RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 2, APRIL-JUNE 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

A REVIEW OF TEXT CLASSIFICATION TECHNIQUES ON LEARNING ALGORITHMS

Mr.A. Ravi and Dr R.Velmurugan
[1] Research Scholar, Presidency College, Chennai
[2] Associate professor, Presidency College, Chennai
gangai_ravi@yahoo.com

Abstract. Text classification plays a strong role in organizing and receiving meaningful information from huge collections of both unstructured and structured data. Considering the immense progress of digital contents across the domains, the manual categorization has become unrealistic, paving the way for automated classification techniques generated by data mining, natural language processing (NLP), and machine learning. Traditional approaches depend heavily on handcrafted features and classical classifiers, but recent advancements in deep learning and graph-based methods have significantly enhanced performance by activating automatic feature extraction and better contextual representation results for the text. This paper provides a comprehensive review of document classification methods, tracking the evolution from rule-based and statistical models the importance of preprocessing, feature engineering, and dataset availability. Moreover, the study identifies unresolved research challenges related to classification effectiveness, model comprehensibility, and scalability of the proposed approaches.

Keywords: Natural language processing, preprocessing, Text classification.

1 Introduction

The growth of digital information in online and offline sources in the reason for the inevitable utility of text classification research. Text mining is needed to turn raw text into relevant insights as enormous amounts of unstructured data generated daily. Text mining uses NLP, Data Mining, and Machine Learning to find patterns, correlations, and meaningful information in text. The inability of humans to search through all accessible documents led to the development of document classification. Automatic categorizing of documents could greatly simplify this process. Text categorization categorizes documents into predetermined groups. The concept of classification is broad and has numerous uses outside information retrieval (IR). Text classification is used in spam detection, sentiment analysis, obscenity detection, email

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

sorting, and topic-specific or vertical searches. The Paper [1] denotes text classification as a key NLP task with applications in multiple domains and focuses on graph convolution network (GCN)-based approaches. compares their performance on benchmark datasets, highlights, strengths and limitations, and outlines future challenges and research directions. The research Paper [2] provides a comprehensive review of text classification from 2014–2022, comparing traditional and deep learning approaches across tasks, datasets, and metrics, while highlighting strengths, limitations, and future directions. This paper [3] talks about current representation issues in long document classification, provides a comprehensive analysis of Transformer-based solutions, and reviews evaluation strategies with a focus on suitable baselines and benchmark datasets.

This paper [4] states about the impact of digital transformation in the public sector, showing how technologies like AI, block chain, chat bots, cloud, and IoT enhance efficiency, transparency, service delivery, and modern governance. The paper [6] reviews challenges in handling unstructured data and highlights semantic approaches for improving text document classification. The research paper [7] reviews deep learning for long documents, encompassing classification, summarization, sentiment, with a model taxonomy and datasets highlights key challenges. The paper [8] takes on the hierarchical text classification, covering traditional and recent methods, datasets, and hierarchy-aware evaluation, and benchmarks models against non-hierarchical baselines. The paper [9] surveys graph neural network methods for text classification, covering graph construction, learning, datasets, and evaluation, and highlights pros, cons, and future directions. The Paper [10] reviews Arabic Text Classification (ATC), categorizing studies by topics, tasks, and processing phases, highlights challenges, applications, and benchmark needs, and suggests future research directions. The Paper [11] deals the Survey Handwritten Text Recognition (HTR) for French historical documents, reviews techniques, datasets, accuracies, and commercial systems, and highlights state-of-the-art methods and future research directions.

2 Related works

This Paper [12] discusses about Reviews deep learning approaches for multi-label learning (MLC), covering DNNs, transformers, auto encoders, CNNs, and RNNs, highlights challenges like high-dimensional data, label correlations, and partial labels, and offers a comparative analysis, open research problems, and future directions. The Paper [13] deals with the text document classification and text mining techniques, emphasizing representation methods,

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

machine learning approaches, and open challenges. The paper [14] explains about embedding-based approaches to text classification, reviewing feature representation advances, effective technique–embedding combinations, and highlighting challenges such as multi-label classification, cost-effectiveness, and the use of knowledge graphs. This Paper [15] discusses challenges in classifying unstructured text data, reviews term-based vs. semantic approaches, and highlights algorithms, techniques, and stages of text document classification. The Paper [16] discusses about the recent advances in text classification, contrasting traditional feature-based methods with deep learning approaches. The paper [17] enunciates the classification based on semantic enriched for subject categorization. The article [18] discusses traditional, static, and contextualized word embeddings, highlighting BERT's effectiveness and future research directions in improving word representation. The Paper [19] deals with unstructured data management through document classification, noting the drawbacks of Bag-of-Words and n-gram models, and uses neural networks as a superior approach for effective feature extraction and contextual word representation. This paper [20] elaborates recent advances in applying deep learning to historical document analysis, reviewing tasks, models, datasets, and highlighting emerging research trends and future directions.

