSum) were empirically compared.

4.1.4 Sentiment-oriented query-focused approach:-

This approach uses preprocessing of the document collection, the analysis of the text similarity, the sentiment analysis, and, finally, the formulation of the multi-objective optimization problem. The multi-objective optimization approach has been applied to solve the problem and it was the first time that this approach was opted for summarizing text. The Query-focused Sentiment-Oriented Multi-Objective Crow Search Algorithm (QSO-MOCSA) was designed, implemented and tested for solving the problem. It is a metaheuristic population-based crow search method which has been adapted for the query-focused sentiment-oriented extractive multi-document text summarization task.

4.1.5 Topic-based vector space model and semantic measure approach:-

This approach is based on the similarity of sentences with the topic word embedded in the input text. In the proposed technique, the sentences which are closer (or similar) to the topic words of the given document are included in the summary. It consists of four steps namely; Preprocessing, Vector Generation, Relevance finding, Ranking and summary generation; in that order.

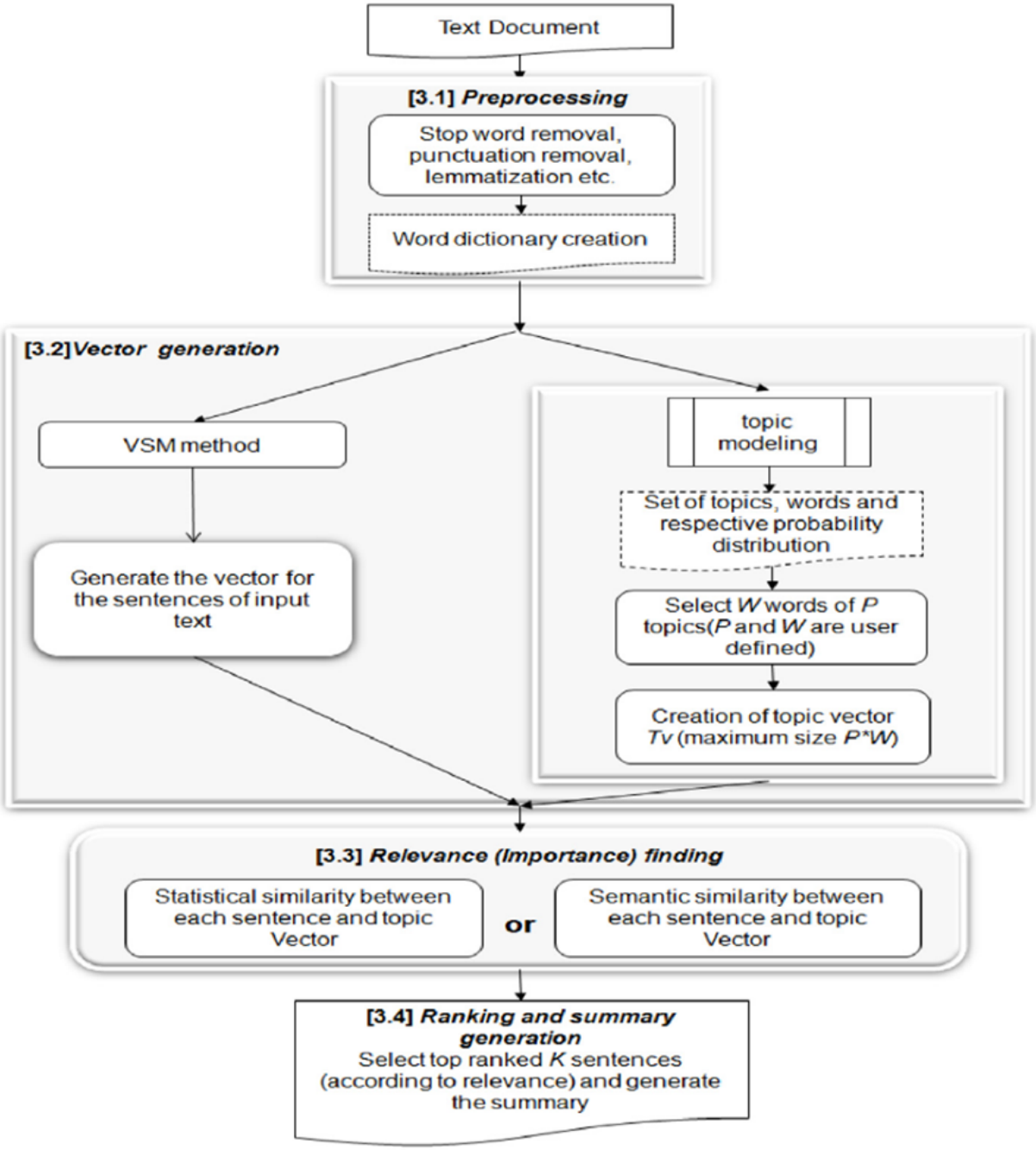


Figure 4.1.5.1: This flowchart explains the basic workflow of this approach

4.1.6 Firefly algorithm:-

After observing the results from meta-heuristic approaches, the researchers utilized the performance of multi-document summarization and an algorithm was formed that uses topic relation factor, cohesion factor and readability as fitness functions. Here are the following steps:(i) Pre-Processing, (ii) Document Representation (iii) Summary Scoring/Fitness Function and (iv) Utilization of Firefly Algorithm. The Firefly algorithm is used and finally, after the maximum iterations, the final summary has to be generated. The main features focused on were coverage, non-redundancy and relevancy to generate a better summary.

4.1.7 Automatic keyword extraction:-

Here automatic summarization was tackled by clustering sentences. The Silhouette index was applied as a fitness function in the GA to evaluate the quality of the groups. In addition, an LDA model is incorporated in our approach, not only to build a vectorial space model but also to find the most representative sentence in each cluster formed.
First, each document is separated into sentences that are considered the document's basic units. Next, the binary individuals of the GA represent the sentences of a certain document, where the algorithm provides the best tentative solutions of clusters. Finally, the key sentences of the clustering are selected, based on the LDA topics, as part of the summary. This process is repeated for each document in the collection.

