SmartDoc: Intelligent Text Search & Classification

INFORMATION RETRIEVAL PROJECT REPORT

Submitted by

Aman Sagar

2110110934



Department of Computer Science and Engineering

November, 2024

Acknowledgement

I am sincerely grateful to **Shiv Nadar University** for granting me the opportunity to work on this project, which has significantly contributed to my academic and professional development. This experience has not only deepened my understanding of key concepts but also strengthened my practical skills in the field.

I would like to extend my deepest appreciation to **Prof. Sonia Khetarpaul** for her invaluable guidance, insightful advice, and continuous encouragement throughout the course of this project. Her mentorship and constructive feedback were instrumental in refining my approach and achieving the project's objectives. Without her support, the successful completion of this project would not have been possible.

I am also thankful to the faculty members and resources provided by the university, which played a crucial role in facilitating my research and enabling me to overcome various challenges during this endeavour. The knowledge and experience gained through this project have been invaluable and will undoubtedly shape my future pursuits.

Table of content

1. **Acknowledgment**
2. **Abstract**
3. **Introduction**  
   1.1. Background  
   1.2. Importance of Topic  
   1.3. Motivation  
   1.4. Contribution
4. **Literature Review / Related Work**
5. **Novelty in My Approach**
6. **Objective**
7. **Proposed Model**
8. **Methodology**
9. **Experimentation and Results**
10. **Conclusions and Limitations**
11. **References**

Abstract

This project presents **SmartDoc**, an AI-powered system designed for efficient text classification and intelligent document search. Utilizing a custom transformer-based approach with DistilBERT, the system is capable of categorizing text into predefined domains, such as business, sports, and technology, while also offering advanced search capabilities. The solution leverages techniques like tokenization, embedding, and extractive question answering, enabling users to classify documents, search within text, and retrieve context-specific answers from uploaded PDFs or Word documents. Implemented with a user-friendly interface using Streamlit, **SmartDoc** offers enhanced information retrieval and text analysis capabilities, making it a powerful tool for both academic research and business applications. This project emphasizes custom solutions for preprocessing, classification, and semantic search, providing a scalable and efficient approach for handling text data.

1. Introduction
   1. Background

The rapid growth of digital content has made it increasingly difficult to manually sift through vast amounts of unstructured data. Traditional information retrieval (IR) methods rely heavily on keyword matching, which often leads to inaccurate or incomplete results, especially when the search query is complex or requires a deeper understanding of context. Recent advances in natural language processing (NLP), particularly with transformer-based models like BERT and its lighter variants like DistilBERT, have revolutionized the way we process and understand text. These models excel in tasks such as text classification, question answering, and semantic search, enabling more accurate and context-aware information retrieval.

* 1. Significance of the Topic

Efficient information retrieval is vital for multiple fields, including academic research, business intelligence, content management, and more. The sheer volume of information available online and in various organizational databases calls for intelligent systems capable of quickly identifying relevant content. Additionally, the ability to classify documents into predefined categories and search them semantically is crucial for improving decision-making processes, enhancing user experiences, and saving time. This project seeks to contribute to this evolving field by building a system that not only classifies text but also allows for intelligent searching and direct question answering from within documents.

* 1. Motivation

The motivation behind this project stems from the need for more powerful and adaptable tools that can handle the challenges posed by diverse and unstructured text data. Existing search engines and document classification systems are often limited in their ability to process complex queries or understand the context in which a search term is used. By incorporating AI-driven techniques such as transformer models and custom-built information retrieval components, this project aims to provide a more accurate, efficient, and user-friendly solution. The ability to search for specific information within documents and receive contextually relevant answers can save users significant time and effort in retrieving valuable insights.

1.4 Contributions

In this project, SmartDoc, an AI-powered system integrating text classification, semantic search, and question answering, was independently developed. The contributions include the design and implementation of a custom transformer-based classifier using DistilBERT to categorize documents into predefined categories such as business, sports, and technology. An intelligent search mechanism was also developed that identifies search terms within documents, corrects typos, and extracts contextually relevant sentences. Additionally, an extractive question answering feature was integrated, allowing users to input questions and retrieve direct, context-specific answers from the text. Finally, an interactive and user-friendly interface was created using Streamlit, enabling seamless document upload, classification, and searching. This comprehensive tool enhances the efficiency and accuracy of document classification and information retrieval.

