iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66731

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Enhanced Deep Learning Approaches for Text Classification: A Comprehensive Review

Divya Arvind Bansod[1], Prof. Rahul Bhandekar[2], Prof. Rahul Nawkhare[3]
[1]*PG Scholar, Department of Computer Science & Engineering, Wainganga College of Engineering & Management, Nagpur*
[2]*Project Guide, Department of Computer Science & Engineering, Wainganga College of Engineering & Management, Nagpur*
[3]*Head of Department, Department of Computer Science & Engineering, Wainganga College of Engineering & Management, Nagpur*

Abstract: ***Deep learning models have outperformed traditional machine learning approaches in a variety of text classification tasks, such as sentiment analysis, news categorization, question answering, and natural language inference. Text classification is a widely employed technique in natural language processing, encompassing applications like spam detection, news categorization, information retrieval, sentiment analysis, and intent assessment. Conventional text classifiers utilizing machine learning methods suffer from issues like sparse data, dimensionality challenges, and limited generalization capabilities. In contrast, deep learning-based classifiers address these shortcomings, eliminating the need for intricate feature extraction processes while offering enhanced learning capabilities and increased prediction accuracy. For instance, convolutional neural networks (CNNs). This article outlines the text classification process, with a specific focus on the deep learning models utilized in the context of text classification.***

Keywords: Deep learning model; Text classification; Natural Language Processing; CNN; RNN. etc.

I. INTRODUCTION

Artificial intelligence has seen rapid development in recent years, significantly transforming our lives. Natural Language Processing (NLP) is a captivating and challenging branch of artificial intelligence that encompasses syntactic and semantic analysis, information extraction, text classification, machine translation, information retrieval, and dialogue systems. Among these applications, text classification is notably widespread and serves various purposes, such as identifying spam, categorizing news articles, retrieving information, analyzing emotions, and recognizing intent. The text classification process comprises several stages: text preprocessing, text feature extraction, and the creation of a classification model. Initially, due to the specific nature of text structure, preprocessing is essential. This typically involves tasks like removing stop words and, in certain cases, handling language-specific requirements, such as word segmentation in Chinese text. Subsequently, feature engineering is applied to preprocessed text to extract essential features that capture the text's characteristics, establishing a connection between these features and the classification task. Lastly, a classification model is built, with a primary focus in this paper on leveraging deep learning techniques. Text classification undeniably stands as one of the most widely employed natural language processing technologies. Traditional text classifiers based on machine learning methods exhibit shortcomings like dealing with sparse data, managing high-dimensional feature spaces, and displaying limited generalization abilities. In contrast, classifiers based on deep learning networks significantly mitigate these issues, eliminating the need for intricate feature extraction processes while offering robust learning capabilities and improved prediction accuracy. For instance, the Convolutional Neural Network (CNN) serves as a prominent example of such deep learning architecture. As shown in fir.1, this paper delves into the text classification process, with a specific emphasis on the utilization of deep learning models.

Fig.1. Text Classification System

©IJRASET: All Rights are Reserved

II. OBJECTIVE

The objective of this paper is centered on the text classification process with a specific focus on employing deep learning models. This paper aims to:

1) Automatically ascertain the class or category to which a given piece of text belongs.
2) Address the challenge of overfitting and lengthy training times associated with CNN-based text classification models.
3) Explore methods to create a text classification system using Deep Learning Techniques.
4) Train models on synthesized data and improve sparsely populated datasets using CNN.
5) Enhance capabilities related to information extraction, summarization, and text retrieval.

III. LITERATURE SURVEY

In recent years, advancements in Internet technology have been remarkable, and the proliferation of mobile apps has led to increased utilization, resulting in a more complex landscape of online information. Certain online social apps have become deeply ingrained in people's lives, offering a wealth of information resources. The Internet has now become an integral part of daily life, leading to a significant surge in the volume of data people encounter online. This has spurred research in the direction of extracting valuable insights from this vast data.

Text classification (TC) [1] plays a crucial role in extracting valuable information from textual data. Its primary objective is to analyze textual features and determine the appropriate category for a given piece of text. As the field of text classification continues to evolve, its applications have grown considerably, spanning areas like text information filtering [2], public sentiment analysis [2], and search engines [3][4].

