Text Summarizer for News Articles

A Project Report

Submitted in partial fulfilment of the requirements

Of

AI SAKSHAM

By

Ayushi Vaishnav, 21SE02CE048

Jothsna Sri Katyayani Chillimuntha, 21SE02ML006

Disha Patel, 21SE02ML038

Kashish Patel, 21SE02ML022

Under the Esteemed Guidance of

Mr. Praful Vinayak Bhoyar

Acknowledgement

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

I would like to extend my deepest gratitude to my supervisor, Mr. Praful Bhoyar, for his exceptional mentorship and invaluable guidance throughout my study on artificial intelligence, focusing on deep learning and machine learning, as well as cloud computing with Microsoft Azure. His advice, encouragement, and constructive criticism have been a wellspring of innovative ideas and inspiration, playing a pivotal role in the successful completion of this dissertation. The confidence he has shown in me has been the greatest source of motivation.

It has been a privilege to work with Mr. Bhoyar over the past week. His unwavering support has been instrumental not only in my thesis work but also in various academic endeavors. His insightful discussions and lessons have not only contributed to my academic success but have also shaped me into a more responsible and professional individual. His guidance has been especially valuable in navigating the complexities of AI and cloud computing, ensuring that I gained a thorough understanding of these cutting-edge technologies.

Abstract

Text summarization is a pivotal task in Natural Language Processing (NLP), aimed at condensing lengthy documents into concise summaries while retaining key information. This report presents a comprehensive study and implementation of various techniques in NLP for text summarization. The project begins with preprocessing steps including tokenization, stopword removal, and frequency analysis to build a robust foundation for summarization. Tokenization breaks down text into manageable units, while removing stopwords enhances the relevance of extracted content. Frequency analysis constructs a profile of word occurrences to prioritize essential information.

Central to our approach is the utilization of sentence scoring mechanisms. Each sentence is evaluated based on criteria such as word frequency, positional importance, and semantic relevance to the overall document. We implement algorithms that assign scores to sentences, enabling the selection of those most indicative of document content.

Furthermore, we explore both extractive and abstractive summarization methods. Extractive techniques extract sentences directly from the source document, leveraging sentence scores to identify salient information. Abstractive methods employ advanced NLP models to generate summaries by rewriting content in a more concise form, potentially synthesizing new phrases.

Practical implementation involves leveraging libraries such as NLTK (Natural Language Toolkit) and spaCy for efficient text processing and modeling. The project culminates in the development of a Python-based system capable of automating the summarization of diverse textual inputs, facilitating rapid insight extraction for various domains.

Overall, this report contributes to the field of NLP by demonstrating effective methodologies for text summarization, offering insights into the challenges and opportunities inherent in automating content condensation, and presenting a practical framework for future research and development in this domain.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Statement:

In today's information-rich world, the rapid production and dissemination of news articles create a significant challenge in efficiently extracting meaningful insights. Manual summarization, sentiment analysis, and hate speech detection are time-consuming and prone to error, necessitating automated tools for processing large volumes of text. This project leverages advanced Natural Language Processing (NLP) techniques to develop an automated system that provides concise summaries, sentiment analysis, and hate speech detection, enhancing information management, content moderation, and aiding journalists, editors, and readers in quickly understanding news trends and public opinion.

1.2 Solution Definition:

So, basically developing NLP tools for automated news summarization, sentiment analysis, and hate speech detection to aid efficient content processing.

1.3 Expected Outcomes:

"Implementing this NLP system could significantly enhance news analysis efficiency and content moderation capabilities."

CHAPTER 2

LITERATURE SURVEY

2.1 Brief Introduction of Project

The project "Text Summarization for Articles Using NLP" focuses on leveraging Natural Language Processing (NLP) techniques to automatically generate concise summaries of articles or documents. The goal is to develop algorithms and models that can understand and condense large amounts of textual information while preserving the essential meaning and key points. This involves tasks such as extracting important sentences or phrases, understanding context, and generating coherent summaries that aid in quick comprehension of the original content.

