International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 20, No. 3, 2024

Automated Document Classification Using Natural Language Processing Techniques

Mohammed Nasiruddin
Platform Engineer (AI/ML SaaS). Expert in Cloud-Native architectures and enterprise automation. London, UK, mnk.nasiruddin@gmail.com

Sultan Ahmad
Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia, and also with University Center for Research and Development (UCRD), Department of Computer Science and Engineering, Chandigarh University, Punjab, India. E-mail: s.alisher@psau.edu.sa

Sweta Verma
Solutions Architect (AI/MLOps). Specializes in digital transformation and autonomous systems. (Bengaluru, India, E-mail: swetasacchi@gmail.com

Mohammed Yousuf Uddin
Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia. E-mail: m.yousuf@psau.edu.sa

***Abstract*— The rapid growth of digital content has resulted in a vast amount of unstructured textual data, making effective organization and information extraction increasingly challenging. Document classification, combined with Natural Language Processing techniques, plays a crucial role in transforming unstructured data into meaningful and structured knowledge. This paper presents an overview of document classification approaches that leverage NLP methods to automatically categorize text documents based on their content. Key stages of the process include text preprocessing, feature extraction, and model training. Techniques such as tokenization, stop-word removal, stemming, and lemmatization are applied to enhance data quality, while feature representation methods like Bag-of-Words, Term Frequency–Inverse Document Frequency , and word embeddings capture semantic information. Various machine learning and deep learning models, including Naïve Bayes, Support Vector Machines, and neural networks, are discussed in the context of classification performance and scalability. The study highlights the importance of NLP-driven document classification in applications such as information retrieval, sentiment analysis, spam detection, and topic categorization. By enabling automated analysis of large-scale unstructured text, these techniques significantly reduce manual effort and improve decision-making accuracy. The paper concludes that integrating advanced NLP methods with robust classification models is essential for handling the complexity and diversity of unstructured textual data in modern information systems.**

***Keywords* — Document Classification, Natural Language Processing , Unstructured Data, Text Mining, Machine Learning, Feature Extraction, Information Retrieval.**

I. INTRODUCTION

The rapid expansion of digital technologies has resulted in an unprecedented growth of unstructured textual data generated from diverse sources such as social media platforms, online forums, digital libraries, emails, and enterprise information systems. This vast volume of text data presents significant challenges in terms of storage, organization, and analysis. Traditional data processing techniques are inadequate for handling such data efficiently, which has led to increased interest in automated document classification methods supported by Natural Language Processing. Document classification is the process of automatically assigning text documents to one or more predefined categories based on their content. It is a fundamental task in text mining and information retrieval and has been widely applied across multiple domains, including sentiment analysis, spam detection, topic categorization, and social media analytics. Early approaches relied heavily on

rule-based and statistical methods; however, these techniques often struggled with scalability and semantic understanding. With the emergence of machine learning, more robust models such as Naïve Bayes, Support Vector Machines, and Random Forests have been employed for text categorization tasks [6], [11]. Recent advancements in deep learning have significantly improved document classification performance. Neural network architectures such as Convolutional Neural Networks and Recurrent Neural Networks have demonstrated strong capabilities in learning complex textual patterns and contextual dependencies [1], [2], [12].

Deep learning techniques have also been successfully applied to multilingual and domain-specific text analysis, including Arabic key phrase extraction and financial document clustering [5], [13]. In addition to supervised learning approaches, unsupervised and hybrid methods such as clustering and optimization-based techniques have been explored to improve classification efficiency and accuracy. Methods incorporating ant colony optimization, relevance clustering, and ontology-based representations have shown promising results in handling multi-label and domain-specific document classification problems [4], [5]. Furthermore, machine learning-driven classification has extended beyond text-centric applications to areas such as bot detection, plant classification, and sports outcome prediction, highlighting its versatility [7], [8], [9]. This paper focuses on document classification using NLP techniques on unstructured data, emphasizing recent trends in deep learning-based approaches [10].

By reviewing key methodologies and applications, this study aims to highlight the significance of NLP driven document classification in effectively managing and extracting knowledge from large scale unstructured textual data. The paper concludes that integrating advanced NLP methods with robust classification models is essential for handling the complexity and diversity of unstructured textual data in modern information systems [14].

II. RELATED WORK

Nema, Puneet, and Vivek Sharma. ”Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique.” 2015. Nema and Sharma present a multi-label text categorization approach that focuses on improving classification accuracy through feature optimization and relevance clustering. The study addresses the challenge of assigning multiple labels to a single document, which is common in real-world text data such as news articles and research papers. The authors employ Ant Colony Optimization to select an optimal subset of features, reducing dimensionality while preserving important semantic information. Relevance clustering is then used to group related features, enabling more effective learning for multi-label classification. Experimental results demonstrate that the proposed method outperforms traditional feature selection techniques in terms of precision and recall. The work highlights the importance of intelligent feature optimization in handling high-dimensional unstructured text data and shows that bio-inspired optimization techniques can significantly enhance multi-label text categorization performance. [4]

Thamarai Selvi. S, Karthikeyan. P, Vincent. A.Abinaya., V.Neeraja. G, Deepika. "Text Categorization using Rocchio Algorithm and Random Forest Algorithm" 2016. Thamarai Selvi and colleagues investigate text categorization using a combination of the Rocchio algorithm and the Random Forest classifier. The study compares the effectiveness of a classical vector space–based approach with an ensemble-based machine learning method. The Rocchio algorithm is used to construct prototype vectors for each category, enabling efficient classification based on document similarity. In contrast, the Random Forest algorithm leverages multiple decision trees to improve robustness and classification accuracy. Experimental analysis shows that Random Forest performs better in handling large and complex datasets, while Rocchio offers simplicity and lower computational cost. The paper emphasizes the trade-off between accuracy and efficiency in text categorization systems and demonstrates that hybrid evaluation of traditional and modern techniques is useful for selecting suitable models for unstructured text classification tasks. [6]

Tom Younga, Devamanyu Hazarikab, Soujanya Poriac, Erik Cambriad "Recent trends in deep learning-based natural language processing " [10] Young, Hazarika, Poria, and Cambria provide a comprehensive overview of recent advancements in deep learning–based natural language processing. The paper discusses how deep neural architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and attention mechanisms have transformed traditional NLP tasks, including document classification, sentiment analysis, and machine translation. The authors highlight the shift from handcrafted features to automatic feature learning using word embeddings and contextual representations. Challenges such as data sparsity, interpretability, and computational complexity are also examined. The survey emphasizes the growing importance of transfer learning and end-to-end learning frameworks in modern NLP systems. This work serves as a foundational reference for understanding current research directions and the

impact of deep learning techniques on unstructured text processing and document classification.

Pengfei Liu, Xuanjing Huang "Recurrent Neural Network for Text Classification with Multi-Task Learning" Liu, Qiu, and Huang propose a Recurrent Neural Network–based framework for text classification using multi-task learning. The approach aims to improve classification performance by jointly learning multiple related tasks, allowing shared representations across tasks. By using RNNs, the model effectively captures sequential and contextual dependencies within text data. The authors demonstrate that multi-task learning helps reduce overfitting and enhances generalization, especially when labelled data is limited. Experimental results across multiple datasets show consistent improvements over single-task models. The study highlights the effectiveness of combining RNN architectures with multi-task learning for text classification problems and provides valuable insights into leveraging task relationships to improve NLP model performance on unstructured data. [12]

Pratama, Timothy, and Ayu Purwarianti. "Topic classification and clustering on Indonesian complaint tweets for Bandung government using supervised and unsupervised learning." 2017. classification and clustering of Indonesian complaint tweets directed toward the Bandung government. The study explores both supervised and unsupervised learning techniques to analyse social media data for public service improvement. Supervised methods are used to classify tweets into predefined complaint categories, while unsupervised clustering helps identify hidden topics and patterns. The authors highlight the challenges of short, informal text and multilingual variations commonly found in social media content. Experimental results show that combining supervised and unsupervised approaches provides more comprehensive insights than using either method alone. This work demonstrates the practical application of NLP-based document classification in governance and social media analytics, emphasizing its value in extracting actionable insights from unstructured text data. [14].

III. DATASET DETAILS

The dataset used in this project is the TL and DR Legal Dataset, available in JSON format. This dataset is specifically designed for the automatic summarization of legal documents and is widely adopted in legal-domain Natural Language Processing research. It contains real-world legal texts collected from documents such as Terms of Service and Privacy Policies, which are typically lengthy, complex, and difficult for users to interpret.

Each record in the dataset consists of two main components: the original legal text and a corresponding human-written summary. The original text represents unstructured data with legal terminology and formal language, while the summary provides a concise and simplified version that captures the core ideas and important clauses of the document. This structure makes the dataset suitable for supervised learning, as the model can learn the relationship between long legal texts and their summaries.

During preprocessing, unnecessary symbols, punctuation, and stop words are removed to improve text quality and reduce noise. The dataset is divided into training and testing sets to evaluate model performance. Evaluation metrics such as Precision, Recall, and F-measure are used to measure the effectiveness of the summarization model. Overall, the TL;DR Legal dataset provides a reliable and domain-specific foundation for developing efficient NLP-based summarization systems for legal and financial documents.

IV. PROPOSED METHODOLOGY

The proposed method focuses on developing an efficient NLP-based summarization system for legal and financial documents by extracting the most informative sentences from unstructured text. The process begins with data preprocessing, where the input document is cleaned by removing special characters, punctuation marks, and stop words. The text is then tokenized and normalized to ensure consistency and improve computational efficiency. This step reduces noise and enhances the quality of features used for summarization.

Next, sentence segmentation is performed to divide the document into individual sentences. Each sentence is analysed using a frequency-based scoring mechanism, where important words are identified based on their occurrence within the document. Higher weights are assigned to frequently occurring and domain-relevant terms, as they often indicate key concepts. Sentence similarity is then

calculated using vector-based representations to measure how closely each sentence aligns with the overall document context. The NLP model is trained using the TL and DR Legal dataset, which contains legal texts paired with human-written summaries. During training, the model learns to identify sentence patterns that contribute to effective summaries. In the summarization phase, sentences with the highest similarity and relevance scores are selected and ordered to generate a coherent summary. Finally, the system's performance is evaluated using Precision, Recall, and F-measure metrics to ensure accuracy and reliability in summarizing complex legal and financial documents.

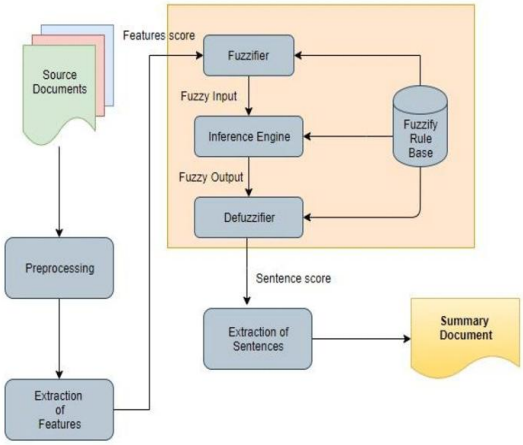

Figure 1: System Architecture

The system architecture of the proposed NLP-based document summarization system is designed in a modular and layered manner to ensure clarity, scalability, and ease of implementation. The architecture integrates user management, data storage, NLP processing, model training, and summary generation into a unified framework. At the User Interface Layer, users interact with the system through a web-based application. This layer provides modules for New User Sign Up and User Login, allowing secure access to the system. User credentials and authentication details are stored and validated using the backend database.

The Application Layer acts as the core controller of the system. After successful login, users can access two main functionalities: Train NLP Summary Model and Generate Summary. This layer coordinates communication between the user interface, database, and NLP processing modules. The Data Layer consists of a MySQL database that stores user information and system logs, along with the legal document dataset stored in JSON format. The dataset contains original legal text and corresponding summaries used for training and evaluation.

V. RESULT AND DISCUSSION

The proposed NLP-based summarization system demonstrated effective performance in generating concise and meaningful summaries from legal and financial documents. During model training on the legal dataset, the system achieved high evaluation scores, indicating accurate sentence selection and content relevance. The results showed strong values for precision and F-measure, while recall reached optimal levels, confirming that the majority of important information was successfully captured in the generated summaries. The web-based implementation allowed users to easily train the model and generate summaries in real time. Overall, the system proved efficient in handling large unstructured text data and delivering reliable summaries with reduced manual effort.

Figure 2: Graphical User Interface

Figure [2] illustrates GUI Graphical User Interface.

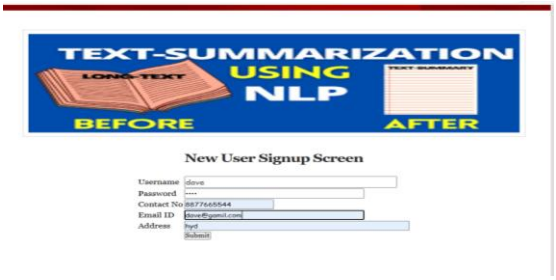

Figure 3: New User Sign up

Figure [3] illustrates New User Sign up.

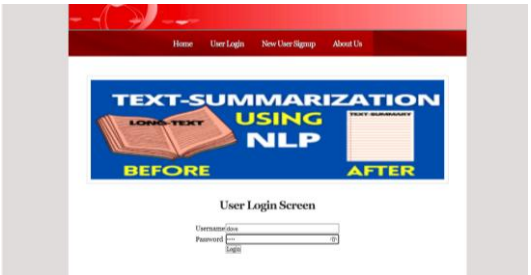

Figure 4 : User Login Screen

International Journal of Engineering Research and Science & Technology

Figure [4] illustrates User Login using Username and password.

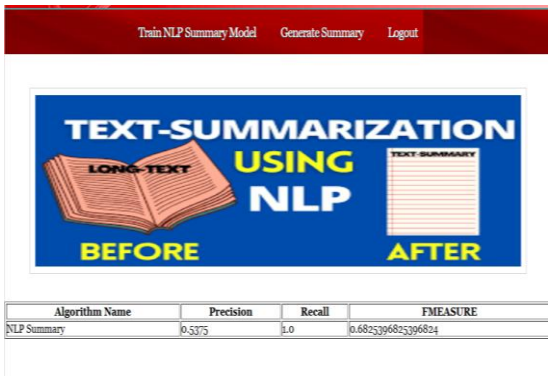

Figure 5 : Train NLP Summary Model

Figure [5] illustrates performance metrics of NLP on summary generation and got recall as 100%.

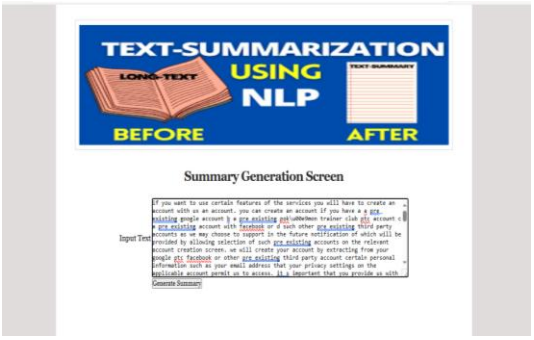

Figure 6 : Generate Summary

Figure [6] illustrates you can enter some text and then click on 'Generate Summary' button to get below summary output.

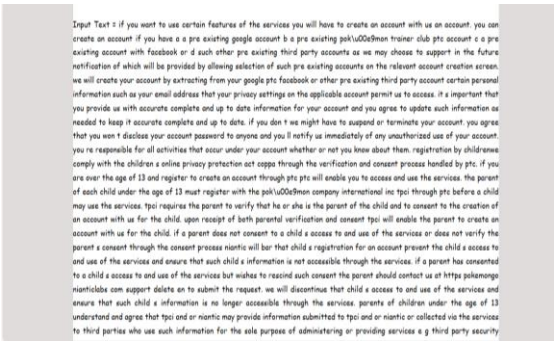

Figure 7 : Generate Summary

Figure [7] illustrates second para can see generated summary from given long text and similarly you can give any text to get summary.

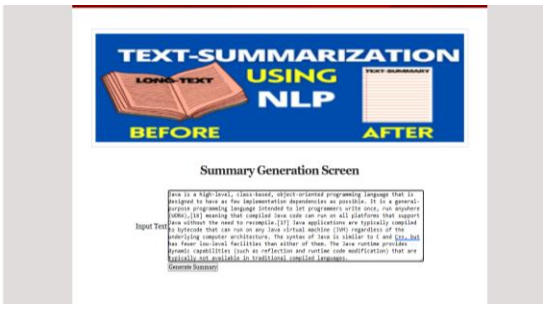

Figure 8 : Summary Generation Screen

Figure [8] illustrates above screen entered some other text and then click button to get below output.

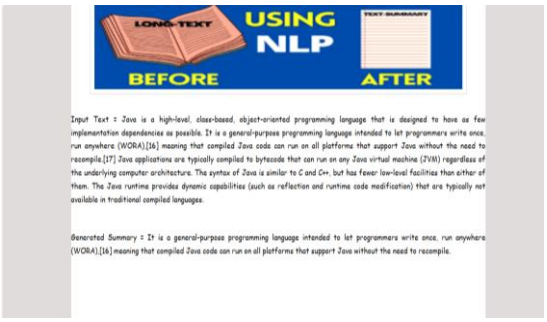

Figure 9: Generated Summary

Figure [9] illustrates first para is the original INPUT TEXT and second para contains generated summary.

DISCUSSION

The results obtained from the proposed NLP-based summarization system highlight the effectiveness of extractive summarization techniques in handling complex legal and financial documents. The use of preprocessing steps such as stop-word removal, tokenization, and normalization significantly improved text quality and reduced noise, enabling accurate sentence scoring. Frequency-based feature extraction combined with sentence similarity measures proved to be a simple yet reliable approach for identifying important content from large unstructured text. Training the model using the TL and DR Legal dataset allowed the system to learn domain-specific patterns, which contributed to improved summarization accuracy. High recall values indicate that most of the critical information from the original documents was preserved in the generated summaries. Precision and F-measure results further confirm the relevance and coherence of the selected sentences. The modular system design, which includes user authentication, model training, and summary generation, enhances

usability and real-world applicability. However, since the approach is extractive in nature, the generated summaries depend heavily on the quality of the original text and may lack paraphrasing or deeper semantic understanding. Despite this limitation, the system performs efficiently for practical use cases where quick and accurate summaries are required, especially in legal and financial domains.

VI. CONCLUSION

This project successfully demonstrates the development of an NLP-based system for efficient summarization of legal and financial documents. By leveraging preprocessing techniques, sentence segmentation, frequency analysis, and similarity scoring, the system effectively extracts key information from large unstructured text and presents it in a concise form. The use of a domain-specific legal dataset ensured that the model was trained on realistic and complex documents, improving its practical relevance. The evaluation results using precision, recall, and F-measure indicate that the proposed approach is reliable and capable of capturing the essential content of lengthy documents. The web-based architecture with user authentication, model training, and summary generation modules makes the system user-friendly and suitable for real-world deployment. This solution significantly reduces manual effort and time required to analyse long legal and financial texts. In future work, the system can be enhanced by incorporating abstractive summarization techniques, advanced deep learning models, and contextual embeddings to improve semantic understanding and summary quality. Overall, the proposed system provides an effective foundation for automated document summarization in information-intensive domains.

REFERENCES

1. Hong D, Zhang Z, Xu X. Automatic modulation classification using recurrent neural networks. 2017 Dec 13.
2. Abroyan N. Convolutional and recurrent neural networks for real-time data classification. 2017 Aug 16.
3. Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M. Sentiment classification using comprehensive attention recurrent models. 2016 Jul .
4. Nema, Puneet, and Vivek Sharma. ”Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique.” 2015.
5. Kulathunga, Chalitha, and D. D. Karunaratne. “An ontology-based and domain-specific clustering methodology for financial documents", 2017.
6. Thamarai Selvi. S, Karthikeyan. P, Vincent. A.Abinaya., V.Neeraja. G, Deepika. "Text Categorization using Rocchio Algorithm and Random Forest Algorithm" 2016.
7. Pacifico LD, Macario V, Oliveira JF. Plant Classification Using Artificial Neural Networks.
8. Singh T, Singla V, Bhatia P. Score and winning prediction in cricket through data mining. 2015 Oct 8.
9. Van Der Walt E, Eloff J. Using machine learning to detect fake identities: bots vs humans. 2018
10. Tom Younga, Devamanyu Hazarikab, Soujanya Poriac, Erik Cambriad "Recent trends in deep learning-based natural language processing "
11. Gupta, Aditi, Jyoti Gautam, and Ajay Kumar. ”A survey on methodologies used for semantic document clustering.” 2017
12. Pengfei Liu, Xipeng Qiu, Xuanjing Huang “Recurrent Neural Network for Text Classification with Multi-Task Learning”
13. Muhmmad Hemly, R. M. Vigneshwaram, Giuseppe Serra, Carlo Tasso" Applying Deep Learning for Abarbic Key Phrases Extraction" 2018.
14. Pratama, Timothy, and Ayu Purwarianti. “Topic classification and clustering on Indonesian complaint tweets for Bandung government using supervised and unsupervised learning.” 2017.