Review Article

International Journal of Advanced Computer Research, Vol 13(63)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2023.1362008

A review and analysis for the text-based classification

Prince Kumar*and Animesh Kumar Dubey
Department of Computer Science, PCST Bhopal, Madhya Pradesh, India

Received: 20-March-2023; Revised: 18-June-2023; Accepted: 21-June-2023
©2023 Prince Kumar and Aminesh Kumar Dubey. This is an open access article distributed under the Creative Commons Attribution (CC BY) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
In the current information-rich era, efficient retrieval and classification of text-based documents have become crucial tasks. With the exponential growth of digital content, the ability to retrieve the most relevant and appropriate documents has become a pressing concern. Effective document retrieval not only saves time and effort but also contributes to enhanced knowledge discovery and decision-making processes. To address these challenges, various text-based classification techniques have been developed and implemented. This paper aims to provide a comprehensive review and analysis of text-based classification techniques. The objectives include evaluating existing methods, identifying their strengths and limitations, and suggesting potential avenues for future research. The paper will analyze various algorithms, feature extraction techniques, and evaluation metrics employed in text-based classification. Additionally, it investigated the impact of different factors, such as document size, language, and domain specificity, on classification performance.

Keywords
Text-based classification, Knowledge discovery, Inherent ambiguity, Extraction mechanism.

1.Introduction
In the current information-rich era, efficient retrieval and classification of text-based documents have become crucial tasks [1, 2]. With the exponential growth of digital content, the ability to retrieve the most relevant and appropriate documents has become a pressing concern for researchers, professionals, and individuals alike [3, 4]. Effective document retrieval not only saves time and effort but also contributes to enhanced knowledge discovery and decision-making processes [5-7]. To address these challenges, various text-based classification techniques have been developed and implemented.

The vast amount of textual data available in diverse domains poses a challenge in organizing and retrieving relevant information [8, 9]. Traditional manual methods of document categorization and retrieval are time-consuming, labor-intensive, and often ineffective due to the sheer volume and complexity of the data [10-12].

As a result, automated approaches for text-based classification have gained significant attention [10]. These approaches leverage machine learning, natural language processing, and information retrieval techniques to automatically categorize and retrieve documents based on their content [13-15].

Text-based classification encounters several challenges. One key challenge is the inherent ambiguity and variability of human language, making it difficult to extract meaningful features and patterns from text data [16–19]. Additionally, the diversity of document formats, languages, and topics further complicates the classification process [17–19]. Furthermore, scalability and efficiency are vital considerations when dealing with large-scale document collections [16–20]. *Figure 1* shows the zcallenges in the existing methods of text-based classification.

The need for efficient document retrieval and categorization motivates researchers to explore and develop advanced text-based classification techniques. By automating the classification process, researchers aim to improve the accuracy, speed, and scalability of document retrieval systems. Additionally, automated classification allows for

*Author for correspondence

targeted information extraction, knowledge discovery, and decision support in various domains, including academia, business, healthcare, and information management.

This paper aims to provide a comprehensive review and analysis of text-based classification techniques. The primary objectives are to evaluate the existing methods, identify their strengths and limitations, and suggest potential avenues for future research. This paper will analyze various algorithms, feature extraction techniques, and evaluation metrics employed in text-based classification. Furthermore, it will investigate the impact of different factors, such as document size, language, and domain specificity, on classification performance.

The remainder of this paper is organized as follows: Section 2 presents a review of the existing literature, highlighting the major approaches and their effectiveness. Section 3 provides the discussion and analysis along with the findings. Finally, Section 4 concludes the paper, summarizing the key contributions and highlighting the significance of text-based classification in modern information retrieval systems.