The Paper [21] discusses the exponential growth of unstructured data and highlights text mining as a valuable tool for knowledge extraction. It emphasizes text classification using predefined categories and addresses the need for incremental learning to handle dynamic databases, reviewing various machine learning algorithms, classifier architectures, and applications in this context. The Paper [24] discusses the growing volume of digital text documents and the challenge of high-dimensional feature space in text classification. It highlights how feature selection and feature extraction techniques reduce dimensionality and improve performance, and further reviews various classifiers for document categorization. The Paper [25] explain about with Question Answering Systems (QASs), tracing their development from restricted-domain systems to modern approaches. It surveys and classifies QASs based on multiple criteria, highlights the current research status, and suggests directions for future work. The paper [26.] discusses a systematic literature review on text classification from 2013 to 2022, encasing 110 studies. It highlights the three main stages of classification (data preparation, classifier training, and evaluation), reviews commonly used technologies and datasets, and identifies the growing dominance of deep learning methods such as RNNs, CNNs, and Transformers in this field. The Paper [27] is a thorough analyzer of text mining

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

techniques for extracting patterns from unstructured and ambiguous data on social networking sites. It reviews recent advances with a focus on classification and clustering approaches for analyzing social media text. The paper [28] explains about document clustering, comparing traditional methods, which ignore word semantics, with semantic clustering that uses meaningful relationships for improved accuracy. It surveys 17 studies, key challenges, tools, ontologies, and algorithms, and uses a system using concept weight with Hierarchical Agglomerative Clustering, Bisecting k-means, and Self-Organized Map Neural Network based on WordNet ontology. The paper [29] narrates about text classification, reviewing methods for organizing digital documents. It summarizes existing studies on document representation, feature selection, data mining, and evaluation techniques illustrating it in a tabular form to show progress in the field.

3 Preprocessing steps

The dataset was collected from Kaggle and GitHub. The applied preprocessing techniques includes converting text to lowercase, removing stop words, stemming, lemmatization, tokenization, and count vectorization. After preprocessing, the data was divided into training and testing sets for applying the methods. It has been observed during the recent years that there has been tremendous progress in automatic text classification, utilizing machine learning methods like Bayesian classifiers, Decision Trees, K-nearest neighbor (KNN), Support Vector Machines (SVMs), Latent Semantic Analysis. Automatic text classification often uses supervised learning techniques, assigning pre-defined category labels based on a training set of tagged documents. Some of the strategies are listed as below:

A. Change the text in to lower case

After collecting the raw data, the data text is converted into a lower case. The texts are changed into correct format.

B. Stop words removal

In NLP, stop words are common words like the, is, in, and, of, that, do not have a significance. Hence, they are often taken out of text before it is processed to make natural language processing jobs faster and more accurate.

C. Stemming

Stemming in NLP is a text preprocessing technique that reduces words to their root or stem, often by removing suffixes, to group related words together.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

D. Lemmatization

Lemmatization is an essential process in Natural Language Processing (NLP) that breaks down words into their simplest form, which is called a "lemma." It is like at the meaning and part of speech (POS) of the word and makes sure that the base form is a legal word, unlike stemming which removes the prefixes and suffixes.

3.1 Tokenization

One of the top priorities given in NLP, the tokenization is the most important. In order to perform this, a text input is broken up into smaller pieces called tokens. The tokens can be letters, words, phrases, or whole lines.

3.2 Feature extraction

The feature of Feature extraction involves, finding and extracting useful variables or qualities from raw data. These created features make the dataset more informative and compact and is used for classification, prediction, and clustering.

3.3 Feature selection

The Feature selection is specifically designed to identify, the most relevant input variables, making models simpler, faster, less overfitted, and easier to interpret by removing irrelevant or redundant features.

