IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Text Summarization Using NLP

Sudarshan Sutar,Indraneel Surve,Mirza Munawwar,Vishal Nanaware, Assistant Prof. Priyanka Dhumal
Department of Computer Engineering Zeal College of Engineering and Research, Pune – 411041.

Abstract: This research introduces a groundbreaking text summarization approach by combining BERT for extractive summarization and GPT for abstractive summarization. The synergy of these models results in a hybrid system that leverages the precision of extraction and the linguistic fluency of abstraction. Experimental results demonstrate the model's efficacy in producing high-quality summaries, showcasing its potential impact on information synthesis across diverse domains.

INTRODUCTION

In today's information age, we are inundated with an ever-expanding volume of textual data. From news articles and research papers to social media posts and legal documents, the sheer quantity of text available can be overwhelming. Amid this data deluge, the need for efficient methods to distill, condense, and extract meaningful information from text has become increasingly critical. This is where Text Summarization using Natural Language Processing (NLP) emerges as a transformative technology.

Text summarization is the process of automatically generating a concise and coherent summary of a longer text, while retaining its essential information and meaning. This technology is a cornerstone of NLP, a field at the intersection of artificial intelligence and linguistics that focuses on enabling computers to understand, interpret, and generate human language. The goal of text summarization is to make large volumes of text more manageable and accessible, catering to the time constraints and information overload faced by individuals, researchers, and organizations.

Text summarization can be broadly categorized into two main approaches: extractive and abstractive summarization. Extractive summarization involves selecting sentences or phrases directly from the source text that are deemed most important or representative of its content. Abstractive summarization, on the other hand, goes a step further by generating summaries that may not be verbatim extracts but convey the same ideas using different words and structures, often resembling a human-authored summary.

DEMAND FOR NATURAL LANGUAGE PROCESSING

The demand for Natural Language Processing (NLP) has surged in recent years, driven by the exponential growth of textual data across digital platforms. With the vast amounts of information available in the form of social media posts, articles, emails, and research papers, NLP plays a pivotal role in extracting meaningful insights from this unstructured text. Enhanced search engines leverage NLP to understand user queries in natural language, providing more accurate and relevant results.

The proliferation of chatbots and virtual assistants in various industries, from customer service to healthcare, relies on NLP for natural language understanding, enabling these systems to interact effectively with users. Sentiment analysis and opinion mining, powered by NLP, allow businesses to gauge and respond to customer sentiments expressed in social media and reviews. Language translation services benefit from NLP, breaking down language barriers and fostering global communication. Additionally, NLP contributes to the development of systems capable of summarizing content and generating human-like text, addressing the need for automated information synthesis in today's data-intensive environment.

For text summary, there are essentially two different methods:

1. Extractive Summarization
2. Abstractive Summarization

EXTRACTIVE SUMMARIZATION

Extractive summarization is a technique in Natural Language Processing (NLP) that involves selecting and combining key sentences or passages directly from the source text to create a concise summary. Unlike abstractive summarization, which generates new sentences to capture the essential meaning, extractive summarization relies on identifying and extracting the most important information already present in the original content. This method often involves algorithms that assign weights to sentences based on various criteria such as relevance, importance, and informativeness. Extractive summarization is particularly useful when the goal is to maintain the original wording and ensure that the summary accurately represents the key points of the source material. While it may lack the creative flexibility of abstractive methods, extractive summarization is valued for its simplicity, transparency, and ability to preserve the intended meaning of the text.

ABSTRACTIVE SUMMARIZATION

Abstractive summarization is a Natural Language Processing (NLP) technique that involves generating a summary of a document by paraphrasing and rephrasing the content in a way that captures the essential meaning while often introducing new language constructs. Unlike extractive summarization, which selects and combines existing sentences from the source text, abstractive summarization aims to create a concise summary that may not necessarily reuse the exact words from the original document.

PROPOSED SYSTEM POSSIBLE ALGORITHMS

1. Bidirectional Representations from Transformers (BERT):

BERT, or Bidirectional Encoder Representations from Transformers, is a revolutionary natural language processing (NLP) model developed by Google. Introduced in 2018, BERT has significantly advanced the field of contextualized word embeddings and language understanding. Unlike traditional NLP models that process words in a unidirectional manner, BERT employs a bidirectional transformer architecture, allowing it to consider both preceding and succeeding words in a sentence simultaneously. This bidirectional context enables BERT to capture rich semantic relationships and contextual nuances within the input text. BERT's pre-training process involves exposing the model to vast amounts of unlabeled text from the internet, allowing it to learn contextual representations of words and phrases. The resulting embeddings exhibit a deep understanding of language semantics, making BERT a powerful tool for various NLP tasks, including text classification, named entity recognition, sentiment analysis, and, notably, text summarization.