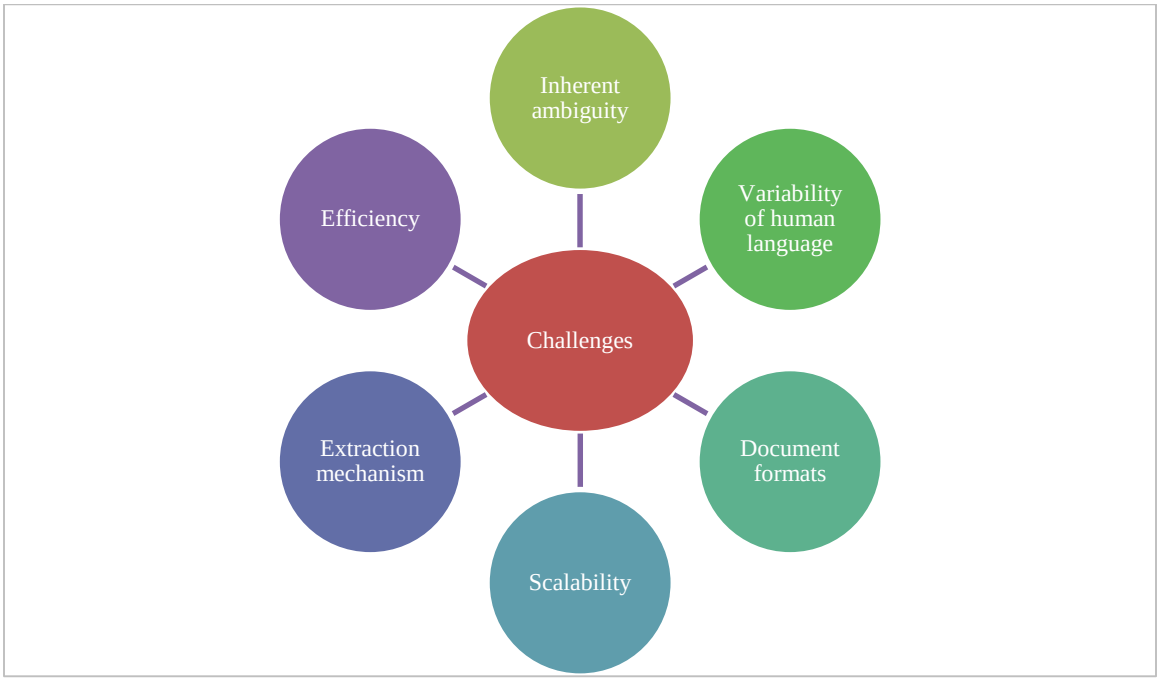

Figure 1 Challenges in the existing methods of text-based classification

2.literature review

In this section, several relevant studies are discussed.

In 2022, Liu et al. [21] employed a pre-trained language model for long text classification. They introduced TextCNN with specialized word embeddings and keyword finetuning, leading to enhanced classification accuracy and expanded keyword data. Experimental results demonstrated significant improvements in accuracy and F1-score across various datasets. Co-training with keyword discovery further improved semantic understanding and classification performance.

In their work, Pathak, and Jain [22] tackled the multilingual abusive comment identification challenge by proposing a solution that incorporates the μBoost ensemble of CatBoost classifier models. The solution achieved a noteworthy mean F1-score of 89.286 on the test data, surpassing the baseline model's F1-score of 87.48.

Wang et al. [23] introduced a text mining method that utilizes semantic framework technology to extract structured information from unstructured defect descriptions. They developed a deep analyzing model specifically for power equipment defects, which offers valuable guidance for equipment upgrading, selection, and maintenance based on historical defect texts. The effectiveness of the proposed method was validated through case studies.

Ma and Pu [24] conducted a study on the relational data search problem involving semi-structured keyword queries. They observed that traditional text indexes encounter performance challenges when dealing with fuzzy text matching. However, they proposed the use of NLP-inspired neural networks to enable predictive index access optimization. The query optimization neural network is trained in an unsupervised fashion, using the text corpus as the training data.

Yu et al. [25] presented a bidirectional encoder-based algorithm for policy text classification, achieving a high F1 value of 93.25% on the test set. This algorithm accurately determines policy fields, enabling efficient analysis and extraction of valuable information from policy text data.