3.4 Label encoding

Label encoding is a basic data preprocessing technique, which is found very reliable to convert categorical data into a numerical format, suitable for machine learning models. Many algorithms cannot process non-numeric values, making encoding a necessary step when working with features such as colors, different types of shapes, subject, and the likes.

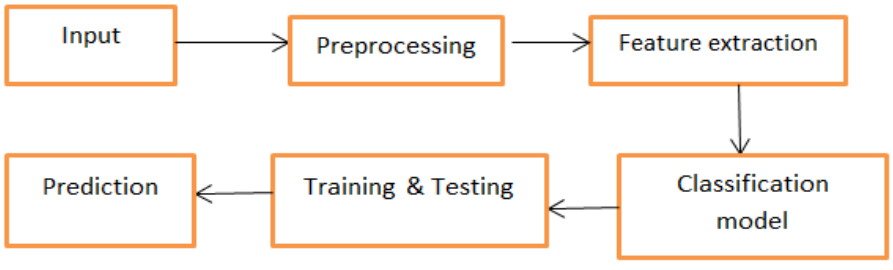

Fig. 1. Flow method for the classification

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

4. Classification techniques

4.1 Bayesian classifiers

Machine learning classification method called as Naive Bayes predicts the data point categories using probability. It assumes all features are independent. Naive Bayes excels at spam filtering, document categorization, and sentiment analysis. The paper [5] deals with multi nominal naïve bayes for classifying document with indexing system.

4.2 Decision Trees

A Decision Tree assists in decision-making by outlining various options and their potential results. It is utilized in machine learning for tasks such as classification and prediction. It has root nodes with different internal nodes and leaf nodes. The paper [22] talks about the automatic document classification using text mining. It proposes a framework that classifies documents into folders based on content, performs sentiment analysis, and also summarizes the datasets. The methodology includes document collection, preprocessing, TF-IDF term weighting, classification using Decision Tree, and evaluation/visualization of results. TF-IDF ranks terms by importance, while the Decision Tree decides the appropriate document folder.

4.3 K-nearest neighbor (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm which predominantly serves for classification, though it is also applicable to regression tasks also. The method operates by identifying the "k" nearest data points (neighbors) to a specified input and formulates predictions based on the predominant class (for classification) or the mean value (for regression). KNN does not make assumptions regarding the underlying data distribution, classifying it as a non-parametric and instance-based learning method. The paper [23] deals with document classification of unstructured text. It enunciates upon the term vector space models for representing documents and addresses the high-dimensionality issue by proposing a term space reduction approach for the KNN algorithm.

4.4 Support Vector Machines (SVMs)

A Support Vector Machine (SVM) is a supervised machine learning algorithm that is commonly employed for tasks such as classification, regression, and outlier detection. The method operates by identifying the optimal decision boundary, referred to as a hyperplane, which effectively distinguishes between data points belonging to different classes. The objective of SVM is to optimize the margin, defined as the distance between the hyperplane and the closest data points from each class. The essential data points that determine the position of the

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

hyperplane are referred to as support vectors. By concentrating on these aspects, SVM develops a strong model that excels in handling intricate datasets. In explicit terms, it has been learned that SVM is establishing the broadest possible boundary between categories, ensuring distinct separation and minimizing misclassification.

4.5 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique used to unveil hidden meanings within the text. It analyzes how words occur across different documents and identifies patterns in their usage. Rather than simply counting word frequencies, LSA focuses on understanding the context and relationships among words. The process involves converting text into a large word–document matrix and then reducing it mathematically to retain only the most significant features. This allows computers to group the related words and documents based on meaning, rather than relying solely on exact word matches. The paper [21] explains about text document classification using a parsimonious CNN. It presents a CNN architecture with parallel 1D convolutional layers (1–5 word windows) leveraging LSA word vectors. The study evaluates the model on balanced and imbalanced datasets, showing that the CNN with LSA vectors outperforms linear classifiers and achieves high classification accuracy, making it a strong baseline for text classification tasks.

4.6 Logistic regression

Logistic regression is a supervised learning approach used in Natural Language Processing (NLP) to classify text, such as spam detection. It uses the sigmoid function to translate numerical representations of text properties like Bag-of-Words word counts into probabilities. This lets the model forecast with the plausibility and categorize inputs. Logistic regression is commonly used in sentiment analysis and document categorization to classify and emotionalize material.

5. Result

Table 1 presents a survey of recent text classification studies, highlighting the methods and datasets used across various application domains. The reviewed literature shows a transition from traditional machine learning techniques to advanced deep learning, graph-based, transformer-based, and LLM-driven approaches.Several studies employed CNN, RNN, and GNN architectures for document classification tasks [1], [2], while logistic regression and kernel-based methods were applied to technical document classification [3]. Deep learning techniques were also adopted for legal document classification [4], Spanish language document

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

classification [5], and fake news detection using BERT [6].Recent research has increasingly focused on graph-grounded learning, interpretable classification, contrastive learning, and word encoding techniques to improve classification performance and semantic understanding [7]–[10]. Advanced approaches such as transformer-based active learning frameworks [11], dual-encoder models for extreme multi-label classification [12], LLM latent-space categorization methods [13], and domain-aware attention mechanisms [14] further demonstrate the evolution of text classification methodologies.The studies utilized a wide variety of datasets, including AI-Khaleej, Reuters, Ohsumed, AG News, POSTURE50K, Wikipedia, Cora, DBPedia, Reddit, Dreaddit, LitCovid, and EURLex-4K [1]–[15].This diversity indicates that modern text classification techniques are being applied across multiple domains such as document management, healthcare, legal analytics, social media analysis, fake news detection, and multi-label classification.

Table 1. Survey on text classification

No	Name of the topic	Published year	Method	Dataset used
1	An effective approach for Arabic document classification using machine learning[1]	2022	CNN	AI-Khaleej
2	Document Classification with Hierarchical Graph Neural Networks[2]	2022	CNN,RNN and GNN	Reuters 90[2] Oshumed[3] AG news[4]
3	Deep learning for technical document classification[3]	2022	Logistic regression or kernel method	Patent documents
4	Multi label legal document classification:A deep learning based approach with label attention & domain specific pre training[4]	2022	State of the art	POSTURE50K Multi-label datasets
5	Document Classification system for the Spanish language[5]	2022	Crawling process, NLP, Deployment process	Google, Wikipedia dataset
6	Overview of the CLEF-2022 CheckThat! Lab: Task 3 on Fake News De-	2022	Bert	English and german subtask

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

	tection[6]			
7	Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting[7]	2023	Graph text model	Cora,Art, Industrial,M.I
8	Interpretable Classification of Wiki-Review Streams [8]	2023	Stream based classification	Media wilki
9	AspectCSE: Sentence Embeddings for Aspect-based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge[9]	2023	multi-aspect embeddings	PwC dataset Wikimedia Dataset
10	Improving text classification via a soft dynamical label strategy[10]	2023	Word encoder	20NG dataset, AG's News dataset, DBPedia dataset, FDCNews dataset, SST-2 dataset
11	Transformer-based active learning for multi-class text annotation and classification[11]	2024	SOAP based framework	i2b2 National Center, Partners Healthcare, and Beth Israel Deaconess Medical Center
12	Dual-Encoders For Extreme Multi-Label Classification[12]	2024	Dual encoder	EURLex-4K and LFAmazonTitles-131K
13	Contextual Categorization Enhancement through LLMs Latent-Space[13]	2024	Hierarchical Navigable Small Worlds (HNSWs) bert	Wikipedia
14.	AttentionDep: Domain-Aware Attention for Explainable Depression Severity Assessment[14]	2025	AttentionDep, a domain-aware attention model	Reddit datasets, Dreaddit dataset
15	One size does not fit all: exploring variable Thresholds for distance-based multi-label text Classification[15]	2025	Word embedding -glove	MLTC LitCovid Reuters

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

Table 2 presents the performance evaluation metrics of various document and text classification approaches reported in recent studies. The results indicate that deep learning and transformer-based models generally achieve superior classification performance compared with traditional methods. Among the reviewed works, CNN-based Arabic document classification [1] and the soft dynamical label strategy [10] achieved the highest accuracies of 98% and 99%, respectively, demonstrating the effectiveness of deep learning architectures for document categorization. Similarly, Hierarchical Graph Neural Networks (GNNs) [2] and the Spanish language document classification system [5] reported accuracies of 97%, highlighting the benefits of graph-based representations and language-specific processing techniques.Transformer-based approaches also showed strong performance. The SOAP-based active learning framework [11] achieved 97% precision, recall, and F1-score, with an overall accuracy of 95%, indicating balanced and reliable classification. In contrast, the BERT-based fake news detection model [6] obtained lower results (Precision = 0.44, Recall = 0.44, F1-score = 0.42), suggesting the increased difficulty of misinformation detection tasks.

Table 2: Performance Evaluation metrics on classification

No	**Name of the topic**	**Method**	**Performance measures**			
			Precision	**Recall**	**F1 score**	**Accuracy**
1	An effective approach for Arabic document classification using machine learning[1]	CNN	-	-	-	98
2	Document Classification with Hierarchical Graph Neural Networks[2]	CNN,RNN and GNN	-	-	-	97
3	Deep learning for technical document classification[3]	Logistic regression or kernel method	-	-	-	0.96
4	Multi label legal document classification: A deep learning based approach with label attention & domain specific pre training[4]	State of the art	0.87	0.66	0.75	-
5	Document Classification system for the Spanish language[5]	Crawling process, NLP, Deployment process	-	-	-	97
6	Overview of the CLEF-2022 CheckThat! Lab:	Bert	0.44	0.44	0.42	0.56

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

	Task 3 on Fake News Detection[6]					
7	Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting[7]	Graph text model	-	-	76	80.08
8	Interpretable Classification of Wiki-Review Streams [8]	Stream based classification	0.84	0.83	0.81	0.96
9	AspectCSE: Sentence Embeddings for Aspect-based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge[9]	multi-aspect embeddings	0.55	0.16	-	-0.55
10	Improving text classification via a soft dynamical label strategy[10]	Word encoder	-	-	-	99
11	Transformer-based active learning for multi-class text annotation and classification[11]	SOAP based framework	0.97	0.97	0.97	0.95
12	Dual-Encoders For Extreme Multi-Label Classification[12]	Dual encoder	0.86	0.91	-	0.91
13	Contextual Categorization Enhancement through LLMs Latent-Space[13]	Hierarchical Navigable Small Worlds (HNSWs) bert	-	-	-	0.74
14.	AttentionDep: Domain-Aware Attention for Explainable Depression Severity Assessment[14]	AttentionDep, a domain-aware attention model	-	-	-	0.91
15	One size does not fit all: exploring variable Thresholds for distance-based multi-label text Classification[15]	Word embedding -glove	67.90	79.67	73.31	0.89

For multi-label classification, Dual-Encoders [12] demonstrated strong effectiveness with 0.86 precision, 0.91 recall, and 0.91 accuracy, while the legal document classification approach [4] achieved an F1-score of 0.75. The graph-grounded pre-training model [7] obtained an F1-score of 76% and 80.08% accuracy, showing the usefulness of graph-enhanced representations in low-resource settings. Overall, the findings suggest that deep learning, transformer-based architectures, graph neural networks, and attention mechanisms consistently

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

provide high classification performance, with accuracies ranging from 74% to 99% depending on the complexity of the dataset and task. The results confirm that advanced neural approaches are highly effective for modern text and document classification applications.Overall, the survey indicates that recent text classification research is increasingly focused on transformer models, attention mechanisms, graph neural networks, and LLM-based techniques, with applications spanning document classification, healthcare, legal analytics, fake news detection, sentiment analysis, and multi-label text categorization. The diversity of datasets further demonstrates the adaptability of modern classification methods across different domains and languages.

6. Conclusion

The comparative study paper is done based on the application of approach, Technique, the data source, feature construction used, and Quantitative evaluation, in the respective papers. It shows the performance compared with the existing method. In this paper we survey and have analyzed compared various classification techniques such KNN, Naïve Bayes, Logistic regression, Latent semantic Analysis, SVM, decision Tree and more. The challenges are investigated and a Novel approach for the documents representation and classification infusing the big data can be proposed as future work.

References

[1.] A. Y. Muaad et al., “An effective approach for Arabic document classification using machine learning,” Glob. Transitions Proc., vol. 3, no. 1, pp. 267–271, 2022, doi: 10.1016/j.gltp.2022.03.003.

[2.] A. Guille and H. Attali, “Document Classification with Hierarchical Graph Neural Networks,” no. i, 2022, [Online]. Available: https://github.com/AdrienGuille/DocGAT

[3.] S. Jiang, J. Hu, C. L. Magee, and J. Luo, “Deep Learning for Technical Document Classification,” IEEE Trans. Eng. Manag., vol. 71, pp. 1163–1179, 2024, doi: 10.1109/TEM.2022.3152216.

[4.] D. Song, A. Vold, K. Madan, and F. Schilder, “Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training,” Inf. Syst., vol. 106, 2022, doi: 10.1016/j.is.2021.101718.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

[5.] L. G. M. Sandoval, L. M. P. Rojas, N. G. A. Cristancho, and C. R. Meneses, "Document Classification System for the Spanish Language," J. Theor. Appl. Inf. Technol., vol. 100, no. 1, pp. 100–112, 2022.

[6.] J. Köhler et al., "Overview of the CLEF-2022 CheckThat! Lab: Task 3 on Fake News Detection," CEUR Workshop Proc., vol. 3180, pp. 404–421, 2022.

[7.] Z. Wen and Y. Fang, "Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting," SIGIR 2023 - Proc. 46th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 506–516, 2023, doi: 10.1145/3539618.3591641.

[8.] S. Garcia-Mendez, F. Leal, B. Malheiro, and J. C. Burguillo-Rial, "Interpretable Classification of Wiki-Review Streams," IEEE Access, vol. 11, no. December, pp. 141137–141151, 2023, doi: 10.1109/ACCESS.2023.3342472.

[9.] T. Schopf, E. Gerber, M. Ostendorff, and F. Matthes, "AspectCSE: Sentence Embeddings for Aspect-based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge," Int. Conf. Recent Adv. Nat. Lang. Process. RANLP, pp. 1054–1065, 2023, doi: 10.26615/978-954-452-092-2_113.

[10.] J. Wang, H. Xie, F. L. Wang, and L. K. Lee, "Improving text classification via a soft dynamical label strategy," Int. J. Mach. Learn. Cybern., vol. 14, no. 7, pp. 2395–2405, 2023, doi: 10.1007/s13042-022-01770-w.

[11.] M. Afzal, J. Hussain, A. Abbas, M. Hussain, M. Attique, and S. Lee, "Transformer-based active learning for multi-class text annotation and classification," Digit. Heal., vol. 10, 2024, doi: 10.1177/20552076241287357.

[12.] N. Gupta, D. Khatri, A. S. Rawat, S. Bhojanapalli, P. Jain, and I. Dhillon, "Dual-Encoders for Extreme Multi-Label Classification," 12th Int. Conf. Learn. Represent. ICLR 2024, pp. 1–27, 2024.

[13.] Z. Bettouche, A. Safi, and A. Fischer, "Contextual Categorization Enhancement through LLMs Latent-Space," 2024, [Online]. Available: http://arxiv.org/abs/2404.16442

[14.] Y. Ibrahimov, T. Anwar, T. Yuan, T. Mutallimov, and E. Hasanov, "AttentionDep: Domain-Aware Attention for Explainable Depression Severity Assessment," pp. 1–15, 2025, [Online]. Available: http://arxiv.org/abs/2510.00706

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

[15.] J. Van Nooten, A. Kosar, G. De Pauw, and W. Daelemans, “One Size Does Not Fit All: Exploring Variable Thresholds for Distance-Based Multi-Label Text Classification,” 2025, [Online]. Available: http://arxiv.org/abs/2510.11160

[16.] Haider Rizvi, Syed Mustafa, Ramsha Imran, and Arif Mahmood. "Text classification using graph convolutional networks: A comprehensive survey." ACM Computing Surveys 57.8 (2025): 1-38.

[17.] Alsammak, Ihab L. Hussein, et al. "Text classification: a comprehensive survey from traditional approaches to deep learning methods." Current and Future Trends on AI Applications: Volume 1 (2025): 247-267.

[18.] Alva Principe, Renzo, Nicola Chiarini, and Marco Viviani. "Long Document Classification in the Transformer Era: A Survey on Challenges, Advances, and Open Issues." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 15.2 (2025): e70019.

[19.] Ferreira, Augusta, and Carlos Santos. "Digital Transformation in Public Sector: Systematic Literature Review." Enhancing Public Sector Accountability and Services Through Digital Innovation (2025): 265-288.

[20.] Dr.B.Lavanya and V.Nirmala, “Multiple Language Document Indexing and Classification with Page Number”, INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY ,2025.

[21.] Kravets, Alla, and Dmitry Semenochkin. "Text Classification Technologies in Document Categorization Systems. A Survey." Advances in Systems Science and Applications 24.2 (2024): 133-165.

[22.] Tsirmpas, Dimitrios, et al. "Neural natural language processing for long texts: A survey on classification and summarization." Engineering Applications of Artificial Intelligence 133 (2024): 108231.

[23.] Zangari, Alessandro, et al. "Hierarchical text classification and its foundations: A review of current research." Electronics 13.7 (2024): 1199.

[24.] Wang, Kunze, Yihao Ding, and Soyeon Caren Han. "Graph neural networks for text classification: A survey." Artificial intelligence review 57.8 (2024): 190.

[25.] Wahdan, Ahlam, Mostafa Al-Emran, and Khaled Shaalan. "A systematic review of Arabic text classification: areas, applications, and future directions." Soft Computing-A Fusion of Foundations, Methodologies & Applications 28.2 (2024).

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

[26.] AlKendi, Wissam, et al. "Advancements and challenges in handwritten text recognition: A comprehensive survey." Journal of Imaging 10.1 (2024): 18.

[27.] Tarekegn, Adane Nega, Mohib Ullah, and Faouzi Alaya Cheikh. "Deep learning for multi-label learning: A comprehensive survey." arXiv preprint arXiv:2401.16549 (2024).

[28.] Ranjan, Nihar M., and Rajesh S. Prasad. "A brief survey of text document classification algorithms and processes." J Data Min Manage 8.1 (2023): 6-11.

[29.] Da Costa, Liliane Soares, Italo L. Oliveira, and Renato Fileto. "Text classification using embeddings: a survey." Knowledge and Information Systems 65.7 (2023): 2761-2803.

[30.] Ranjan, Nihar M., and Rajesh S. Prasad. "A brief survey of text document classification algorithms and processes." J Data Min Manage 8.1 (2023): 6-11.

[31.] Gasparetto, Andrea, et al. "A survey on text classification algorithms: From text to predictions." Information 13.2 (2022): 83.

[32.] Dr.B.Lavanya and V.Nirmala, " A study on semantic enriched document classification using deep learning techniques", Shodhasamhita : Journal of Fundamental & Comparative Research,2022.

[33.] Selva birunda, s., and R. Kanniga Devi. "A review on word embedding techniques for text classification." Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020 (2021): 267-281.

[34.] Panickar, Vishnu, et al. "A brief survey on text document classification." International Journal of Future Generation Communication and Networking 13.3s (2020): 568-575.

[35.] Lombardi, Francesco, and Simone Marinai. "Deep learning for historical document analysis and recognition—a survey." Journal of Imaging 6.10 (2020): 110.

[36.] Gultepe, Eren, Mehran Kamkarhaghighi, and Masoud Makrehchi. "Document classification using convolutional neural networks with small window sizes and latent semantic analysis." Web Intelligence. Vol. 18. No. 3. Sage UK: London, England: SAGE Publications, 2020.

[37.] Noormanshah, W., P. Nohuddin, and Zuraini Zainol. "Document categorization using decision tree: Preliminary study." International journal of engineering & technology 7.4.34 (2018): 437-440.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/2/009

[38.] Moldagulova, Aiman, and Rosnafisah Bte Sulaiman. "Document classification based on KNN algorithm by term vector space reduction." 2018 18th international conference on control, automation and systems (ICCAS). IEEE, 2018.

[39.] Shah, Foram P., and Vibha Patel. "A review on feature selection and feature extraction for text classification." 2016 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE, 2016.

[40.] Mishra, Amit, and Sanjay Kumar Jain. "A survey on question answering systems with classification." Journal of King Saud University-Computer and Information Sciences 28.3 (2016): 345-361.

[41.] Singh, Upendra, and Saqib Hasan. "Survey paper on document classification and classifiers." Int. J. Comput. Sci. Trends Technol 3.2 (2015): 83-87.

[42.] Irfan, Rizwana, et al. "A survey on text mining in social networks." The Knowledge Engineering Review 30.2 (2015): 157-170.

[43.] Naik, Maitri P., Harshadkumar B. Prajapati, and Vipul K. Dabhi. "A survey on semantic document clustering." 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT). IEEE, 2015.

[44.] Jindal, Rajni, Ruchika Malhotra, and Abha Jain. "Techniques for text classification: Literature review and current trends." webology 12.2 (2015).