1. Literature Review

**1. Large Language Models for Information Retrieval: A Survey (Zhu et al., 2023)**  
This comprehensive survey provided a detailed overview of how large language models (LLMs) can be leveraged in information retrieval (IR) systems. The insights from this paper were instrumental in understanding the potential of LLMs to improve text classification and document search tasks. Zhu et al. emphasized the challenges in optimizing LLMs for efficient IR, which guided my project's focus on using lightweight models like DistilBERT to achieve efficient processing without compromising accuracy. The study also informed my approach to integrating search functionalities with advanced language models for better retrieval performance.

**2. Text‐Based Question Answering from Information Retrieval and Deep Neural Network Perspectives: A Survey (Abbasiantaeb & Momtazi, 2021)**  
Abbasiantaeb and Momtazi's work explored the intersection of traditional IR techniques and deep neural networks for question-answering systems. Their survey was invaluable in shaping the question-answering module of my project, helping me understand how neural models can be combined with retrieval techniques to improve response accuracy. This paper provided the theoretical background for incorporating a QA module into my application, using DistilBERT to extract relevant answers from documents uploaded by users.

**3. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper, and Lighter (Sanh, 2019)**  
The research by Sanh on DistilBERT was critical in selecting the model for my project. By distilling BERT into a smaller, faster version, Sanh demonstrated that similar levels of accuracy could be maintained while significantly reducing computational costs. This was particularly relevant given my project's requirement for a fast, real-time text classification model that could be easily integrated into a web application. Sanh’s work guided the model training phase, where I fine-tuned DistilBERT on a specific dataset for document classification.

**4. A Multi-Source Retrieval Question Answering Framework Based on RAG (Wu et al., 2024)**  
Wu et al. introduced a framework combining retrieval-augmented generation (RAG) with question-answering systems to improve retrieval accuracy from multiple sources. Although my project did not implement RAG, the concept of enhancing QA capabilities by leveraging diverse data sources inspired the architecture of my information retrieval system. Wu's framework underscored the value of integrating various retrieval techniques, which influenced the design of my custom search functionality using TF-IDF and cosine similarity.

**5. Cross-Lingual Information Retrieval with BERT (Jiang et al., 2020)**  
Jiang et al.’s exploration of cross-lingual information retrieval using BERT models provided insights into expanding the reach of retrieval systems to multiple languages. Although my current implementation focuses on English-language documents, the methodologies discussed in this paper highlighted opportunities for future enhancements. This research laid the groundwork for potentially incorporating multilingual support in subsequent iterations of my project, aiming to extend its versatility and applicability.

1. Novelty of the approach

The novelty of this project lies in its integration of transformer-based models with a custom-built information retrieval pipeline tailored specifically for document classification, semantic search, and context-aware question answering. While traditional methods and prior research have relied on classic techniques such as keyword matching and shallow machine learning algorithms (e.g., Naive Bayes and Random Forest), these approaches often fall short in capturing the nuanced semantics of natural language, especially within unstructured text data fine-tuned DistilBERT model, this project significantly enhances both accuracy and efficiency in text classification and information retrieval. DistilBERT, a distilled version of BERT, offers faster processing while maintaining high classification accuracy, as explored in research by Sanh et al. (2019) . The integration-tolerant search engine and an extractive question-answering module further distinguishes this system by combining classification and retrieval with real-time, context-aware responses to user queries.

Unlike existing solutions that are limited to keyword-based search or predefined rule-based systems , the custom model here wlly to adapt to the nuances of the dataset, thereby enabling more accurate semantic understanding and retrieval capabilities. Additionally, while various studies have explored multilingual support , this project focuses on optimizing performnglish text processing, ensuring precision and relevance in a specific language domain.

Thus, the project not only advances the state-of-the-art by optimizing the DistilBERT model for domain-specific tasks but also addresses gaps in user interaction by offering a streamlined, interactive web application with enhanced search and classification features.

1. Objectives

The primary motivation behind this project is to simplify and optimize the process of extracting relevant information from large volumes of unstructured text. In today's fast-paced digital world, accessing the right information quickly and efficiently is essential, especially for professionals dealing with extensive data, research papers, or business reports. However, traditional methods of manual document sorting, keyword searches, or scanning for specific content are not only time-consuming but also prone to inaccuracies. This project addresses these limitations by building an intelligent information retrieval system that leverages the capabilities of state-of-the-art transformer models.