4.2 Abstractive Text summarization : -

4.2.1 Frame Semantic Approach: -

[2] (Figure 4.1.2.1.1)When we rephrase the source text , the semantics of the source text might change while summarizing the source text. So propose a framework or a model which helps in keeping the semantics same as the source text also to extract important words from the source text and which may not be in the target text/summary but to give high quality summaries. In the novel FSum model , it uses the frame semantics guide to generate abstractive summaries. They select text through frame selection which selects the important and relevant frames from the source sentence to guide the summary generation by leveraging summary frames and F-to-F relations. They also designed an interaction between the source sentence representation and frame representation which further helps in learning a better semantic representation.(Figure 4.2.1.1) (1) Frame Representation separates frame semantic information F S from the source sentence; (2) Frame Selection selects important Frames F in a given sentence based on its summary; (3) Encoder represents the selected Frames and given sentence through Frame Encoder Hf and Source Encoder, respectively; (4) Interaction Layer combines the Frame representation and sentence representation into an overall semantic representation C; (5) Summary Generation uses the overall preset. Then, one by one, we'll go through each of the six modules.

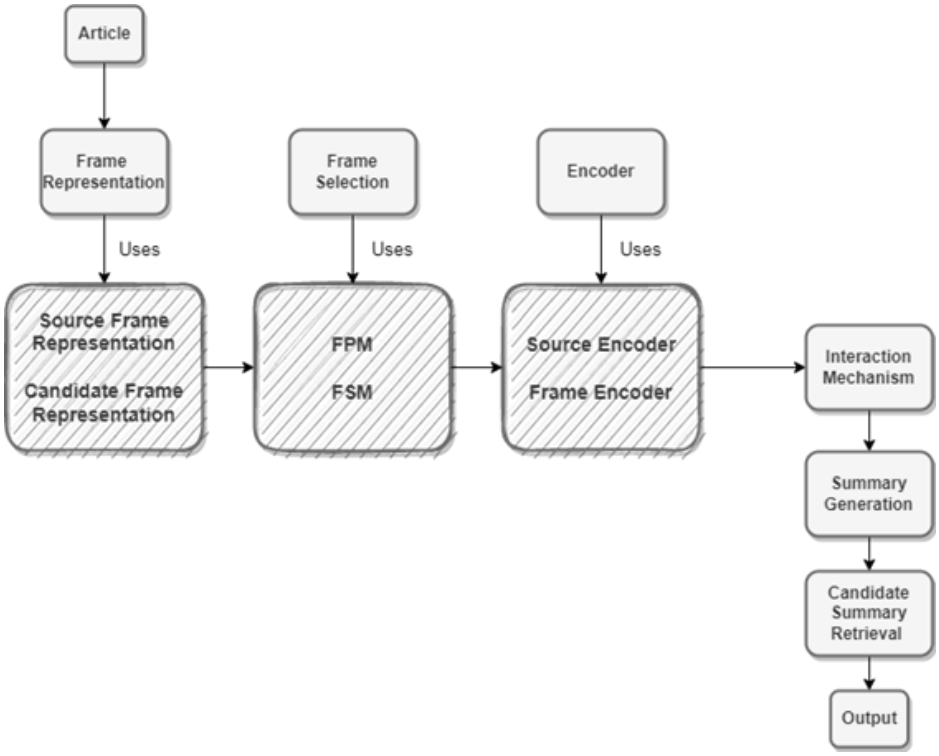


Figure 4.2.1.1 This flowchart explains the method proposed by the Author in [2].

4.2.2 Knowledge-guided unsupervised rhetorical parsing for text summarization :-

An unsupervised Chinese-oriented rhetorical parsing method is first proposed in the paper as it leverages the idea of translation and embeds the Chinese and English texts in the same latent space. The paper also proposed an unsupervised summarization evaluation metric which considers many aspects of how faithful a generated summary is and hyper-parameters were tuned by supervised learning on the golden standard of DC2002 to make it more effective. They also proposed an unsupervised summarization evaluation metric. This evaluation metric considers many aspects of how faithful a generated summary is.

4.2.3 Fuzzy logic :-

The problem was overcome by applying copy machines which can handle words which are out of vocabulary and coverage machines to solve repetition problems. Also, the use of fuzzy logic is utilized with a combination of LSTM and Bi-LSTM. The proposed approach utilizes fuzzy measures and inference to extract textual information from the document to find the most relevant sentences.y.

The method consists of four major steps executed in the following stages:

(i) Text preprocessing,(ii) Feature extraction,(iii) Extractive summarization by fuzzy rules and (iv) Abstractive summarization using Bi-LSTM

4.3 Hybrid: -

4.3.1 Segment augmented with TCN architecture: -

[1] The problem is to present a simple and efficient way to extract important phrases from the source text and make an abstract summary which may or may not contain those words to solve the encoder – decoder is widely used for many sequence generation tasks including text summarization. However, this framework cannot handle some words that are in the source text / input that the model has never seen hence it leads to OOV. Various approaches are given in papers for e.g. A paper presents a joint extractive-abstractive model that can compress important sentences from source text then use these sentences to train the abstractive system. Another approach is to use a separated content selector that produces a binary mask to verify words are in the target source which is basically an improvisation on a previous paper which introduced a system that selects important phrases first then uses encode-decoder models to paraphrase them. Solution presents a simple and efficient extractor architecture for the abstractive summarization system that affects the selection of words from the source text to create summary. They have used segment embedding layers to enrich information for that abstractor, which will help them to increase cohesion selectivity.
Extractor which outputs two binary masks which are focus masks and segment masks responsible for words in target and source in phrases included in the summary.
Positional Context which adds anchor points while combining two separated features space has been applied in image captioning similar to this positional contexts has been added which is a joint feature between encoder outputs and positional embedding. SEGMENT filters the individual words and compactifies the information to yield more salient information from the selected words from the masks generated. Temporal convolutional network (TCN) which is improved upon by the extractor by adding a max – pooling layer to select significant features. At last the abstractor has gated attention which avoids a significant amount of unnecessary attention on unattended elements and allows the model to have more concentration on important parts of the text.

4.3.2 Energy-aware approach:-

Energy-efficient hybrid summarization algorithms are developed by combining previously available algorithms while monitoring their energy consumption as well as the quality of summary obtained. Since there is a trade off between energy consumption and summary quality, the summarization approaches (like LR and LSA) that generally produce better quality summaries consume more energy than others are combined with others which consume less energy and hence strong hybrids are generated.

