International Journal of Innovation Studies

ISSN:2096-2487 | E-ISSN:2589-2975

A MACHINE LEARNING-BASED FRAMEWORK FOR AUTOMATED MULTI-VIEW DOCUMENT STRUCTURE CLASSIFICATION

T.Vani

Associate Professor, Rajeswari Vedachalam Government Arts College, Chengalpattu

email-id: vani.thangapandian@gmail.com

Abstract

In multi-view document classification, various machine learning approaches such as supervised, unsupervised, and semi-supervised different techniques have been applied in existing systems. To effectively categorize document objects, it is essential to first extract background knowledge and metadata from the documents. Different machine learning algorithms contribute distinct classification methods based on features such as short text content, metadata, and heading structures. Typically, an expert can determine whether a document follows a supervised, unsupervised, or semi-supervised learning approach by reading and analyzing its structure. However, this manual process can be time-consuming and prone to ambiguity. To address these challenges, we propose an IDS (Identifying Document Structure) model — a machine learning-based approach for automated identifying document structure classification and tries to categorize according to the document. In this model, keywords are trained using labeled data, while clustering techniques are employed to handle unsupervised data. A combination of both methods is used for semi-supervised classification. We split the dataset into 60% for training and 40% for testing, demonstrating improved classification performance and efficiency compared to existing techniques.

Keywords: supervised, unsupervised and semi supervised

1. Introduction

Document classification plays a vital role in managing and organizing large volumes of digital information by automatically categorizing documents into predefined classes. Traditional approaches rely on supervised, unsupervised, or semi-supervised machine learning techniques to analyze document features such as short text, metadata, and heading structures. While domain experts can often determine the classification type by manually examining the document, this process is time-consuming, prone to errors, and not scalable for large datasets. To address these limitations, we propose the IDS (Identifying Document Structure) model, a machine learning-based approach that automates document classification by combining keyword-based supervised learning with clustering methods for unsupervised data. The model effectively handles semi-supervised scenarios by integrating both approaches, improving accuracy and reducing manual effort. Our experiments use a data split of 60% for training and 40% for testing, showing promising performance improvements over traditional methods. The IDS model can be applied in various real-world scenarios. For example, in digital libraries, it can automatically organize academic research papers by subject area without manual tagging. In enterprise document management, it can classify invoices, contracts, and internal reports into categories such as "financial," "legal," or "technical" for easy retrieval. Web publishing

platforms can use the model to automatically categorize articles into topics like "technology," "health," or "finance." In medical applications, it can classify patient records based on disease type or treatment history, assisting doctors in quickly accessing relevant information. Similarly, in the legal domain, the model can help organize case files by case type (e.g., civil, criminal, or administrative). This automated approach reduces manual workload, enhances classification accuracy, and improves the overall efficiency of document processing systems across multiple domains.

2. Related works

The paper [1] presents a self-supervised learning method that improves remote sensing image classification by learning from unlabeled multimodal data, reducing the need for manual annotation. The paper [2] uses DCASL (Dual-view Cross Attention enhanced Semi-supervised Learning), a method that combines dual-view cross-attention and semi-supervised learning to improve cognitive engagement classification in online education, especially with limited labeled data. The paper [3] develops an artificial sensory system using sensors and neural networks to predict fish freshness with ~99% accuracy, enabling automated quality monitoring in the supply chain. The paper [4] reviews the use of nongenerative AI methods in healthcare, focusing on supervised and unsupervised machine learning techniques. It explains how models like decision trees, SVM, and clustering improve diagnostic accuracy and efficiency. The review also discusses challenges in data quality, model interpretability, and reliability for safe clinical use.The paper [5] uses Multiwavelength Polarimetric LiDAR with machine learning to accurately classify tree species for improved vegetation monitoring.

