A Study of Deep Learning-Based Text Representation and Classification Methods

Wei Nie

Xianyang Normal University Xianyang 712000 China

Abstract: ***The advent of the information age, coupled with the extensive implementation of large-scale informatization initiatives, has triggered an explosive growth of digital text data. This deluge of information presents a paramount challenge: how to efficiently and accurately extract actionable insights and effective knowledge from complex, high-dimensional text corpora. The core of this endeavor lies in the fundamental tasks of textual analysis and categorization. This paper provides a comprehensive elaboration on the persistent problems and corresponding innovative solutions within the critical pipeline of text classification, which is fundamentally underpinned by text representation. Conventional text representation methods, such as Bag-of-Words (BoW) and TF-IDF, while intuitive, often grapple with the "curse of dimensionality," data sparsity, and an inability to capture semantic and syntactic nuances, leading to suboptimal feature selection and diminished representational efficacy. The selection of discriminative and non-redundant text features thus remains a significant challenge. In recent years, the methodological landscape for text representation and classification has diversified considerably, introducing techniques ranging from traditional machine learning models (e.g., SVM, Naive Bayes) to more contemporary deep learning architectures. While these advancements have spurred innovation, they concurrently introduce new challenges, including sensitivity to imbalanced label distributions, which can bias models towards majority classes, and poor generalizability across different domains or datasets. To address these limitations, this paper introduces a novel perspective grounded in the deep learning domain. We systematically explore and evaluate advanced neural architectures—including Convolutional Neural Networks (CNNs) for local feature extraction, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks for sequential context modeling, and Transformer-based models (e.g., BERT) for leveraging contextualized word embeddings. The primary objective is to propose and validate a framework that enhances the robustness of text representation, mitigates the impact of label imbalance through advanced sampling or loss functions, and ultimately improves classification accuracy and generalization capability. By leveraging the hierarchical feature learning and representation power of deep models, this research aims to continuously optimize the acquisition of text information and significantly improve the efficiency and precision of knowledge discovery in the era of big data.***

Keywords: Text Classification, Text Representation, Deep Learning, Feature Extraction, Natural Language Processing (NLP), Imbalanced Data, Generalizability, BERT.

1. THE CURRENT SITUATION OF TEXT REPRESENTATION AND CLASSIFICATION METHODS BASED ON DEEP LEARNING

1.1 Textual representation of the status quo

Text is a collection of large numbers of characters composed of unstructured or semi-structured digital information that cannot be directly identified by a classifier, which requires that we now convert the text representation into a language that can be understood by a computer when classifying text. It is important to note that the translated language must be concise, unified, and recognizable by different algorithms and classifiers. Text content can now be represented either by graphs or vectors. That is, to obtain the content of the text as well as the category information by identifying the translated language.

The main problems in text representation are shallow text representation and deep text representation. Among them, in shallow text representations, there is mainly a problem of semantic missing, which directly results in the classifier being less efficient and less accurate in the recognition process. Whereas deep texts are mostly based on current computational models, the classification and extraction of these texts mostly depend on the selection of artificial features. The classification of text is achieved by adding a corresponding threshold selection to the classification process, but this approach undermines the text's self-learning ability and ignores issues such as uneven label data distribution.

1.2 Text classification status quo

Text categorization mainly refers to the grouping of large amounts of text into one or more categories based on characteristics such as text content, text subject, text attributes, etc. In terms of classification methods, text

classification can be divided into rule-based classification methods and statistical classification methods. The rule-based classification method requires more expertise and a library of rules to support it, but this classification method is not widely applied, and is more suitable for a specialized discipline or a specific field. But the learning method based on statistics is more based on some statistics or relevant laws, to calculate and calculate the samples, and to establish the corresponding data model, to realize the text classification. At the same time, a sample needs to be predicted for classification based on the parameters of the sample prior to classification.

1.3 Deep Learning Status

Deep learning is not a new way of learning, but originated from artificial neural networks, a general term for learning methods based on deep neural networks that solve problems by simulating the cognition muscles of the human brain.

