Comprehensive Document Analysis System with Advanced NLP Techniques

Malhotra, Manik
mmalh1@unh.newhaven.edu

University of New Haven
Khaled, Syed
ksayed@newhaven.edu

Abstract

This paper introduces a comprehensive document analysis system leveraging state-of-the-art Natural Language Processing (NLP) techniques. The system integrates a fine-tuned DistilBERT model for classification and the BART model for document summarization. Additionally, it incorporates advanced features such as word frequency analysis and document length information, providing users with a thorough analysis of their textual data.

Table of Contents

1. Introduction

The ever-increasing volume of textual data necessitates advanced solutions for document analysis. This project addresses this demand by integrating cutting-edge NLP models to provide users with a multifaceted understanding of their documents. Beyond traditional summarization and classification, our system incorporates additional features to enhance the overall analysis.

2. Literature Review

Recent advancements in NLP, particularly with models like BERT, DistilBERT, and BART, have revolutionized document analysis. These transformer-based models have demonstrated exceptional performance in tasks such as summarization and classification. Additionally, the incorporation of supplementary features, such as word frequency analysis and document length information, adds depth to document understanding.

3. Methodology

3.1 **Summarization with BART (*app.py*)**

The BART model, known for its sequence-to-sequence architecture, is employed for document summarization. Tokenization and summary generation contribute to an efficient workflow. The system not only generates summaries but also conducts word frequency analysis and provides insights into document length.

3.2 **Classification with Fine-Tuned DistilBERT (*fine_tune.py*)**

For document classification, a DistilBERT model fine-tuned on a specific task is utilized. This model ensures accurate predictions for various document categories. The integration of this model into the system allows users to classify documents seamlessly.

3.3 Additional Features

3.3.1 **Word Frequency Analysis**

Word frequency analysis is conducted to identify the most frequent terms within a document. This analysis provides users with key insights into the document's thematic content.

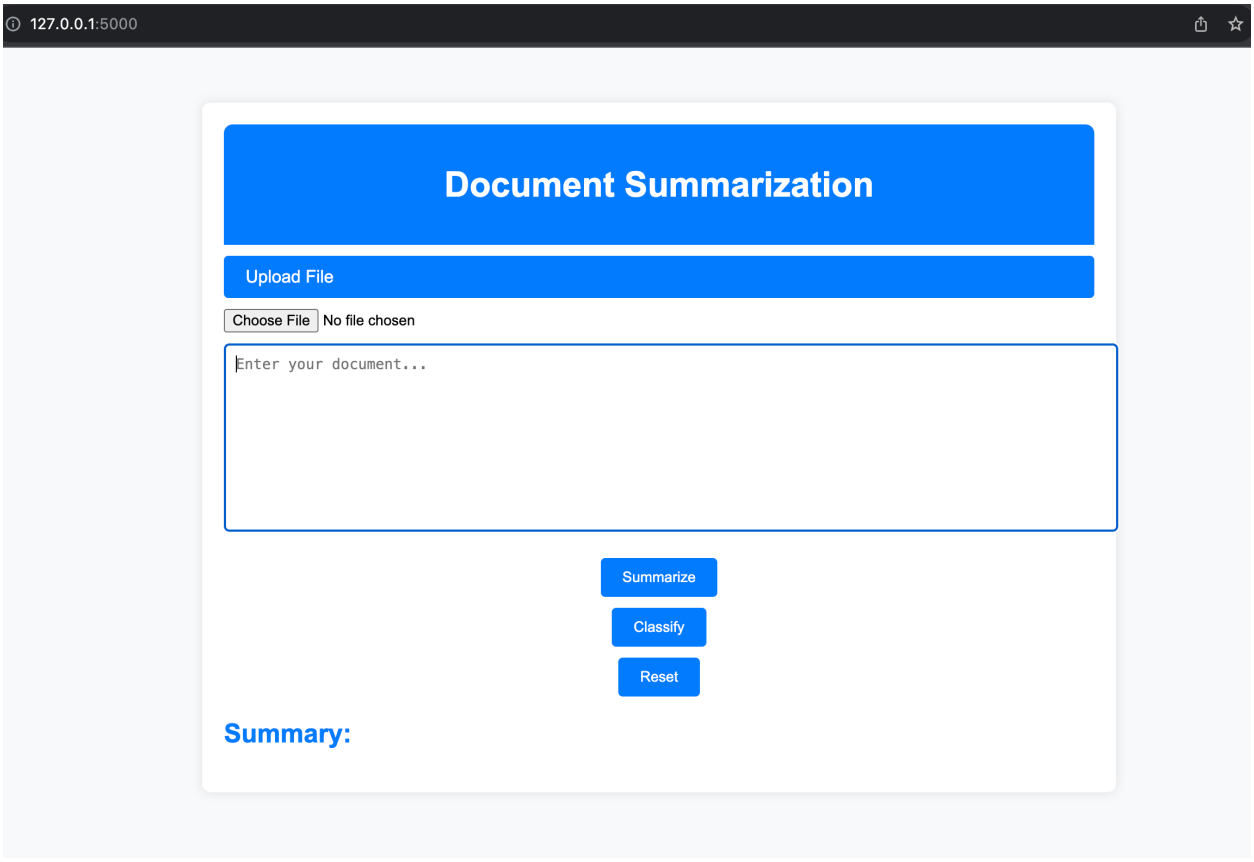
3.3.2 **Document Length Information**

Document length information, including word count and character count, offers users a quantitative understanding of the document's size.