The paper [6] presents a hybrid AI framework using CNN, Transformer, SVM, and K-means for accurate colorectal cancer detection and segmentation, achieving 99% accuracy and improved visualization for clinical use. The paper [7] discuss about machine learning clustering and classification to build and manage U.S. stock portfolios. It shows that combining HAC clustering or Random Forest classification with portfolio optimization outperforms the market, especially during Covid-19. The paper [8] deals a machine learning framework for industrial processes, identifying key inputs and predicting production outcomes. Using Chemical Vapor Deposition (CVD) as a case study, it combines clustering and Shapley analysis to find critical factors, enabling accurate predictions with limited data. The paper [9] applies supervised and unsupervised learning to classify in-service OTDR trace changes in optical fiber networks. The multilayer perceptron achieved accuracy, while Gaussian mixture clustering performed best on single-effect data, improving network monitoring. The paper [10] discusses an unsupervised method for CTC detection that avoids manual labeling, is domain-independent, and robust to imaging variations, offering a cost-effective and transferable solution for distinguishing CTCs from white blood cells. The paper [11] uses a semi-supervised soft-voting ensemble model for TBM rock mass classification, using limited labeled and extensive unlabeled data. It iteratively expands its training set with pseudo labels, achieving higher accuracy and robustness than other methods, while providing reliable confidence estimates. The paper [12] uses a hybrid supervised–unsupervised framework for breast ultrasound classification, achieving improved tumor discrimination without masks or complex preprocessing. The paper [13] discuss about a semi-supervised GAN-based model for image classification, using collaborative training of generators, discriminators, and classifiers to leverage limited labeled and abundant unlabeled data, improving both image generation quality and classification accuracy in complex

environments. The paper [14] uses malmixer, a semi-supervised malware family classifier that achieves high accuracy with sparse training data. Using domain-knowledge-aware data augmentation, it improves few-shot classification performance and demonstrates the effectiveness of lightweight semi-supervised methods for malware classification.

The paper [15] deals a data-driven framework combining active and semi-supervised learning for HVAC fault diagnosis, leveraging unlabeled data to reduce labeling effort and show that active learning can cut labeling costs, integrated strategies enable cost-effective, robust fault detection in building systems. The paper [16] discuss about a semi-supervised learning model for group decision making in social networks with incomplete trust information. Using a cost-sensitive SVM, it incorporates unlabelled decision-makers' preferences to improve classification and constructs a minimum cost consensus model, reducing total consensus costs in urban renewal scenarios. The paper [17] studys robust embedding regression, a robust semi-supervised learning method that handles noisy and redundant data, achieving superior classification and clustering performance. The paper [18] provides a quick reference guide to basic supervised machine learning classification methods, covering classification and regression, their applications, advantages, and limitations. The paper [19] applies similarity analysis using Dynamic Time Warping (DTW) to classify time series in both supervised and unsupervised settings. The paper [20] uses an integrated credit scoring model combining Kohonen's SOM- unsupervised and Random Forest -supervised to enhanced credit risk prediction accuracy.

Problem statement

To categorize a large volume of documents into supervised, unsupervised, and semi-supervised classes, while optimizing system accuracy and enhancing performance.

Objective

To develop a machine learning-based multi-label document classification system that reduces redundancy, improves runtime performance, identifies the structure of data, and categorizes files accordingly.

3. Research Methodology

First, the document is loaded into the system for processing. Once loaded, the preprocessing stage begins, where the raw text undergoes several important steps to prepare it for further analysis. Initially, the text is converted into lowercase to ensure uniformity, preventing the same words in different cases from being treated as distinct features. Next, **stemming** is applied, which reduces words to their root forms by removing prefixes and suffixes. For example, words like "running," "runner," and "runs" are all reduced to the root word "run." This helps in minimizing the vocabulary size and focusing on the essential meaning of words, improving computational efficiency. Alongside stemming, **lemmatization** is performed, which converts words into their base or dictionary form by considering the context and correct part of speech. For instance, "better" becomes "good," and "was" becomes "be." Unlike stemming, lemmatization provides meaningful root words based on the linguistic structure of the sentence. Together, these preprocessing steps — lowercase conversion, stemming, and lemmatization — clean and standardize the text data, enabling the system to extract the most relevant and consistent features from the document. This ensures that further tasks, such as feature extraction, similarity calculation, and classification, are performed more accurately and efficiently.

