Text Summarization and Keyword Extraction using Lexical Analysis and NLP

[1]Abhinav Sarkar, [2]Rishabh Agarwal, [3]Dheerendra Chaudhary, [4]Sandeep Kumar

CSE department
Sharda University
Greater Noida, India.

***Abstract*- The paper is mainly based on time consumption reduction, which is the most required aspect nowadays. This paper provides solution to the problem that the reader/user needs to spend a lot of time in reading articles or papers written about any particular topic which they want to learn about for a particular type of task they want to perform, as for performing any work research about that particular topic is needed and for the requirement analysis phase as those studies helps to relate what will be required for their work, for which the user/reader needs to spend a lot of time finding the documents/papers which are of their use and moreover reading the whole document and research paper to gain information. This process becomes takes a lot of time as reading all those documents and paper is a lengthy work. Which is a major problem as it takes a lot of time in today's world, where everyone is trying to minimize time consumption in every field. The paper is providing the solution to this problem with the help of providing summaries and keywords for a particular document. Summaries can be thought as the shortened form of the document, to which gives the overall idea of the entire document in smaller portion that is in lesser number of words, while the keywords are words which have some specific meaning and helps the reader/user to understand the context of the passage. These two-term used above will solve this major problem by decreasing the time consumption in this research and requirement analysis phase.**

The keyword extraction is used for finding the relevant words, which have a particular meaning and provide idea of the context of particular document, decreasing the amount of time find the relevant document and papers and finding whether a particular document is relevant or not. After finding the relevant document or paper text summarization module is used to shorten the entire document without changing the content of the that particular document and reading the number of words present in the document, providing the content of the document is short reducing the amount of time required for information gathering as now the user will have to read short documents/ less words the gain the same amount of information that they will gain by reading he same document, reducing the amount of time required more. This work approaches the issue of automatic keyword extraction from papers as a supervised learning challenge. It can be claimed that a lexical chain represents the semantic content of a section of text because it holds a group of words from a text that are semantically related to one another. Lexical chains have been employed extensively in text summarization, but their application to the problem of keyword extraction has not yet been properly explored. In this study, a lexical chain-based keyword extraction method is introduced, and encouraging results are produced.

Keywords: **Keyword extraction; Lexical chains; Natural language processing; Machine learning.**

I. Introduction

Data has recently increased quickly across all industries, including journalism, social media, banking, education, etc. Due to the abundance of data, an automatic summarizer is required that can condense the data, particularly textual material in the original document, without sacrificing any important objectives. In recent years, text summarization has become a significant topic of study. Reviewing previous research on the text summarising method is helpful for conducting additional study in this area. Since the process of text summarising heavily depends on keyword extraction, recent literature on automatic keyword extraction and text summarization is presented in this work. This body of work discusses many approaches to text summarization and keyword extraction. It also touches on several databases utilised in conjunction with evaluation matrices for text summarization across several areas. Finally, it briefly highlights problems and difficulties in study that researchers have encountered, along with possible future directions.

Knowledge production has expanded dramatically in a variety of fields, including news, social media, education, finance, etc. There is a need for an automatic summarizer that can condense the textual data within the original document while maintaining the integrity of the information due to the variety of material that we are exposed to. In recent years, text summarization has become a crucial topic of study. Since the process of text summarization heavily relies on keyword extraction, this paper provides recent research on automatic keyword extraction and text summarization. The numerous techniques for text summarization and keyword extraction are discussed in this literature. Finally, it briefly explores the potential future course that these processes could follow.