2. ANALYSIS OF TEXT CLASSIFICATION METHODS

The arrival of the information age, is both an opportunity and a challenge, in the face of the vast and complex and massive resources, in order to be able to effectively manage and use these information resources, it is necessary to classify these information resources, which also makes the content-based information retrieval has become a concern area. Among them, text classification and text retrieval techniques are important ways of sorting and mining information. That is, by setting a pre-existing category, and then judging the attribution of the text based on its content. Text classification and text retrieval both require the processing, understanding, information organization and management of text content on the basis of natural language, so that content information can be more widely used.

A text classification problem is a problem of predicting an unknown sample class by carefully learning from a sample of a known class. For known text classification methods, they are primarily based on the learning and classification results of classifiers. The research on the classification method is also aimed at improving the efficiency and accuracy of classification. Especially in text classification, text can also be divided into single-label and multi-label categories according to the category to which the text tag belongs. In a single label category, a mature classifier can be used to accomplish this task, However, in multi-label classification, due to the uneven distribution of some data, there are many label categories, making it difficult for simple classifiers to meet multi-label requirements, so it is necessary to research newer classification methods to make the search and classification of text resources more scientific and rational. Tian et al. (2025) proposed a cross-attention multi-task learning framework for ad recall optimization, achieving superior performance in digital advertising through business intelligence innovations [1]. In financial risk management, Wang et al. (2025) designed an AI-enhanced intelligent system for multinational supply chains, integrating predictive analytics and real-time monitoring [2]. Xie et al. (2024) advanced legal text classification using Conv1D-based models, achieving high accuracy in multi-class citation analysis [3]. Medical AI saw progress with Chen et al. (2023), who introduced generative text-guided 3D vision-language pretraining for unified medical image segmentation [4]. Xu (2025) developed CivicMorph, a generative model for public space design optimization [5], while Tu (2025) proposed ProtoMind, combining neural architecture search (NAS) with SIP message modeling for smart regression detection [6]. Industrial monitoring was enhanced by Xie and Liu (2025) through InspectX, leveraging OpenCV and WebSocket for real-time analysis [7]. In 3D vision, Peng et al. (2025) presented 3D Vision-Language Gaussian Splatting at ICLR, enabling dynamic scene reconstruction [8]. Wang (2025) improved clinical trial forecasting via transformer-augmented survival analysis [9]. For LLM optimization, Liu et al. (2025) proposed hybrid-grained pruning to enhance adaptive model efficiency [10]. Zhou (2025) applied swarm intelligence to UAV path planning for precision pesticide spraying [11], while Tan et al. (2024) optimized fault diagnosis using CI-JSO-based densely connected networks [12]. Marketing strategies were explored by Zhuang (2025), who analyzed real estate digital transformation through evolutionary logic [13]. Han and Dou (2025) integrated hierarchical graph attention networks with multimodal knowledge graphs for user recommendations [14]. Finally, Yang (2025) applied the Prompt-BioMRC model to intelligent medical consultation systems [15].

2.1 Classical Text Classification Methods

The multi-classifier integrated learning method is a more widely accepted and commonly used method of classification in recent years. However, with the continuous development of the level of information technology, the classification method of text has also shown a more diversified development trend. With the increasing complexity of information resources, the traditional classical text categorization can not meet the requirements of

text representation and categorization. At present, there are several common classification methods, such as Naive Bayesian inference of phylogeny, neighborhood algorithm, decision tree and ensemble learning seal.

2.2 Naive Bayes

Plain Bayesian is a Bayesian method based on a simple assumption. The main element of this hypothesis is that the different characteristics of the hypothetical sample are not related to the classification effects of the sample content. The basic idea of plain Bayes is to derive the range from probability estimation and then the probability from computation. However, for items already given to be classified, they need to be classified according to specific numerical values, i.e. which numerical value is larger is classified into which category.

2.3 KNN algorithm

KNN algorithm is also known as the proximity algorithm, the core of this algorithm is to find from the training set to be classified items and the most similar to the required classification of a number of texts, The classification of the remaining categories is further determined on the basis of the classification of this text, and if the classifications are more similar, they are determined to be in the same category, and the sample also belongs to that category. Therefore, the average values of a number of samples are very important and can even be said to directly affect the classification results. However, at the same time, the problem of sample content imbalance can be resolved better by calculating the values of several samples, comparing the proximity of these samples with the judgment values, and finally classifying them.

