(ISSN: 2582-5526) www.thevoiceofcreativeresearch.com

The Voice of Creative Research

Vol. 7 & Issue 2 (April 2025)



https://doi.org/10.53032/tvcr/2025.v7n2.23

Research Article

A Review of Document Classification Techniques Using Machine Learning and Deep Learning

Sanjay M. Pardhi
[a]Department of Computer Science, University of Mumbai, Vidyanagari, Mumbai, India
Research Scholar, [b]Department of Computer Science,
Kirti M. Doongursee College of Arts, Science & Commerce, Dadar, Mumbai, India
sanjaypardhi@mu.ac.in

Sampada M. Margaj
Kirti M. Doongursee College of Arts, Science & Commerce,
Department of Computer Science, Mumbai, India
sampada.vyom@gmail.com

Abstract
The study shows different machine learning and natural language processing techniques are used to address fully automated text classification of extensive datasets. The research looks at multiple studies which employ probabilistic models with deep learning approaches and established machine learning methods to identify documents. The discussion evaluates target model advantages against disadvantages while exploring future development paths in order to resolve the need for highly accurate scalable classification systems. This research evaluates how transformer-based models recently developed will affect classification model outcomes.

Keywords: Document, Classification, Deep Learning, RNN, BERT

1. Introduction

Extended digital content requires efficient document classification as a necessary tool for information retrieval and knowledge management and decision-making systems. Conventionally studies used machine learning techniques like k-Nearest Neighbors (kNN), Support Vector Machines (SVM) and Naïve Bayes (NB) are applied to classify various documents according to Sebastiani (2002). These models experience difficulties when dealing with huge unstructured information. The processing of text for classification tasks using deep learning models advanced through recurrent neural networks (RNN) and convolutional neural networks (CNN) as well as transformers especially BERT according to (Devlin *et al.,* 2018 &

Vaswani *et al.,* 2017). The complex models enhance their performance through attention-based procedures while using contextual embeddings.

The adoption of deep learning models in text categorization becomes easier because of growing massive dataset access and powerful computing platforms. The future development of neural networks must handle three key obstacles which are adequate labelled datasets together with high computational capabilities alongside better methods to understand complex neural architectures (Minaee *et al.,* 2021). This research investigates multiple studies that examine document classification solutions to explain key approaches and their relationship to each other. This exploration discusses both emerging developments that benefit document categorization tasks together with their effects on these problems especially through transfer learning with unsupervised learning approaches.

2. Review of Document Classification Techniques

In order to conduct a thorough analysis of document classification methods, we divide them into five main categories: unsupervised classification using transformers, probabilistic models for multilingual classification, machine learning-based classification, deep learning methods, and NLP-based classification.

Reference No	Methodology	Dataset & Application	Strengths	Limitations	Future Directions
[5]	Naïve Bayes with TF-IDF	IT research papers	Effective TF-IDF representation	No comparison with other classifiers, no scalability analysis	Incorporate deep learning models, benchmark with other classifiers
[6]	RNN with NLP preprocessing	Research papers	Superior handling of sequential data	No computational efficiency analysis, lacks transformer comparison	Compare RNNs with transformers, improve computational efficiency
[7]	Probabilistic mixture model	Multilingual web pages	Efficient language detection	Limited handling of code-switching, lacks deep learning comparison	Integrate BERT for improved detection, expand language support
[8]	SVM, Naïve Bayes, kNN	Scientific & news articles	Compares multiple classifiers	No multi-label classification, lacks deep learning comparison	Implement CNNs/Transformers, enable multi-label classification

[9]	BERT embeddings with K-Means	Thesis manuscripts	Robust text representation, web application	High computational cost, lacks traditional model comparison	Various text datasets
[10]	CNN, RNN, Transformer models	Various text datasets	Covers multiple deep learning models	No multi-label classification, lacks deep learning comparison	Implement CNNs/Transformers, enable multi-label classification
[3]	Transformer-based pretraining	Large-scale text datasets	Achieves state-of-the-art results	Requires large computational resources	Optimize pretraining efficiency, reduce memory requirements
[2]	Transformer-based NLP model	Machine translation and classification	Eliminates recurrence, improves parallelization	Lacks interpretability in model decisions	Improve explainability, refine training efficiency

3. Discussion

The examined research depicts the development of document classification techniques through complex deep learning formats to simple machine learning platforms. SVM together with Naïve Bayes exhibit excellent structured data performance though they traditionally have trouble interpreting context-based relations (Sebastiani 2002; Aggarwal & Zhai 2012). Deep learning models specifically CNNs and Transformers use their substantial computation needs to achieve outstanding classification results (Mikolov *et al.,* 2013) (Vaswani *et al.,* 2017). The transformer-based models BERT and GPT along with their ability to implement contextual word embeddings and transfer learning have greatly advanced text classification effectiveness according to expertise (Howard & Ruder, 2018; Devlin et al., 2018). Despite their difficulty in processing large datasets while costing high amounts of money and presenting understanding challenges for models, many obstacles remain to be solved. Research on multilingual categorization proves to be challenging especially when handling papers which use many languages (Schwenk & Li, 2018). Model development research needs to create efficient and accurate categorization models with lightweight operation capabilities (Raffel *et al.,* 2020). A deeper understanding of deep learning classifiers emerges through explainable strategies which are necessary to establish trust among users in automated classification systems (Samek *et al.,* 2019).

4. Conclusion

The review analyzes various document classification strategies in detail to present their advantages and shortcomings. The better performance of RNNs and BERT compared to traditional approaches does not negate ongoing scalability and dataset variety and computational efficiency issues. Future research aiming to optimize deep learning model practical application should work on model enhancements and multi-label classification and compare the results with current techniques. The development of neural network-pairing approaches together with probabilistic methods as dual classification systems should be studied because they show potential to heighten both efficiency and accuracy levels. The implementation of ideal classification models requires real-world integration of theoretical progress to achieve practical success.

References

1. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1–47. https://doi.org/10.1145/505282.505283
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 5998–6008). Curran Associates, Inc.
3. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805
4. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021). Deep learning-based text classification: A comprehensive review. ACM Computing Surveys (CSUR), 54(3), 1–40. https://doi.org/10.1145/3439726
5. Baygin, M. (2018). Classification of text documents based on naïve Bayes using n-gram features. Journal of Information Science and Engineering, 34(4), 987–1002.
6. Deshmukh, R., Patil, M., Bhosale, R. (2019). A document classification using NLP and recurrent neural network. International Journal of Computer Applications, 181(5), 23–29. https://doi.org/10.5120/ijca2019918562
7. Schwenk, H., Li, X. (2018). A corpus for multilingual document classification in eight languages. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).
8. Cheng, Y. (2019). Document classification based on convolutional neural network and hierarchical attention network. Neural Networks, 110, 56–64. https://doi.org/10.1016/j.neunet.2019.07.004
9. Adhikari, A., Ram, A., Tang, R., Lin, J. (2019). DocBERT: BERT for document classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 961–969. https://doi.org/10.18653/v1/D19-1094
10. Huang, X., Paul, M. J. (2018). Examining temporality in document classification Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 394–404. https://doi.org/10.18653/v1/N18-1036

11. Aggarwal, C. C., Zhai, C. (2012). Mining text data. Springer. https://doi.org/10.1007/978-1-4614-3223-4
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 3111–3119).
13. Howard, J., Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031
15. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
16. Samek, W., Wiegand, T., Müller, K. R. (2019). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296. https://arxiv.org/abs/1708.08296