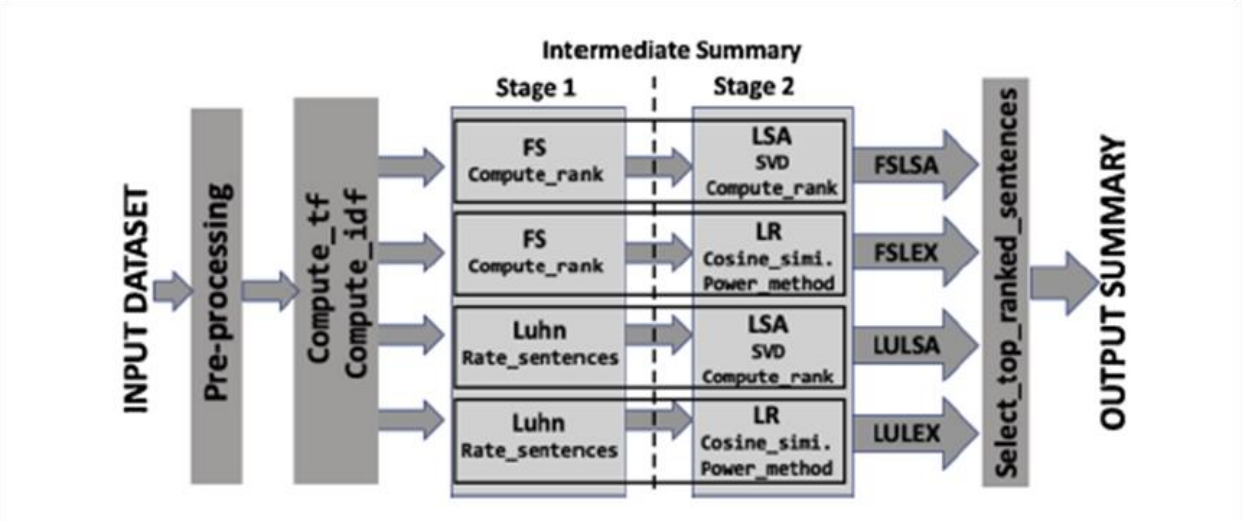


Figure 4.3.2.1: Explains the method to develop hybrid summaries in [].

4.3.3 Multidimensional approach:-

In the proposed method, an external, internet knowledge database is being used for analyzing the document and creating a semantic relationship between the concepts in the document which is an entirely new solution to the concerned problem. The solution consists of four steps, i.e., document creation, document summarization, creating semantic relations and Data analysis using OLAP.

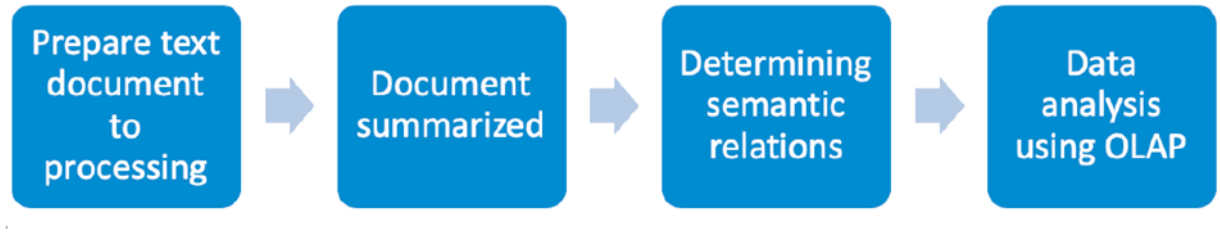


Figure 4.3.3.1: Concept of multidimensional text data generation in this approach

4.3.4 Link Discovery:-

The proposed solution is to provide a generic multilingual approach that allows verbalization of LS into understandable natural language. These natural language descriptors can even be understood by lay users. Extension of previous work (Ahmed et al., 2019) by proposing a generic multilingual approach that allows verbalization of LS in many languages, i.e., converts LS into understandable natural language text. We ported our LS verbalization framework into German and Spanish, in addition to the English language. Our adequacy and fluency evaluations show that our approach can generate complete and easily understandable natural language descriptions even by lay users.

5. Evaluation Method

Appropriate evaluation procedures are necessary for summarization systems. For summary evaluation, a variety of methodologies such as the Pyramid method, ROUGE, and others are utilized. The numerous assessment methods utilized in this survey study are presented in this section. Table lists the many types of automatic summary evaluations as well as the methods that are utilized to perform them. ROUGE, for example, is a widely used and popular collection of automated assessment metrics. It consists of a software that assesses summaries automatically and is utilized by the majority of the systems assessed, including. The number of similar words between a particular summary and a set of reference summaries is counted using the ROUGE technique. As a result, it aids in the automated assessment of the summary issue. ROUGE-L estimates the ratio between the longest common subsequence of two summaries and the size of the reference summary. ROUGE-N counts the number of N-gram units shared by a given summary and a group of reference summaries, where N is the number of N-Grams. ROUGE-S determines the fraction of common skip bigrams in a single summary and a group of summaries. ROUGE-W It's an improvement over the simplest longest common subsequence method. ROUGE-SU is a combination of ROUGE-S and ROUGE-1 that adds a counting word called unigram to ROUGE-S.