Caron [26] examined the vulnerability of modern financial text mining methods to shortcut learning. The study found that out-of-distribution performance estimates were consistently weaker than in-distribution estimates, resulting in misleading evaluations. Preprocessing techniques, such as entity removal and vocabulary filtering, were proposed to mitigate the impact of shortcut learning.

Sun et al. [27] identified limitations of traditional learning algorithms for text classification, including unclear text features, long training periods, and the disregard for word order. They proposed a bidirectional encoder-based algorithm to auto-categorize technology information text, improving the accuracy of science and technology information classification. The results showed substantial enhancements in accuracy, recall, and F1-score, indicating its effectiveness in Chinese text classification.

In 2023, Umer et al. [28] assessed the efficacy of word representation techniques and convolutional neural networks (CNN) in text classification. Their model employed FastText word embeddings on both benchmark and non-benchmark datasets, yielding promising results and highlighting the potential of FastText for text classification.

In 2023, Shi et al. [29] introduced a POS-aware and layer ensemble transformer neural network, called PL-Transformer, to enhance natural language processing tasks. By incorporating parts-of-speech information and leveraging outputs from multiple encoder layers using correlation coefficient attention, PL-Transformer achieved improved text classification performance, including a 3.95% increase in accuracy.

In 2023, Chandran et al. [30] introduced TopicStriKer, a model that integrates unsupervised topic modeling and supervised string kernels for text classification. By utilizing co-occurring topic words and topic proportions, the algorithm reduces the document corpus to a topic-word sequence. This combined approach enhances accuracy and reduces training time.

Table 1 shows the review discussion on the latest papers along with the method, ad advantages and limitations.

Table 1 Review discussion on the latest papers along with the method, ad vantages and limitations

S. No.	References	Method	Advantages	Limitations
1	[31]	Neural network-based methods	By identifying latent topics, researchers can gain a deeper understanding of the reasons behind consumers' sentiments, which can be valuable for marketing research and decision-making.	These models may struggle to capture fine-grained sentiment information.
2	[32]	Pretrained transformer-based models	These pretraining strategies contribute to enhancing the effectiveness of pretrained models in specific domains and topics.	The effectiveness of pretrained models and pretraining strategies can vary depending on the specific task and dataset.
3	[33]	Latent Dirichlet Allocation	The proposed method has the potential to enhance the performance of news classification tasks.	Depending on the size and complexity of the text data, the efficiency and computational requirements of the method may vary.
4	[34]	Pretraining prefix template	By incorporating label word embeddings and filtering redundant information, the method aims to improve the quality of the text representation used for classification.	Processing long texts and calculating cosine similarities can be computationally expensive, especially with large-scale datasets.

S. No.	References	Method	Advantages	Limitations
5	[35]	Modelling label dependencies	Minor additional computational and memory overheads.	The performance and applicability of the method to other domains or types of data should be further investigated.

3.Discussion and analysis

Neural network-based methods have been employed to identify latent topics in text data, providing valuable insights for marketing research and decision-making [31]. However, these models may struggle to capture fine-grained sentiment information, which can limit their effectiveness in certain applications.

Pretrained transformer-based models have shown promise in various domains and topics, with pretraining strategies contributing to their enhanced effectiveness [32]. However, the performance of pretrained models and pretraining strategies can vary depending on the specific task and dataset, highlighting the need for careful evaluation and selection.

The use of Latent Dirichlet Allocation has been proposed as a method to enhance the performance of news classification tasks [33]. By leveraging the generated topic-document matrix, the method aims to improve the retrieval and classification of relevant news articles. However, the efficiency and computational requirements of the method may vary depending on the size and complexity of the text data being processed.

The introduction of a pretraining prefix template in text classification aims to improve the quality of text representation by incorporating label word embeddings and filtering redundant information [34]. While this approach shows promise in enhancing classification performance, the computational cost of processing long texts and calculating cosine similarities can be significant, particularly with large-scale datasets.

Modelling label dependencies in multi-label classification tasks has been explored, with minor additional computational and memory overheads [35]. The method offers the potential to capture both co-occurrences and rare label subsets, improving the understanding of relationships between labels. However, further investigation is needed to assess its performance and applicability to other domains or types of data.

