S3 Data Summarization and Visualization

Prof. Amit Narote[1], Prathamesh Jawale[2], Frason Francis[3], Krishna More[4]
[1,2,3,4]*Dept. of Information Technology, Xavier Institute of Engineering, Mumbai, India*

Abstract— Collecting and managing data has always been a difficulty, but with the arrival of AI and deep learning, the amount of data created has increased tremendously, resulting in Big Data. The sheer volume of unstructured data, which includes items like music, video, photos, and social network data, might make it challenging to exploit the information efficiently. The amount of text data available from various sources has increased dramatically. This amount of literature is an excellent source of information and expertise that must be adequately summarized to be useful. It becomes difficult for corporations to handle and process these papers promptly. To efficiently manage large amounts of text files, it is critical to evaluate and categorize the data. With this solution, we strive to create a platform that will provide quick and reliable summarization using various pre-trained natural language processing algorithms and visualization so that domain experts can get insights into the text data efficiently by fetching data from S3 buckets.

Keywords—S3 bucket, Cloud, Big Data, Natural Language, Transformer, Amazon Web Services.

I. INTRODUCTION

The proposed research focuses on the development of a novel method for summarizing multiple documents related to a particular topic. The goal is to reduce the effort and time needed to search for relevant information by generating a summary from various sources. The method employs two primary techniques: extractive summarization and sentiment analysis.

Extractive summarization involves identifying the most relevant sentences or phrases from a document and combining them to create a summary that captures the main points of the original text. This method is commonly used in natural language processing and is effective in generating informative and concise summaries.

In addition to extractive summarization, the proposed method also utilizes sentiment analysis. Sentiment analysis is a computational method that utilizes machine learning to analyze text and identify the feelings, viewpoints, and attitudes conveyed in the text. In the context of document summarization, sentiment analysis can help identify any dissimilar opinions or conflicting viewpoints among the input documents.

The method begins by fetching various document URLs related to the topic of interest. The system then generates individual summaries for each document using extractive summarization techniques. Finally, a unique summary is generated from all the first-level summaries, which reflects the main points of the input documents while accounting for any differences in opinion or attitude.

The proposed method's performance is compared to that of other existing methods to determine its effectiveness in generating informative and accurate summaries.

Overall, the proposed research presents a promising approach to generating summaries from documents that can save time and effort in searching for relevant information while maintaining the accuracy and relevance of the original content.

A. Extractive summarization

To create a system-generated summary, one approach is to select important parts or sentences from the source text based on linguistic and statistical features and then combine them to produce a condensed version. The selection of important sentences is determined by factors such as language usage and statistical significance.

B. Abstractive summarization

Systems can produce novel phrases, which may involve rephrasing or incorporating new vocabulary that wasn't present in the original text. The process of generating an abstractive summary is more difficult compared to an extractive one because it requires the model to first comprehend the document and then express that understanding in a condensed format, potentially using original words and phrases. Abstractive summarization possesses more advanced capabilities such as generalization, paraphrasing, and

integrating real-world knowledge. Extractive approaches, on the other hand, have been the focus of much research due to their simplicity in defining rules that select significant sentences, resulting in a grammatically correct and coherent summary. However, these approaches may not perform well on lengthy and intricate texts, as they are more limited in scope.

II. RELATED WORK

In the field of natural language processing (NLP), text summarization has been a topic of interest. Researchers have proposed various methods for summarizing text, including unsupervised learning, abstractive and extractive summarization, and technique-based methodologies. Pratibha Devi Hosur [2]. An approach to automatic text summarization is to utilize unsupervised learning methods and implement the Lesk algorithm to produce a summary. In the case of summarizing lengthy texts in a time-efficient manner, the focus should be on utilizing extractive summarization methods. Deepali K Gaikwad [3,4,5]. discuss the extraction of the necessary information from long text documents to form a summary, mainly focusing on abstractive and extractive methods. Regular patterns are useful for extracting keywords in text summarization. Neelima Bhatia [6]. Explore different methodological approaches for summarizing text, such as the frequency-based approach that utilizes terms, diagram-based techniques, time-based strategies, division and combination-based strategies, semantic dependency strategies, theme-based approaches, talk-based methods, latent semantic-based methods, and methods that depend on lexical chains or fuzzy logic. [10,11,12]. They also discuss the considerable efforts made in the area of summarizing units and numerous archive outlines. Historically, most of the research in text summarization has focused on extractive methods, where important sentences or passages from the source document are selected and reproduced in summary [8,9]. However, human summaries tend to be abstractive, meaning they rephrase the original content in their own words instead of simply reproducing sentences from the source. Recently, there has been some research into abstractive summarization using machine learning techniques, which has been briefly summarized by Sarker. [13].

