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.2023.52248

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

Text Summarizer Using SpaCy in NLP

Sangita Lade[1], Tanishka Pawar[4], Sanskruti Morey[2], Parth Mule[3], Siddhi Rajeshirke[5]
Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract: Using machine learning and natural language processing techniques to summarize the huge volume of text data and give a short summary. To develop a system which identifies contexts of a document and give the best possible summarized text with the help of SpaCy in natural language processing.
Keywords: NLP, Machine Learning, SpaCy, Summarization Text Summary.

I. INTRODUCTION

Due to the massive volume of textual content that is generated on the Internet and in numerous archives of news stories, scientific papers, legal documents, etc., text summarization is becoming much more crucial. With the enormous amount of textual content, manual text summarizing takes a lot of time, effort, money, and even becomes impracticable. The three types of Text summarization techniques are extractive, abstractive, and hybrid. The extractive method chooses the key phrases from the source documents and then concatenates them to create the summary. The abstractive approach generates the summary using sentences that are distinct from the original sentences after representing the input documents in an intermediary form.

We went through various research papers and found out insights about our topic. The creation of a graph that represents the text in order to create a ranking mechanism based on graphs that uses voting or recommendations to advance a vertex. It is used for sentence and keyword extraction. Comparison of the support vector machine and artificial neural network algorithms for text classification of news articles. A technique that separates each message into conversational threads after extracting and mapping the semantic content, social interactions, and timestamp in a three-dimensional representation. Filling gaps between the limitations of current NLP technologies and variations in social media use and communication due to geography, culture, and age.

Therefore, generating a system for providing a summary of the enormous amount of text material, machine learning and natural language processing techniques will be used. To use SpaCy for natural language processing to create a system that recognizes contexts in a document and provides the best possible summarization.

II. METHODOLOGY/EXPERIMENTAL

The numerous aspects of artificial text summarization can be broadly divided into multiple methods based on certain traits, such as output-based learning algorithms, single or multiple documents , specific or general purpose.

Extractive Method – Extractive summarization is one of the techniques that includes summarizing data using a score system. When compared to sentences with words of lower value, it gives sentences with key words a greater value mark. One group of these powerful phrases is picked inside the text's confines. Extraction and expectation, which are both necessary for extracting and grouping, are two crucial components for completing this strategy. Words and sentences are displayed as the appropriate summary based on their score. Text processing is the automated process of text analysis and manipulation. It receives the text as input, processes it, and then produces the desired result; it might be widely employed in many organizational considering that product teams may learn from user input to computerize customer support. In this case, the text's words and tokens stand in for distinct, categorical properties.

SpaCy is a free, open-source library accurate for advanced Natural Language Processing (NLP) via Python. For processing the text, spaCy offers a variety of built-in capabilities that make it a useful tool for language modelling and text processing.

Document pre-processing – The input documents we receive might not be in the formal English language and might even contain sounds words, stops, newlines, undesired spaces, and other special characters. To ensure that you only receive the document's useful portions, execute the following operations on the input file. All line breaks are eliminated. Eliminate all corner brackets and special numerals. Remove all commas, excess spaces, and repetitive sentences.

Removal of stop words – In this step, all subtitles will be removed from your input based on your language of origin. These stop words don't offer accurate details about a certain context. As it builds a collection of emotions like "is," "am," and "who" to make a picture, it does not provide any information about this feeling.

Tokenisation – Dividing into tokens. Any single unit is referred to as a token. Any information in the software that is relevant to the machine or the individual.

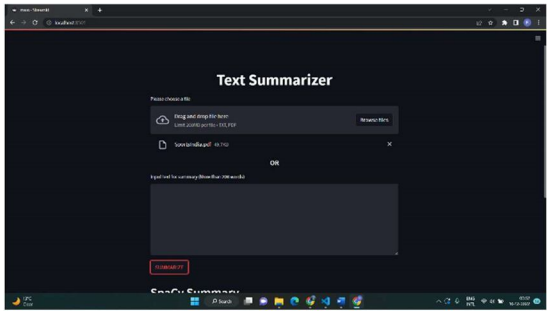
Word Tokenization – When every word in the text is given a word score, the entire text is separated into individual words.in keeping with its count.

