Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

A Review of Text Classification Based on ML & Data Mining Algorithms

Ashraf Atam Mustafa[1], Adnan Mohsin Abdulazeez[2]
ashrafatam96@gmail.com,adnan.mohsin@dpu.edu.krd
[1]Akre University for Applied Science, Technical College of Informatics, Akre, Department of Information Technology, Akre, Kurdistan Region, Iraq
[2]Presidency of Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq

Article Information

Submitted : 15 May 2024
Reviewed : 27 May 2024
Accepted : 15 Jun 2024

Keywords

Text Classification, ML, DM, Bayesian Classifiers, SVM.

Abstract

In the digital era, the field of text classification has experienced transformative growth through the application of Machine Learning (ML) and Data Mining (DM) algorithms. This review traces the evolution from traditional data mining methods to sophisticated ML strategies that significantly enhance the analysis and categorization of textual data. We discuss pivotal technologies including Bayesian classifiers, Support Vector Machines (SVM), and contemporary advances such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The integration of Natural Language Processing (NLP) techniques is highlighted for their critical role in enriching semantic analysis capabilities, a necessity for effective text classification. Additionally, the paper addresses challenges like handling high-dimensional data, dealing with imbalanced datasets, and confronting ethical issues such as bias and privacy in automated systems. By synthesizing the latest research, this review identifies current gaps, proposes practical solutions, and forecasts future trends in text classification to support ongoing research and application across various sectors.

A. Introduction

The As we navigate the digital age, the sheer volume of text generated daily across various platforms presents both challenges and opportunities for automated processing and analysis. Text classification, a pivotal task within the realms of machine learning (ML) and data mining (DM), serves as a fundamental method for organizing, managing, and deriving insights from text data. This technique underpins many applications, from filtering spam in email inboxes to automating customer service responses and enabling sentiment analysis in social media feeds[1][2][3].This literature review provides a comprehensive examination of the methodologies employed in text classification, focusing on the evolution and application of both traditional data mining techniques and modern machine learning algorithms. Initially, the review outlines the historical context and the theoretical foundations of text classification, setting the stage for a deeper exploration of various classification strategies[4][5].From the early use of Bayesian classifiers to the more recent deployment of deep learning models, this review highlights the progression and sophistication of technologies over time. It scrutinizes the effectiveness of algorithms like Decision Trees, Support Vector Machines (SVM), and Random Forests in handling textual data. Additionally, it delves into the transformative impact of neural network architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have redefined benchmark standards by their ability to capture contextual and sequential information in text[6][7][8].Moreover, the review addresses the integration of natural language processing (NLP) techniques with ML and DM algorithms, which enhances the semantic processing capabilities necessary for more nuanced classifications. Issues such as the handling of high-dimensional data, dealing with imbalanced datasets, and the challenges of semantic ambiguity are discussed. The review also considers the ethical implications of automated text classification, including bias, privacy, and security[9][10].By synthesizing contemporary research and case studies, this literature review aims to present a critical analysis of the state-of-the-art methods in text classification. It seeks to identify gaps in current research, suggest practical solutions to unresolved problems, and predict future directions in the development of more robust and intelligent text classification systems.The ultimate objective of this review is to furnish researchers, practitioners, and policymakers with the insights necessary to harness the power of text classification tools effectively and ethically, thereby contributing to the advancement of knowledge management and information retrieval in various sectors[11][12][13][14].

B. Background Theory

B.1 Text Classification

Classification serves as a cornerstone in machine learning (ML) and data mining (DM), providing essential capabilities for organizing and interpreting the vast amounts of textual data generated daily. Its significance is underscored by its wide range of applications, from sentiment analysis to automated document sorting, which facilitate efficient information retrieval and decision-making processes.

B.2 Historical Evolution of Text Classification

The field began with the use of simple statistical models, such as Bayesian classifiers, which applied probability theory to predict the category of text based on its features. As computational capabilities expanded, the field evolved, embracing more complex machine learning techniques.

B.3 Techniques in Text Classification

Support Vector Machines (SVM): SVMs have played a pivotal role by constructing hyperplanes in high-dimensional spaces, effectively categorizing texts with greater accuracy.

Convolutional Neural Networks (CNNs): These models have been instrumental in capturing spatial hierarchies, enhancing the ability to process and classify image-based or spatially structured text data.

Recurrent Neural Networks (RNNs): RNNs excel in handling data where sequences or temporal dependencies are crucial, such as in processing sentences or paragraphs where context evolves over time.

B.4 Role of Natural Language Processing (NLP)

The integration of NLP techniques has significantly enhanced text classification systems by enabling deeper semantic processing. This allows for a more nuanced understanding and classification of texts, addressing complexities such as semantic ambiguity and context sensitivity.

Word Embeddings: Techniques like word embeddings have transformed how machines understand human language, representing words as vectors of real numbers that capture their meanings, relationships, and the contextual nuances within a text corpus.

B.5 Challenges with High-Dimensional Data

Text classification often involves handling high-dimensional data, a challenge due to the vast number of unique words and phrases. This issue has been mitigated through innovative dimensionality reduction techniques and advanced regularization methods, which help manage the complexity and prevent overfitting.

B.6 Ethical Considerations in Text Classification

As text classification technologies are increasingly applied in sensitive areas—from filtering news feeds to automating legal decisions—the importance of developing transparent and fair models has grown. The field is actively researching ways to mitigate bias and ensure privacy, aiming for systems that make ethically sound decisions based on balanced and representative data.