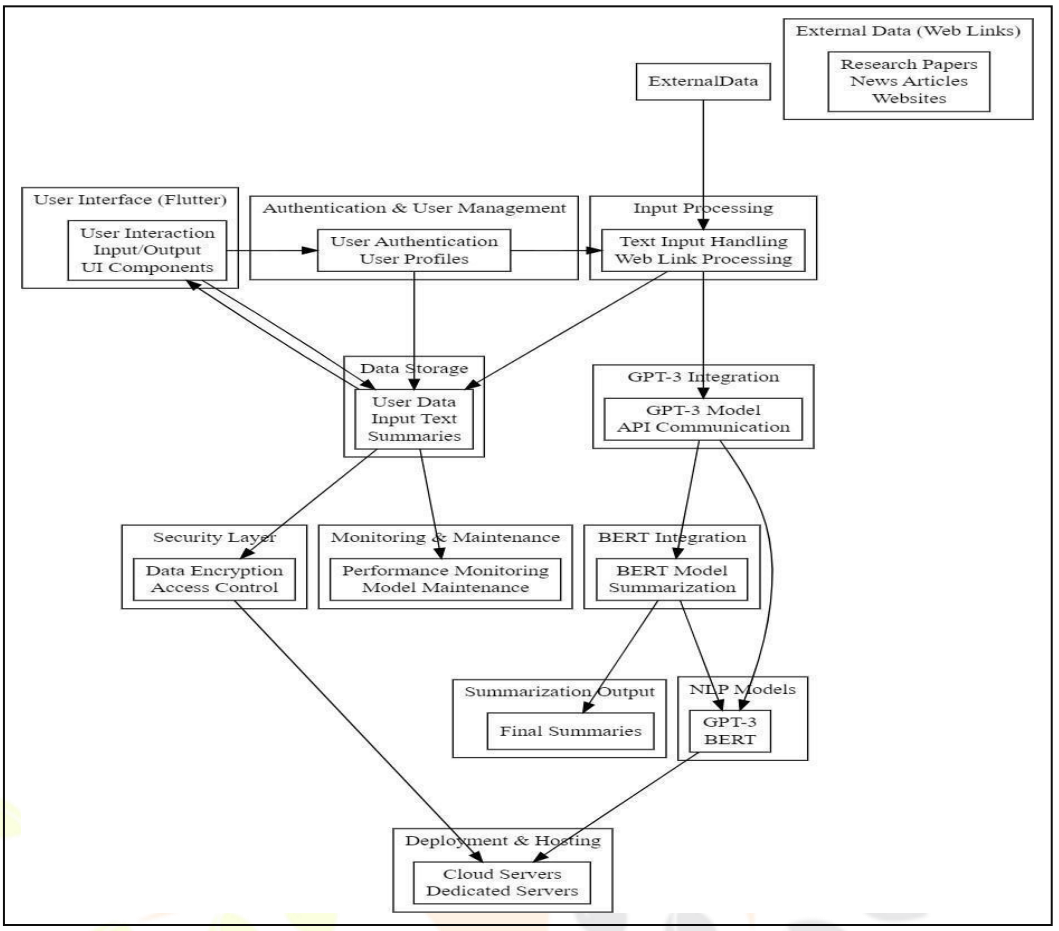
BERT's success can be attributed to its attention mechanism, which allows the model to assign different weights to different words in a sentence based on their contextual relevance. This attention to context enables BERT to generate contextually rich embeddings, making it effective in capturing the intricate meanings and relationships within sentences. BERT's versatility extends beyond its pre-training; fine-tuning the model on specific downstream tasks allows it to adapt to various applications with relatively small amounts of task-specific labeled data. The contextual embeddings produced by BERT have become a cornerstone in advancing the state-of-the-art performance across a wide range of NLP benchmarks and applications, making it a pivotal component in the development of sophisticated and context-aware language models.