Junwei Ge et. al. 2020 , In this paper, a novel text classification algorithm is presented, which leverages the neural topic model ProdLDA and a convolutional neural network (CNN). To begin, text data is first represented using a word vector model. Next, the convolutional neural network is applied to capture fine-grained features from the high-dimensional text data, while the neural topic model ProdLDA is employed to extract potential topic-related features. A connection layer is then established to integrate these text features, and the final classification layer is processed. Simultaneously, a new method for introducing topic-related features during the feature extraction process is introduced. Empirical results demonstrate the algorithm's effectiveness in enhancing text classification performance.

YA Chen et. al. 2023, In this paper, the language used on social network platforms exhibits a structural front-to-back dependency, and directly converting Chinese text into vectors results in very high dimensionality, leading to lower accuracy in existing text classification methods. Consequently, this study introduces a deep learning model that combines an ultra-deep convolutional neural network (UDCNN) based on big data and a long short-term memory network (LSTM). The UDCNN's deep architecture is harnessed to extract features for text vector classification. The LSTM is employed to retain historical information, capturing the contextual dependencies within lengthy texts. Additionally, word embedding is incorporated to transform the text into lower-dimensional vectors. The study conducts experiments using the Sogou corpus from social network platforms and the University HowNet Chinese corpus. The research outcomes reveal that in comparison to models like CNN + rand and LSTM, the neural network deep learning hybrid model significantly enhances the accuracy of text classification.

Qi Wang et. al. 2021, In this research work, Text classification stands as a significant research area within natural language processing. Deep learning-based text classification offers enhanced efficiency and accuracy in comparison to traditional manual methods. Nevertheless, the learning process can be extensive and intricate. To aid the understanding of more researchers, this paper provides a comprehensive overview of deep learning for text classification. The initial section of this paper presents the preprocessing steps involved in text classification. Subsequently, the second part provides an in-depth exploration of various viable approaches to deep learning for text classification. The third part outlines the model testing methodology. In the fourth section, a summary and analysis of the strengths and weaknesses of these methods are presented, providing a foundational reference for future research.

Vidushi Garg et.al. 2020, In this paper, A novel technique is introduced for the offline detection of handwritten characters utilizing deep neural networks. In the current era, training deep neural networks has become more accessible due to the availability of extensive data and ongoing algorithmic advancements. Handwritten Text Recognition (HTR) serves as an automated means of transcribing documents through computer processes. HTR typically employs two primary approaches, hidden Markov models, and Artificial Neural Networks (ANNs). The proposed HTR system is built upon ANNs. Preprocessing methods are employed to enhance input images, simplifying the task for the classifier. These techniques encompass contrast normalization and data augmentation, effectively expanding the dataset size.

©IJRASET: All Rights are Reserved

The classifier consists of Convolutional Neural Network (CNN) layers, which extract features from the input image, and Recurrent Neural Network (RNN) layers to transmit information throughout the image. The RNN generates an output matrix containing a probability distribution over the characters at each image position. Decoding this matrix using the connectionist temporal classification operation produces the final text output.

Xiaojing Fan et.al. 2021, This study investigates text classification using deep learning techniques. To address concerns related to overfitting and the time-consuming nature of training CNN text classification models, a Sparse Dropout Convolutional Neural Network (SDCNN) model is developed. Experimental findings demonstrate that, when compared to the conventional CNN model, SDCNN exhibits superior classification performance. Specifically, SDCNN achieves a classification accuracy of 98.96% and a precision of 85.61%, underscoring its advantages in addressing text classification challenges.

Jingjing Cai et. al. 2018, In this paper, Text classification represents a highly prevalent technology in the field of natural language processing. Its applications encompass a wide array of tasks, including spam detection, news categorization, information retrieval, sentiment analysis, and intent recognition. Traditional text classifiers that rely on machine learning approaches exhibit limitations such as dealing with sparse data, experiencing dimensionality issues, and struggling with generalization. In contrast, classifiers built on deep learning networks effectively address these limitations. They circumvent the need for complex feature extraction processes and offer strong learning capabilities, resulting in improved prediction accuracy. One such example is the Convolutional Neural Network (CNN) [1]. This paper provides an overview of the text classification process, with a primary focus on the utilization of deep learning models in this context.

