doi.org/10.54473/IJTRET.2022.6405
International Journal of Trendy Research in Engineering and Technology
Volume 6 Issue 4 August 2022
ISSN NO 2582-0958

TEXT SUMMARIZATION USING NLP

[1] ChetanaVaragantham, [1]J.SrinijaReddy, [1]UdayYelleni, [1]MadhumithaKotha, [2]P.VenkateswaraRao
[1]Final year Students, [2]Associate Professor, Department of Computer Science and Technology,
ACE Engineering College, Hyderabad, India

Received 09 May 2022 Received in revised form 18 May 2022 Accepted 19 May 2022

ABSTRACT

This Project represents the work related to Text Summarization. In this paper, we present a framework for summarizing the huge information. The proposed framework depends on highlight extraction from the internet, utilizing both morphological elements and semantic data. Presently, where huge information is available on the internet, it is most important to provide improved ways to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large document of text. There are plenty of text materials available on the internet. So, there is a problem of searching for related documents from the number of documents available and absorbing related information from it. In essence to figure out the previous issues, automatic text summarization is very much necessary. Text Summarization is the process of identifying the most important and meaningful information in an input document or set of related input documents and compressing all the inputs into a shorter version while maintaining its overall objectives.

Keywords: Machine Learning, Text Summarization, Natural Language Processing (NLP) ,Clustering, Tokens

I. INTRODUCTION

In this paper, we present a framework for Text Summarization. The proposed framework depends on summarizing the text from the internet, utilizing both morphological elements and semantic data. The length of text data is increasing, and people have less time to read those data. Internet, media, and other data sources have a huge dump of data and hence a system is required for generating easier and short forms of data. So, a tool is required for the users, which would ease the effort for them to read the entire text or matter. Such systems or tools would be beneficial and a great time saver for the users. Hectic schedules made it impossible for everyone to read and access the information from News information, biographical information, or from other journals. Reliable and easier information are needed to be efficient. With summaries, People can make productive decisions instantly. The motivation here is to build such a tool which is efficient and creates summaries automatically. Natural Language Processing (NLP) is an area of automated cogitation where PCs probe, understand and get importance from human language in a radiant and useful manner. By implying NLP, designers can arrange and build information to carry out tasks like programmed rundown, interpretation, named element acceptance, relationship production, judgment investigation, discourse acceptance, and point subdivision.

Aside from similar word processor tasks that work with a message like a simple positioning of images, NLP reflects on the various get even construction of language: a few words make a declaration, a few declarations make a sentence, and, at last, sentences disclose thoughts, John Reeling who is an NLP master at software as solution company“ Meltwater Group”, communicated in How Natural Language Processing assists Uncover Social Media Sentiment. By decomposing the language for its importance, NLP frameworks play a long complete useful representation, for example, regulating punctuation, altering conversation over to message, and accordingly interpreting between dialects. NLP is used to disintegrate the text, allowing machines to how people communicate. This human-PC alliance empowers fair applications like programmed message outlines, judgment investigation, subject subdivision, named element acceptance, grammatical features classification, relationship production, stemming, and the endless limit from there.

II. LITERATURE SURVEY

Automated text summarization and the approaches of single document and multi- documents text summarizations have been discussed based on requirements extractive summarization is report by Tanni [1]

Patil et al in their paper 'Automatic text summariser' have designed and constructed an algorithm that can summarize a document by extracting key text and modifying this extraction using a thesaurus[2]. Mainly it is to reduce the size ,maintain coherence

In 'Text Summarization: A Review' by Biswas et al have reviewed text summarization by using various technologies and methodologies in creating a coherent summary that includes the key points of the original input document.[3]

An article by Andhale and Bewoor ' An overview of Text Summarization techniques' gives an overview survey on both extractive and abstractive approaches.[4]

Janjanam and Reddy in their paper 'Text Summarization: An Essential Study' discussed the abstractive text summarization approaches and the state of art machine learning models used to summarize single and multi-documents and eventually lead to large document summarization.[5]

Awasthi et al ,[6] explained the study of extractive and abstractive text summarization methods .They have used linguistic and statistical characteristics to calculate the implications of sentences. The objective of their work is to have less repetition and accurate summary.

III.EXISTING SYSTEMS

The text summarizations involves two approaches which are extractive and abstractive summarizations including summarization of single document and multi documents based on the requirements of extractive and abstractive summarizations. There are many methodologies and techniques used in the process of text summarization where the main objective is to reduce the size of the output by maintaining the coherence and accurate meaning of the original input[7-9].

IV.PROPOSED SYSTEM

The objective is to automate the summarization of documents. The proposed system works on the extractive summarization-based approach. The system calculates the frequency weightage of each word in the sentence in the entire document and also authenticates for the parts of speech of the words then allocates a total score for the sentences. The system then incorporates the clustering technique for extracting the final summary sentences.

The advantage of this approach is that in this, the Clustering phase the clusters are formed based upon the sentence scores and are segregated into lowest and highest weighted sentences from which the final phase provides the output based upon the highest scored clusters which give meaningful and efficient summaries.

k-means clustering is an approach of quantization vector, initially from signal processing,. Its objective is to divide and observe into k clusters where the cluster belongs to each observation with the closest mean cluster centroid or cluster centers . It is helping as a prototype of the cluster, which leads to the division of the data space into Voronoi cells. As a result. k-means clustering reduces intra-cluster differences, and the regular Euclidean distances would have the toughest Weber issues hence we squared Euclidean distances and the squared errors are optimized by mean.

The mathematical formula of Euclidean distance formula is

$$d=\sqrt{[(x2 - x1)^2 + (y2 - y1)^2]}$$

V. WORKFLOW

Pre-processing Step: Pre-processing is a process that is done before the translation. The document or set of related documents is the input to the summarizer system. The document should be shifted into a sack of words or phrases of the document. The pre-processing step includes Natural Language Processing (NLP) phases like sentence segmentation, sentence tokenization, stop word removal, and stemming. Once the pre-processing is done, the word frequency and reverse documents frequency values are calculated for every token.

Sentence segmentation: Sentence segmentation is the process of dividing a string of written language into its unit or module sentences. In languages such as English and some other languages, punctuation utilized, particularly the symbols such as full stop and period characters are sensible estimations.

Tokenization: Tokenization is the process of dissecting sentences into a course of discrete tokens that are adapted by the spaces and that can be used for additionally refining and comprehending. Tokens can be discrete words, keywords, phrases, identifiers, etc. In the process of tokenization, tokens or words are segregated by the white space, the punctuation marks, or the line breaks. The white space or the punctuation marks are likely or unlikely to be entangled depending on the needs.

Stop Word Removal: Stop Words are the words that occur frequently in the language. Deletion of Stop Word is the process of removing words like "the", "to", "are", "is", etc. And stop words are removed to benefit support phrase search.

Stemming: Stemming is the process of reducing the operationally related forms or intentionally related forms of words to their stem form, common base form or root form – generally a written word form that may help to increase the coverage of Natural Language Processing (NLP) utilities.

Feature Extraction: Feature extraction is the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It complies with better results than simply applying machine learning straight to the raw data.

Clustering Technique: Clustering is a process that involves the classification of data points. K Means is an unsupervised learning algorithm, which assembles the data to form sentences. When the set of data points is given then, the clustering algorithm can be used to classify each data point into a particular class. An algorithm will be generated that contains a clustering machine learning technique i.e., K means.

Summary Generation: The summary of the text document will be generated using two techniques, namely:

- The clustering technique and
- The clustering technique cascade with K-means

Summarization of the clustered documents is done based on the ranking and scoring in order to get the brief summaries.