2. Generative Pre-trained Transformer (GPT):

Generative Pre-trained Transformer (GPT) is an innovative natural language processing (NLP) model developed by OpenAI. Unveiled in 2018, GPT has garnered attention for its ability to generate coherent and contextually relevant human-like text. Positioned as a transformer-based autoregressive language model, GPT utilizes a multi-layer architecture with self-attention mechanisms. Its unique feature lies in its pre-training methodology, where the model is exposed to diverse internet text without task-specific labels, allowing it to learn the intricacies of language structure, context, and semantics. This pre-training, combined with a large number of parameters, endows GPT with a robust understanding of language, enabling it to perform a variety of downstream tasks, including text completion, question-answering, language translation, and text summarization.

One of GPT's key strengths is its capacity for autoregressive text generation. During fine-tuning on specific tasks, GPT refines its language generation capabilities to produce contextually appropriate and coherent responses. The model excels at predicting the next word in a sequence given the preceding context, making it adept at generating human-like responses and completing text prompts. GPT's effectiveness in natural language understanding and generation has led to the development of subsequent versions, each with increased model size and capabilities, solidifying GPT as a groundbreaking technology in the realm of NLP and contributing significantly to the progress of language models capable of understanding and producing nuanced, contextually appropriate text.

PROPOSED SYSTEM BLOCK DIAGRAM



EXPECTED OUTCOME

Text summarization using models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) typically aims to produce concise and coherent summaries of longer pieces of text. The expected outcome of text summarization using these models includes:

[1] Conciseness: Summaries generated by BERT and GPT models are expected to be concise and to the point. They should capture the essential information from the input text while eliminating redundant or irrelevant details.

[2] Coherence: The generated summaries should maintain a coherent and logical flow of information. Sentences in the summary should be well-structured and connected, making it easy for readers to follow the main points.

[3] Preservation of Key Information: BERT and GPT models are designed to understand the context and semantics of the input text. As a result, the expected outcome includes the preservation of key information, facts, and ideas from the original text in the summary.

[4] Grammatical Correctness: Summaries should be grammatically correct and free from errors. BERT and GPT models, being pre-trained on large text corpora, have strong language generation capabilities and can produce fluent summaries.

[5] Abstraction: Depending on the type of summarization (extractive or abstractive), the models may either extract sentences directly from the input text (extractive) or generate novel sentences that convey the same meaning (abstractive). Abstractive summarization using models like GPT aims to provide more human-like summaries by rephrasing and reorganizing information.

CONCLUSION

Text summarization has emerged as a crucial application of Natural Language Processing (NLP), addressing the growing challenges of information overload and the need for efficient content distillation. Leveraging BERT and GPT in this context has yielded promising results.

In summary, text summarization using NLP represents a valuable technology for coping with the information-rich digital landscape. It empowers individuals and organizations to manage, understand, and utilize vast amounts of textual data efficiently. As NLP techniques continue to advance, the future of text summarization holds the potential for even more sophisticated and tailored approaches, further enhancing its role in information management and accessibility.

ACKNOWLEDGMENT

We appreciate Ms Priyanka Dhumal , our guide, for her unwavering support and direction. We were able to effectively complete our job thanks to her coaching. Additionally, we would like to express our gratitude to the institution, division, and specific people for their unwavering support and direction during this study. The success of this initiative would not have been achieved without their assistance.

REFERENCES

[1] Dilawari, M. U. G. Khan, S. Saleem, Zahoor-Ur-Rehman and F. S. Shaikh, "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space," in IEEE Access, vol. 11, pp. 23557-23564, 2023, doi: 10.1109/ACCESS.2023.3249783.

[2] Á. Hernández-Castañeda, R. A. García-Hernández and Y. Ledeneva, "Toward the Automatic Generation of an Objective Function for Extractive Text Summarization," in IEEE Access, vol. 11, pp. 51455-51464, 2023, doi: 10.1109/ACCESS.2023.3279101.

[3] D. Yadav, R. Katna, A. K. Yadav and J. Morato, "Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey," in IEEE Access, vol. 10, pp. 133981-134003, 2022, doi: 10.1109/ACCESS.2022.3231016

[4] P. Mahalakshmi and N. S. Fatima, "Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques," in IEEE Access, vol. 10, pp. 18289-18297, 2022, doi: 10.1109/ACCESS.2022.3150414.