3.1 Proposed method-IDS(Identifying Document Structure)

After the text has been preprocessed, feature extraction is used to find important parts of it, like keywords, metadata, heading structures, and short text snippets that show what the document is about. After these features are removed, they are sent through feature selection. This process keeps the most useful and informative features and gets rid of the useless or unnecessary ones to make the model work better and make the computations easier. Label encoding turns categorical document labels into numerical code so that machine learning algorithms can use them. This is done in supervised learning. The IDS model is then used. This model uses machine learning to look at the document's structure and content traits and figure out what class it belongs to.

After that, similarity measurement is used to figure out how close two documents are to each other based on the extracted features. This helps find documents or changes that are connected. Using similarity scores, clustering sorts papers into groups that have things in common. This can help with both unsupervised and semi-supervised classification tasks. The dataset is then split into training and testing sets, with training data making up about 60% of the dataset and testing data making up the other 40%. This lets the IDS model learn trends from the training data and use testing data to make sure it is correct. Lastly, the system makes predictions on new or unseen documents and automatically sorts them into groups based on what it has learned. This makes it easier to organize, find, and handle documents.In label encoding assigned class0 – supervised learning, class1- unsupervised learning, class2- semi supervised learning.The prediction of data assign the class which type of structure follows the document.

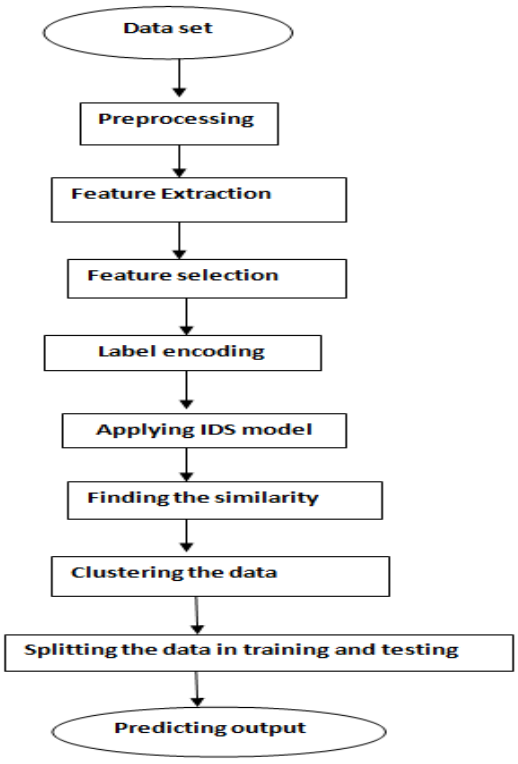

Figure 1. IDS (Identifying document structure) architecture model

3.2 Dataset

The dataset collected from the kaggle dataset. In the table1, Predicted for the following dataset are **class 0** represents **supervised learning** where datasets like "Identifying the author," "Analyzing Student academic Trends," and "Titanic dataset" contain labeled data with known outcomes, enabling algorithms to learn direct input-output mappings; **class 1** stands for **unsupervised learning** seen in "Country Socio-Economic Indicators," "Market Basket data," and "Fake job prediction," which comprise unlabeled data, prompting algorithms to discover underlying patterns or clusters without predefined answers; **class 2** refers to **semi-supervised learning**, exemplified by "OCR document text recognition" and "Paragraph dataset," where both labeled and large amounts of unlabeled data are present, allowing models to leverage the structure in unlabeled data to boost performance even with limited labeled samples, effectively bridging the gap between supervised and unsupervised approaches by combining the strengths of both paradigm.

Table 1: Predicting classes in dataset

Sno	File name	Source	Predicted class-IDS model
1	Identifying the author	kaggle	0
2	Analyzing student academic trends	Kaggle	0
3	Titanic dataset	Kaggle	0
4	Country socio-econonmic Indicators	Kaggle	1
5	Market baseket data	Kaggle	1
6	Fake job predidtion	Kaggle	1
7	OCR document text recognition	Kaggle	2
8	Paragraph dataset	Kaggle	2

4. Result

The table2 summarizes the performance of different datasets in terms of Precision, Recall, and F1-score for classification tasks. Quality and Accuracy scores for most datasets were good, with most being above 0.90 for both. Overall, the Accuracy is 0.96, which means the model worked very well. The Macro Average (0.94 Precision, 0.91 Recall, and 0.93 F1-score) shows how well the model performed on average across all datasets. The Micro Average (0.95 Precision, 0.92 Recall, and 0.94 F1-score), on the other hand, looks at the total number of instances and shows slightly better performance because bigger datasets have a bigger effect. Overall, the model shows that it can reliably and consistently classify different types of data.