B.7 Conclusion and Future Research Directions

The landscape of text classification is continually evolving, driven by technological advancements and the ever-increasing complexity of data. Ongoing research and thoughtful application are essential to address these challenges, ensuring that text classification tools remain effective, fair, and relevant in various applications[15][16][17][18][19].

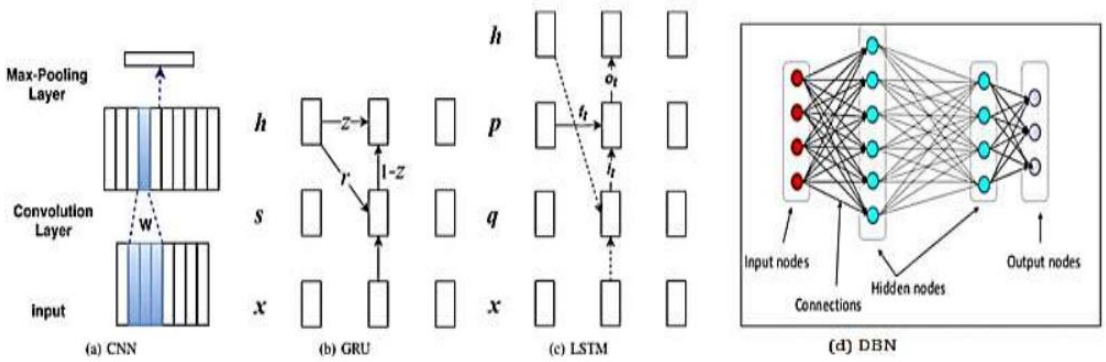

Figure 1. Four Typical DNN architecture

C. Literature Review

Dong et al [20]. proposed a text classification model integrating label embedding with a Self-Interaction Attention Mechanism, employing the BERT model for enhanced feature extraction. Despite its innovative approach and improved classification accuracy over existing methods, the study is limited by its focus on theoretical performance without extensive real-world application testing. From my perspective, while the approach shows potential, practical validation across diverse datasets and real-world scenarios would substantiate its effectiveness and adaptability, encouraging broader adoption in practical applications. This could significantly bridge the gap between theoretical research and practical utility in NLP tasks.

J. Zhang and Liu [21] introduced an advanced text classification model utilizing a Gated Graph Neural Network with an Attention mechanism, set within a framework of Coupled P Systems. Their approach innovatively integrates attention-based feature extraction and contextual semantic analysis, which shows improved classification accuracy across multiple datasets. However, the study focuses mainly on theoretical simulations and controlled datasets, which may not fully reflect the complexities of real-world data. The model shows promise, particularly in its novel use of graph neural networks and attention mechanisms; however, its real-world effectiveness remains untested. Extensive practical testing and refinements are essential to establish its utility and potential as a robust tool in natural language processing.

D. Zhang [22] introduced a text classification model that combines the Time Correlation Principle with Rough Set Theory to enhance literary text feature classification and information extraction. The model leverages temporal associations and rough set approximations to improve the accuracy and efficiency

of literary text processing. Despite its novel integration and demonstrated improvements in classification tasks, the model's primary limitation lies in its reliance on theoretical simulations and controlled datasets, which may not fully reflect the complexities and variability of real-world data. In my opinion, while the theoretical foundation and initial results are promising, the model would benefit significantly from testing and validation in diverse, real-world environments to establish its practical effectiveness and adaptability. This approach would ensure that the model can handle the variety and unpredictability inherent in real-world applications, thus making it a more robust and valuable tool for natural language processing tasks.

Onita [23] explored the integration of active and transfer learning techniques for text classification, aimed at improving efficiency with limited data. The study tests various selection criteria, including random, uncertainty sampling, and active transfer selection, across multiple datasets. The results demonstrate that the combined approach of active and transfer learning significantly outperforms traditional methods in terms of label efficiency. However, the research predominantly leverages theoretical simulations and controlled data, which might not fully represent real-world complexities. Personally, I see this study as a compelling advancement in resource-efficient machine learning, highlighting the potential for significant improvements in text classification tasks with sparse data. Further empirical validation could strengthen its applicability, ensuring it addresses the nuanced challenges of diverse, real-world datasets.

Ansari et al [24]. introduced a novel feature selection method using a Genetic Algorithm (GA) to optimize text classification in natural scene images, aimed at addressing challenges like noise and diverse text styles. This approach leverages a Support Vector Machine (SVM) classifier and uses an average F-Score as a fitness function, achieving notable improvements in classification accuracy over standard techniques. However, the study primarily relies on theoretical enhancements without extensive real-world dataset validation, which may limit its practical applicability. From my perspective, while the innovative use of GA for feature optimization in text classification shows promise, especially in adapting to diverse and challenging conditions, its effectiveness and robustness could be further validated by expanding tests to more varied real-world scenarios and integrating with other deep learning architectures, potentially making it a more comprehensive solution for scene text recognition.

Wang et al [25]. explored the classification of proactive personality using text mining based on Weibo text and short-answer questions. They employ machine learning algorithms to analyze texts, demonstrating that combining these two data sources enhances the predictive accuracy for proactive personality traits. The research highlights the model's efficacy in distinguishing individuals with varying levels of proactive personality, utilizing metrics such as accuracy and sensitivity. However, the study primarily relies on controlled datasets and simulated scenarios, which might not capture the complexities of real-world data. The approach is innovative, but further validation in diverse real-life settings is

essential to establish its practical relevance and robustness, ensuring it can effectively handle real-world variability in text data.