Muhammad Zulqarnain et. al. 2020, In this paper, Text classification serves as a foundational task in numerous domains within natural language processing (NLP), encompassing semantic classification of words, sentiment analysis, question answering, and dialog management. This study delves into an examination of three key deep learning architectures for text classification tasks: Deep Belief Neural (DBN), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). These three primary types of deep learning structures are extensively explored to address various classification objectives. DBNs exhibit remarkable learning capabilities for extracting highly distinctive features and are well-suited for general purposes. CNNs are considered proficient at discerning the positional relationships between various features, while RNNs excel at modeling sequential, long-term dependencies. This paper presents a systematic comparison of DBN, CNN, and RNN in the context of text classification tasks. Subsequently, research experiments are employed to showcase the results of these deep models. The primary objective of this paper is to furnish foundational insights into deep learning models and determine which model is most suitable for the specific task of text classification.

IV. ANALYSIS OF LITERATURE SURVEY

In prior studies, the supply of labeled data for text classification has been notably limited. This limitation stems from the sensitivity and time-consuming nature of data collection, requiring the involvement of pediatricians or experienced caregivers. Additionally, text classification involves intricate acoustic and prosodic phenomena associated with both short-term and long-term influences. In this paper, the goal is to develop an algorithm for the automatic detection of text classification. Text classification, as a concept, seeks to categorize a given document, with the outcome of the classification typically falling into binary or multiple categories. The fundamental stages of text classification encompass text preprocessing, text feature extraction, and the construction of a classification model.

V. RESEARCH METHODOLOGY

A. *Workflow Diagram*

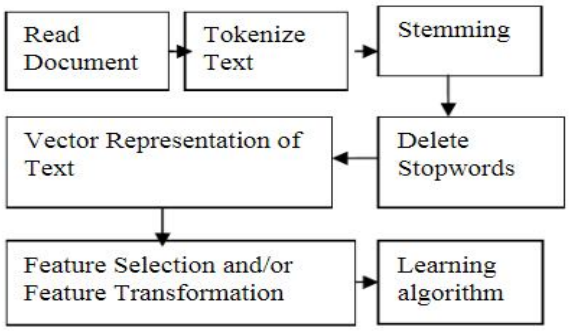


Fig.2. The workflow of a machine learning-based model for facial expression

B. Working

1) As shown in fig.2, Developing a document classifier is a task akin to various other Machine Learning challenges, with the primary concern revolving around document representation.
2) A distinctive aspect of text categorization is the potential for an extensive number of features, encompassing unique words or phrases, which can easily number in the tens of thousands. This poses significant challenges when employing advanced learning algorithms in text categorization, necessitating the use of dimension reduction methods.
3) Two alternatives exist: either selecting a subset of the original features or reshaping the features into novel forms, essentially creating new features as functions of the existing ones. Both approaches will be explored.
4) Subsequent to these stages, a Machine Learning algorithm can be employed. Certain algorithms have demonstrated superior performance in Text Classification tasks and are commonly employed, including Support Vector Machines.
5) This paper will offer a concise overview of recent adaptations to learning algorithms tailored for Text Classification applications. Various methods for assessing the performance of machine learning algorithms in Text Classification will also be discussed.

VI. ADVANTAGES

1) Deep learning architectures offer huge benefits for text classification because they perform at super high accuracy with lower-level engineering and computation.
2) The percentage of texts that were categorized with the correct tag.
3) Real-time analysis
4) Consistent criteria
5) Scalability.

VII. DIS- ADVANTAGES

1) System is complex.
2) Systems are costly to implement.

VIII. CONCLUSION

In this endeavor, we present and provide an overview of text classification preprocessing, the associated computational techniques, and testing methodologies. We also examine the strengths and weaknesses of relevant models. Notably, the CNN model has the ability to capture crucial textual content. It has gained significant popularity in long text classification models due to its high parallelizability. The model employs multiple channels, each utilizing filters of various sizes, along with Max pooling to select the most impactful and lower-dimensional high-dimensional classification features. Following this, a depth-based dropout mechanism is employed for extracting text features in the fully connected layer, ultimately yielding the classification results.