	Corpus / Dataset Used	Metrics used to evaluate	Number of Documents	Advantage of dataset	References
1	CNN Daily Mail	ROUGE metrics including ROUGE – 1, ROUGE - 2, ROUGE – L	287,113 train samples, 13,368 validation samples, and 11,490 test samples	This dataset has approximately 28 sentences per document in the training set and 3-4 sentences in the reference summaries. Also the average word count per document is 802.	1. Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, *28*. 2. Nallapati, R., Zhai, F., & Zhou, B. (2017, February). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*.
2	Annotated English Gigaword DUC 2004	ROUGE metrics [51]], including ROUGE-1, ROUGE-2 and ROUGE-L. F-score-based ROUGE metric on Gigaword data [52,5,53,54] and recall-based ROUGE metric on DUC 2004 data [52,53,54,55]	Gigaword has over 3.8M of sentence-headline pairs as the training set, 189k pairs as the development set, and 2000 pairs as the test set. DUC consists of 500 news articles from the New York Times and Associated Press Wire services, and each article has 4 different human-generated reference summaries.	The Gigaword dataset pairs the first sentence in the news article and its headline as the summary with heuristic rules . DUC 2004 has articles from New York Times and Associated Press Wire services and each article has 4 different human – generated reference summaries.	1. Napoles, C., Gormley, M. R., & Van Durme, B. (2012, June). Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)* (pp. 95-100). 2. Over, P., Dang, H., & Harman, D. (2007). DUC in context. *Information Processing & Management*, *43*(6), 1506-1520.
3	CSPubSum AllPubSDum BioPubSum	Rouge Toolkit .	CSPubSum dataset, which included over 10,000 training and over 150 test articles. BioPubSum had 8070 training articles and 2690 test articles, whereas	CSPubSum is the only dataset available that has been tuned to automatic highlight extraction and is available for research purposes. Also the datasets , beyond the article full-text, the collections	1. Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

	Corpus / Dataset Used	Metrics used to evaluate	Number of Documents	Advantage of dataset	References
			AllPubSum had 198 training articles and 66 test articles.	contain the abstracts and from 3 to 6 highlights per article provided by the respective authors.	
4	1. DailyMail 2. NYT 3. DUC 4. TeMario 5. SicommNet	ROUGE metrics including ROUGE – 1, ROUGE - 2, ROUGE – L	CNN has 92579, DailyMail has 219506, NYT has 167,223, DUC2002 has 567 new articles each , TeMario has 100 Portuguese columns, SicommNet has 1000 scientific articles.	These datasets have been commonly used for text summarization because of its being one of the most comprehensive datasets for this task. CNN/Daily has been used especially for abstraction since the goal summaries highlight sentences that are more suitable for abstraction. The DUC dataset satisfies the need for labeled sentences. It is relatively smaller than CNN/Dailymail. However, it has the advantage of having manually generated extracts.	1.. Sandhaus, E. (2008). New York Times corpus: Corpus overview. LDC catalog entry LDC2008T19. 2. Hachey, B., Murray, G., & Reitter, D. (2005, October). The embra system at duc 2005: Query-oriented multi-document summarization with a very large latent semantic space. In Proceedings of the Document Understanding Conference (DUC) 2005, Vancouver, BC, Canada. 3. Pardo, T. A. S., & Rino, L. H. M. (2003). TeMario: a corpus for automatic text summarization. NILC Tech. Report NILC-TR-03-09. 4. Cabanac, G., Chandrasekaran, M. K., Frommholz, I., Jaidka, K., Kan, M. Y., Mayr, P., & Wolfram, D. (2016, June). Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). In Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) (pp. 1-5).

	Corpus / Dataset Used	Metrics used to evaluate	Number of Documents	Advantage of dataset	References
5	1. DUC02 2. TAC11	ROUGE metrics including ROUGE – 1, ROUGE - 2, ROUGE – L	The DUC02 dataset contains 567 English-language news items. Every news item was produced by two human specialists, allowing us to compare the system's summaries to those provided by humans. The TAC11 dataset comprises writings in Arabic, Czech, French, Greek, Hebrew, and Hindi, among other languages. Each language includes a collection of 100 papers that cover ten distinct themes, with each topic containing ten documents that have certain common event sequences.	The DUC02 dataset was selected to measure the effectiveness of the proposed approach. Moreover, this dataset includes human-generated summaries that could be used to compare the capability of the proposed EATS algorithm with human skills. The TAC11 dataset was selected to prove that our methods can be applied to different languages.	No particular datasets were cited in the paper.
6	RDF datasets	RDF datasets: they were used for representing highly interconnected data and describing model information.		RDF is the easiest, most powerful and expressive standard designed by now.	Köpcke, H., Thor, A., & Rahm, E. (2009). Comparative evaluation of entity resolution approaches with fever. *Proceedings of the VLDB Endowment*, *2*(2), 1574-1577.
7	ABT-BUY	The ABT-BUY dataset was used for study and experimental purposes.	The dataset contains 1081 entities from abt.com and 1092 entities from buy.com as well as a gold standard with 1097 matching record pairs.		Köpcke, H., Thor, A., & Rahm, E. (2009). Comparative evaluation of entity resolution approaches with fever. *Proceedings of the VLDB Endowment*, *2*(2), 1574-1577.

	Corpus / Dataset Used	Metrics used to evaluate	Number of Documents	Advantage of dataset	References
8	DUC-2003	The dataset used for testing the automatic text summarizer (The average recall, precision and the F-score were calculated for each ROUGE score)	2004 corpus having 500 document and summary pairs	DUC 2003 will incorporate focus of various sorts to reduce variability and better model real tasks	Over, P. (2003). An introduction to DUC 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of Document Understanding Conference 2003*.
9	SogouCA	SogouCA was used in the model as it was crawled and provided by Sogou Labs from dozens of Chinese news websites, news reports and reviews. SogouCA leveraged "url" information, which helped in categorizing documents into 15 corresponding documents	a benchmark **dataset** that contains 500K **document** pages with fine-grained token-level	In terms of **data** size, **Sogou**-QCL is far larger than the two **datasets**, which is a main **advantage** to serve the train- ing of deep neural networks.	Wang, F., Zhou, Y., & Lan, M. (2016, November). Dimensional sentiment analysis of traditional Chinese words using pre-trained Not-quite-right Sentiment Word Vectors and supervised ensemble models. In *2016 International Conference on Asian Language Processing (IALP)* (pp. 300-303). IEEE.
10	RST-DT datasets	The commonly used evaluation metric for text summarization is ROUGE. ROUGE evaluates n-gram co-occurrences between summary pairs.	The Rhetorical Structure Theory (RST) Discourse Treebank consists of 385 Wall Street Journal articles from the Penn Treebank	It has a large number of articles which help in evaluation and comparison.	Huber, P., & Carenini, G. (2019). Predicting discourse structure using distant supervision from sentiment. *arXiv preprint arXiv:1910.14176*.
11	TAC 2008	71 submitted runs of this dataset were automatically evaluated with the ROUGE and BE metrics, NIST assessors manually evaluated only 57 of the submitted runs.	In the TAC 2008 summarization track, the main task was to produce two 100-word summaries from two related sets of 10 documents	The TAC 2008 datasets are used since they provide large data collections for testing. Main focus for evaluation is on Opinion Summarization Track in which there are 25 topics and an average of 24 documents. There is a summary length constraint of 250 words on this track too.	TAC 2008 - Dang, H. T., & Owczarzak, K. (2008). Overview of the tac 2008 opinion question answering and summarization tasks. In Proc. of the First Text Analysis Conference (Vol. 2).