N. Shi et al [26]. developed a text classification model using a Chaotic Neural Oscillatory Long Short-Term Memory (CNO-LSTM) structure, integrating chaotic dynamics to enhance learning efficiency and model response times. Their model, validated across multiple datasets, shows improvements in accuracy and computational efficiency over standard LSTM models. The primary limitation, however, is the model's reliance on simulation data, which may not accurately represent real-world complexities. From my perspective, the model's innovative approach to incorporating chaotic dynamics is promising for improving neural network performance. Yet, extensive testing in real-world applications is necessary to fully gauge its effectiveness and to adapt it further for practical use, ensuring it meets the varied demands of dynamic data environments.

W. Zhang et al [27]. developed a robust text classification method employing virtual adversarial training, which enhances model robustness by generating adversarial texts that are readable and contextually appropriate. The method, based on the continuous bag-of-words model, manipulates perturbation direction vectors to produce adversaries that remain interpretable by humans. While it achieves higher robustness against adversarial attacks and maintains classification accuracy across various datasets, the approach primarily focuses on theoretical simulations without extensive real-world testing. This limitation could affect its practical applicability. Personally, I believe that while the approach is innovative and shows promise in increasing the robustness of text classification models, real-world application and further empirical testing are crucial for validating its effectiveness in practical scenarios.

Akhter et al [28]. presented a text classification model employing a Single-layer Multisize Filters Convolutional Neural Network (SMFCNN) to handle Urdu text documents. Their innovative method, validated across diverse datasets, demonstrates enhanced classification accuracy and efficiency by incorporating filters of varying sizes to capture a wide range of text features. However, the study is primarily grounded in simulations and controlled datasets, which may not fully reflect the intricacies and variability of real-world data scenarios. The model demonstrates significant theoretical promise, but its practical efficacy and adaptability remain untested in real-world environments. Extending the application and testing of this model across more varied and natural datasets is crucial to substantiate its effectiveness and readiness for broader practical deployment.

Q. Meng et al [29]. delve into the realm of electric power audit text classification by employing a Multi-Grained Pre-Trained Language Model, specifically designed for this task. Their model, EPAT-BERT, leverages pre-trained word and entity level tasks to understand and classify electric power-related texts more effectively, demonstrating superior performance compared to existing models. However, their study mainly utilizes controlled experimental data, which

may not fully capture the complexity of real-world scenarios. In my opinion, while the innovative approach of integrating domain-specific pre-training tasks shows significant potential in enhancing text classification, the model's real-world applicability needs to be validated across more diverse and dynamic environments to ascertain its effectiveness and scalability. This validation is crucial for transitioning from theoretical models to practical tools in the industry.

Gu et al [30]. presented a novel text classification method that utilizes Graph Neural Networks (GNNs) with a Multi-Granular Topic-Aware Graph approach, significantly enhancing text classification by integrating topic-awareness into text graphs. Their methodology improves the propagation and classification accuracy by introducing topic nodes, which help to strengthen class-aware representation learning and mitigate heterophily caused by polysemous words. However, the study's reliance on controlled datasets for validation limits its assessment of real-world applicability. In my opinion, while the proposed method innovatively addresses issues in text classification and shows promising results, further validation in more diverse and dynamic real-world settings is crucial to evaluate its practical effectiveness and readiness for deployment in varied applications.

Peng and Huo [31]. developed a few-shot text classification method that enhances text classification by combining Wide and Deep Attention Bidirectional Long Short Time Memory (WDAB-LSTM) with a prototypical network. This method optimizes feature extraction by enhancing text preprocessing and employing a sophisticated model to manage distance measurements effectively. Despite its innovative approach, the model primarily tests on simulated environments, which may not adequately capture the challenges present in real-world applications. From my perspective, while the method shows promise in improving classification accuracy, further validation in varied real-world scenarios is crucial to ensure its efficacy and adaptability across different text classification tasks. This validation is vital to moving from theoretical development to practical, deployable technology.

Zhao et al [32]. developed a graph convolutional network based on multi-head pooling (MP-GCN) for short text classification. This approach addresses the challenges of sparse features and limited training data typical in short text classification by using a novel pooling strategy that evaluates and selects important nodes without the need for pre-trained embeddings. While the model demonstrates enhanced classification performance across multiple benchmarks, its reliance on structural data and lack of pre-training may limit its applicability in diverse real-world scenarios. The model's innovative pooling mechanism represents a significant advancement in efficiently handling sparse text data. Further validation in varied applications and the incorporation of context-rich pre-trained embeddings could broaden its practical utility, ensuring robustness and adaptability across different text classification tasks.

Gong et al [33]. developed a Hierarchical Graph Transformer-Based Deep Learning Model for large-scale multi-label text classification. Their model

introduces a graph-based document modelling combined with a hierarchical transformer architecture, which significantly enhances feature extraction and classification accuracy by preserving the logical and hierarchical structure of text. Despite its advancements, the model's reliance on benchmark datasets may not entirely capture real-world complexities. Personally, I see the innovative combination of graph structures and transformer techniques as a promising step toward more nuanced text classification. However, broader validation across more diverse real-world datasets is crucial to ascertain its practical effectiveness and scalability, ensuring it can handle the nuanced demands of various applications.

She et al [34]. have developed a joint learning model that integrates BERT, Graph Convolutional Network (GCN), and a multi-attention mechanism for event text classification and event assignment. Their approach effectively harnesses the strengths of each component to address the unique challenges of classifying and assigning event texts within the context of Chinese government hotlines. While their model demonstrates improvement over several baseline methods and offers a sophisticated technique for handling complex data structures, it primarily relies on controlled dataset evaluations. While their method is innovative and technically robust, broader and more diverse real-world testing is necessary to validate its effectiveness and ensure its adaptability to various practical scenarios beyond the initial governmental context.