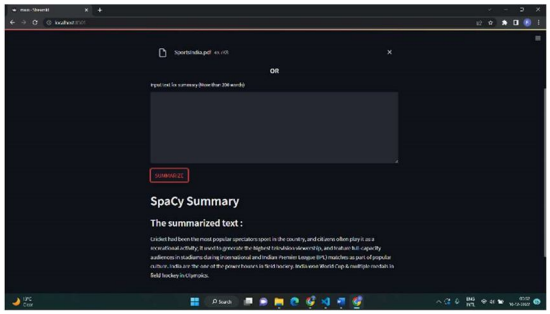
Extraction of important sentences – Method for confirming the text's value in the scroll is required. Calculate the frequency of each word in the text that has already been read. The weight of each word is determined by dividing its frequency by its maximum frequency. Go over each key phrase for the input that was given. The sentence's score is determined by summing the weighted frequencies of the terms it contains. Involve removing the "n" statement from the token list and sorting the set's token list in descending order based on points.

Final summary generation – The top score sentences are finally joined together. This is in the form for a list. The list is converted to a summary paragraph and passed to the main program for final display of the output.

III. RESULTS AND DISCUSSIONS

We successfully created a Text summarizer using SpaCy with the streamlit host. The summarizer can be used to summarize text documents, pdf documents and chat threads as well as direct text and obtain an efficient summary through NLP.





IV. FUTURE SCOPE

The current system works for text data in the form of different documents, it can be further improvised to accepting inputs in form of audio files and extracting speech text from videos for summarization.

V. CONCLUSION

Text summarization is a fascinating academic subject with numerous practical applications in industry. Summaries are helpful in a variety of downstream applications, such as news summaries, reporting, news summaries, and headlines, by condensing enormous volumes of information into brief bursts. Therefore, using spaCy in natural language processing the system identifies text of a document and gives the best possible summary.

VI. ACKNOWLEDGEMENT

We extend our warm gratitude to our respected Prof. Sangita Lade ma'am, who guided us at every step for the completion of our project. It seems appropriate to say thank you in the end, rather than at the beginning, because it is the omega of the project which she has helped us to bring about. And last but not the least we thank each one of us group members for providing insights that greatly assisted the development of our project work to the final outcome.

REFERENCES

[1] M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroni, and R. Leonardi, "A free Web API for single and multi-document summarization," ACMInt.Conf.ProceedingSer.,vol.Part F1301, 2017, doi: 10.1145/3095713.3095738.

[2] A.T.SardaandM.Kulkarni,"TextSummarization using Neural Networks and Rhetorical Structure Theory," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 49–52, 2015, doi: 10.17148/IJARCCE.2015.4612.