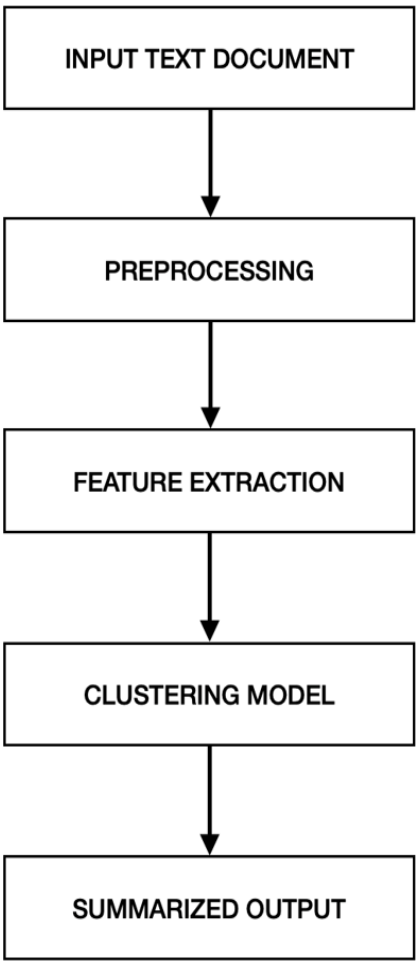

Fig 1: work flow

SAMPLE GRAPH OF DISTRIBUTED PROBLEMS IN THE PROCESS OF TEXT SUMMARIZATION

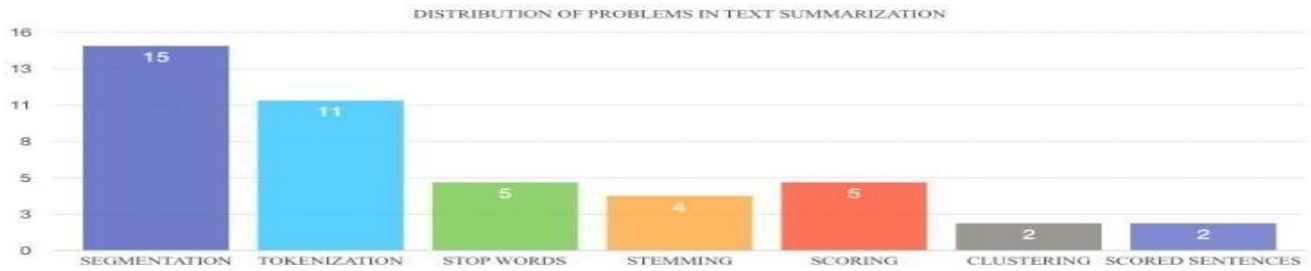

Fig 2: Sample Graph

As mentioned in the above graph we are showing the changes in the distribution of problems in text summarization which involves segmentation, tokenization, stop words, stemming, scoring, and clustering scored sentences in the form of a graph.

VI.RESULTS

As mentioned in the above procedure first we take the input as the document and by using k means clustering we will summarize the text as output.