Kutbi [35] presented a preprocessing approach using Named Entity Recognition (NER) to enhance text classification while preserving privacy. By replacing named entities with their type categories in texts, this method improves classification accuracy and reduces feature dimensionality, without removing the entities or allowing them to become data noise. Although the technique shows an increase in classifier performance and privacy protection, it relies heavily on the accuracy of the NER system used, which may not consistently identify or categorize entities correctly in diverse datasets. This approach is innovative in integrating privacy considerations into text preprocessing, but its effectiveness across different languages and more informal text genres needs thorough evaluation to confirm its versatility and reliability in varied applications.

Yan et al [36]. proposed a network-based Bag-of-Words (BoW) model that improves text classification by incorporating the structural and semantic relationships among words. Their model leverages network attributes to capture deeper contextual information, which allows for more effective differentiation of text meanings. While the model shows improved efficiency and performance over traditional methods, it heavily depends on the quality of network construction and the proper interpretation of complex network attributes. This approach significantly enriches text representation, but its practical application requires careful calibration and testing across varied datasets to fully validate its robustness and scalability in real-world scenarios.

Thaminkaew et al [37]. presented a framework called PLAML for few-shot multi-label text classification that incorporates a novel approach using prompt-

based label-aware techniques. This model improves classification performance by integrating discrete verbalizers and dynamic threshold mechanisms, addressing the complexities of multi-label classification with few data instances. While this approach demonstrates advancements over existing methods in handling sparse and multi-dimensional data, the reliance on simulated validation limits its tested efficacy in real-world scenarios. The innovative application of prompt-based techniques in multi-label settings shows considerable promise; however, extensive real-world testing and iterative refinement are essential to ensure its effectiveness and adaptability in practical deployments.

Jun et al [38]. presented a novel text classification method that incorporates weighted negative supervision at the classifier layer to enhance rating classification accuracy. This technique modifies traditional classifier training by emphasizing the magnitude of differences between labels, thereby improving the model's ability to distinguish between them more effectively. While the model achieves significant improvements across various datasets and in different languages, its primary limitation is its heavy reliance on theoretical enhancements without extensive real-world testing, which may not adequately represent real-world complexities. The innovative approach requires extensive testing across varied real-world scenarios to validate its effectiveness and ensure practical applicability.

Lan et al [39]. proposed a stacked residual recurrent neural network with cross-layer attention (SRCLA) designed to enhance text classification by leveraging deeper feature extraction and attention mechanisms. Their model, tested on various text classification tasks, shows superior performance, especially in capturing semantic nuances by effectively filtering and supervising lower-level features with higher-level outputs. However, the study's emphasis on benchmark datasets limits exposure to practical, diverse data scenarios. From my perspective, the methodology is robust and innovative, but real-world applications and further empirical testing are essential to validate its effectiveness across different operational environments, ensuring it can consistently perform well in practical deployments.

Liu et al [40]. proposed a document-relational Graph Convolutional Network (GCN) model for text classification that enhances classification accuracy by incorporating document-document relationships through cumulative term frequency-inverse document frequency (TF-IDF) values. Their approach differs from traditional models that primarily utilize document-word relations, offering a more nuanced understanding of document interconnections. Although their model shows promise in improving classification accuracy across various datasets, it relies significantly on controlled data environments which may not fully capture real-world text complexities. The novel integration of document-document relations represents a promising advancement; however, further validation in diverse real-world scenarios and adjustments to the model to handle such complexities are necessary to make this approach more robust and widely applicable.

Fiok et al [41]. introduced a novel text classification method tailored for long texts using a Text Guide approach based on feature importance. This method selectively truncates texts to a predefined limit while maintaining significant information, showing effectiveness in handling long documents with recent language models like Longformer. However, the method's validation predominantly occurs through simulations that do not fully mimic the intricacies of real-world data. The Text Guide method demonstrates potential for improving long text classification with lower computational costs, but its practical effectiveness requires comprehensive testing across various real-world scenarios to confirm its adaptability and reliability in practical applications.

L. Meng [42] developed a CNN-BiLSTM text classification algorithm to enhance information management in smart tourism, integrating Internet of Things (IoT) services. This model effectively categorizes sentiment in tourism reviews, improving the precision and reliability of data annotation. However, its evaluation primarily on controlled datasets may not fully illustrate its performance under real-world complexities. In my opinion, while the model demonstrates notable potential in refining smart tourism services through advanced text analysis, further validation in diverse and dynamic real-world settings is crucial to assess its true effectiveness and adaptability, ensuring it can reliably support the evolving needs of the tourism industry.

Alshalif et al [43]. developed an Alternative Relative Discrimination Criterion (ARDC) feature ranking technique for text classification, aiming to enhance the accuracy of feature ranking by accounting for term frequency across categories. Their approach, compared with standard methods like Information Gain and Pearson Correlation Coefficient, shows superior performance in precision, recall, and accuracy across multiple datasets. However, the validation primarily relies on theoretical metrics without extensive real-world application testing, which could impact the generalizability of the results. The effectiveness and robustness of ARDC in improving feature ranking across diverse real-world scenarios need further testing to confirm its practical application and ability to meet deployment demands.