Key phrases can be assigned by authors to their documents; these key phrases may or may not appear in the content. In automatic keyphrase extraction, the phrases that are most representative of a document are chosen as keyphrases for that document. As a result, text-based phrases are the only ones that can be used by automatic keyphrase extraction methods. The more generic variant

of keyphrase extraction is keyphrase generation, which creates and assigns keyphrases for the document rather than choosing phrases from it. To underline that keyphrases might consist of more than one word, we exclusively extract keywords in this study and only focus on "keywords" as opposed to "keyphrases". According to our opinion, a keyword in a text should be semantically related to its terms. A portion of the words (word senses) in a text are included in the lexical chain for that text. The lexical chain's words have semantic ties to one another. There are different sizes of lexical chains that can be found in a text. Each lexical chain may contain a varied number of words and the number of semantic relationships between those words. How well a lexical chain captures the semantic content of the text can be determined by its size and coverage. Therefore, we think that a keyword should be chosen from a lexical chain of words that conveys the semantic substance of the text the majority of the text's semantic content. In this study, we describe a keyword extraction method that chooses keywords for a text based on attributes based on lexical chains.

Automated text summarization and keyword extraction are closely related processes. The most illustrative sentences from the text are taken out for text summarising. In keyword extraction, the text is represented by the most suggestive terms. In order to identify a pattern denoting importance in a text, factors including word frequencies, cue phrases, position in text, lexical chains, and discourse structure are used in both of these difficulties. In this study, we investigate the impact of lexical chains in the context of supervised machine learning for the problem of keyword extraction. Features from the lexical chains of words are used in this learning exercise. We focus on the keyword extraction issue because the WordNet ontology only allows us to create lexical chains for words, not phrases instead of keyphrase extraction.

Although we tested a variety of classifiers, including Naive Bayes, the decision tree induction technique produced the best results. Because of this, we have employed NLP to model the keyword extraction issue as a learning activity. With two separate sets of characteristics, we applied NLP. We simply utilised the text features in our baseline system (without using any feature based on the lexical chains of words). In the second instance, in addition to the characteristics used in - the baseline system, NLP was utilised using features based on lexical chains. The outcomes of these two iterations are then compared. When we employed the lexical chain- based characteristics, the results were better. We begin by outlining the relevant work on lexical chains and keyword extraction. The formation of lexical chains and their definition follow We examine the outcomes of our keyword extraction method after explaining the lexical chain-based features that are used in our keyword extraction system.

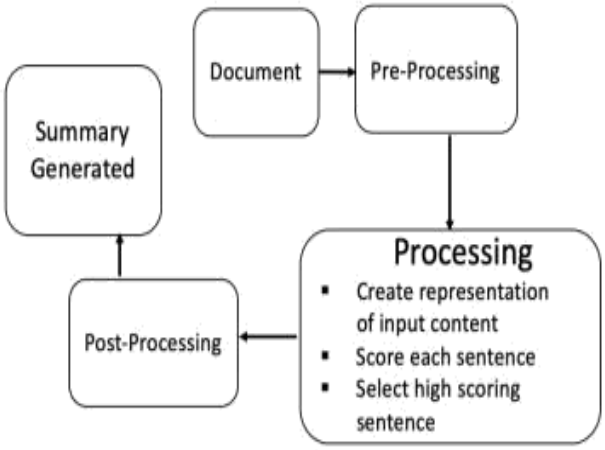


Fig 1. Extractive Text Summarization Architecture, Adopted from [5]

II. Literature Survey

They took the keywords from a Chinese microblog, and to do so, they used three features: a graph model, a semantic space, and the word locations. wherein they applied the technique The user's micro blog API must be downloaded in the first stage. We perform pre-processing in the second phase by cleaning up the data, word- segmenting, POS-tagging, and stopping words. In the third phase, we build a graph model to extract the keywords based on the co-occurrence of the words, assign a sequence number to the words based on where they are, and calculate the weight of the words using the Score method. The fourth phase involves creating a semantic space based on a topic. TFIDF is used to identify and calculate statistical weight. In the fifth phase, we take into account still another factor—word location—and compute the rank value, concluding that a number with a smaller location will be ranked higher. [1]

The structure approach and graph creation are the main topics of this study. The strategy utilized in this paper is structure-based; we build a graph model and identify the themes and events that spike in frequency. Twitter tweets are divided into homogeneous and heterogeneous graphs during the clustering of topics process. To locate the users in a homogeneous graph, we apply the OSLOM algorithm. interaction. To create a series of tweets ranked with a number for heterogeneous rank, we apply the rankclus

algorithm. Finally, using Python to extract the meaning for each tweet, we connect all of the tweets with the same name. In the future, several graph models may be used to define various sorts of events and to define the events themselves. [2]