The goal was to create a system that could automatically classify documents into predefined categories, streamlining the process of organizing and managing text-based data. By using DistilBERT, the project aims to provide users with a high level of accuracy in classifying content, thus minimizing human effort in document categorization. Additionally, the project introduces an enhanced search feature that allows users to locate specific content within documents through a custom search engine that goes beyond simple keyword matching. This search functionality is further augmented by typo correction, ensuring users can retrieve information even when minor errors are present in the input query.

Furthermore, the integration of a question-answering module is designed to assist users in extracting precise information from documents based on natural language questions. This feature is especially useful in scenarios where quick answers are needed from lengthy reports or text-heavy documents, enabling more informed and efficient decision-making. By combining these elements, the project not only enhances the accessibility of information but also improves the overall user experience by making document management more interactive, efficient, and user-friendly.

4.Proposed Model

The proposed model for this project is a comprehensive information retrieval and document classification system powered by the DistilBERT transformer model. This system is designed to classify documents into predefined categories, provide a semantic search functionality, and support question answering based on document content.

The core components of the proposed model include:

1. **Document Preprocessing**: Before feeding text data into the model, it undergoes preprocessing to improve consistency and input quality. This includes tokenization, stopword removal, and lemmatization.
2. **Text Classification with DistilBERT**: The system leverages a fine-tuned DistilBERT model for document classification. The model is trained on labeled data to categorize documents into predefined classes like "World," "Sports," "Business," and "Sci/Tech." This enables automatic categorization of newly uploaded documents based on their content.



1. **Custom Information Retrieval**: A custom search function is implemented using a TF-IDF (Term Frequency-Inverse Document Frequency) model. The user can input a query, and the system will retrieve the most relevant sentences from the documents using cosine similarity, providing a focused search experience.
2. **Question Answering with DistilBERT**: The system integrates a question-answering module, utilizing a fine-tuned version of DistilBERT trained on extractive question answering tasks. When a user submits a question, the system identifies relevant sections within the document and extracts the most probable answer.
3. **Interactive Web Interface**: The model is deployed as an interactive web application using Streamlit, where users can upload documents, classify them, perform searches, and ask questions—all via an intuitive interface.

This model aims to provide an efficient, easy-to-use solution for document classification and semantic search, integrating state-of-the-art transformer models for a streamlined user experience.

5.Methodology

The project was structured into two main phases:

1. **Model Training**: Developing a robust classifier using various machine learning models, eventually leading to the selection of a transformer-based model for optimal performance.
2. **Web Application Development**: Building an interactive platform using Streamlit to utilize the trained model for document classification, search, and question-answering functionalities.

**5.1. Phase 1: Model Training**

In this phase, the focus was on creating a highly accurate document classification model. The process involved selecting and fine-tuning various machine learning models and implementing the best one for the classification task.

**Preprocessing in Model Training**:  
To prepare the dataset for model training, extensive preprocessing was performed:

* **Tokenization**: The text was split into smaller components such as words or subwords to streamline the analysis.
* **Stopword Removal**: Common, non-informative words were filtered out to reduce noise and improve the model’s focus on meaningful content.
* **Lemmatization**: Words were reduced to their base or root form (e.g., converting "running" to "run") to standardize variations.

After preprocessing, for the DistilBERT model, an additional step of generating **embeddings** was performed. This conversion into dense vector representations allowed the model to understand contextual relationships within the text. This was unique to DistilBERT and not applied to other initial models.

**Models Evaluated**:  
Initially, several machine learning algorithms were explored:

* A **Custom Neural Network**: Designed to classify text, but it lacked the contextual depth needed for more nuanced categories.
* **Naive Bayes**: A simple, efficient model but limited in handling complex text relationships.
* **Random Forest**: An ensemble learning method that showed better accuracy but was still outperformed by transformer-based models.

Ultimately, the **DistilBERT transformer model** was selected after fine-tuning on the preprocessed dataset. The model demonstrated superior contextual understanding and classification accuracy compared to traditional models.