2.4 Decision Tree

A decision tree is a method of calculation based on rule prediction, which is to compute a large amount of text to generate the corresponding data, and to purposely categorize the data to find some valuable information for the decision maker to make the most correct decision. The basic idea of a decision tree is that the structure of the tree is used to categorize all data records, in particular that each internal node in the tree represents a set of records under a certain condition, and to establish different branches of value based on the recorded fields. The advantage of this design is the ability to continuously repeat the lower nodes and branches of the resume under each branch, thus constructing a decision tree from top to bottom. In a decision tree, the nodes of the leaves are the categories of the sample, and the decision tree classification process begins at the root of the tree. Classify different nodes based on their characteristics, test the sample properties of different nodes, and compare the resulting values with the branch nodes, continuously moving to find the branch node that meets all the criteria.

In text classification, the advantages of decision trees are obvious, one is that the computation difficulty is low, and it is easy to use in everyday classification. Second, unrelated data can be processed relatively quickly. At the same time, however, the downside of decision trees is that they tend to ignore correlations between attributes in the data set, which creates certain coincidences, resulting in inaccurate data and classifications.

2.5 Integrated learning

Integrated learning is also known as multiple learning or classifier combinations. Integrated learning mainly involves invoking simple classification algorithms to obtain multiple different classifiers, and combining several classifiers using decision optimization and coverage optimization. This aims to use these classifiers to improve the differences between the overall models and generalization performance, and to improve the overall performance of the classification system.

In integrated learning classification, the lower the classification strength of individuals and the lower correlation between individuals, the stronger the generalizability of the integrated learner and vice versa. So integration learning is simply the generation of base classifiers and the merging of base classifier.

3. ANALYSIS OF TEXT CLASSIFICATION TECHNIQUES

3.1 Tag Text Classification Techniques

The classification technology of text has been widely used in people's daily life after a long period of development. Especially in traditional classification problems, where each text can belong to only one label, the most similar

label can be determined based on the properties and content of the sample. But as the content of the information becomes more complex and richer, the form of the label is becoming more and more diverse. Data information is also increasingly complex, and a single label can no longer accurately describe text content and text properties, so it is often necessary to establish corresponding subsets of tags to express text content, which creates the problem of multi-label classification.

3.1.1 Tag relevance classification

Multi-label classification can be seen as a broader and more complex method of classification resulting from the extension of a single label classification. At the same time, multi-label classification is more detailed and complex for the interpretation of text, the main reason being the larger output space of the tags. In multi-label classification, the more a collection of labels, the larger the combination. However, in the classification process, label sets or label combinations with too large numbers have certain limitations in computation and classification efficiency and quality. Therefore, to be able to solve this problem better, we distinguish the label algorithm according to the correlation of the label, namely the first-step method, the second-step method and the higher-step method.

A first-step approach means that each label is treated independently. That is, the currently common method of label classification treats the classification task as multiple diclassification tasks, although the two tasks are independent of each other in the computation process. It can improve the accuracy and scientificity of classification to a certain extent, but it cannot solve the problem of uneven data distribution because it does not take into account the dependencies between labels, and some labels are still not computationally well classified.

The second-order method focuses more on the correlation between the two labels, but it cannot include the correlation of all the labels. In a sense, it is still essentially about obtaining the corresponding numerical value by focusing on the correlation between two labels, For the label of other relevance is still considered to be independent of each other, although compared with the first-order method has a certain improvement and optimization, but due to the high computational difficulty, it is difficult to calculate large-scale data information processing, corresponding learning problems.

3.2 Plane classification

The flat classification method can successfully complete the task of text classification using classic machine learning algorithms directly. But these same machine learning algorithms still have problems with data tilt and data sparseness when they encounter large-scale classification problems, resulting in a reduced ability to classify. The data bias is mainly due to the large gap between text categories, If one category is used as a positive sample and the number of negative samples far exceeds the positive sample, there is a data tilt, while data sparseness occurs because the length of the samples is different, resulting in a large number of short text vectors being too sparsely represented.