They developed an approach for keyword extraction in order to address issues like high variance and lexical variants, as well as to contrast the present method with earlier ones. For the issue of lexical variation, where words that sound differently but have the same meaning are concerned, the technique we employ is blind to this fact. As a result, we employed two techniques, both of which include Brown clustering: -in this, we group words that have the same meaning, such as "no," "noo," and others, into clusters before identifying the feature for each cluster. [3]

High relevance keyword extraction (HRKE), a tool created for Bayesian text classification, allows for the extraction of keywords during the classification stage without the usage of pre-classification procedures. The facility extracts the keywords using a posterior probability value. The HRKE employs a Bayesian classification strategy, where the first step is to extract words from a text that comprises a list of words. This list is created by assuming that the text is n words long, and the posterior probability is then determined. then the weights for the words are assigned using the TFIDF method. [4]

In this paper, a measure that ranks words is proposed. This method makes use of Shannon's entropy, which distinguishes between intrinsic and extrinsic mode, meaning that the words in the text identify the author's purpose while the irrelevant words appear at random. This approach works well when dealing with a single document about which no prior knowledge is available. The concept behind intrinsic and extrinsic is that meaningful words are grouped together. The extracted words are ranked based on the entropy difference. Entropy difference is used to calculate the mean, mode, and median, and the ED metric performed well in terms of ordering the words. Future applications of this type of work include several types of natural language processing. [5]

The brain has been the subject of numerous studies that have been widely applied worldwide. It is important to utilize the resources effectively and efficiently. There are numerous neuro- informatics sites and centers that are utilized to communicate knowledge and resources relating to the brain among other nations. There are many created platforms with their own keywords that reflect the essential terms. The primary benefit of the previously described keywords is that they aid in categorizing both the main content and the resources. Thus, a technique that automatically identifies the most crucial terms is defined. [6]

The author of this work has examined and contrasted the effectiveness of three alternative algorithms. First, an explanation of the various text summarizing methods. Important keywords are extracted using extraction-based approaches and added to the summary. Three keyword extraction methods, TextRank, LexRank, and Latent Semantic Analysis (LSA), were utilised for comparison. Python code is used to demonstrate and execute three algorithms. The ROUGE 1 is employed to assess how successful the extracted keywords are. The performance of the algorithms was assessed by comparing the output to handwritten summaries. The Text-Rank Algorithm ultimately provides a superior result than the other two-algorithms.[7]

In this study, the author suggests a method for creating abstractive summaries from extractive summaries using the WordNet ontology. Multiple documents, including text, PDF, and Word files, were utilised. The author first covered a variety of text summarising methods before going over step-by-step procedures for text summary of several documents. The experiment's results are compared to those of currently available online extraction tools and to those of human-written summaries, and it is clear that the suggested system produces good results. Finally, the author suggested that in the future, the accuracy of the summarization might be improved by contrasting this abstractive system with other abstractive systems. [8]

The author of this study report suggests two techniques for producing general text summaries by ranking and selecting sentences from the primary text sources. The first technique ranks the sentence relevance and assigns relevance scores to sentences using information retrieval (IR) techniques. The second method makes use of the latent semantic analysis (LSA) method, which is based on latent semantic indexing (LSI), to determine the phrases' semantic relevance in order to create summaries. To create the text summary, the author employs the Singular Value Decomposition (SVD). Additionally, the paper's author walks readers through each stage of the SVD-based approaches. The approaches used for this aim offer general, abstractive summaries. The outcomes are then contrasted with the summaries created by people. It produces abstractive summaries that are more human-like. The author suggested researching different machine learning techniques in the future to enhance the quality of generic text summarization. [9]

With a few small adjustments, we have applied the strategy outlined in Silber and McCoy (2002). The WordNet database is re-indexed in (Silber and McCoy, 2002) to make accessing it easier. This re-indexing was skipped in favour of flat relationship lists for each word (up to 3 levels of depth). Using this method, we were able to find a relationship between two-word senses more quickly and check for relationships in linear time. The first step in creating lexical chains is to locate all nouns in the text. Since noun relations convey more information about the subject and are frequently used as nouns in text, we have exclusively used noun relations. [10]

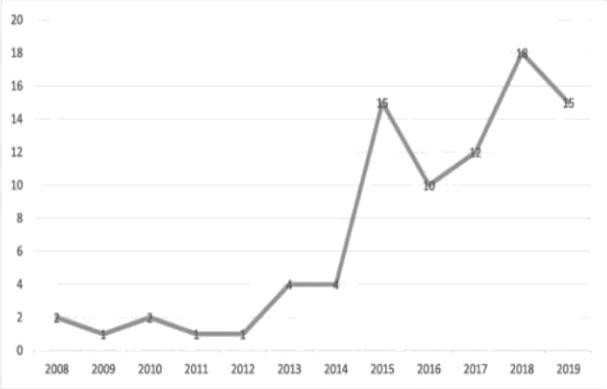
There have been numerous research using this learning method for text summarization, and the training uses a labelled dataset. In an effort to detect negative events, Aramaki et al. attempted to implement a simple system of medical text summarization using supervised learning. They also looked into the forms of information that were helpful in identifying negative events. He employs the SVM Classification to separate bad occurrences from other ones. The method used by Kamal et al. to automatically

summarise medical publications using supervised learning. The decision tree C4.5 has been selected as the base learner in the machine learning method known as baggin. Riadh and Ahmed's team used AdaBoost to present summaries of Arabic texts using a supervised learning approach.[11]

There are no training recommendations for this type of learning technique. By extracting phrases using K-Means in an independent domain utilising a supervised learning approach, René et almethod .'s produces automatic text summarization. Similar concepts (sentences) can be grouped together using this technique. The best sentence from each cluster should then be used to create the summary. Summaries of automatic text in bilingual (Hindi and English) languages are presented by Shasi et al. utilising unsupervised deep learning. Researchers employed the Boltzmann machine to create shorter summaries without losing crucial information in order to increase the overall accuracy of their findings. [12]

almethod .'s produces automatic text summarization. Similar concepts (sentences) can be grouped together using this technique. The best sentence from each cluster should then be used to create the summary. Summaries of automatic text in bilingual (Hindi and English) languages are presented by Shasi et al. utilising unsupervised deep learning. Researchers employed the Boltzmann machine to create shorter summaries without losing crucial information in order to increase the accuracy of their findings. [13]

Text summarizing with many documents involves gathering a lot of information from various sources of documents and reducing it to the document's key points or concepts. The term "multi document summarization" can also refer to material gathered from several sources or a summary of papers that discuss the same subject. For researchers, handling relevant features from a single document to multiple documents is a significant challenge. A multi-document text summarising experiment by John et al. produced results that outperformed the state-of-the-art in terms of recall and precision.[14]



Graph 1. Number of Research Papers published in every year [2008-2019]

According to Graph 1, it was found that the paper which discussed about text summarization have increased years by years. Total of 85 research papers are there which discuss about text summarization from 2008 to 2019.

III. METHODOLOGY

Here we are going to use text summarization using Natural Language Processing (NLP), basically the spacy library. Discussion about the text modulation a text summarization along with the spacy library. Firstly, the tokenization is done for the whole provided document, that is all words are classified into different types of tokens as show by fig. 2 here.

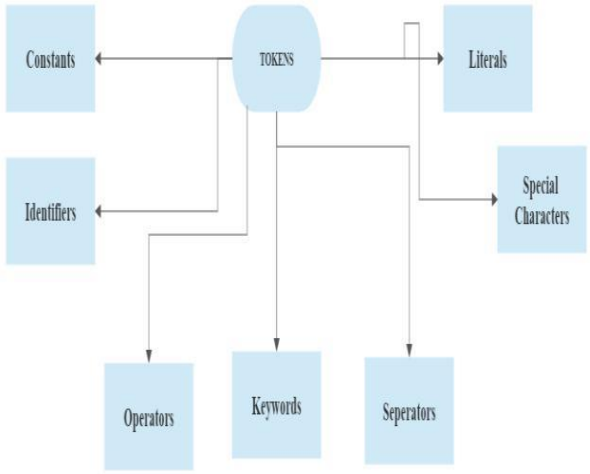


Fig 2. Types of Tokens

As shown in fig.3, The words present in the document are converted into tokens, after tokenization is done the word frequency table is created to analyze the word count. After that table is normalized, normalization is done when a dataset with features is provided of varying scale(magnitude). This is done to maintain the underlying distribution of data (not disturbing the ranges encompassed by the features of varying scale. It's been observed to improve the performance / accuracy of models, there are probably a lot of reasons why but the one that logically makes sense to me is if you have two features, one is on a scale of 0 to 1, and the other is on a scale of 0 to 1000, the variation in the second feature accounts for the majority of the variation in the dataset and an ML model might overvalue feature two and undervalue feature 1, as we humans would do, but this is not how data works in reality. Once the normalization is done sentence tokenization is done where all the words present in a particular sentence their optimal frequencies are added and the total frequency is known as the sentence tokenization. That total frequency of a particular sentence which is created by adding the optimal frequency of all the words is known as the sentence score.

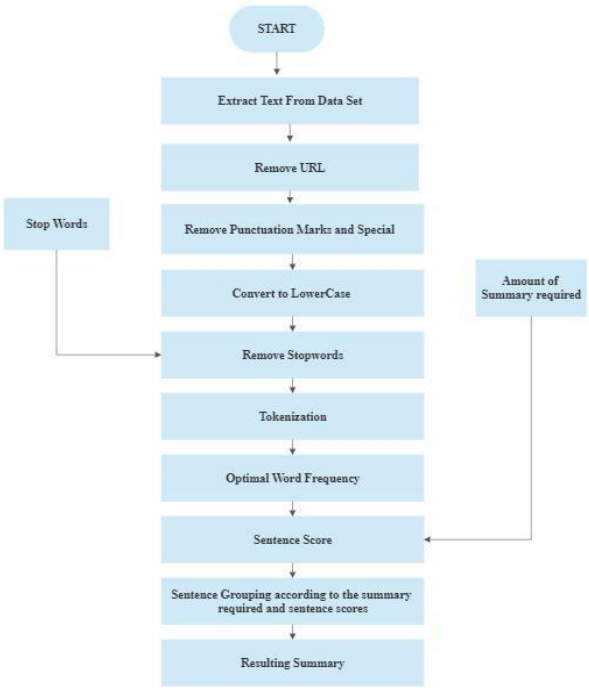


Fig 3. Flowchart for the process of Text Summarization

This sentence score is used for creating the summary according the amount of summary user wants, before this optimal frequency of a particular word is calculated by dividing frequency of all the words with the frequency of word which has highest frequency. Automatic summarization is needed as it reduces the amount of time required for searching the document summarizes some reason the selection process easier automatic summarization improves the effectiveness of indexing and automatic summarizes and algorithms are less bash than the human Sam Rogers and first analyzed some reads are useful in Christian answering systems as they provide personalized information and using automatic are semi- automatic summarization system enables commercial abstract services to increase the number of text documents they are able to process at resume a summarization is also available for a human resources department for a fasting screening and there are so many others there are so many other it wanted a advantages of text summarization how to text what are the types of the text summarization text which are present. Here have tried to use the Natural Language processing model of small version with the help of lexical analysis in order to create summary of the text provided, the input is given in the form of text or document as a file or as a direct input and the model processes that paragraph using lexical analysis and which is taken as the summary. Once the sentences are broken in the form of tokens then then count of all the sentences are counted in the form of list too. Then the words with the highest count is taken as the optimal value and the count of all the other words are divided by the frequency of that word to get the optimal frequency. Once the optimal frequency is provided the word tokenization is done. This shows the importance of all the words in the particular input provided. Then the word tokens of all the sentences are summed up in order to find the sentence priority. The sentences are then used for considering which sentence should be included in the summary which should not be in the summary now the clustering of all the sentences together is done using the clustering. Summarization can be done based on the input types and then finally it can do a single document summarizations or multi-document summarization that is the summarization can be based on a single documents or multiple documents.

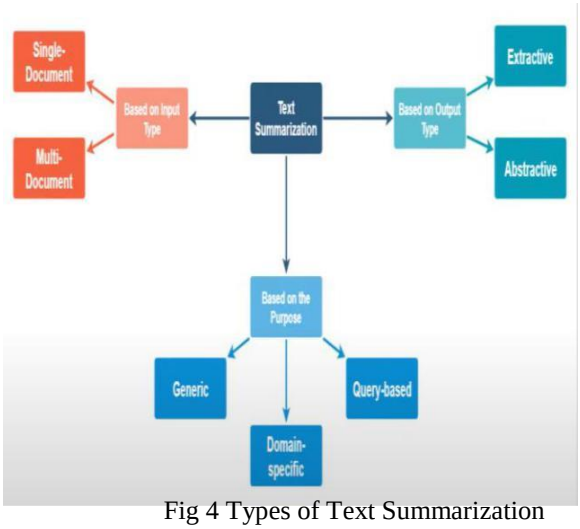


Fig 4 Types of Text Summarization

As shown in fig. 4, The text summarization can be based on output types like extractive or abstractive. Extractive- summarization means it will give you the summary by selecting few sentences from the overall the overall sentences. While the Abstract Summarization creates the summary with the abstract of overall text document. The spacy library used here has three models small, medium and large here small model is used any of the model can be used for this purpose without any problem. At first the text cleaning should be done where all the stop words from the document are removed along with all the punctuations, all these punctuations along with the stop words like the, it, are stored into the list. The NLP Model is called by using the Spacy library. The whole document is passed into the model and the tokens created from the model are placed into a list. The list of tokens also contains the stop words and punctuations which needs to be removed, so these tokens list is compared with the list of all the punctuations and stop words along with the new line created earlier and are removed. Once these stop words and punctuations are removed the model finds the most important lines from the given document. This is done by using token list and creating word frequency count from this word frequency count the word with the highest frequency is used to count the optimal frequency of tokens. if any key is being introduced the first time the word of that occurrence will be equal to one but if after the first time it is being introduced for second and the third time then it will just increment one in already present are the key in word frequency. The optimal frequency is calculated by dividing the word frequency with the highest frequency. This Optimal frequency is used for sentence tokenization and finding the priority of sentences and creating summary according to the amount of summary required and at last all those sentences which comes under the provided criteria are clustered to form summary.

IV. RESULT

Automatic text summarization is an old challenge but the current research direction diverts towards emerging trends in biomedicine, product review, education domains, emails and blogs. This is due to the fact that there is information overload in these areas, especially on the World Wide Web Automated summarization is an important area in NLP (Natural Language Processing) research. It consists of automatically creating a summary of one or more texts. The purpose of extractive document summarization is to automatically select a number of indicative sentences, passages, or paragraphs from the original document. Text summarization approaches based on Neural Network, Graph Theoretic, Fuzzy and Cluster have, to an extent, succeeded in making an effective summary of a document. Both extractive and abstractive methods have been researched, as shown in fig 4. Most summarization techniques are based on extractive methods. Abstractive method is similar to summaries made by humans. Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate into the domain specific areas.

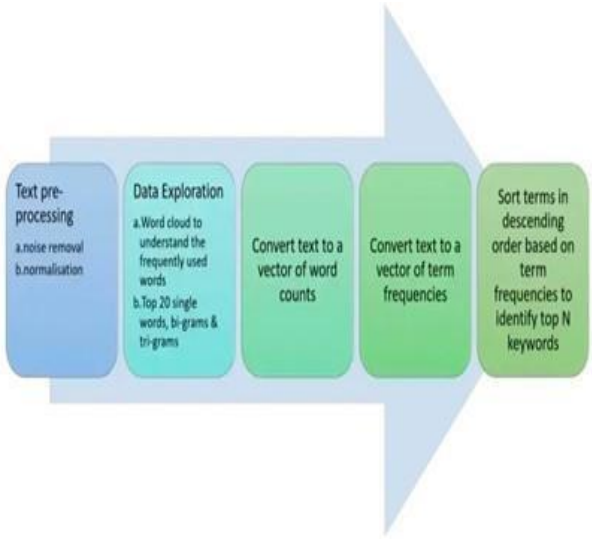


Fig 5. Steps involved in Text Summarization, adopted from [6]