	Corpus / Dataset Used	Metrics used to evaluate	Number of Documents	Advantage of dataset	References
12	1. WikiData 2. DBpedia	Python and external library: NLTK, SPARQLWrapper, Mysql-Connector, Sumy, Pandas, PyPDF2 and also OWL/RDF - SPARQL EndPoint and Mysql	Extracted from Wikipedia's large dataset.	These datasets are used because they are readily available to everyone and have a large amount of data that can be experimented on. DBpedia extracts structured data from the infoboxes in Wikipedia, and publishes them in RDF and a few other formats.Wikidata provides a secondary and tertiary database of structured data that everyone can edit.	Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85.; Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722-735). Springer, Berlin, Heidelberg.

Comparison of base papers

In this section, we compare our selected base papers on the basis of their advantages and disadvantages as well as the approach used in these papers.

	Pros	Cons	Name of the Approach	Paper
1	Due to the help of position awareness, SEGMENT can capture long-range dependencies	The Segment with Gated attention can potentially benefit the model but there is a slight decrease in performance.	Segment augmented with TCN architecture	Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
2	The model can generate better quality summaries by integrating Frame semantics information.	The differences between the proposed method and ground truth are relatively small.	A novel Frame semantic guided network (FSum)	Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
3	The model outperforms all traditional and BERT-based summarization approaches, particularly when the number of selected phrases is between 3 and 5. (i.e., the most common number of requested highlights).	The model is constrained by the number of selected phrases as it increases and the performance of the model decreases.	Article ranking based approach for annotating articles.	Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.
4	The summaries generated were similar to human generated summaries or at least comparable.	The domain-based summaries have not been generated or tested hence the result for that is unclear; also the performance of summarization on the same is unknown.	LSTM model based hierarchical structure for processing semantic and syntactic features of text for summarization.	Chen, Y. C., & Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.
5	The keyword-selection process allows a more accurate detection of the representative sentences of the documents because these words tend to be contained in the key sentences.	Even though the approach is language and domain independent, the LDA algorithm showed much more results on some specific languages than others.	Automatic Summarization by clustering sentences and utilizing Genetic Algorithms	Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. *Advances in neural information processing systems*, *28*.

Comparison of base papers

In this section, we compare our selected base papers on the basis of their advantages and disadvantages as well as the approach used in these papers.

	Pros	Cons	Name of the Approach	Paper
6	They also introduce the first version of a neural-based LS verbalization approach trained by automatically generated verbalization from our template-based LS approach. In addition, they summarize the produced multilingual verbalization using a selectivity-based LS summarization approach.	Their evaluation shows that when the LS is more complex and contains different operators, the fluency of our approach decreases.	Multilingual Verbalization and Summarization for Explainable Link Discovery	Ahmed, A. F., Sherif, M. A., Moussallem, D., & Ngomo, A. C. N. (2021). Multilingual Verbalization and Summarization for Explainable Link Discovery. *Data & Knowledge Engineering*, *133*, 101874.
7	The experimental result was evaluated using a ROUGE score. The proposed FbTS algorithm showed a higher ROUGE-1 and ROUGE-2 score than the other nature-inspired ones, genetic algorithm and particle swarm optimization.	Due to many calculations there is a chance of error in the work, it takes time and has to process many things.	Multi-document extractive text summarization based on firefly algorithm	Tomer, M., & Kumar, M. (2021). Multi-document extractive text summarization based on firefly algorithm. *Journal of King Saud University-Computer and Information Sciences*.
8	This method can be applied to different languages.(The TAC11 dataset was selected to prove that our methods can be applied to different languages) None of the procedures introduced in this paper require a priori information to generate vectors.	The main disadvantage of bag-of-words methods is that context information is lost. Each index has advantages and disadvantages for different datasets.	Language-independent extractive automatic text summarization based on automatic keyword extraction	Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., & Millán-Hernández, C. E. (2022). Language-independent extractive automatic text summarization based on automatic keyword extraction. *Computer Speech & Language*, *71*, 101267.
9	It leveraged the idea of translation and designed a novel attention-based sequence-to sequence model for rhetorical relation identification. Then the subroutine-based ATS model can accept different word length	Using the RS tree in the algorithm takes time for large sentences, and in future rs trees will be utilized for better working.	Knowledge-guided unsupervised rhetorical parsing for text summarization	Hou, S., & Lu, R. (2020). Knowledge-guided unsupervised rhetorical parsing for text summarization. *Information Systems*, *94*, 101615.

Comparison of base papers

In this section, we compare our selected base papers on the basis of their advantages and disadvantages as well as the approach used in these papers.

	Pros	Cons	Name of the Approach	Paper
	limit or summarization ratio and provide content-balanced results based on RS-tree			
10	The integration of fuzzy with deep learning models not only improves the results but also reduces the time required to train the abstractive model. The idea is to overcome the limitations of extractive and abstractive summarization.	The limitation of abstractive summarization is that it needs large databases to train and training consumes lots of time and the limitation of extractive summarization is that it just selects the important sentences from the summary.	Improving Text Summarization using Ensembled Approach based on Fuzzy with LSTM	Tomer, M., & Kumar, M. (2020). Improving text summarization using ensembled approach based on fuzzy with LSTM. *Arabian Journal for Science and Engineering*, *45*(12), 10743-10754.
11	The major strong point of the proposed solution, i.e., QSO-MOCSA, is that it outperforms the other existing summarization methods in the scientific literature. Furthermore, it focuses on the analysis of the polarities of the sentences from a document collection, also considering their sentiment scores. Therefore, it becomes possible to produce a summary that includes the most relevant sentences for the user's query, also having a similar sentiment orientation.	Since the algorithm does a lot of things to make the summary perfect, it uses a lot of energy. So, it is energy inefficient.	Sentiment-oriented query-focused text summarization addressed with a multi-objective optimization approach.	Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2021). Sentiment-oriented query-focused text summarization addressed with a multi-objective optimization approach. Applied Soft Computing, 113, 107915.
12	The major advantage of the proposed solution is that it minimized the energy consumption by around 50–90% with degradation of summary quality by nearly 5–40% based on the selection of different configuration parameters. It gives a promising solution for developing a sustainable energy-aware summarization	The main disadvantage of this approach is that even though the summarization process consumes less energy, still a small percentage of summary quality degradation is observed or vice versa in some cases.	Sustainable text summarization over mobile devices using hybrid summarization algorithms.	Hazra, K., Ghosh, T., Mukherjee, A., Saha, S., Nandi, S., Ghosh, S., & Chakraborty, S. (2021). Sustainable text summarization over mobile devices: An energy-aware approach. Sustainable Computing: Informatics and Systems, 32, 100607.

