iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51650

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Data Summarization Web Application

Sakshi Jawale[1], Pranit Londhe[2], Rushikesh Kolekar[3], Sarika Jadhav[4], Prajwali Kadam[5]
Information Technology Department, Parvatibai Genba Moze College of Engineering

Abstract: ***As there is an increase in the usage of digital applications, the availability of data generated has increased to a tremendous scale. Data is an important component in almost every domain where research and analysis are required to solve problems. It is available in a structured or unstructured format. Therefore, to get corresponding data as per the application purpose, easily and quickly from different sources of data on the internet, an online content summarizer is desired. Summarizers make it easier for users to understand the content without reading it completely. Abstractive Text Summarizer helps in defining the content by considering the important words and helps in creating summaries that are in a human-readable format. The main aim is to make summaries in such a way that they should not lose their context. Various Neural Network models are employed along with other machine translation models to bring about a concise summary generation. This paper aims to highlight and study the existing contemporary models for abstractive text summarization and also to explore areas for further research.***
Keywords: Data, Abstractive, Summarization, Neural Network, NLP Techniques

I. INTRODUCTION

The amount of data available on the internet is increasing at an alarming rate. this data, on the internet, is unstructured. this unorganized form of data has made it difficult for a user to locate specific content. eventually, the user spends time and resources to locate such relevant content. hence, a content summarizer would be desirable to get a gist of an article and to check if the article answers the user's query before the user proceeds to read the article in detail. automatic text summarization is the elementary methodology used for any text summarization system. automatic text summarizer has many applications in the industry: news aggregators, blogs, product descriptions, etc. the text summarization technique paved the way for the development of summarization models. initially, the text summarization was developed considering basic parameters and measures.

Text summarization can be classified as Abstractive and Extractive:

1) *Extractive Summarization:* Extractive summarization finds out the important sections of the content and creates a subset of sentences from the sentences present in the original document. It does not add new words to the existing content and cannot combine two or more sentences to compact the content. Based on the number of documents used as sources, summarization is further classified as single-sourced and multi-sourced. Extractive Summarization works based on combining the words or phrases from the corpus for the summary.
2) *Abstractive Summarization:* Abstractive summarization, on the other hand, analyses the whole content to reproduce the original content in a new and optimized way using advanced natural language techniques. The newly generated content is shorter and more importantly, conveys the most critical information of the original content. Abstractive summaries also generate fluent sentences that are grammatically correct, unlike extractive methods, which may lead to disfluent sentences.

II. LITERATURE SURVEY

Automatic text summarization techniques such as cluster-based, template-based, ontology-based, semantic graph-based methods, machine learning, fuzzy logic, and neural network approach have proven to be very useful over 50 years up to now. Automatic text summarization approaches are twofold extractive and abstractive summarization. Extractive summarization techniques extract important sentences or phrases from the original text documents based on the weight of statistical and linguistic features and produce a summary without changing the original text. Abstractive summarization generates the summary by understanding the source text using linguistic methods and generating new sentences by improving the focus of a summary while reducing the redundancy rate. Abstractive summarization is more complex because it requires deeper analysis of source document(s) and concept-to-text generation. These techniques can be further classified as structured and semantic-based approaches. Semantic-based approaches provide more coherent, information rich, and well-structured abstractive summaries than structured-based approaches. Due to the semantic representation, semantic graph-based approaches have been proven to be very successful over other methods for producing multi document abstractive summaries.