III. COMPARATIVE ANALYSIS OF ALGORITHMS

A. XLNET

XLNet is a high-performance language model that has demonstrated remarkable performance in various natural language processing tasks, including text summarization. It employs the Transformer architecture, a widely used NLP framework, and is pre-trained on massive amounts of unannotated text data using a modified language modeling objective. This objective predicts the next token in a sequence based on the preceding ones. However, what distinguishes XLNet is its ability to take into account all possible permutations of the input sequence during training, resulting in more accurate modeling of bidirectional context compared to previous models.

For text summarization, XLNet can be fine-tuned on a significant number of summarization examples to learn how to produce summaries that encapsulate the critical information in the input text. During fine-tuning, the model is trained to generate a summary based on the input text, and its parameters are adjusted using gradient descent to minimize the difference between the predicted and target summaries. XLNet-based summarization models have proven to surpass previous state-of-the-art models in multiple benchmark datasets, such as the CNN/Daily Mail and Gigaword datasets.

XLNet-based summarization models are particularly efficient in generating summaries that are informative and easy to read, which makes them suitable for various applications, such as news, document, and social media summarization.

B. GPT2

GPT-2 is a powerful language model developed by OpenAI that can perform various natural language processing tasks, including text summarization. It is based on the Transformer architecture and is trained on a large amount of unlabeled text data using a language modeling objective. For text summarization, GPT-2 can be fine-tuned on a large corpus of examples to generate a summary that captures the main points of the input text. GPT-2 can generate summaries in a way that includes new sentences that convey the same meaning as the original text. It has achieved

impressive results on multiple summarization tasks, including the CNN/Daily Mail and XSum datasets., GPT-2 has been shown to generate summaries that are both informative and fluent, which makes it suitable for a wide range of applications, such as news summarization, document summarization, and social media summarization. One of the key advantages of GPT-2 for text summarization is that it can generate summaries of variable length, which allows it to generate summaries that are tailored to the length requirements of different applications. GPT-2 can also be fine-tuned on specific domains to generate summaries that are more relevant to a particular application, such as scientific research papers or legal documents.

C. BERT

BERT (Bidirectional Encoder Representations from Transformers) is an open-source machine learning framework that has transformed the field of natural language processing (NLP). Developed by Google, it utilizes a deep-learning architecture called Transformers. BERT's major breakthrough is its ability to understand the context of language. Unlike traditional NLP models, where words are viewed as separate units of meaning, BERT considers the context in which they are used, as the meaning of a word can vary depending on the context. This is made possible by the bidirectional nature of the Transformers, and a self-attention mechanism that allows the model to focus on various parts of the input sequence while processing it. Pre-trained on a vast dataset of text from Wikipedia, BERT was trained to predict missing words in a sentence using surrounding context, allowing it to learn patterns and structures in natural language. BERT can then be fine-tuned for specific NLP tasks such as text classification, question answering, and language translation, by training the model on a smaller dataset specific to the task. The fine-tuned model can subsequently be used to predict on new data.

IV. PROPOSED METHODOLOGY

Fig. 1. Flowchart of Proposed Solution

The project begins with uploading a pdf document that is in text-structured data format. Although only a single pdf document can be uploaded. Now making use of python script and boto3 library. The document gets uploaded to the Amazon S3 bucket. Moreover, this pdf is fetched to get the text data from the document.