INPUT

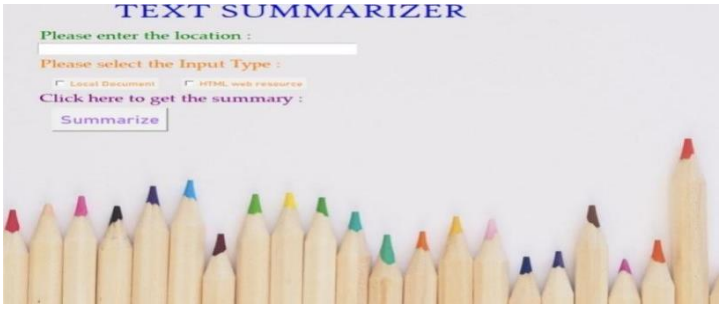

Fig 3: UserInterface

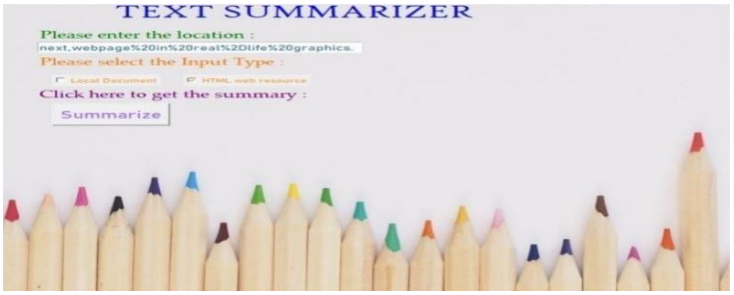

Fig 4: Input Document Linked OUTPUT

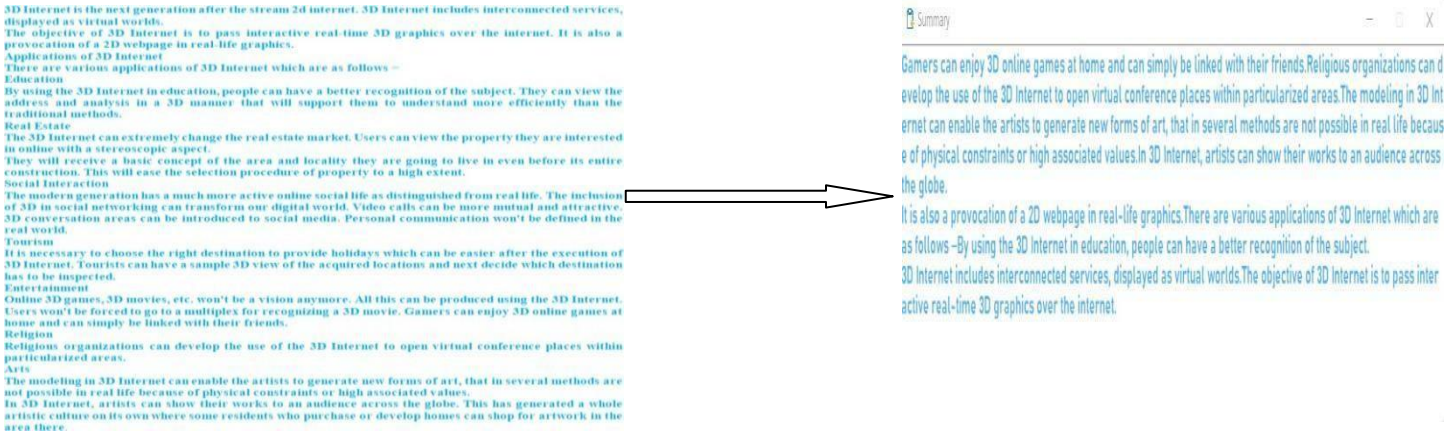

Fig 5: Summarized Output

VII.CONCLUSION

The estimate at which the information has been growing due to the World Wide Web has created issues for which they need to develop structured and exact summarizations. Even though research on summarization has begun about 55 years ago, there is still a vast path to explore in this area. Over decades, observation has been carried away from summarizing scientific documents to news articles, Emails , blogs, and advertisements. The two extractive and abstractive techniques have been ventured, upon the application available. mostly, abstractive summarization needs hefty machinery for language production and is tough to reproduce or expand to the broader area. In comparison, easy extraction of sentences has generated acceptable results in extensive applications. The Project has carried out its purpose thereby decreasing the input textual data to more compact reduced summarized results.

ACKNOWLEDGEMENT

We would like to thank our guide Dr. P. Venkateswara Rao and Mrs. Soppari Kavitha for their continuous support and guidance. Due to their guidance, we can complete our project successfully. Also, we are extremely grateful to Dr. M. V. VIJAYA SARADHI, Head of the Computer Science and Engineering, Ace Engineering College for his support and invaluable time.

REFERENCES

[1] A. T. Al-Taani, "Automatic text summarization approaches," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017, pp. 93-94, doi: 10.1109/ICTUS.2017.8285983.

[2] A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul and V. Bhatnagar, "Automatic text summarizer," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1530-1534, doi:10.1109/ICACCI.2014.6968629.

[3] S. Biswas, R. Rautray, R. Dash and R. Dash, "Text Summarization: A Review," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018, pp. 231-235, doi: 10.1109/ICDSBA.2018.00048.

[4] N. Andhale and L. A. Bewoor, "An overview of Text Summarization techniques," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860024.

[5] P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi:10.1109/ICCIDS.2019.8862030.

[6] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi:10.1109/ICICT50816.2021.9358703.

[7] H. T. Le and T. M. Le, "An approach to abstractive text summarization," 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), 2013, pp. 371-376, doi:10.1109/SOCPAR.2013.7054161.

[8] P. R. Dedhia, H. P. Pachgade, A. P. Malani, N. Raul and M. Naik, "Study on Abstractive Text Summarization Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-8, doi:10.1109/ic-ETITE47903.2020.087.

[9] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.