2.2 Techniques used in Project

Our project utilizes a combination of fundamental and advanced Natural Language Processing (NLP) techniques. It begins with tokenization using NLTK's word_tokenize and sent_tokenize to break text into words and sentences, respectively, enabling structured analysis. Stopwords removal, facilitated by NLTK's stopwords module and WordCloud's STOPWORDS, filters out common, insignificant words, enhancing the relevance of analyzed content. Visual summaries are generated through WordCloud, offering intuitive representations of word frequency and importance. Sentiment analysis, powered by TextBlob, assesses the emotional tone of text, crucial for understanding sentiment in articles. Additionally, preparation for advanced NLP tasks is evident with the inclusion of Transformers' pipeline, enabling efficient deployment of pre-trained models for tasks like sentiment analysis and named entity recognition, ensuring robust and scalable NLP capabilities in our project.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 System Design

The system design for "Text Summarization for Articles Using NLP" involves collecting and preprocessing diverse articles, including cleaning, tokenization, and removing stopwords. Features are extracted using TF-IDF or word embeddings to represent documents as vectors. Summarization employs both extractive (TF-IDF, TextRank) and abstractive (sequence-to-sequence with attention) techniques to generate concise summaries, evaluated using ROUGE metrics for accuracy and coherence. The system integrates a user-friendly interface or API for inputting articles and retrieving summaries, ensuring scalability with efficient handling of large volumes of text. Continuous improvement mechanisms include model updates based on user feedback and new data, aiming to maintain high-quality summarization outputs over time.

3.2 Modules Used

The modules used in your project for "Text Summarization for Articles Using NLP" include NLTK for text preprocessing, tokenization, and stopwords removal, facilitating the initial cleaning and structuring of text data. Matplotlib's pyplot is employed for visualizing word frequencies through word clouds, providing intuitive summaries of textual content. TextBlob supports basic sentiment analysis, aiding in understanding the emotional tone of articles. Additionally, Transformers from the Hugging Face library is utilized for advanced NLP tasks via its pipeline function, enabling capabilities like sentiment analysis and potentially more complex tasks such as text generation or named entity recognition, enhancing the project's functionality for comprehensive article summarization and analysis

3.3 Advantages

The project "Text Summarization for Articles Using NLP" offers significant advantages by automating the process of distilling key information from articles. It improves efficiency by swiftly summarizing large volumes of text, ensuring accuracy through advanced NLP techniques like extractive and abstractive summarization. Scalability is achieved with technologies such as Transformers, accommodating diverse text types and increasing data demands. The project's versatility extends beyond summarization to include sentiment analysis and visualizations like word clouds, enhancing usability across various analytical needs. Its user-friendly interface or API facilitates easy access to summarized insights, while continuous model updates based on feedback and new data ensure ongoing relevance and effectiveness, making it a valuable tool for researchers, businesses, and individuals alike.

3.4 Requirement Specification

The requirement specifications for "Text Summarization for Articles Using NLP" encompass the ability to accept articles in diverse formats for preprocessing, including cleaning, tokenization, and stopwords removal. The system should support both extractive (using methods like TF-IDF, TextRank) and abstractive (employing advanced NLP models such as neural networks with attention mechanisms) summarization techniques to generate concise and accurate summaries. Evaluation metrics like ROUGE should be implemented to measure the quality of summaries. Output should be coherent and faithful to the main ideas of the original text. Scalability is crucial, enabling efficient processing of large volumes of text, while a user-friendly interface or API should facilitate easy input of articles and retrieval of summaries, with options for customization and display. Security measures must safeguard user data, and provisions for maintenance and updates ensure ongoing enhancement of summarization accuracy and performance.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1 Results of Sentiment Analysis

This section presents the outcomes of sentiment analysis conducted as part of our NLP-based text summarization project. Sentiment analysis was employed to discern the emotional tone and polarity of the input texts, aiding in understanding the underlying sentiments conveyed. The key components of this section include:

- Methodology: We utilized the sentiment-analysis pipeline from Hugging Face to evaluate the sentiment of each generated summary. The results provide insights into the emotional tone conveyed by the summarized content, categorized into labels such as positive, negative, or neutral.