J. Shi et al [44]. introduced two quantum-inspired deep neural networks for text classification, leveraging complex-valued word embeddings to enhance interpretability and performance. Their approach, rooted in quantum mechanics, offers a novel perspective on handling linguistic complexities. However, the practical applicability and scalability of these theoretical models in diverse real-world scenarios remain largely untested. The effectiveness of this quantum-inspired methodology outside of controlled experimental settings needs thorough validation to ensure its integration into everyday text processing tasks.

Yao et al [45]. introduced a neural network-based model for multi-label text classification that leverages label co-occurrence to enhance performance. Their method represents each class as a column vector in a class embedding matrix,

innovatively linked to the label co-occurrence matrix to ensure classes with high co-occurrence are closely embedded. Despite the improvements in handling label correlations, the model primarily relies on public datasets for evaluation, which may not fully capture real-world complexities. In my opinion, while the approach significantly contributes to the accuracy of multi-label classification, its effectiveness needs to be further validated in diverse, real-world scenarios to confirm its adaptability and robustness across different applications.

Steur and Schwenker [46]. presented a Capsule Network (CapsNet) model enhanced by routing-by-agreement for text classification, which optimizes the signal propagation through layers by focusing on highly agreed predictions. This innovative method leverages the expressiveness of distributed entity representations, aiming to improve the robustness and interpretability of neural network predictions. However, the adoption of CapsNets in practical applications is limited by the complexity of tuning and training such networks, as well as by their computational demands. In my opinion, while CapsNets present a promising avenue for advancing text classification, their practical deployment requires further simplification and optimization to accommodate real-world computational constraints and data variability.

Alemayehu and Fang [47] proposed a submodular optimization framework to address imbalances in text classification datasets using data augmentation. This innovative framework selects synthesized items that maximize both the likelihood of correctly representing their labels and the diversity among the items. Although their method shows improvements in classifier performance on real-world datasets, its primary limitation is the potential overfitting due to high diversity in the selected items. In my opinion, while the approach significantly enhances model training with imbalanced data, additional research is needed to optimize the balance between diversity and accuracy, ensuring the practical effectiveness of the augmentation in various real-world applications.

Ameer et al [48]. developed a framework for multi-label emotion classification on code-mixed text data, utilizing a comprehensive dataset that includes English and Roman Urdu. Their method leverages state-of-the-art machine learning techniques to accurately identify multiple emotions from text, significantly enhancing the understanding of mixed-language content. Despite these advances, the study's reliance on a specifically tailored dataset may not universally represent all types of code-mixed text, potentially limiting its broader applicability. The research offers valuable insights into multi-label emotion classification in a bilingual context, highlighting the importance of further exploration and validation across diverse code-mixed datasets to confirm the model's effectiveness and adaptability in various linguistic environments.

Hao et al [49]. introduced a joint representation approach for short text classification using compositional loss. This model incorporates label embeddings and a novel loss function that balances cross-entropy and triplet loss to better classify texts by exploiting the similarity between text representations and label

embeddings. This method shows enhanced performance on ambiguous texts by making the text representation both close to its corresponding label and distant from other labels. However, the complexity of implementing this new loss function and the dependency on the precision of label embeddings might limit its application. Personally, while I find the approach promising for increasing classification accuracy, especially in scenarios involving ambiguous texts, the practical deployment requires rigorous testing and optimization to ensure it performs well across diverse real-world datasets.

D. Comparison

Ref.	Model/Tech Used	Dataset Used	Limitations	Key Results
[20]	Self-Interaction Attention with BERT	Not specified	Lack of real-world application testing	Improved classification accuracy
[21]	Gated Graph Neural Network with Attention	Multiple datasets	Mainly theoretical simulations	Improved accuracy across datasets
[22]	Time Correlation Principle with Rough Set Theory	Not specified	Theoretical simulations and controlled datasets	Enhanced feature classification
[23]	Active and transfer learning techniques	Multiple datasets	Theoretical simulations and controlled data	Significantly outperforms traditional methods
[24]	Genetic Algorithm with SVM	Natural scene images	Limited real-world dataset validation	Notable improvements in classification accuracy
[25]	Text mining on Weibo text and short-answer questions	Controlled datasets	Simulated scenarios	Enhanced predictive accuracy for personality traits
[26]	Chaotic Neural Oscillatory LSTM	Multiple datasets	Reliance on simulation data	Improvements in accuracy and efficiency
[27]	Virtual adversarial training	Various datasets	Theoretical simulations without extensive real-world testing	Higher robustness against adversarial attacks
[28]	Single-layer Multisize Filters CNN	Diverse datasets	Grounded in simulations and controlled datasets	Enhanced classification accuracy and efficiency
[29]	Multi-Grained Pre-Trained Language Model	Experimental data	Controlled experimental data	Superior performance in specific text classification
[30]	Graph Neural Networks with Multi-Granular Topic-Aware Graph	Controlled datasets	Limited real-world validation	Improved text classification accuracy
[31]	WDAB-LSTM with prototypical network	Simulated environments	Not adequately testing in real-world applications	Improved feature extraction and classification