**5.2. Phase 2: Web Application Development**

Once the DistilBERT model was trained and evaluated, the next phase was to deploy it within a user-friendly application to make its capabilities accessible.

**Preprocessing in Web Application**:  
While similar preprocessing steps (like tokenization, stopword removal, and lemmatization) were applied, this stage focused on handling text from **uploaded documents** (PDFs or DOCX files). Unlike the model training phase, embeddings were not generated here since the pre-trained DistilBERT model was directly used for predictions.

**Implementation Details**

The web application was developed using Streamlit and included the following features:

1. **Document Classification**: Users can upload documents, which are preprocessed and classified into categories (e.g., "World," "Sports," "Business," "Sci/Tech") using the fine-tuned DistilBERT model. This provides users with immediate insights into the content of their files.
2. **Search Functionality and Information Retrieval**: A custom search engine was integrated, utilizing **TF-IDF vectorization** and **cosine similarity**. Users can search for specific terms within the uploaded documents. The search system was optimized to handle typos and deliver accurate results by highlighting relevant sections.
3. **Question-Answering Module**: Leveraging the capabilities of DistilBERT, the application allows users to input questions related to the content of the uploaded documents. The system extracts precise answers based on the context of the query, making it highly effective for information retrieval tasks.
4. Experiments and Results

**6.1. Datasets**

For this project, a news dataset was used to train and evaluate the DistilBERT model for document classification and semantic search. The dataset contains multiple news articles, each labeled with a specific category or class. Below is an overview of the dataset structure:

* **Class Index**: Each article is assigned a class index, representing the category of the article. For example, articles about business and finance are labeled under a specific index, while articles about politics or technology are given another.
* **Title**: The title of each article serves as the document identifier. Titles typically offer a concise description of the article's content.
* **Description**: The description field provides a brief summary or excerpt from the article, which serves as the primary content for classification and search tasks.

**6.2. Evaluation Metrics**

To benchmark the models, the following metrics were used:

* **Accuracy**: Measures the proportion of correctly predicted instances among the total instances.
* **Precision**: Indicates the accuracy of positive predictions, i.e., the proportion of true positive predictions among all positive predictions.
* **Recall (Sensitivity)**: Reflects the ability of the model to identify all relevant instances, calculated as the proportion of true positives among all actual positives.
* **F1-Score**: The harmonic mean of precision and recall, providing a single metric that balances both.

**6.3. Performance of Traditional Machine Learning Models**

Three traditional models were initially explored for classification: a **Neural Network**, **Naive Bayes**, and **Random Forest**.

* **Neural Network**:
  + **Accuracy**: 87.89%
  + Classification report

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Precision | Recall | F1-score |
| Class - 1 | 0.90 | 0.88 | 0.89 |
| Class - 2 | 0.95 | 0.93 | 0.95 |
| Class - 3 | 0.84 | 0.86 | 0.84 |
| Class - 4 | 0.81 | 0.87 | 0.85 |

* + Demonstrated decent precision and recall but was less effective in distinguishing between closely related categories.
* **Naive Bayes**:
  + **Accuracy**: 88.49%
  + Classification report

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Precision | Recall | F1-score |
| Class - 1 | 0.90 | 0.88 | 0.89 |
| Class - 2 | 0.94 | 0.96 | 0.95 |
| Class - 3 | 0.86 | 0.83 | 0.84 |
| Class - 4 | 0.84 | 0.87 | 0.85 |

* + Achieved high precision but lower recall, indicating that while it was good at making positive predictions, it missed out on several actual positives.
* **Random Forest**:
  + **Accuracy**: 87.70%
  + Classification report

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Precision | Recall | F1-score |
| Class - 1 | 0.90 | 0.88 | 0.89 |
| Class - 2 | 0.91 | 0.96 | 0.93 |
| Class - 3 | 0.85 | 0.83 | 0.84 |
| Class - 4 | 0.85 | 0.84 | 0.84 |

* + Showed balanced precision and recall but was computationally intensive, making it less practical for real-time applications.

Despite achieving satisfactory accuracy scores, these models exhibited limitations in handling complex text data, leading to lower F1-scores, especially in edge cases.