Comparison of base papers

In this section, we compare our selected base papers on the basis of their advantages and disadvantages as well as the approach used in these papers.

	Pros	Cons	Name of the Approach	Paper
	approach that minimizes energy efficiency while giving a good quality summary.			
13	With the help of the proposed method's combined vector approach, it is shown that the sentence covering most of the topics representing the input document is assigned a higher rank. As a result, even more, topics can be covered in fewer sentences, which satisfies the summary's completeness property. This is the main advantage of the proposed approach.	As most of the redundancy is removed from the summaries produced, there is a chance that some important data can be removed too instead of redundant data.	Text summarization using topic-based vector space model and semantic measure.	Belwal, R. C., Rai, S., & Gupta, A. (2021). Text summarization using topic-based vector space model and semantic measure. Information Processing & Management, 58(3), 102536.
14	The use of summary before displaying the results to the user solves the problem of having duplicate sentences by removing duplicate sentences and merging the summarized sentences into one text.	This solution is burdened with a long processing time at the initial stages of the proposed solution.	Multidimensional approach to text summarization	Janaszkiewicz, P., & Różewski, P. (2019). The method of multidimensional approach to text summarization. Procedia Computer Science, 159, 2189-2196.
15	NATSUM performs better than other summarization approaches including extractive summarization approaches. It also performs better than the multi-document entity-focused extractive summarization tested.	This algorithm takes time to generate summaries and consumes relatively more energy than other summarization algorithms.	NATSUM: Narrative abstractive summarization through cross-document timeline generation	Barros, C., Lloret, E., Saquete, E., & Navarro-Colorado, B. (2019). NATSUM: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, *56*(5), 1775-1793.

Conclusion and Future work:-

With this survey, we have attempted to give a comprehensive overview of the most prominent recent methods for automatic text summarization. We have outlined the connection to early approaches and have contrasted approaches in terms of how they represent the input, score sentences and select the summary. We have shown various classifications of text summarization techniques and systems and their technicality and also compared them with various other older as well as newer text summarization techniques. The recent summarization techniques have not only made the retrieval process of information easier for us, it has also helped us save a lot of time in these times where there is an overflow of information on the internet. In this survey, we have done an in-depth analysis of all the base papers that we selected for the process.

We studied the various methods of text summarization and did this survey. But a lot of flaws can be found in even the latest of techniques. For instance, in a hybrid summarization system, better base algorithms can be used and included. Likewise, there are some or the other flaws with all the summarization techniques which we surveyed in this paper. These flaws include either time complexity or quality of summaries. At other times, there are problems with multi-document summarization. All these flaws can be corrected at some later point of time when further work will be done on these techniques.

The future work is to improve these text summarization techniques to such a level that they resemble human techniques and interactions.

[1] Nguyen, M. P., & Tran, N. T. (2021). Improving Abstractive Summarization with Segment-Augmented And Position-Awareness. *Procedia Computer Science*, *189*, 167-174.

[2] Guan, Y., Guo, S., Li, R., Li, X., & Zhang, H. (2021). Frame semantics guided network for abstractive sentence summarization. *Knowledge-Based Systems*, *221*, 106973.

[3] Cagliero, L., & La Quatra, M. (2020). Extracting highlights of scientific articles: A supervised summarization approach. *Expert Systems with Applications*, *160*, 113659.

[4] Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2020). Candidate sentence selection for extractive text summarization. *Information Processing & Management*, *57*(6), 102359.

[5] Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., & Millán-Hernández, C. E. (2020). Extractive automatic text summarization based on lexical-semantic keywords. *IEEE Access*, *8*, 49896-49907.

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[7] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

[9] Chen, Y. C., & Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

[10] Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. *Advances in neural information processing systems*, *28*.

[11] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

[12] Tu, Z., Lu, Z., Liu, Y., Liu, X., & Li, H. (2016). Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*.

[13] Zhou, Q., Yang, N., Wei, F., & Zhou, M. (2017). Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*.

[14] Wang, K., Quan, X., & Wang, R. (2019). BiSET: Bi-directional selective encoding with template for abstractive summarization. *arXiv preprint arXiv:1906.05012*.

[15] Zhao, W., Peng, H., Eger, S., Cambria, E., & Yang, M. (2019). Towards scalable and reliable capsule networks for challenging NLP applications. *arXiv preprint arXiv:1906.02829*.

[16] Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A generative model for category text generation. *Information Sciences*, *450*, 301-315.

[17] Hsu, W. T., Lin, C. K., Lee, M. Y., Min, K., Tang, J., & Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

[18] Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

[19] Cao, Z., Wei, F., Li, W., & Li, S. (2018, April). Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.

[20] Gollapalli, S. D., & Caragea, C. (2014, June). Extracting keyphrases from research papers using citation networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 28, No. 1).

[21] Nikolov, N. I., Pfeiffer, M., & Hahnloser, R. H. (2018). Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

[22] Balabantaray, R. C., Sahoo, D., Sahoo, B., & Swain, M. (2012). Text summarization using term weights. International Journal of Computer Applications, 38(1), 10–14. Barrios, F., Lopez, F., Argerich, L., & Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. Argentine symposium on artificial intelligence.

[23] Gong, Y., & Liu, X. (2001, September). 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* (pp. 19-25).