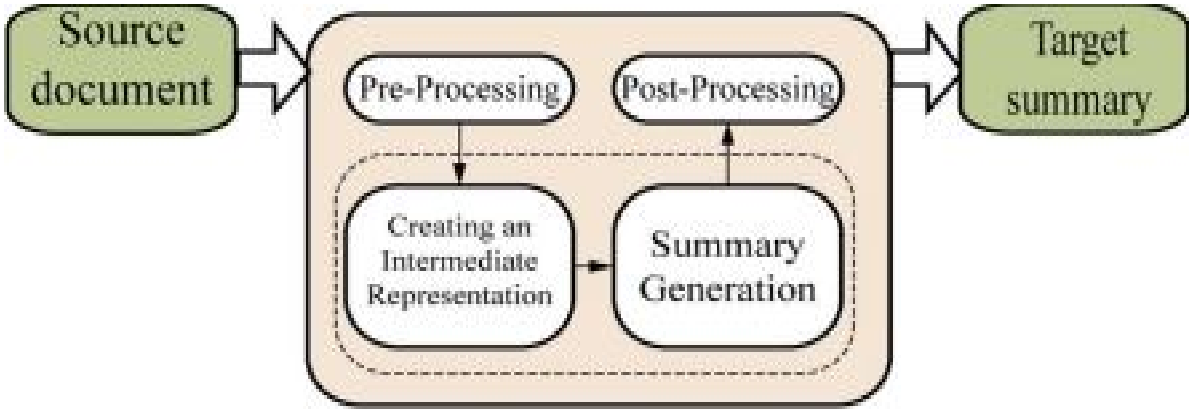
©IJRASET: All Rights are Reserved

Semantic-based methods are not able to achieve similar performance to deep learning approaches. Machine learning approaches to generate automatic summaries have greatly improved compared to the primitive text summarization methods, but they cannot be combined with the background information to obtain higher level abstraction and cannot scale to the requirement of large data sets. To understand a text well enough, we need some background knowledge. Some researchers were inspired by those ideas and used knowledge graphs to generate automatic text summaries. Knowledge graphs capture domain specifics and add rich and explicit semantics to infer additional knowledge and encapsulate a large amount of knowledge for human and machine consumption. Graph-based approaches use ranking functions comprising one or more sentence weighting features to determine the relevancy of each edge for inclusion in the summary.

According to the literature, we have identified that traditional word co-occurrence measures like TF-IDF give more importance to the words that are more frequent in the document. Hence, a pure ranking algorithm can be less effective because ranking alone cannot determine what kinds of features are selected for the summary. But, increasing the set of attributes to include semantic properties and topological graph properties in the ranking function yields statistically significant improvement for the abstractive summaries. Knowledge graphs lead to information overload, and hence proper summarization techniques need to be explored. Techniques in data mining are widely used to explore data from various perspectives and uncover unexpected relationships between pieces of text and summarize them into useful information. Therefore, computers use text mining to discover valuable and interesting information and knowledge techniques. Among these techniques, Association Rule from unstructured textual data in large volumes using different Mining (ARM) can be used in discovering the relationships for the decision-making process and discovering correlations between multidimensional features in the text documents.

III. IMPLEMENTATION DETAILS OF A MODULE

The source document is taken from the internet or from manually entering text paragraphs to get a summary. After receiving the text, that text is pre-processed. Pre-processing involves:



A. *Removal of High-Frequency Words*

This is done by comparing each word from the document to the list of high-frequency words. High-frequency words are those words that come a greater number of times in the text. The advantage of this process is not only the non-significance words are removed but also the size of the document can be reduced to 30 to 50%.

B. *Suffix Stripping*

In this step, each word is handled from the output of the first step. If any word is having a suffix, then the suffix gets removed and the word is converted to its original form.

C. *Detecting Equivalent Stems*

After suffix stripping, we have a list of words. Only one occurrence of the word is kept in the list. If two words have the same underlying steam then they refer to the same concept and should be indexed as such.

After that based on the keywords received, the system generates a new text summary that captures the most relevant information. And this summary is given to the user after post-processing of the text. In post-processing, similar meaning lines are removed.

IV. CONCLUSION & LIMITATIONS

Our website gives an accurate summary to the user based on the functions that we have built. Our website, too, has certain boundaries. When the text input on our website is shorter than 1000 words, our website gives an accurate summary to users. Our website offers a summary of articles based on a selection of scales. The selection of scale is just for testing reasons. When we pick 100 scales, that is the actual summary.