[32]	Multi-head pooling Graph Convolutional Network	Multiple benchmarks	Lack of pre-trained embeddings and reliance on structural data	Enhanced performance in short text classification
[33]	Hierarchical Graph Transformer	Benchmark datasets	Limited exposure to real-world complexities	Improved feature extraction and classification
[34]	BERT with GCN and multi-attention mechanism	Controlled datasets	Primarily relies on controlled evaluations	Improved classification and event assignment
[35]	Named Entity Recognition for text preprocessing	Not specified	Reliance on NER system accuracy	Improved classification accuracy and privacy protection
[36]	Network-based Bag-of-Words model	Not specified	Quality of network construction	More effective differentiation of text meanings
[37]	PLAML framework for few-shot multi-label classification	Simulated validation	Limited tested efficacy in real-world scenarios	Improved performance with sparse data
[38]	Weighted negative supervision method	Various datasets	Theoretical enhancements without extensive real-world testing	Enhanced rating classification accuracy
[39]	Stacked residual recurrent network with cross-layer attention	Various tasks	Emphasis on benchmark datasets	Superior performance in semantic nuances capture
[40]	Document-relational Graph Convolutional Network	Various datasets	Controlled data environments	Improved classification accuracy
[41]	Text Guide method for long texts	Simulations	Simulations do not mimic real-world data	Effective handling of long documents
[42]	CNN-BiLSTM for smart tourism	Controlled datasets	Evaluation may not reflect real-world performance	Improved sentiment categorization in tourism reviews
[43]	Alternative Relative Discrimination Criterion for feature ranking	Multiple datasets	Theoretical metrics without real-world testing	Superior performance in precision, recall, and accuracy
[44]	Quantum-inspired deep neural networks	Not specified	Limited testing in diverse real-world scenarios	Enhanced interpretability and performance
[45]	Neural network model for multi-label text classification	Public datasets	May not fully capture real-world complexities	Improved handling of label correlations
[46]	Capsule Network model	Not specified	Complexity of tuning and training	Improved robustness and interpretability of predictions
[47]	Submodular optimization framework for data augmentation	Real-world datasets	Potential overfitting due to high diversity	Enhanced model training with imbalanced data

[48]	Multi-label emotion classification for code-mixed text	English and Roman Urdu	Limited representativeness of the dataset	Accurate identification of multiple emotions
[49]	Joint representation approach for short text classification	Not specified	Complexity and dependency on label embeddings precision	Enhanced performance on ambiguous texts

E. Discussion

This review underscores the significant advancements in text classification driven by machine learning and data mining algorithms, revealing an evolution from traditional classifiers to sophisticated neural architectures integrated with NLP techniques. Despite these technological strides, challenges such as high-dimensional data management, dataset imbalance, and ethical concerns like bias and privacy remain prevalent. Addressing these requires enhanced dataset diversity to better capture real-world complexities, the incorporation of cross-disciplinary approaches for more nuanced models, and the development of frameworks to mitigate biases and ensure transparency. Future research should also explore unsupervised learning to leverage the abundance of unlabelled data and investigate emerging technologies like quantum computing to revolutionize classification strategies. By focusing on these areas, the next generation of text classification models can achieve greater accuracy, efficiency, and ethical compliance, significantly impacting various sectors and enriching societal well-being.

F. Recommendations

To enhance the practical applicability and robustness of text classification models, it is crucial to prioritize real-world testing and diversification of datasets. Many of the reviewed studies, while innovative in theoretical development and simulation environments, exhibit limitations in their testing scenarios, which do not fully encapsulate real-world complexities. To bridge this gap, future research should focus on extensive validation across diverse, natural datasets to confirm the effectiveness of these models under varied conditions typical of practical deployments. Additionally, integrating multimodal data sources could provide a more holistic view of text classification challenges and potential solutions.For models heavily reliant on theoretical enhancements, such as those using quantum-inspired networks or complex neural architectures like Capsule Networks, simplification and optimization for real-world computational demands are essential. This approach would not only make these models more accessible but also increase their adaptability across different operational environments.Furthermore, to address the issue of data sparsity and improve the robustness of models against adversarial attacks, the development of novel techniques that incorporate unsupervised or semi-supervised learning should be explored. These techniques can leverage unlabeled data, which is abundant and often underutilized, to enhance learning efficiency and model generalization.Finally, the field would benefit from a closer collaboration between academia and industry to ensure that the models developed are not only theoretically sound but also practically viable. This partnership could lead to the

creation of standardized benchmarks that more accurately reflect real-world applications, facilitating the transition of text classification models from research prototypes to practical tools.

G. Conclusion

In conclusion, this review has charted the evolution of text classification within machine learning (ML) and data mining (DM), from basic classifiers like Bayesian methods and SVMs to advanced techniques such as CNNs and RNNs, enriched by Natural Language Processing (NLP). Despite considerable progress, challenges remain in handling high-dimensional data, dataset imbalance, and ethical concerns such as bias and privacy. Future research should enhance model robustness, expand unsupervised learning applications, and ensure ethical compliance. The collaboration between academia and industry is crucial to translate theoretical advancements into practical applications, optimizing text classification tools for broader societal benefits.

H. References

[1] M. M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," Expert Systems with Applications, vol. 106. Elsevier Ltd, pp. 36–54, Sep. 15, 2018. doi: 10.1016/j.eswa.2018.03.058.

[2] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," International Journal of Research in Marketing, vol. 36, no. 1, pp. 20–38, Mar. 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[3] P. Siebers, C. Janiesch, and P. Zschech, "A Survey of Text Representation Methods and Their Genealogy," IEEE Access, vol. 10, pp. 96492–96513, 2022, doi: 10.1109/ACCESS.2022.3205719.

[4] X. Liu, S. Dai, G. Fiumara, and P. De Meo, "An adversarial training method for text classification," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 8, Sep. 2023, doi: 10.1016/j.jksuci.2023.101697.

[5] D. Ehring, P. Ferraz-Doughty, J. Luttmer, and A. Nagarajah, "A first step towards automatic identification and provision of user-specific knowledge: A verification of the feasibility of automatic text classification using the example of standards," in Procedia CIRP, Elsevier B.V., 2023, pp. 1103–1108. doi: 10.1016/j.procir.2023.02.183.