Now the text data is extracted from the pdf and then it is processed to form a token concerning the document stored in the Amazon S3 bucket. This token is used to access the pdf document in S3 and to extract the text data from it. This extracted text data is preprocessed so that it can be visualized and summarized. Preprocessing includes numerous techniques like Lower Casing, Tokenization, Punctuation Mark Removal, Stop Word Removal, Stemming, and Lemmatization of the words extracted from the pdf document.This preprocessed data is used for visualization of the data then it is sent to the summarization models for getting a summary.There are various methods used to visualize the text data but for the use case, we have done Word length histogram, Word length probability plot, Word Cloud, and Top Unigrams.We preferred not to use the traditional topic modeling as we were able to attain better visualization and results from the techniques we used and analyze the text data in a better way.Then for the summarization, we decided to use the transformer models for summarization by changing the input and output dimensions.For this, we have used the pertained transformer models which are BERT, GPT-2, and

XLNET.To get the best result we compared the accuracy and performance of the models based on the test dataset and get the best result out of it.This is how the final model for summarization was finalized and used for the summary generation.

V. EVALUATION METRICS

A.ROUGE

(Recall-Oriented Understudy for Gisting Evaluation): ROUGE is a group of metrics utilized to assess the excellence of a summary. It gauges the similarity between the summary created and the reference summary in regards to n-gram overlap, recall, and precision.

B.BLEU (Bilingual Evaluation Understudy)

While BLEU is typically used to evaluate machine translation systems, it can also be utilized for evaluating the effectiveness of text summarization. Its focus is on measuring the accuracy of n-gram overlap between the generated summary and the reference summary.

C. F1-score

The F1-score is a metric used to evaluate the comprehensive effectiveness of a summarization system. It is calculated as the harmonic mean of precision and recall, and its value falls within the range of 0 to 1.

D. Recall

The metric of recall gauges the proportion of pertinent details in the reference summary that can also be found in the generated summary.

E. Precision

Precision is a metric that calculates the proportion of information in the generated summary that is significant in the context of the reference summary.

VI. OUTPUT

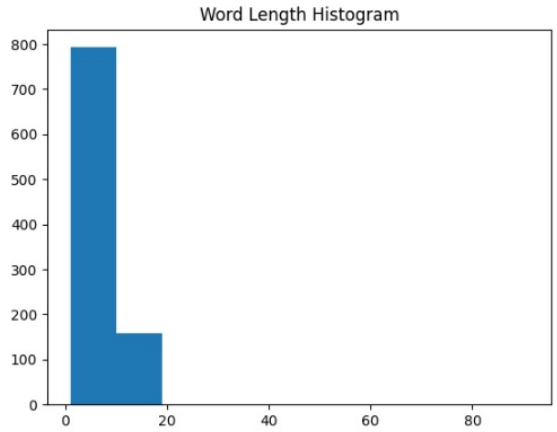

Fig. 2. Word Length Histogram

A. Word length histogram

A word length distribution is a visual representation of how frequently words of different lengths appear in a text or dataset. The horizontal axis displays the length of the word, while the vertical axis displays the frequency of words with that length. This histogram is useful in determining the most common word lengths and understanding the overall distribution of word lengths in the text.

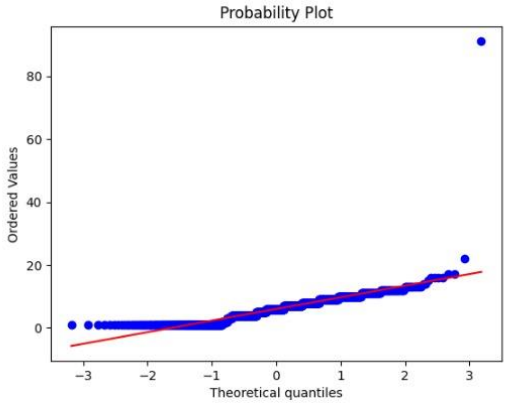

Fig. 3. Probability Plot

B. Word length probability plot

It is a graphical representation of the probability distribution of word lengths in a given text or dataset. It shows the probability of the occurrence of words of different lengths. This plot can help in identifying the probability of the occurrence of words of a particular

length and how it compares to the overall distribution.

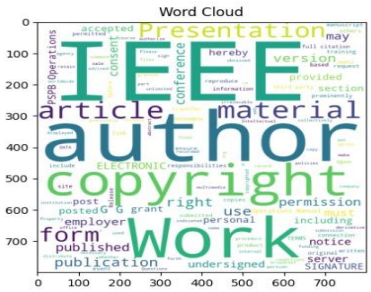

Fig. 4. Word Cloud

C. Word Cloud

It is a graphical representation of the most frequent words in a given text or dataset. The size of each word in the cloud represents its frequency of occurrence. This visualization technique can help in identifying the most common words in the text and can be used for exploratory analysis or data visualization.