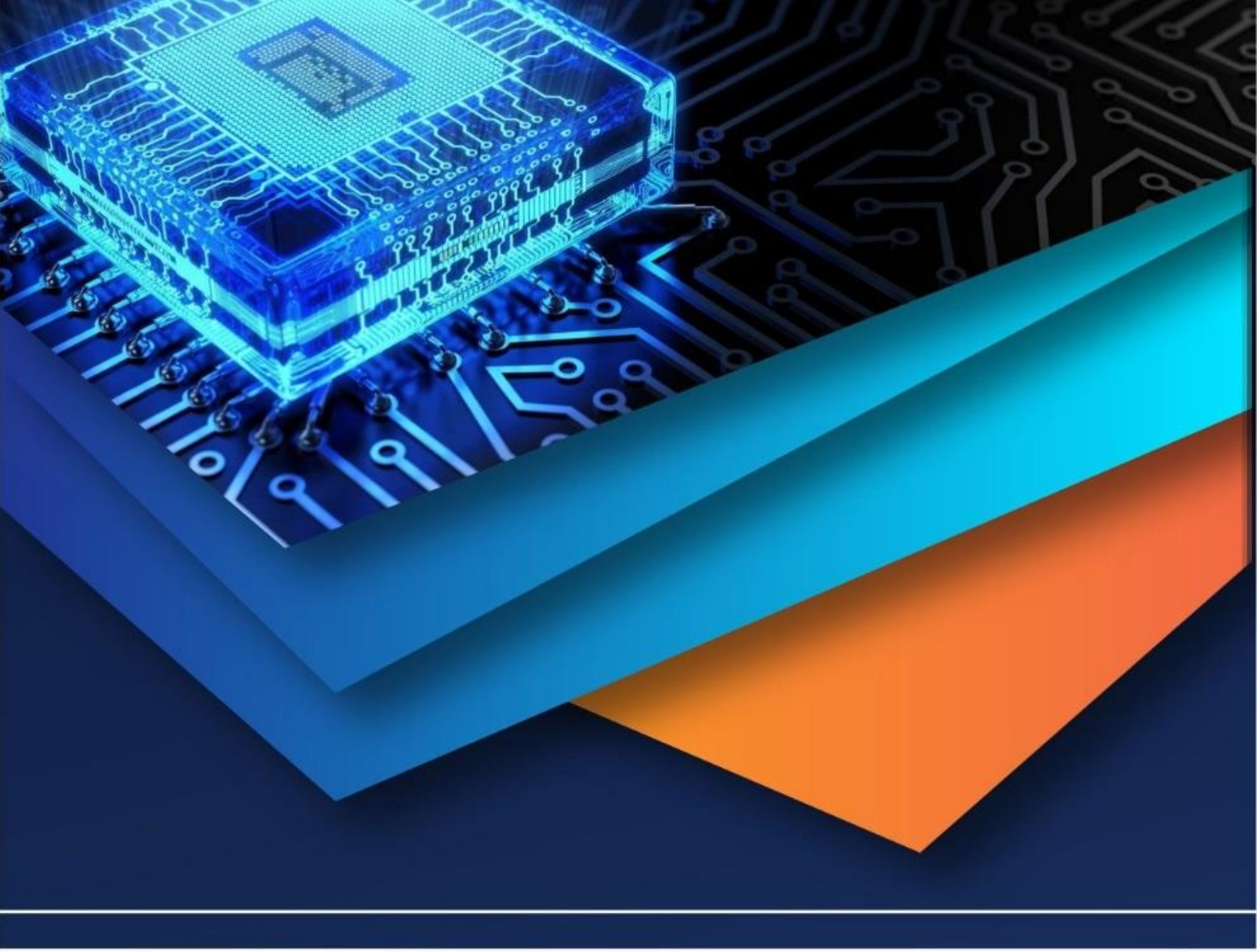
[3] G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Riss, and H. O. Cabral, "Automatic text document summarization based on machine learning," DocEng 2015 - Proc. 2015 ACM Symp. Doc. Eng., pp. 191–194, 2015, doi: 10.1145/2682571.2797099.

[4] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 20184thInt.Conf.Web Res. ICWR 2018,pp.128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[5] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.

[6] M. Moradi, G. Dorffner, and M. Samwald, "Deep contextualized embeddings for quantifying the informative content in biomedical text summarization," Comput. Methods Programs Biomed., vol. 184, p. 105117, 2020, doi: 10.1016/j.cmpb.2019.105117.

[7] A. Jain, D. Bhatia, and M. K. Thakur, "Extractive Text Summarization Using Word Vector Embedding," Proc. - 2017 Int. Conf. Mach. Learn. DataSci.MLDS 2017,vol.2018-Janua, pp. 51– 55, 2018, doi: 10.1109/MLDS.2017.12.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY