

Machine Learning-Driven Document Classification and Automated File Organization Using Text Analysis

Surabhi C D¹, Suraksha A S², Utkarsh Kashyap³, Yashaswini C H⁴, Mr. Madhu C K⁵

¹(4MC23CS164) CSE & Malnad College of Engineering (Hassan)

²(4MC23CS165) CSE & Malnad College of Engineering (Hassan)

³(4MC23CS178) CSE & Malnad College of Engineering (Hassan)

⁴(4MC23CS186) CSE & Malnad College of Engineering (Hassan)

⁵Assistant Professor, CSE & Malnad College of Engineering (Hassan)

Abstract - The issue of digital document management is a growing challenge which we see in that many organizations are producing large sets of reports, invoices, records, and academic papers at a constant rate. Traditional methods which include the use of folder structures and manual organization are becoming inefficient as document volume increases. This paper presents a Machine Learning-driven framework for automated document classification and file organization using Natural Language Processing techniques.

Key Words: Machine Learning, Document Classification, TF-IDF, Text Mining, Natural Language Processing, Automated File Organization, Support Vector Machine.

1. INTRODUCTION

In the few years we have seen a lot of digital information in schools, hospitals, companies and other businesses. We are always adding reports, invoices, legal papers, school work and research papers to our files. As these files get bigger and more complicated it becomes hard to manage them by hand.

The old way of managing documents is to use folders and file names.. This way does not really understand what the documents mean or what is inside them so it is hard to find and organize them.

Machine Learning and Natural Language Processing are ways to automatically sort and organize documents. We can make the text better by removing words breaking it down into small parts and simplifying it.

We can also turn text into numbers that computers can understand using things like TF-IDF and N-gram analysis. Then we can use Machine Learning to categorize documents with things, like Support Vector Machine, Logistic Regression and Naïve Bayes. Machine Learning is used to make this process easier. We use Natural Language Processing and Machine Learning to make our digital files more organized.

2. LITERATURE SURVEY

The field of document management has gotten a lot of attention because there are so many more digital documents now. This is happening in organizations and schools and hospitals and big companies. People who do research have tried a lot of Machine Learning and Natural Language Processing techniques. They want to make it better for computers to understand what documents are about and to help people find documents they need and to make systems that can organize documents on their

own. Intelligent document management is really important. It can help with things, like text classification and document retrieval and automated organization systems. Intelligent document management systems are getting better because of Machine Learning and Natural Language Processing techniques.

1. Machine Learning for Text Classification

Machine Learning algorithms play an important role in document classification systems. Sayar Ul Hassan et al. [1] conducted a comparative study on text classification models such as Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, Random Forest, and K-Nearest Neighbor. Their findings showed that supervised learning algorithms, particularly SVM and Logistic Regression, achieved better classification performance when combined with TF-IDF vectorization.

Similarly, Xiaoyu Luo [2] proposed an efficient English text classification framework and observed that SVM achieved higher precision, recall, and F1-score compared to traditional approaches. The study highlighted the suitability of SVM for handling textual datasets.

2. Importance of Text Preprocessing Techniques

Text preprocessing plays a major role in improving classification accuracy. Ikonomakis et al. [3] emphasized that preprocessing techniques such as tokenization, stop-word removal, punctuation filtering, stemming, and lemmatization significantly improve textual quality and reduce unnecessary noise in datasets.

Agarwal and Mittal [4] further explained that supervised classification techniques perform better when feature engineering and preprocessing methods are properly applied. Their study also highlighted that poor preprocessing may reduce the effectiveness of Machine Learning algorithms.

3. Feature Extraction and Contextual Understanding

Feature extraction methods convert textual information into numerical representations suitable for Machine Learning models. Researchers have explored techniques such as TF-IDF vectorization, Bag-of-Words representation, cosine similarity, and N-Gram analysis to improve contextual understanding of text documents.

Studies by Ali et al. [5] and Abbas et al. [6] demonstrated that TF-IDF and metadata-based approaches significantly improve document categorization accuracy. Their research also highlighted the importance of supporting multiple file formats such as PDF, DOCX, and TXT documents.

4. Intelligent File Organization Systems

Recent studies have moved beyond simple text classification towards intelligent file management systems. Abbas et al. [6] proposed automated file labeling techniques for heterogeneous file repositories using Machine Learning. Their system

improved retrieval efficiency by automatically assigning labels and categories to files.

Ahmed et al. [8] also proposed an accurate document classification model using Machine Learning methods and demonstrated improved prediction reliability through feature optimization and confidence-based analysis.

5. Challenges in Existing Systems

Although existing systems show promising results, several limitations still exist. Many systems mainly focus on classification accuracy while lacking automated file organization capabilities. Scalability issues, dependency on training data quality, and support for heterogeneous file formats remain challenging problems in current frameworks [9][10].

Researchers have also identified computational complexity and real-time document handling as important limitations, particularly for large-scale repositories [11][12].

Table 1: Literature Survey of Document Classification Systems

Paper Title	Methods	File Types	Accuracy
Analytics of ML Algorithms for Text Classification	SVM, RF, NB, LR	TXT	90–97%
Efficient English Text Classification	SVM, TF-IDF	TXT	High Precision
Intelligent File Classification System	SVM, Ensemble ML	PDF, DOCX, TXT	96.4%
Hybrid Text Classification Approach	Hybrid ML, TF-IDF	TXT	92–95%
Automated File Labeling Using ML	SVM, Meta-data Analysis	PDF, PPTX, TXT	94–96%
Document Classification Using ML	NB, KNN, Neural Network	TXT	92%

3. OBSERVATIONS AND INSIGHTS FROM THE LITERATURE

After reviewing various research works related to document classification, text mining, and intelligent file organization, several important observations were identified. Existing Machine Learning-based document classification systems have shown promising results in improving text categorization, but certain limitations still remain in practical implementation.

1. Limitations in Existing Document Management Systems

Document management systems usually work with folders. File names to keep things organized. This way of doing things is easy to use. It does not really understand what the documents are about. It just looks at the names. Where they are. This can make it hard to find what you need and keep your documents in order especially when you have a lot of them.

There is another problem with these systems. When you use them to sort your documents they do not always put them away in the place. Many people have tried to make these systems better at sorting documents using Machine Learning models. Not many of these systems can actually put the documents in the right folders after they have been sorted. Document management systems have this problem. Document management

systems need to be better, at organizing documents.

2. Challenges in Text Classification

Research studies highlighted that text preprocessing plays a critical role in classification performance. Poor-quality text, unwanted symbols, punctuation noise, and inconsistent document formatting can negatively affect prediction results.

Documents stored in formats such as PDF, DOCX, and TXT often differ in structure and extraction quality. During literature review, it was observed that PDF files occasionally produce noisy outputs due to formatting irregularities, whereas DOCX files generally provide cleaner and more structured content for analysis.

Another challenge identified is dependency on training data quality. Machine Learning models tend to perform well only when datasets are properly labeled, balanced, and preprocessed. Insufficient or poor-quality datasets may reduce classification accuracy and prediction reliability.

3. Key Insights for the Proposed System

The literature survey indicates that combining Natural Language Processing (NLP) and Machine Learning (ML) can significantly improve document classification accuracy. Techniques such as TF-IDF vectorization and N-Gram analysis help convert textual information into meaningful numerical representations, improving contextual understanding of documents.

Research findings also suggest that supervised learning algorithms such as Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes consistently provide reliable performance in text categorization tasks. Their ability to handle textual datasets efficiently makes them suitable for intelligent document classification systems.

Based on these observations, the proposed framework is designed not only to classify documents accurately but also to automate file organization. The system integrates text preprocessing, feature extraction, classification, confidence score analysis, and category-wise folder management into a single workflow for improved efficiency.

4. Relevance to the Proposed Framework

The literature clearly highlights the growing need for an intelligent and automated document management system. By supporting multiple file formats and reducing manual effort, the proposed framework aims to improve document accessibility, minimize classification errors, and simplify large-scale file organization.

The integration of Machine Learning and NLP techniques also helps improve retrieval efficiency and makes document management more practical for organizations handling large volumes of digital files.

4. PROPOSED SYSTEM AND PRELIMINARY METHODOLOGY

The proposed system is designed to improve document classification and automate file organization using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The framework supports multiple document formats such as PDF, DOCX, and TXT, making it flexible for different textual datasets.

The process begins with document upload, where files are imported into the system. Since documents vary in format, extraction tools such as *PyPDF2* for PDF files and *python-docx* for DOCX files are used to convert content into readable text.

This allows the framework to process documents effectively regardless of format.

The review of existing research clearly shows that using Natural Language Processing and Machine Learning can greatly improve how well documents are classified.

Natural Language Processing and Machine Learning are really good at helping us classify documents.

We found that methods like TF-IDF and N-Gram analysis and machine learning algorithms such as SVM and Logistic Regression and Naïve Bayes consistently work well for categorizing text.

These methods are very useful for Natural Language Processing and Machine Learning.

These findings led to the design of a system that aims to not classify documents accurately but also organize files automatically.

This system brings together steps: preparing text extracting key features classifying documents analyzing confidence scores and managing folders by category all, in one process.

The system uses Natural Language Processing and Machine Learning to do all these things.

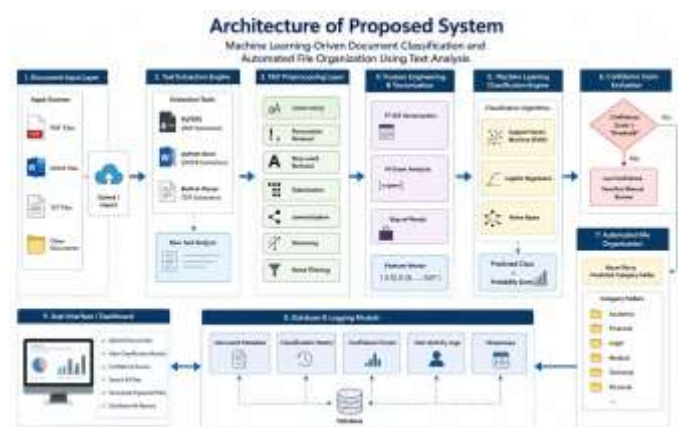


Figure 1: Architecture of Proposed System

Figure 1 illustrates the architecture of the proposed system. The framework includes document input, text extraction, NLP preprocessing, feature engineering, Machine Learning-based classification, confidence score evaluation, automated file organization, and database logging. This structured workflow improves document retrieval efficiency and minimizes manual effort in file management.

During preliminary testing, it was observed that DOCX documents generally produced cleaner outputs compared to PDF files, which occasionally introduced formatting inconsistencies and unwanted symbols. These findings highlight the importance of preprocessing and feature engineering in improving classification performance.

The framework also maintains metadata such as classification history, confidence scores, timestamps, and activity logs to support efficient document tracking and future retrieval.

5. EXPECTED OUTCOMES

The new system is going to make document management by automatically figuring out what is in a document and putting it in the right place. This means people will not have to spend

much time sorting documents by hand. The system can save a lot of time and effort this way.

The system uses Machine Learning and Natural Language Processing to get better at putting documents into the categories. It looks at the words in the documents to understand what they are about. The system uses things like TF-IDF and N-Gram analysis to get an idea of what is in the documents.

The system can also work with kinds of documents like PDF and DOCX and TXT files. This makes it useful for schools and companies and hospitals and government offices that have a lot of documents to manage.

The system is also going to make sure that documents are organized in a way. It does this by putting documents into folders based on what is in them. This reduces mistakes that people might make and makes it easier to find documents when you need them.

The system is also going to have a way to check if it's really sure about what it is doing. If it is not sure it will ask for help to make sure that documents are not put in the place. This means that documents will be organized correctly and will be easy to find when you need them. The document management system will be very helpful, for managing documents.

6. ADVANTAGES OF PROPOSED SYSTEM

The proposed system offers several advantages in intelligent document management:

- Reduces manual effort involved in document classification and organization.
- Improves document retrieval efficiency through category-wise storage.
- Supports multiple document formats such as PDF, DOCX, and TXT.
- Enhances classification accuracy using Machine Learning algorithms.
- Reduces human error during document handling and sorting.
- Provides scalable support for managing large document repositories.
- Improves semantic understanding using Natural Language Processing techniques.
- Enables confidence-based analysis for better prediction reliability.
- Can be extended for enterprise-level intelligent document management systems.

7. FUTURE ENHANCEMENT

The proposed framework is a solution for classifying and organizing documents.. There are some things that can be done to make it better in the future.

One thing that can be done is to use transformer-based Natural Language Processing models and deep learning approaches. Using models like BERT and neural network-based

classifiers can help the framework understand what documents are about and classify them accurately especially for complex documents.

Another thing that would be helpful is to add Optical Character Recognition (OCR) functionality. This means the framework can look at scanned and handwritten documents and make sense of them which would be really useful for real-world document repositories.

It would also be an idea to make the framework work with cloud-based storage. This way people can get to the documents from anywhere and from any device.

The framework should also be able to classify documents in different languages. A lot of organizations have to deal with documents, in languages so it would be really helpful if the framework could handle this.

The document classification framework may also be able to learn on its own. This means the model gets better over time because it is getting feedback from users and new examples to learn from. This would help the framework keep classifying documents over time.

REFERENCES

1. S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, 2022.
2. X. Luo, "Efficient English text classification using selected machine learning techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, 2021.
3. M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Transactions on Computers*, vol. 4, no. 8, pp. 966–974, 2005.
4. B. Agarwal and N. Mittal, "Text classification using machine learning methods: A survey," *International Conference on Soft Computing for Problem Solving*, pp. 701–709, 2016.
5. M. Ali et al., "Machine learning for text document classification-efficient classification approach," *Journal of Intelligent Systems*, vol. 14, no. 2, pp. 122–130, 2024.
6. S. Abbas et al., "Automated file labeling for heterogeneous files organization using machine learning," *Journal of Information Processing Systems*, vol. 18, no. 5, pp. 912–925, 2022.
7. K. Sethi, A. Gupta, G. Gupta, and V. Jaiswal, "Comparative analysis of machine learning algorithms on different datasets," *International Journal of Computer Applications*, vol. 162, no. 8, pp. 12–18, 2017.
8. A. Ahmed et al., "An accurate model for text document classification using machine learning techniques," *Expert Systems with Applications*, vol. 210, pp. 118–127, 2025.
9. T. Taha et al., "A comprehensive survey of text classification techniques and their research applications," *Artificial Intelligence Review*, vol. 57, no. 1, pp. 45–72, 2024.
10. A. Abatcha et al., "Document classification using machine learning," *International Journal of Advanced Computer Science*, vol. 11, no. 2, pp. 88–96, 2022.
11. R. Jindal, "Techniques for text classification: Literature review and current trends," *International Journal of Computer Science Trends and Technology*, vol. 3, no. 6, pp. 120–128, 2015.
12. A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 273–292, 2019.