Table 2: performance metrics in IDS model

Sno	File name	Precision	Recall	F1-score
1	Identifying the author	0.96	0.95	0.94
2	Analyzing student academic trends	0.92	0.85	0.84
3	Titanic dataset	0.95	0.95	0.95
4	Country socio-econonmic Indicators	0.89	0.84	0.92
5	Market baseket data	1.00	0.95	0.97
6	Fake job predidtion	0.94	0.92	0.94

7	OCR document text recognition	0.96	0.96	0.97
8	Paragraph dataset	0.90	0.86	0.93
	Accuracy			0.96
	Macro average	0.94	0.91	0.93
	Micro average	0.95	0.92	0.94

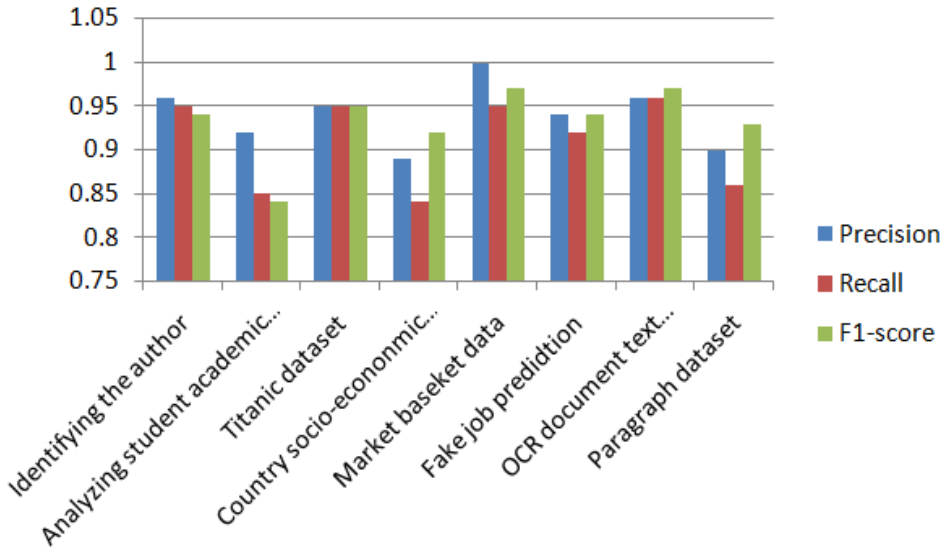

Fig.1. Performance metrics on dataset

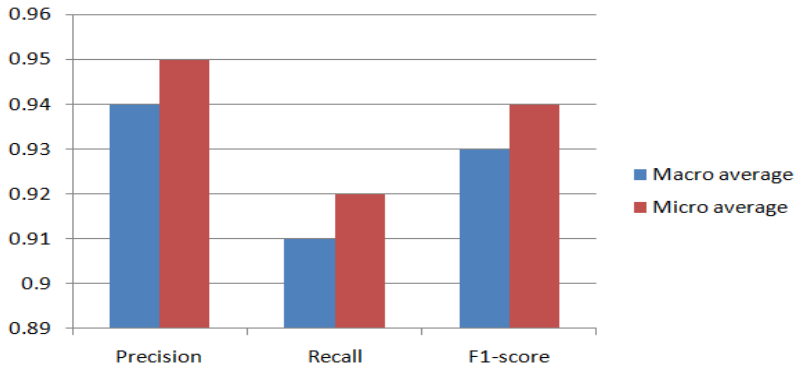

Fig.2. Macro average and Micro average in metrics

5. Conclusion

In this study, we proposed the **Identifying Document Structure (IDS)** model for automated document classification, addressing the limitations of manual analysis in multi-view document classification. By combining machine learning for supervised learning and clustering techniques for unsupervised data, the model effectively categorizes documents based on metadata, keywords, and structural features.The performance evaluation, as shown in the tabular results, demonstrates high precision, recall, and F1-scores across multiple datasets, with an overall **accuracy of 0.96**. These results confirm that the IDS model improves classification efficiency and accuracy over existing manual and traditional machine learning techniques. It offers an effective solution for large-scale document classification, significantly reducing human effort and ambiguity, and providing scalable performance across different document types.