[6] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, " Survey on Text Classification Algorithms: From Text to Predictions," Information (Switzerland), vol. 13, no. 2, Feb. 2022, doi: 10.3390/info13020083.

[7] B. M. Hsu, "Comparison of supervised classification models on textual data," Mathematics, vol. 8, no. 5, May 2020, doi: 10.3390/MATH8050851.

[8] R. Majeed and A. M. Abdulazeez, "Electrocardiogram Classification Based on Deep Convolutional Neural Networks: A Review," Fusion: Practice and Applications, vol. 3, no. 1, pp. 43–53, 2021, doi: 10.5281/zenodo.4642772.

[9] M. Orosoo et al., "Performance analysis of a novel hybrid deep learning approach in classification of quality-related English text," Measurement: Sensors, vol. 28, Aug. 2023, doi: 10.1016/j.measen.2023.100852.

[10] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," IEEE Access, vol. 11, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[11] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification," ACM Computing Surveys, vol. 54, no. 3. Association for Computing Machinery, Jun. 01, 2021. doi: 10.1145/3439726.

[12] B. A. Muhammad, R. Iqbal, A. James, and Di. Nkantah, "Comparative Performance of Machine Learning Methods for Text Classification," in 2020 International Conference on Computing and Information Technology, ICCIT 2020, Institute of Electrical and Electronics Engineers Inc., Sep. 2020. doi: 10.1109/ICCIT-144147971.2020.9213788.

[13] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[14] O. Ahmed and A. Brifcani, "Gene Expression Classification Based on Deep Learning," in 4th Scientific International Conference Najaf, SICN 2019, Institute of Electrical and Electronics Engineers Inc., Apr. 2019, pp. 145–149. doi: 10.1109/SICN47020.2019.9019357.

[15] Y. Zhou, "A Review of Text Classification Based on Deep Learning," in ACM International Conference Proceeding Series, Association for Computing Machinery, Apr. 2020, pp. 132–136. doi: 10.1145/3397056.3397082.

[16] J. Joy, S. M., Federal Institute of Science And Technology. Department of Computer Science and Engineering, Institute of Electrical and Electronics Engineers. Kerala Section, and Institute of Electrical and Electronics Engineers, Proceedings, 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA) : July 02-04, 2020, Federal Institute of Science And Technology (FISAT), Cochin, Kerala, India.

[17] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, "A comparative review on deep learning models for text classification," Indonesian Journal of Electrical Engineering and Computer Science, vol. 19, no. 1, pp. 325–335, 2020, doi: 10.11591/ijeecs.v19.i1.pp325-335.

[18] B. He, L. Zhu, X. Wang, H. Zhang, and J. Shi, "Research on Text Classification based on Deep Learning," Scientific Journal of Technology, vol. 4, p. 2022.

[19] H. Wu, Y. Liu, and J. Wang, "Review of text classification methods on deep learning," Computers, Materials and Continua, vol. 63, no. 3. Tech Science Press, pp. 1309–1321, Apr. 01, 2020. doi: 10.32604/CMC.2020.010172.

[20] Y. Dong, P. Liu, Z. Zhu, Q. Wang, and Q. Zhang, "A Fusion Model-Based Label Embedding and Self-Interaction Attention for Text Classification," IEEE Access, vol. 8, pp. 30548–30559, 2020, doi: 10.1109/ACCESS.2019.2954985.

[21] J. Zhang and X. Liu, "A Gated Graph Neural Network With Attention for Text Classification Based on Coupled P Systems," IEEE Access, vol. 11, pp. 72448–72461, 2023, doi: 10.1109/ACCESS.2023.3295572.

[22] D. Zhang, "A Novel Text Classification Model Combining Time Correlation Principle and Rough Set Theory," IEEE Access, vol. 11, pp. 135797–135810, 2023, doi: 10.1109/ACCESS.2023.3332909.

[23] D. Onita, "Active Learning Based on Transfer Learning Techniques for Text Classification," IEEE Access, vol. 11, pp. 28751–28761, 2023, doi: 10.1109/ACCESS.2023.3260771.

[24] G. J. Ansari, J. H. Shah, M. C. Q. Farias, M. Sharif, N. Qadeer, and H. U. Khan, "An Optimized Feature Selection Technique in Diversified Natural Scene Text for Classification Using Genetic Algorithm," IEEE Access, vol. 9, pp. 54923–54937, 2021, doi: 10.1109/ACCESS.2021.3071169.

[25] P. Wang et al., "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text," IEEE Access, vol. 8, pp. 97370–97382, 2020, doi: 10.1109/ACCESS.2020.2995905.

[26] N. Shi, Z. Chen, L. Chen, and R. S. T. Lee, "CNO-LSTM: A Chaotic Neural Oscillatory Long Short-Term Memory Model for Text Classification," IEEE Access, vol. 10, pp. 129564–129579, 2022, doi: 10.1109/ACCESS.2022.3228600.

[27] W. Zhang, Q. Chen, and Y. Chen, "Deep Learning Based Robust Text Classification Method via Virtual Adversarial Training," IEEE Access, vol. 8, pp. 61174–61182, 2020, doi: 10.1109/ACCESS.2020.2981616.

[28] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network," IEEE Access, vol. 8, pp. 42689–42707, 2020, doi: 10.1109/ACCESS.2020.2976744.

[29] Q. Meng et al., "Electric Power Audit Text Classification With Multi-Grained Pre-Trained Language Model," IEEE Access, vol. 11, pp. 13510–13518, 2023, doi: 10.1109/ACCESS.2023.3240162.