3.3 Hierarchical classification

For multi-labels, the large-scale classification needs need to be achieved through hierarchical classification. Information retrieval, as a basic requirement of the user, is the first factor to be considered before each classification method is calculated. Generally, information retrieval is based on the sorting and classification of a large number of information documents, and through the interrelationship between texts, a multi-level and multi-structure classification system is established, thereby further improving the retrieval speed of users. Compared to label classification and plane classification, the advantage of hierarchical classification is that it can greatly improve the accuracy of text classification, and it can improve the accurateness of text classes by establishing different levels of systems. In this process, the more relevant label categories can be grouped into a large category, and the large categories can then be differentiated, thereby achieving the purpose of categorizing the different levels of text content, and better achieving the requirements of classification and the scientificity and accuracy of classification.

4. CONCLUSION

Both text representation and text classification under deep learning require specific computational methods and classification methods, especially in hierarchical structures, and it is necessary to extract corresponding features from different classifications to solve corresponding classification problems. This provides a more efficient and

accurate text model and classification model for the classification of text. Therefore, in the context of deep learning, this paper analyzes in detail the current state of text representation and the method of text classification in order to better improve the efficiency and accuracy of text sorting and make text classification more scientific.

REFERENCES

[1] Q. Tian, D. Zou, Y. Han and X. Li, "A Business Intelligence Innovative Approach to Ad Recall: Cross-Attention Multi-Task Learning for Digital Advertising," 2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Shenzhen, China, 2025, pp. 1249-1253, doi: 10.1109/AINIT65432.2025.11035473.

[2] Wang, Zhiyuan, et al. "An Empirical Study on the Design and Optimization of an AI-Enhanced Intelligent Financial Risk Control System in the Context of Multinational Supply Chains." (2025).

[3] Xie, Y., Li, Z., Yin, Y., Wei, Z., Xu, G., & Luo, Y. (2024). Advancing Legal Citation Text Classification A Conv1D-Based Approach for Multi-Class Classification. Journal of Theory and Practice of Engineering Science, 4(02), 15–22. https://doi.org/10.53469/jtpes.2024.04(02).03

[4] Chen, Yinda, et al. "Generative text-guided 3d vision-language pretraining for unified medical image segmentation." arXiv preprint arXiv:2306.04811 (2023).

[5] Xu, Haoran. "CivicMorph: Generative Modeling for Public Space Form Development." (2025).

[6] Tu, Tongwei. "ProtoMind: Modeling Driven NAS and SIP Message Sequence Modeling for Smart Regression Detection." (2025).

[7] Xie, Minhui, and Boyan Liu. "InspectX: Optimizing Industrial Monitoring Systems via OpenCV and WebSocket for Real-Time Analysis." (2025).

[8] Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., Chen, C. and Wu, Z., 3D Vision-Language Gaussian Splatting. In The Thirteenth International Conference on Learning Representations.

[9] Wang, Y. (2025). Efficient Adverse Event Forecasting in Clinical Trials via Transformer-Augmented Survival Analysis.

[10] Liu, Jun, et al. "Toward adaptive large language models structured pruning via hybrid-grained weight importance assessment." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 18. 2025.

[11] Zhou, Dianyi. "Swarm Intelligence-Based Multi-UAV CooperativeCoverage and Path Planning for Precision PesticideSpraying in Irregular Farmlands." (2025).

[12] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Highly Reliable CI-JSO based Densely Connected Convolutional Networks Using Transfer Learning for Fault Diagnosis.

[13] Zhuang, R. (2025). Evolutionary Logic and Theoretical Construction of Real Estate Marketing Strategies under Digital Transformation. Economics and Management Innovation, 2(2), 117-124.

[14] Han, X., & Dou, X. (2025). User recommendation method integrating hierarchical graph attention network with multimodal knowledge graph. Frontiers in Neurorobotics, 19, 1587973.

[15] Yang, J. (2025, July). Identification Based on Prompt-Biomrc Model and Its Application in Intelligent Consultation. In Innovative Computing 2025, Volume 1: International Conference on Innovative Computing (Vol. 1440, p. 149). Springer Nature.