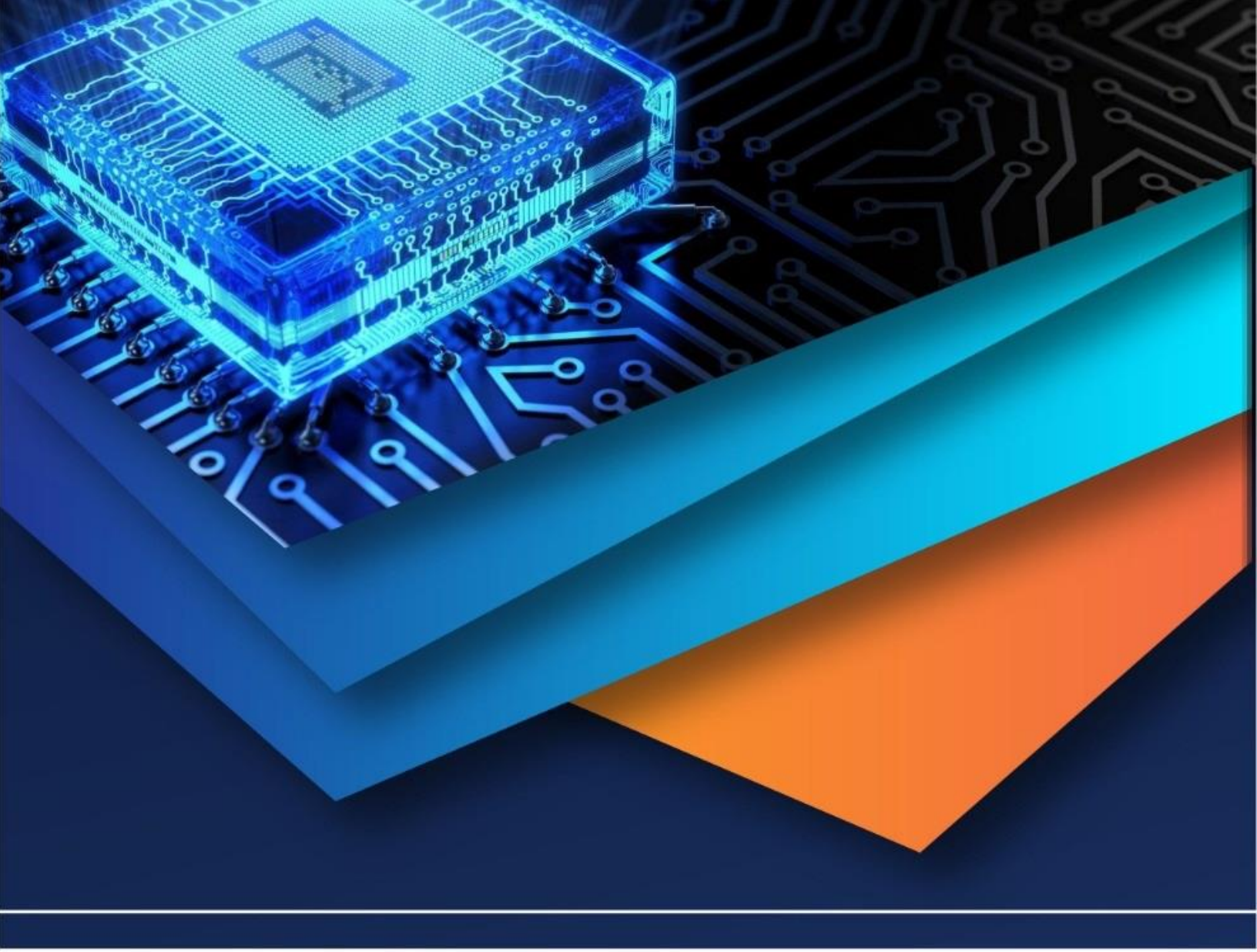
V. FUTURE SCOPE

Our future ambitions include adding a file upload mode to the website, writing summaries with no word count restrictions, linking the website with other platforms, providing customization choices, and improving summary accuracy.

REFERENCES

[1] 2020 @IEEE, A Survey on NLP based Text Summarization for Summarizing Product Reviews by Ravali Boorugu, Dr. G. Ramesh
[2] 2020 @IEEE, Study on Abstractive Text Summarization Techniques by Parth Rajesh Dedhia, Hardik Pradeep Pachgade, Aditya Pradip Malani, Nataasha Raul, Meghana Naik
[3] 2021 @IEEE, Multimedia Data Summarization Using Joint Integer Linear Programming by Sidhant Allawadi, Ritika, Vivek Rana, Minni Jain
[4] 2021 @IEEE, Techniques and Research in text summarization- a Survey by Manish Shinde, Disha Mhatre, Gaurav Marwal
[5] 2020 @IEEE, Topic-Centric Unsupervised Multi- DocumentSummarization of Scientific and News Articles by Amanuel Alambo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, Michael Raymer
[6] 2020 @IEEE, Abstractive Web News Summarization Using Knowledge Graphs by M.V.P.T.Lakshika, H.A.Caldera, W.V.Welgama

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)