[24] Brdiczka, O., & Chu, M. K. (2011). *U.S. Patent No. 8,086,548*. Washington, DC: U.S. Patent and Trademark Office.

[25] Aliguliyev, R. M. (2006, December). A novel partitioning-based clustering method and generic document summarization. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* (pp. 626-629). IEEE.

[26] Fang, C., Mu, D., Deng, Z., & Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, *72*, 189-195.

[27] Mohamed, M., & Oussalah, M. (2019). SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, *56*(4), 1356-1372.

[28] Nallapati, R., Zhai, F., & Zhou, B. (2017, February). Summarunner: A recurrent neural network-based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*.

[29] Hassan, M., & Hill, E. (2018, September). Toward automatic summarization of arbitrary java statements for novice programmers. In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 539-543). IEEE.

[30] Cardinaels, E., Hollander, S., & White, B. J. (2019). Automatic summarization of earnings releases: attributes and effects on investors' judgments. *Review of Accounting Studies*, *24*(3), 860-890.

[31] Afsharzadeh, M., Ebrahimpour-Komleh, H., & Bagheri, A. (2018, April). Query-oriented text summarization using sentence extraction technique. In *2018 4th international conference on web research (ICWR)* (pp. 128-132). IEEE.

[32] Sakhadeo, A., & Srivastava, N. (2018). Effective extractive summarization using frequency-filtered entity relationship graphs. *arXiv preprint arXiv:1810.10419*.

[33] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

[34] Charitha, S., Chittaragi, N. B., & Koolagudi, S. G. (2018, August). Extractive document summarization using a supervised learning approach. In *2018 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)* (pp. 1-6). IEEE.

[35] Sinha, A., Yadav, A., & Gahlot, A. (2018). Extractive text summarization using neural networks. *arXiv preprint arXiv:1802.10137*.

[36] García-Hernández, R. A., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., & Cruz, R. (2008, October). Text summarization by sentence extraction using unsupervised learning. In *Mexican International Conference on Artificial Intelligence* (pp. 133-143). Springer, Berlin, Heidelberg.

[37] Fernández-González, D., & Gómez-Rodríguez, C. (2020, April). Discontinuous constituent parsing with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 7724-7731).

[38] Li, C., Xu, W., Li, S., & Gao, S. (2018, June). Guiding generation for abstractive text summarization based on key information guide network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 55-60).

[39] Chen, G., Ye, D., Xing, Z., Chen, J., & Cambria, E. (2017, May). Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International joint conference on neural networks (IJCNN)* (pp. 2377-2383). IEEE.

[40] Chaturvedi, I., Ong, Y. S., Tsang, I. W., Welsch, R. E., & Cambria, E. (2016). Learning word dependencies in text by means of a deep recurrent belief network. *Knowledge-Based Systems*, *108*, 144-154.

[41] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

[42] Li, L., Liu, W., Litvak, M., Vanetik, N., & Huang, Z. (2019). In conclusion not repetition: Comprehensive abstractive summarization with diversified attention based on determinantal point processes. *arXiv preprint arXiv:1909.10852*.

[43] Cagliero, L., & La Quatra, M. (2020). Extracting highlights of scientific articles: A supervised summarization approach. *Expert Systems with Applications*, *160*, 113659.

[44] Gollapalli, S. D., & Caragea, C. (2014, June). Extracting keyphrases from research papers using citation networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 28, No. 1).

[45] Collins, E., Augenstein, I., & Riedel, S. (2017). A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

[46] Kim, M., Singh, M. D., & Lee, M. (2016). Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization. *arXiv preprint arXiv:1607.00718*.

[47] Nikolov, N. I., Pfeiffer, M., & Hahnloser, R. H. (2018). Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

[48] Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, *129*, 200-215.

[49] Ledeneva, Y., García-Hernández, R. A., & Gelbukh, A. (2014, April). Graph ranking on maximal frequent sequences for single extractive text summarization. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 466-480). Springer, Berlin, Heidelberg.

[50] Bando, L. L., Lopez, K. R., Vidal, M. T., Ayala, D. V., & Martinez, B. B. (2007, September). Comparing four methods to select keywords that use n-Grams to generate summaries. In *Electronics, Robotics and Automotive Mechanics Conference (CERMA 2007)* (pp. 724-728). IEEE.

[51] Lin, C. Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics* (pp. 150-157).

[52] Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

[53] Chopra, S., Auli, M., & Rush, A. M. (2016, June). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 93-98).

[54] Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

[55] Gao, Y., Wang, Y., Liu, L., Guo, Y., & Huang, H. (2020). Neural abstractive summarization fusing by global generative topics. *Neural Computing and Applications*, *32*(9), 5049-5058.

[56] Sanchez-Gomez, J. M., Vega-Rodríguez, M. A., & Pérez, C. J. (2021). Sentiment-oriented query-focused text summarization addressed with a multi-objective optimization approach. Applied Soft Computing, 113, 107915.

[57] Hazra, K., Ghosh, T., Mukherjee, A., Saha, S., Nandi, S., Ghosh, S., & Chakraborty, S. (2021). Sustainable text summarization over mobile devices: An energy-aware approach. Sustainable Computing: Informatics and Systems, 32, 100607.

[58] Belwal, R. C., Rai, S., & Gupta, A. (2021). Text summarization using topic-based vector space model and semantic measure. Information Processing & Management, 58(3), 102536.

[59] Janaszkiewicz, P., & Różewski, P. (2019). The method of multidimensional approach to text summarization. Procedia Computer Science, 159, 2189-2196.

[60] He, T., Chen, J., Gui, Z., & Li, F. (2008). CCNU at TAC 2008: Proceeding on Using Semantic Method for Automated Summarization Yield. In *TAC*.

[61] Varma, V., Pingali, P., Katragadda, R., Krishna, S., Ganesh, S., Sarvabhotla, K., ... & Bharadwaj, R. G. (2009, November). IIIT Hyderabad at TAC 2009. In *TAC*.

[62] Lin, Z., Hoang, H. H., Qiu, L., Ye, S., & Kan, M. Y. (2008). NUS at TAC 2008: Augmenting Timestamped Graphs with Event Information and Selectively Expanding Opinion Contexts. In *TAC*.