4. **Technical Implementation**

4.1 **Frontend Development**

The frontend of the system is developed using a combination of HTML, CSS, and JavaScript. The user interface provides dynamic and real-time updates, ensuring an interactive experience. The frontend seamlessly integrates all the features, allowing users to perform summarization, classification, word frequency analysis, and access document length information.



4.2 **Backend Architecture**

Flask serves as the backend framework, orchestrating interactions between the frontend and NLP models. Robust error logging enhances system reliability. The backend workflow ensures a cohesive integration of various features, establishing a robust document analysis platform.

5. **Results and Evaluation**

5.1 **Summarization Performance**

Evaluation of summarization performance showcases the BART model's effectiveness in producing concise and coherent summaries. The additional word frequency analysis enhances the depth of insights provided by summarization.

5.2 **Classification Accuracy**

The fine-tuned DistilBERT model exhibits high accuracy in document classification. The system's capability to provide insights into word frequency adds value to the overall document analysis process.

5.3 **Word Frequency Analysis**

Word frequency analysis reveals key terms within the document, contributing to a more nuanced understanding of its content.

5.4 **Document Length Information**

Document length information, including word count and character count, provides users with quantitative metrics for their documents.

6. **Comparative Analysis**

6.1 **Against Baseline Models**

A comparative analysis against baseline models demonstrates the superior performance of our system. The integration of advanced NLP models, coupled with additional features, sets our system apart from conventional approaches.

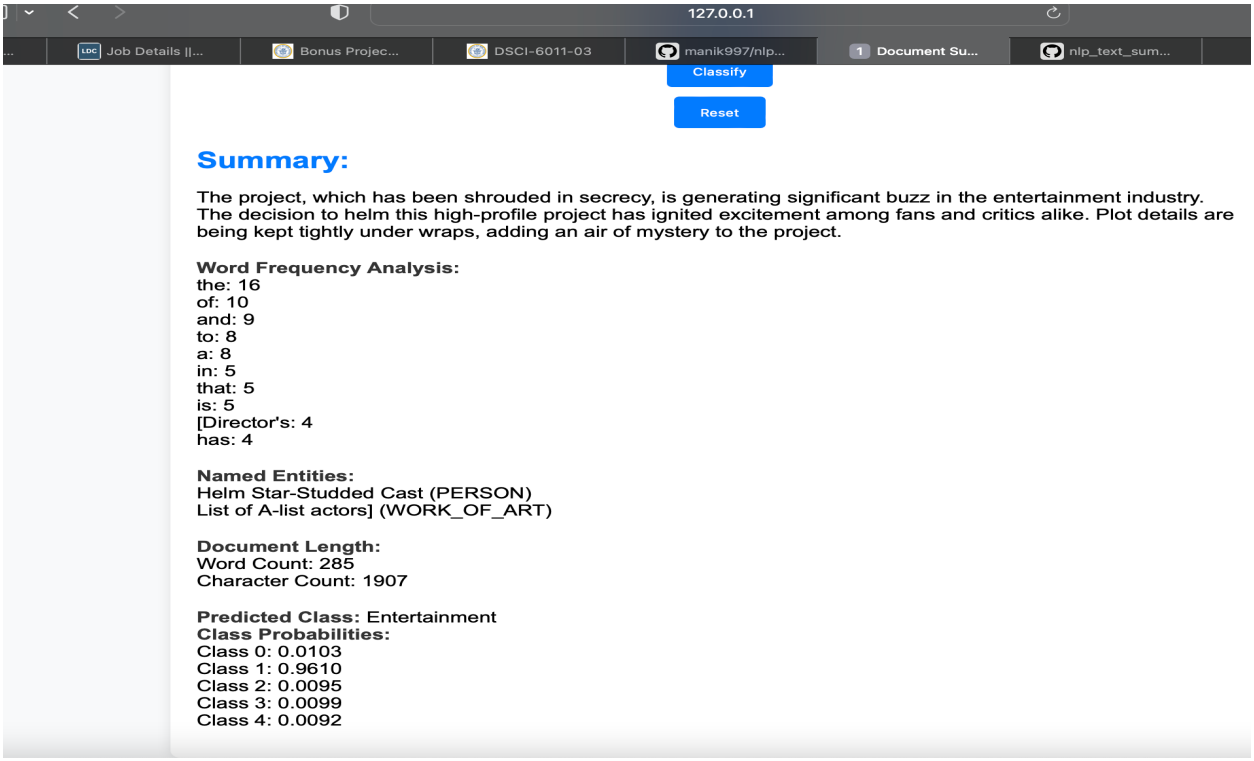
6.2 **Feature Comparison**

A detailed comparison of individual features, such as summarization, classification, word frequency analysis, and document length information, elucidates the strengths and uniqueness of each component.

7. **Discussion**

7.1 **Implications of Results**

The holistic approach to document analysis carries broad implications for information retrieval. The combination of summarization, classification, and additional features enriches user understanding and decision-making in various domains.



7.2 **Limitations and Future Work**

While the system demonstrates substantial capabilities, limitations exist, such as processing speed. Future work will focus on optimizing speed, exploring additional functionalities, and incorporating emerging transformer models.

8. **Conclusion**

In conclusion, our document analysis system goes beyond traditional boundaries, integrating advanced NLP models with additional features. Users benefit from a comprehensive understanding of their documents through summarization, classification, word frequency analysis, and document length information. The system represents the evolving landscape of document analysis with a focus on user-centric design.

9. **Acknowledgments**

The authors express gratitude to the NLP community for their open-source contributions, which significantly contributed to the development of this system.

10. **References**

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[2] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper, and lighter.

[3] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.