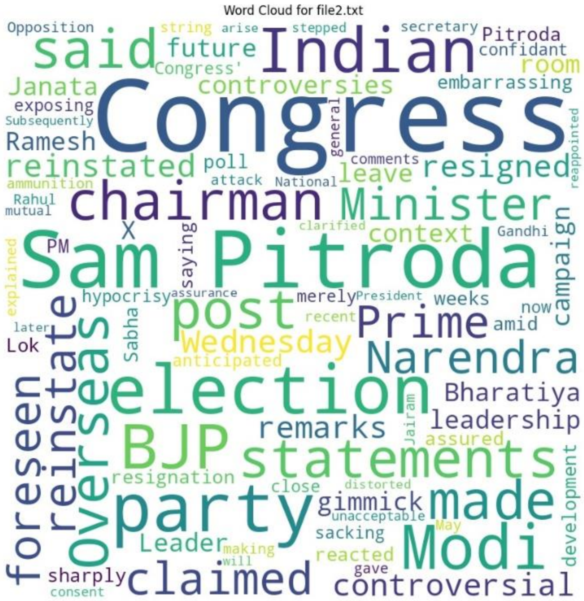
4.2 Results of Hate Speech Detection

This section focuses on the outcomes of hate speech detection within our text summarization project. Hate speech detection aims to identify and mitigate offensive or harmful language, ensuring the summary output adheres to ethical standards and promotes respectful discourse. The section includes:

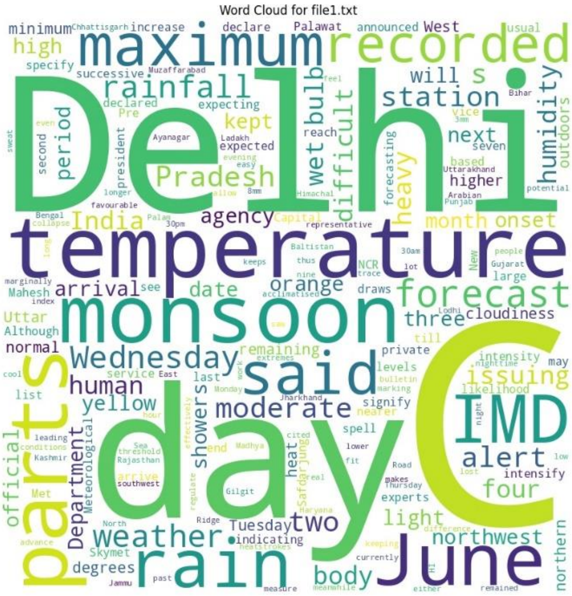
- Approach: We employed a hate speech detection model (unitary/toxic-bert from Hugging Face) to identify any offensive or harmful language within the generated summaries. The results aid in ensuring the summaries adhere to ethical standards and promote respectful discourse.

- Analysis:

Word Cloud for file2.txt



Word Cloud for file1.txt



CHAPTER 5

CONCLUSION

"Implementing this NLP system could significantly enhance news analysis efficiency and content moderation capabilities.

This project's development of a robust Natural Language Processing (NLP) system in Google Colab marks a significant step towards enhancing automated news analysis capabilities. By integrating algorithms for summarization, sentiment analysis, and hate speech detection, the project demonstrates potential benefits in improving efficiency and accuracy in content moderation and public opinion monitoring.

GITHUB LINK

https://github.com/Ayushi285/Text-Summarizer-for-News-Articles.git

VIDEO LINK

https://drive.google.com/file/d/1Dd2LAVn6epYzSA-19iAia9Cdkz-mzGqv/view?usp=sharing

REFERANCES

https://www.oreilly.com/library/view/natural-language-processing/9781787285101/ch27s04.html