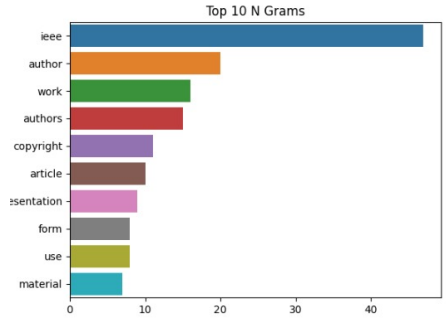

Fig. 5. Top Unigrams

D. Top Unigrams

These refer to the most frequent single words in a given text or dataset. Identifying the top unigrams can help in understanding the most commonly used words and can be useful in tasks such as text classification and sentiment analysis.

VII. RESULT

Type	Model	R	P	f	Score
Rogue-1	BERT	0.2456	1.0	0.39436	0.7987
	XLNET	0.3333	1.0	0.49999	0.8292
	GPT2	0.2719	1.0	0.42758	0.8053
Rouge - 2	BERT	0.1773	0.97	0.29999	0.798
	XLNET	0.1799	0.97	0.30368	0.8292
	GPT2	0.1902	0.97	0.31827	0.8053

Table. 1. Results of ROGUE evaluation over test text data

The following result was obtained by testing the model performance on the test text data. The overall Score was obtained from the various text data. Wherein, Evaluation of ROUGE-1 and ROUGE-2 was carried out for each model.

VIII. CONCLUSION

In this paper, we have described a general overview of text summarization and visualization. By pursuing this project we are trying to reduce the hefty work of domain experts to go through all the tedious documentation sourced from cloud platforms. This would also help in the data retention policy of various organizations where the backup of data can be stored on-premises and relevant summaries and visualization can be obtained to save storage and costs. In the future, we plan to continue researching and improving the quality of text processing and NLP/API models to enhance text summarization. Additionally, we aim to study the impact of other techniques on summarization and find various ways to integrate our projects with various

REFERENCE

[1] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," Computational linguistics, vol. 28, pp. 399- 408, 2002.

[2] Pratibha Devihosur, Naseer R. "Automatic Text Summarization Using Natural Language Processing" International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 08, Aug-2017

[3] Deepali K. Gaikwad, C. Namrata Mahender, "A Review Paper on Text Summarization", "International Journal of Advanced Research in Computer and Communication Engineering". Vol.5, Issue 3, March 2016.

[4] G. Vijay Kumar, M. Sreedevi, NVS Pavan Kumar, "Mining Regular Patterns in Transactional Databases using Vertical Format"."International Journal of Advanced Research in Computer Science", Volume 2, Issue 5, 2011. [5] G. Vijay Kumar and V. Valli Kumari, "Sliding Window Technique to Mine Regular Frequent Patterns in Data Streams using

Vertical Format", IEEE International Conference on Computational Intelligence and Computing Research, 2012.
[6] Neelima Bhatia and Arunima Jaiswal, "Automatic Text Summarization and its Methods-AReview", 6th International Conference. Cloud System and Big Data Engineering, 2016.
[7] Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6)
[8] El-Shishtawy T, El-Ghannam F. Keyphrase-based Arabic summarizer (KPAS). arXiv Prepr arXiv:1206.5384. 2012.
[9]Azmi A. Al-thankyou S. Ikhtasir—a user-selected compression ratio Arabic text summarization system. In: 2009 international conference on natural language processing and knowledge engineering. Dalian: IEEE; 2009. p. 1–7.
[10]AlSanie W. Towards an infrastructure for Arabic text summarization using rhetorical structure theory. M.Sc. Thesis, Dept. of Computer Science, King Saud University, Riyadh, Saudi Arabia. 2005.
[11]Douzidia FS, Lapalme G. Lakhas, an Arabic summarization system. In: Proceedings of 2004 document understanding conference (DUC2004). Boston: NIST; 2004. p. 260–73.
[12]Haboush A, Al-Zoubi M. Arabic text summarization model using clustering techniques. World Comput Sci Inf Technol J. 2012;2:62–7.
[13]Sarker IH, Kayes ASM, Watters P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J Big Data. 2019;6:57.