Overall, these advancements in text classification techniques offer promising opportunities for improving the accuracy and efficiency of text-based tasks. However, careful consideration should be given to the specific requirements and characteristics of the task at hand to select the most suitable approach.

4.Conclusion

This paper has provided a comprehensive review and analysis of text-based classification techniques, aiming to evaluate existing methods and identify their strengths and limitations. Various algorithms, feature extraction techniques, and evaluation metrics employed in text-based classification have been analyzed. The impact of factors such as document size, language, and domain specificity on classification performance has also been investigated. The findings of this review shed light on the current state of text-based classification and highlight potential avenues for future research. By understanding the strengths and limitations of existing methods, researchers can further advance the field and develop more effective and efficient approaches. Continued research and development in this area will contribute to further advancements in knowledge discovery, decision-making processes, and information management.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Gasparetto A, Marcuzzo M, Zangari A, Albarelli A. A survey on text classification algorithms: From text to predictions. Information. 2022; 13(2):83.

[2] Wang Y, Wang C, Zhan J, Ma W, Jiang Y. Text FCG: Fusing contextual information via graph learning for text classification. Expert Systems with Applications. 2023:119658.

[3] Chen X, Cong P, Lv S. A long-text classification method of Chinese news based on BERT and CNN. IEEE Access. 2022; 10:34046-57.

[4] Bayer M, Kaufhold MA, Reuter C. A survey on data augmentation for text classification. ACM Computing Surveys. 2022; 55(7):1-39.

[5] Qasim R, Bangyal WH, Alqarni MA, Ali Almazroi A. A fine-tuned BERT-based transfer learning approach for text classification. Journal of Healthcare Engineering. 2022.
[6] Ma Y, Liu X, Zhao L, Liang Y, Zhang P, Jin B. Hybrid embedding-based text representation for hierarchical multi-label text classification. Expert Systems with Applications. 2022; 187:115905.
[7] Muñoz S, Iglesias CA. A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. Information Processing & Management. 2022; 59(5):103011.
[8] Dubey AK, Kushwaha GR, Shrivastava N. Heterogeneous data mining environment based on dam for mobile computing environments. Information Technology and Mobile Communication. 2011:144.
[9] Mohammed A, Kora R. An effective ensemble deep learning framework for text classification. Journal of King Saud University-Computer and Information Sciences. 2022; 34(10):8825-37.
[10] Khataei Maragheh H, Gharehchopogh FS, Majidzadeh K, Sangar AB. A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification. Mathematics. 2022; 10(3):488.
[11] Dubey AK, Shandilya SK. A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques. In 2010 5th international conference on industrial and information systems 2010 (pp. 207-212). IEEE.
[12] Zhou H. Research of text classification based on TF-IDF and CNN-LSTM. In journal of physics: conference series 2022 (p. 012021). IOP Publishing.
[13] Zhang H, Zhang X, Huang H, Yu L. Prompt-based meta-learning for few-shot text classification. In proceedings of the 2022 conference on empirical methods in natural language processing 2022 (pp. 1342-57).
[14] Yang X, Li Y, Li Q, Liu D, Li T. Temporal-spatial three-way granular computing for dynamic text sentiment classification. Information Sciences. 2022; 596:551-66.
[15] Li Q, Peng H, Li J, Xia C, Yang R, Sun L, Yu PS, He L. A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST). 2022; 13(2):1-41.
[16] Zhao H, Xie J, Wang H. Graph convolutional network based on multi-head pooling for short text classification. IEEE Access. 2022; 10:11947-56.
[17] Yang D, Kim B, Lee SH, Ahn YH, Kim HY. AutoDefect: defect text classification in residential buildings using a multi-task channel attention network. Sustainable Cities and Society. 2022; 80:103803.
[18] Dubey AK, Kapoor D, Kashyap V. A review on performance analysis of data mining methods in IoT. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(73):193.
[19] William P, Badholia A, Patel B, Nigam M. Hybrid Machine Learning Technique for Personality Classification from Online Text using HEXACO Model. In2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) 2022 Apr 7 (pp. 253-259). IEEE.
[20] Shelke N, Chaudhury S, Chakrabarti S, Bangare SL, Yogapriya G, Pandey P. An efficient way of text-based emotion analysis from social media using LRA-DNN. Neuroscience Informatics. 2022: 100048.
[21] Liu L, Wu Y, Yin L, Ren J, Song R, Xu G. A method combining text classification and keyword recognition to improve long text information mining. In 7th IEEE International Conference on Data Science in Cyberspace (DSC) 2022 (pp. 242-8). IEEE.
[22] Pathak M, Jain A. µBoost: An Effective Method for Solving Indic Multilingual Text Classification Problem. In eighth international conference on multimedia big data (BigMM) 2022 (pp. 96-100). IEEE.
[23] Wang H, Cao J, Lin D. Deep analysis of power equipment defects based on semantic framework text mining technology. CSEE Journal of Power and Energy Systems. 2019; 8(4):1157-64.
[24] Ma L, Pu KQ. Neural network accelerated tuple search for relational data. In 2022 IEEE 23rd international conference on information reuse and integration for data science (IRI) 2022 (pp. 81-2). IEEE.
[25] Yu B, Deng C, Bu L. Policy text classification algorithm based on bert. In 11th international conference of information and communication technology (ICTech)) 2022 (pp. 488-91). IEEE.
[26] Caron M. Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift. In2022 IEEE International Conference on Big Data (Big Data) 2022 Dec 17 (pp. 3486-3495). IEEE.
[27] Sun JW, Bao JQ, Bu LP. Text classification algorithm based on TF-IDF and BERT. In 2022 11th international conference of information and communication technology (ICTech)) 2022 (pp. 1-4). IEEE.
[28] Umer M, Imtiaz Z, Ahmad M, Nappi M, Medaglia C, Choi GS, Mehmood A. Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications. 2023; 82(4):5569-85.
[29] Shi Y, Zhang X, Yu N. PL-Transformer: a POS-aware and layer ensemble transformer for text classification. Neural Computing and Applications. 2023; 35(2):1971-82.
[30] Chandran NV, Anoop VS, Asharaf S. Topicstriker: A topic kernels-powered approach for text classification. Results in Engineering. 2023; 17:100949.
[31] Alantari HJ, Currim IS, Deng Y, Singh S. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer

reviews. International Journal of Research in Marketing. 2022; 39(1):1-9.
[32] Guo Y, Ge Y, Yang YC, Al-Garadi MA, Sarker A. Comparison of pretraining models and strategies for health-related social media text classification. In Healthcare 2022 (p. 1478). MDPI.
[33] Shao D, Li C, Huang C, Xiang Y, Yu Z. A news classification applied with new text representation based on the improved LDA. Multimedia Tools and Applications. 2022; 81(15):21521-45.
[34] Chen J, Lv S. Long Text Truncation Algorithm Based on Label Embedding in Text Classification. Applied Sciences. 2022; 12(19):9874.
[35] Ozmen M, Zhang H, Wang P, Coates M. Multi-relation message passing for multi-label text classification. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2022 (pp. 3583-7). IEEE.

Prince Kumar is doing M.Tech in Computer Science, PCST ,RGPV, Bhopal (MP) and he has completed Graduation with Mathematics Honours and MCA from NIT DURGAPUR. His area of interest are Data Mining, Optimization, Machine Learning and Artificial Intelligence.

Email: princekarna@gmail.com

Animesh Kumar Dubey is working as Assistant professor with the department of Computer Science and Engineering, at Patel College of Science and Technology, Bhopal, India. He has completed his Bachelor of Engineering (B.E.) and M.Tech. degree with Computer Science Engineering from Rajeev Gandhi Technical University, Bhopal (M.P.). He has more than 15 publications in reputed, peer-reviewed national and international journals and conferences. His research areas are Data Mining, Optimization, Machine Learning, Cloud Computing and Artificial Intelligence.
Email: animeshdubey123@gmail.com