[63] Li, W., You, O., Hu, Y., & Wei, F. (2008, November). PolyU at TAC 2008. In *TAC*.

[64] Kumar, S., & Chatterjee, D. (2008). IIT Kharagpur at TAC 2008: Statistical Model for Opinion Summarization. In *TAC*.

[65] Abdi, A., Shamsuddin, S. M., & Aliguliyev, R. M. (2018). QMOS: Query-based multi-documents opinion-oriented summarization. *Information Processing & Management*, *54*(2), 318-338.

[66] Mothe, J., Chrisment, C., Dousset, B., & Alaux, J. (2003). DocCube: Multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology*, *54*(7), 650-659.

[67] Park, B. K., Han, H., & Song, I. Y. (2005, August). XML-OLAP: A multidimensional analysis framework for XML warehouses. In the International *Conference on Data Warehousing and Knowledge Discovery* (pp. 32-42). Springer, Berlin, Heidelberg.

[68] Tao, F., Zhang, C., Chen, X., Jiang, M., Hanratty, T., Kaplan, L., & Han, J. (2015). Doc2cube: Automated document allocation to text cube via dimension-aware joint embedding. *Dimension*, *2016*, 2017.

[69] Zhang, D., Zhai, C., & Han, J. (2009, April). Topic cube: Topic modeling for olap on multidimensional text databases. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (pp. 1124-1135). Society for Industrial and Applied Mathematics.

[70] Varma, V., Pingali, P., Katragadda, R., Krishna, S., Ganesh, S., Sarvabhotla, K., ... & Bharadwaj, R. G. (2009, November). IIIT Hyderabad at TAC 2009. In *TAC*.

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

[72]A.-C.N. Ngomo, S. Auer, Limes - A time-efficient approach for large-scale link discovery on the web of data, in: IJCAI, 2011

[73]D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR, 2015.

[74]G. Doddington, Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: Proceedings of HLT, 2002, pp. 138–145.

[75]Ngomo, A. C. N., & Auer, S. (2011, June). LIMES—a time-efficient approach for large-scale link discovery on the web of data. In *Twenty-Second International Joint Conference on Artificial Intelligence*

[76]Pohl, A. (2010, November). The polish interface for linked open data. In Proceedings of the ISWC (pp. 165-168).

[77]Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., ... & Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, *40*(14), 5755-5764.

[78]Sherif, M. A., Ngonga Ngomo, A. C., & Lehmann, J. (2017, May). W ombat–a generalization approach for automatic link discovery. In *European Semantic Web Conference* (pp. 103-119). Springer, Cham.

[79]Rautray, R., & Balabantaray, R. C. (2018). An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA. *Applied computing and informatics*, *14*(2), 134-144

[80]Rautray, R., & Balabantaray, R. C. (2018). An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA. *Applied computing and informatics*, *14*(2), 134-144

[81]Verma, P., & Om, H. (2019). MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization. *Expert Systems with Applications*, *120*, 43-56.

[82]Lee, J. S., Hah, H. H., & Park, S. C. (2017). Less-redundant text summarization using ensemble clustering algorithm based on GA and PSO. *Wseas Transactions On Computers*, *16*.

[83]Peng, L., Zhu, Q., Lv, S. X., & Wang, L. (2020). Effective long short-term memory with fruit fly optimization algorithm for time series forecasting. *Soft Computing*, *24*(19), 15059-15079.

[84]Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.

[85]Belkebir, R., & Guessoum, A. (2015). A supervised approach to arabic text summarization using adaboost. In *New contributions in information systems and technologies* (pp. 227-236). Springer, Cham.

[86]Fattah, M. A., & Ren, F. (2008). Automatic text summarization. *World Academy of Science, Engineering and Technology*, *37*(2), 192.

[87]García-Hernández, R. A., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., & Cruz, R. (2008, October). Text summarization by sentence extraction using unsupervised learning. In *Mexican International Conference on Artificial Intelligence* (pp. 133-143). Springer, Berlin, Heidelberg.

[88]Narayan, S., Cohen, S. B., & Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

[89]Suanmali, L., Binwahlan, M. S., & Salim, N. (2009, August). Sentence features fusion for text summarization using fuzzy logic. In *2009 Ninth International Conference on Hybrid Intelligent Systems* (Vol. 1, pp. 142-146). IEEE.

[90]Witte, R., & Bergler, S. (2007, May). Fuzzy clustering for topic analysis and summarization of document collections. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 476-488). Springer, Berlin, Heidelberg.

[91]Dixit, R. S., & Apte, S. S. (2012). Improvement of text summarization using fuzzy logic based method. *IOSR Journal of Computer Engineering (IOSRJCE)*, *5*(6), 5-10.

[92]Chopade, H. A., & Narvekar, M. (2017, November). Hybrid auto text summarization using deep neural network and fuzzy logic system. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 52-56). IEEE.

[93]Sahba, R., Ebadi, N., Jamshidi, M., & Rad, P. (2018, June). Automatic text summarization using customizable fuzzy features and attention on the context and vocabulary. In *2018 World Automation Congress (WAC)* (pp. 1-5). IEEE.

[94]Deepa, R., Konshi, J., Haritha, A., & Shobini, K. Automatic Text Summarization System.

[95]Lamsiyah, S., El Mahdaouy, A., Espinasse, B., & Ouatik, S. E. A. (2021). An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. Expert Systems with Applications, 167, 114152.

[96] Rani, R., & Lobiyal, D. K. (2021). A weighted word embedding based approach for extractive text summarization. Expert Systems with Applications, 186, 115867.

[97] Mojrian, M., & Mirroshandel, S. A. (2021). A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA. Expert systems with applications, 171, 114555.

[98] Abdi, A., Hasan, S., Shamsuddin, S. M., Idris, N., & Piran, J. (2021). A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion. Knowledge-Based Systems, 213, 106658.

[99]Rautray, R., & Balabantaray, R. C. (2018). An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA. Applied computing and informatics, 14(2), 134-144

[100] Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, *47*(1), 1-66.

[101] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

[102] Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data* (pp. 43-76). Springer, Boston, MA.

[103] Hasan, K. S., & Ng, V. (2014, June). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1262-1273).