[30] Y. Gu, Y. Wang, H. R. Zhang, J. Wu, and X. Gu, "Enhancing Text Classification by Graph Neural Networks With Multi-Granular Topic-Aware Graph," IEEE Access, vol. 11, pp. 20169–20183, 2023, doi: 10.1109/ACCESS.2023.3250109.

[31] J. Peng and S. Huo, "Few-shot Text Classification Method Based on Feature Optimization," Journal of Web Engineering, vol. 22, no. 3, pp. 497–514, Jul. 2023, doi: 10.13052/jwe1540-9589.2235.

[32] H. Zhao, J. Xie, and H. Wang, "Graph Convolutional Network Based on Multi-Head Pooling for Short Text Classification," IEEE Access, vol. 10, pp. 11947–11956, 2022, doi: 10.1109/ACCESS.2022.3146303.

[33] J. Gong et al., "Hierarchical Graph Transformer-Based Deep Learning Model for Large-Scale Multi-Label Text Classification," IEEE Access, vol. 8, pp. 30885–30896, 2020, doi: 10.1109/ACCESS.2020.2972751.

[34] X. She, J. Chen, and G. Chen, "Joint Learning with BERT-GCN and Multi-Attention for Event Text Classification and Event Assignment," IEEE Access, vol. 10, pp. 27031–27040, 2022, doi: 10.1109/ACCESS.2022.3156918.

[35] M. Kutbi, "Named Entity Recognition Utilized to Enhance Text Classification While Preserving Privacy," IEEE Access, vol. 11, pp. 117576–117581, 2023, doi: 10.1109/ACCESS.2023.3325895.

[36] D. Yan, K. Li, S. Gu, and L. Yang, "Network-Based Bag-of-Words Model for Text Classification," IEEE Access, vol. 8, pp. 82641–82652, 2020, doi: 10.1109/ACCESS.2020.2991074.

[37] T. Thaminkaew, P. Lertvittayakumjorn, and P. Vateekul, "Prompt-Based Label-Aware Framework for Few-Shot Multi-Label Text Classification," IEEE Access, vol. 12, pp. 28310–28322, 2024, doi: 10.1109/ACCESS.2024.3367994.

[38] Z. Jun, Q. Longlong, S. Fanfan, H. Yueshun, T. Hai, and H. Yanxiang, "Rating Text Classification with Weighted Negative Supervision on Classifier Layer," Chinese Journal of Electronics, vol. 32, no. 6, pp. 1304–1318, Oct. 2023, doi: 10.23919/cje.2021.00.339.

[39] Y. Lan, Y. Hao, K. Xia, B. Qian, and C. Li, "Stacked Residual Recurrent Neural Networks with Cross-Layer Attention for Text Classification," IEEE Access, vol. 8, pp. 70401–70410, 2020, doi: 10.1109/ACCESS.2020.2987101.

[40] C. Liu, X. Wang, and H. Xu, "Text Classification Using Document-Relational Graph Convolutional Networks," IEEE Access, vol. 10, pp. 123205–123211, 2022, doi: 10.1109/ACCESS.2022.3221820.

[41] K. Fiok et al., "Text Guide: Improving the Quality of Long Text Classification by a Text Selection Method Based on Feature Importance," IEEE Access, vol. 9, pp. 105439–105450, 2021, doi: 10.1109/ACCESS.2021.3099758.

[42] L. Meng, "The Convolutional Neural Network Text Classification Algorithm in the Information Management of Smart Tourism Based on Internet of Things," IEEE Access, vol. 12, pp. 3570–3580, 2024, doi: 10.1109/ACCESS.2024.3349386.

[43] [43] S. A. Alshalif et al., "Alternative Relative Discrimination Criterion Feature Ranking Technique for Text Classification," IEEE Access, vol. 11, pp. 71739–71755, 2023, doi: 10.1109/ACCESS.2023.3294563.

[44] J. Shi et al., "Two End-to-End Quantum-Inspired Deep Neural Networks for Text Classification," IEEE Trans Knowl Data Eng, vol. 35, no. 4, pp. 4335–4345, Apr. 2023, doi: 10.1109/TKDE.2021.3130598.

[45] J. Yao, K. Wang, and J. Yan, "Incorporating Label Co-Occurrence into Neural Network-Based Models for Multi-Label Text Classification," IEEE Access, vol. 7, pp. 183580–183588, 2019, doi: 10.1109/ACCESS.2019.2960626.

[46] N. A. K. Steur and F. Schwenker, "Next-Generation Neural Networks: Capsule Networks with Routing-by-Agreement for Text Classification," IEEE Access, vol. 9, pp. 125269–125299, 2021, doi: 10.1109/ACCESS.2021.3110911.

[47] E. Alemayehu and Y. Fang, "A Submodular Optimization Framework for Imbalanced Text Classification with Data Augmentation," IEEE Access, vol. 11, pp. 41680–41696, 2023, doi: 10.1109/ACCESS.2023.3267669.

[48] I. Ameer, G. Sidorov, H. Gomez-Adorno, and R. M. A. Nawab, "Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods," IEEE Access, vol. 10, pp. 8779–8789, 2022, doi: 10.1109/ACCESS.2022.3143819.

[49] M. Hao, W. Wang, and F. Zhou, "Joint representations of texts and labels with compositional loss for short text classification," Journal of Web Engineering, vol. 20, no. 4, pp. 669–688, May 2021, doi: 10.13052/jwe1540-9589.2035.