**6.4. Performance of Transformer-Based Model (DistilBERT)**

To address the shortcomings of traditional models, a **DistilBERT transformer model** was implemented. The model was trained using preprocessed data, where tokenization, stopword removal, and lemmatization were applied. In addition, embeddings were generated during the training phase to capture deeper semantic relationships.

* **DistilBERT Model**:
  + **Accuracy**: 93%
  + Classification report

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Precision | Recall | F1-score |
| Class - 1 | 0.95 | 0.94 | 0.94 |
| Class - 2 | 0.98 | 0.98 | 0.98 |
| Class - 3 | 0.92 | 0.88 | 0.90 |
| Class - 4 | 0.89 | 0.93 | 0.91 |

* + **Precision**: Higher compared to traditional models.
  + **Recall**: Significantly improved due to the model's contextual understanding.
  + **F1-Score**: Notably higher, demonstrating its ability to handle nuanced text categories effectively.

The results highlight that the DistilBERT model outperformed traditional methods in every metric, demonstrating superior accuracy and generalization capabilities, especially for complex text inputs.

**6.5. Classification Report Summary**

Below is a summary of the performance metrics for each model. The detailed classification tables include **precision, recall, F1-score**, and **support** for each category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| |  | | --- | | Neural Network | | |  | | --- | | 88% | | |  | | --- | | 0.90 | | |  | | --- | | 0.88 | | |  | | --- | | 0.87 | |
| Naive Bayes | 88% | 0.90 | 0.89 | 0.85 |
| Random Forest | 88% | 0.90 | 0.90 | 0.86 |
| DistilBERT | 94% | 0.94 | 0.94 | 0.94 |

The transformer-based approach (DistilBERT) was ultimately chosen for integration into the web application due to its superior performance.

1. Conclusion and Limitations

**7.1. Conclusion:**

The project successfully achieved its objective of building an efficient and interactive document classification and search system by leveraging the DistilBERT transformer model. Through a detailed evaluation, it was demonstrated that DistilBERT outperformed traditional machine learning models like Neural Networks, Naive Bayes, and Random Forest in terms of accuracy, precision, and F1-score. This model's superior understanding of contextual semantics made it the optimal choice for integrating into a real-time web application using Streamlit. The final application can categorize uploaded documents, retrieving relevant information through a custom search function, and supporting a question-answering module to interact with users' queries. These functionalities aim to streamline document analysis, providing users with an intuitive interface for efficient content exploration.

**7.2. Limitations:**

While the project has achieved its core goals, there are several limitations that highlight areas for future improvement:

1. **Lack of Multilingual Support**:
   * Currently, the system only supports English text. Extending support to multiple languages would greatly enhance its usability across diverse user bases and content types. Future iterations could leverage multilingual transformer models like XLM-R or mBERT for better global applicability.
2. **Search and Information Retrieval Enhancements**:
   * The existing search mechanism relies on TF-IDF vectorization combined with cosine similarity, which may not be as robust for long, complex queries. Incorporating neural search methods, such as dense retrieval using sentence embeddings or semantic search with models like SBERT, could improve search accuracy and relevance.
3. **Improvement in Question-Answering Module**:
   * The current QA module uses DistilBERT, which, while effective, is limited in handling complex or ambiguous questions. More advanced models like GPT-based architectures or fine-tuning a model specifically for extractive question answering could enhance the system's ability to generate precise answers.
4. **Scalability and Performance**:
   * As the document corpus increases, the current preprocessing and search pipeline may become slow, especially with larger text files. Implementing more efficient indexing techniques (e.g., FAISS or ElasticSearch) could help improve scalability.
5. References

[1] Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. "Large language models for information retrieval: A survey." *arXiv preprint arXiv:2308.07107* (2023).

[2] Abbasiantaeb, Zahra, and Saeedeh Momtazi. "Text‐based question answering from information retrieval and deep neural network perspectives: A survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, no. 6 (2021): e1412.

[3] Sanh, V. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

[4] Wu, Ridong, Shuhong Chen, Xiangbiao Su, Yuankai Zhu, Yifei Liao, and Jianming Wu. "A Multi-Source Retrieval Question Answering Framework Based on RAG." *arXiv preprint arXiv:2405.19207* (2024).

[5] Jiang, Zhuolin, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. "Cross-lingual information retrieval with BERT." *arXiv preprint arXiv:2004.13005* (2020).