As in the fig 5, firstly the text pre-processing is done which consist of noise removal and normalization Then tokenization is done for the whole provided document. The words present in the document are converted into tokens, after tokenization is done the word frequency table is created to analyze the word count. After that table is normalized, normalization is done when a dataset with features is provided of varying scale(magnitude). This is done to maintain the underlying distribution of data (not disturbing the ranges encompassed by the features of varying scale. It's been observed to improve the performance / accuracy of models, there are probably a lot of reasons why but the one that logically makes sense to me is if you have two features, one is on a scale of 0 to 1, and the other is on a scale of 0 to 1000, the variation in the second feature accounts for the majority of the variation in the dataset and an ML model might overvalue feature two and undervalue feature 1, as we humans would do, but this is not how data works in reality. Once the normalization is done sentence tokenization is done where all the words present in a particular sentence their optimal frequencies are added and the total frequency is known as the sentence tokenization. That total frequency of a particular sentence which is created by adding the optimal frequency of all the words is known as the sentence score. This sentence score is used for creating the summary according the amount of summary user wants, before this optimal frequency of a particular word is calculated by dividing frequency of all the words with the frequency of word which has highest frequency.

Here have tried to use the Natural Language processing model of small version with the help of lexical analysis in order to create summary of the text provided, the input is given in the form of text or document as a file or as a direct input and the model processes that paragraph using lexical analysis and which is taken as the summary. Once the sentences are broken in the form of tokens then then count of all the sentences are counted in the form of list too. Then the words with the highest count is taken as the optimal value and the count of all the other words are divided by the frequency of that word to get the optimal frequency. Once the optimal frequency is provided the word tokenization is done. This shows the importance of all the words in the particular input provided. Then the word tokens of all the sentences are summed up in order to find the sentence priority.

The sentences are then used for considering which sentence should be included in the summary which should not be in the summary now the clustering of all the sentences together is done using the clustering. summarization can be done based on the input types and then finally it can do a single document summarizations or multi-document summarization That is the summarization can be based on a single documents or multiple documents and the text summarization can be based on output types like extractive or obstructive. Extractive summarization means it will give you the summary by selecting few sentences from the overall the overall sentences, according to the need as shown in fig 6, the summary was set to 30 percent and the model provided exact 30 percent summary, consisting of 30 percent of the total word count that were in the original document.

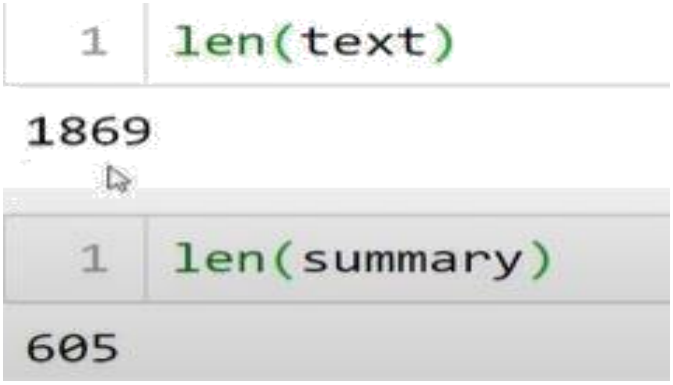


Fig 6. Result showing total number of words in original text and in summary

V. CONCLUSION AND FUTURE WORK

Text summarization is very helpful for users to extract only needed information in stipulated time. In this area, considerable amount of work has been done in the recent past. Due to lack of information and standardization lot of research overlap is a common phenomenon. Users can quickly extract only the information they need by using text summary. There has recently been a significant amount of effort in this area. Numerous instances of study overlap are widespread due to a lack of knowledge and uniformity. Since 2012, comprehensive reviews