References

1. Xue, Zhixiang, et al. "Multimodal self-supervised learning for remote sensing data land cover classification." *Pattern Recognition* 157 (2025): 110959.
2. Liu, Shiqi, et al. "Dual-view cross attention enhanced semi-supervised learning method for discourse cognitive engagement classification in online course discussions." *Expert Systems with Applications* 278 (2025): 127339.
3. Saeed, Rehan, et al. "Supervised learning-based artificial senses for non-destructive fish quality classification." *Biosensors and Bioelectronics* 267 (2025): 116770.
4. Pantanowitz, Liron, et al. "Nongenerative artificial intelligence in medicine: advancements and applications in supervised and unsupervised machine learning." *Modern Pathology* 38.3 (2025): 100680.
5. Hu, Zhong, and Songxin Tan. "Supervised and unsupervised machine learning approaches for tree classification using multiwavelength airborne polarimetric lidar." *Smart Agricultural Technology* 11 (2025): 100872.
6. Raju, Akella S. Narasimha, et al. "Colorectal cancer detection with enhanced precision using a hybrid supervised and unsupervised learning approach." *Scientific Reports* 15.1 (2025): 3180.
7. Salah, Ayari, and Gatfaoui Hayette. "A meta-analysis of supervised and unsupervised machine learning algorithms and their application to active portfolio management." *Expert Systems with Applications* 271 (2025): 126611.
8. Papavasileiou, Paris, et al. "Integrating supervised and unsupervised learning approaches to unveil critical process inputs." *Computers & Chemical Engineering* 192 (2025): 108857.
9. Tremblay, Christine, et al. "Supervised and unsupervised learning for classifying changes in optical time domain reflectometer traces." *Journal of Optical Communications and Networking* 17.9 (2025): D118-D124.
10. L. An *et al*., "Unsupervised Classification for Circulating Tumor Cells," in *IEEE Access*, vol. 13, pp. 85669-85681, 2025, doi: 10.1109/ACCESS.2025.3564012.
11. Zeng, Shaoxiang, et al. "Semi-supervised ensemble model for TBM rock mass classification." *Tunnelling and Underground Space Technology* 162 (2025): 106632.
12. Song, Mingue, and Yanggon Kim. "Optimizing proportional balance between supervised and unsupervised features for ultrasound breast lesion classification." *Biomedical Signal Processing and Control* 87 (2024): 105443.
13. Hu, Jiyu, Haijiang Zeng, and Zhen Tian. "Applications and Effect Evaluation of Generative Adversarial Networks in Semi-Supervised Learning." *arXiv preprint arXiv:2505.19522* (2025).
14. Li, Jiliang, et al. "Malmixer: Few-shot malware classification with retrieval-augmented semi-supervised learning." *2025 IEEE 10th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2025.
15. Fan, Cheng, et al. "Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance." *Applied Energy* 356 (2024): 122356.

16. Feng, Mengying, et al. "Social relation-driven consensus reaching in large-scale group decision-making using semi-supervised classification." *Information Fusion* 104 (2024): 102160.
17. Bao, Jiaqi, et al. "Robust embedding regression for semi-supervised learning." *Pattern Recognition* 145 (2024): 109894.
18. Alnuaimi, Amer FAH, and Tasnim HK Albaldawi. "An overview of machine learning classification techniques." *BIO Web of Conferences*. Vol. 97. EDP Sciences, 2024.
19. Corliss, David J. "Similarity Analysis: Classification of Time Series Data Using Supervised and Unsupervised Learning." *Intelligent Computing-Proceedings of the Computing Conference*. Cham: Springer Nature Switzerland, 2025.
20. Xu, Tianyi. "Credit risk assessment using a combined approach of supervised and unsupervised learning." *Journal of Computational Methods in Engineering Applications* (2024): 1-12.