on automatic keyword extraction and text summarization, particularly in the context of India, have not been published. We therefore believed that the survey study examining recent work in text summarizing and keyword extraction may inspire the research community to close some significant research gaps. Using the perspectives of automatic keyword extraction, text databases, the summary

process, summarization approaches, and evaluation matrices, this paper reviews the literature on recent work in text summarization. Several crucial research questions in Since 2012, exhaustive review paper is not published on automatic keyword extraction and text summarization especially in Indian context. Therefore, we thought that, the survey paper covering recent work in keyword extraction and text summarization may ignite the research community for filling some important research gaps. This

paper contains the literature review of recent work in text summarization from the point of views of automatic keyword extraction, text databases, summarization process, summarization methodologies and evaluation matrices. Some important research issues in the area of text summarization are also highlighted in the paper.

This paper describes a keyword extraction method to investigate the benefits of using lexical chain features in keyword extraction. According to the results that are obtained, the lexical chain features improve the precision significantly in the keyword extraction process. Although lexical chains have been used in different application domains, to the best of our knowledge we are the first to use lexical chains in the keyword extraction problem.In this paper, we have tried to extract keywords only since WordNet does not contain too many word phrases. We are working on how to extract key-phrases in addition to keywords. In order to able to do this, we are investigating ways of putting phrases into lexical chains. In fact, some phrases are already present in WordNet. For the phrases that are not in WordNet, a possible approach is to use the head noun of the phrase to represent that phrase and put the head noun in the lexical chains.

REFERENCES:

1. A.R. Kulkarni . An Automatic Text Summarization Using Lexical Cohesion And Corelation Of Sentences. In International Journal Of Research In Engineering And Technology - June 2014.
2. Ayush Aggarwal , Chhavi Sharma, Minni Jain, Amita Jain From Delhi Technical University, Delhi. Semi Supervised Graph Based Keyword Extraction Using Lexical Chains And Centrality Measures.
3. Ben Hachey And Claire Grover . Sentence Extraction For Lega Text Summarization. In University Of Edinburgh School Of Informatics- 2017.
4. Bharathi Mohan Gurusamy , R. Prasanna Kumar . Lattice Abstraction-Based Content Summarization Using Baseline Abstractive Lexical Chaining Process. In International Journal Of Information Technology–October 2022.
5. Farhanaaz, Sanju. V M.- F. (2002). An Exploration On Lexical Analysis. In International Confrence On Electrical, Electronics, And Optimization Techniques (Iceeot) – 2017.
6. Gonenc Ercan, Ilyas Cicekli. Using Lexical Chain For Keyword Extraction. UploadedBy Gonenc Ercan On 4th November 2019.
7. Meru Brunn , Yllias Chali , Christopher J. Pinchak. Text Summarization Using Lexical Chains. September 2012.
8. Murali Krishna Rvv , Ch. Satyananda Reddy . Extractive Text Summarization Using Lexical Association And Graph Based Text Analysis. In Book: Computational Intelligence In Data Mining – Volume 1 – December 2016.
9. Rasim Alguliyev, Ramiz Alguliyev. Experimental Investigating The F-Measure As Similarity Measure For Automatic Text Summarization. In Applied And Computational Mathematics- January 2007.
10. Ravishek Kumar Singh. Autoregressive Nlp Model For Text Summarization And Analysis. September 2021.
11. Rohit Parimoo , Rohit Sharma , Nimish Jain . A Review On Text Summarization Techniques.
12. Santosh Kumar Bharti , Korra Sathya Babu & Sanjay Kumar Jena. Automatic Keyword Extraction For Text Summarization.
13. Vaikunta Pai T, A. Jayanthila Devi & P.S. Aithal. A Systematic Literature Review Of Lexical Analyzer Implementation Techniques In Compiler Design. In International Confrence On Computer Science And Electronics Forum(Iccse)-2020.