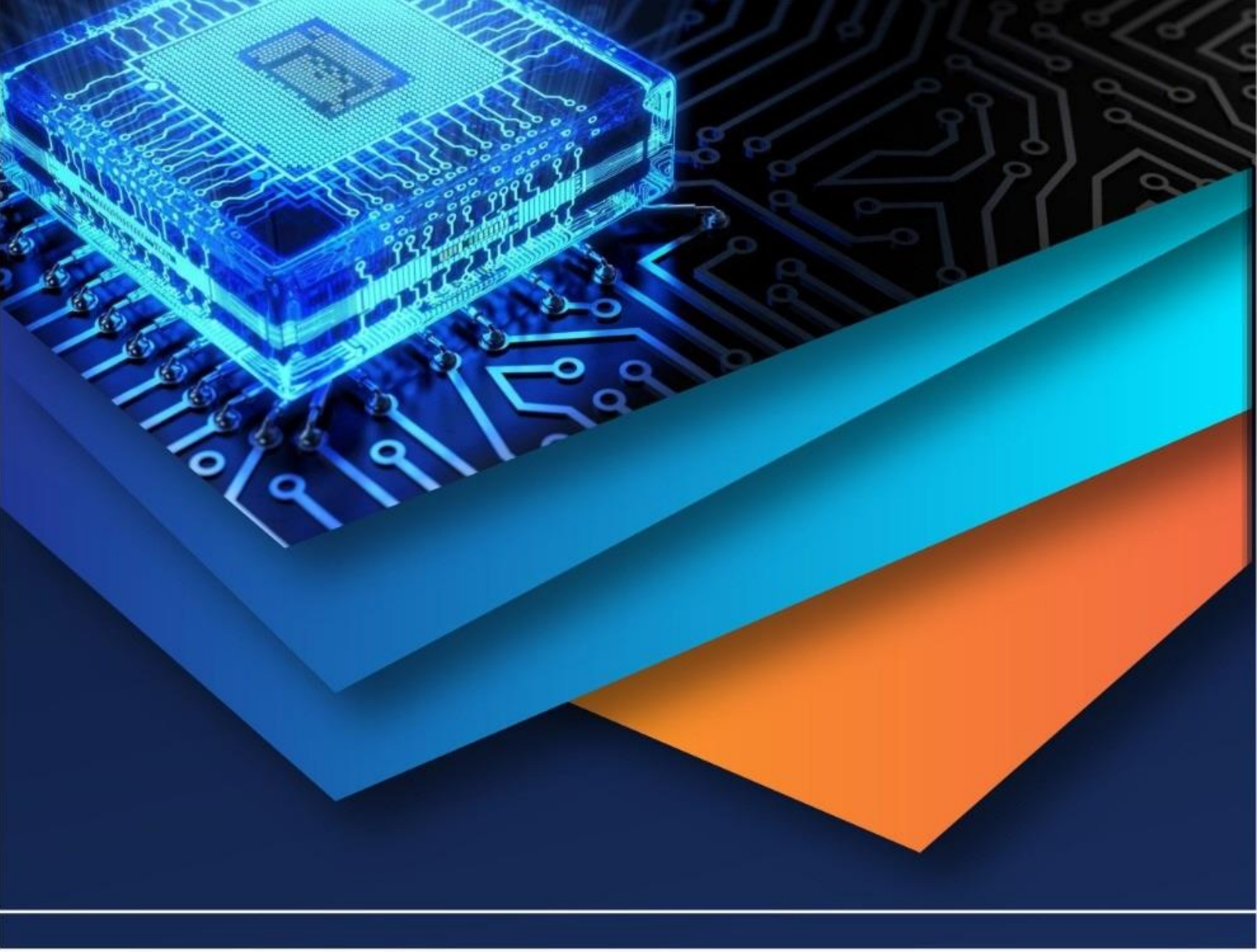
REFERENCES

[1] Kim Y. "Convolutional Neural Networks for Sentence Classification." Computer Science, 1408.5882 (2014).
[2] Kalchbrenner N, Grefenstette E, Blunsom P. "A convolutional neural network for modeling sentences." Computer science, 655-665 (2014).
[3] Fan H, Xia G S, Hu J, et al. "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery." Remote Sensing, 7(11): 14680-14707 (2015).
[4] Alfaro C, Cano-Montero J, J Gómez, et al. "A multi-stage method for content classification and opinion mining on weblog comments." Annals of Operations Research, 236(1): 197-213 (2016).
[5] Jiang M, Liang Y, Feng X, et al. "Text classification based on deep belief network and softmax regression." Neural Computing & Applications, 29, 61–70 (2018).
[6] Cao F, Chen B. "New architecture of deep recursive convolution networks for super-resolution." Knowledge-Based Systems, 178: 98-110 (2019).
[7] Cunha W, Canuto S, Viegas F, et al. "Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling." Information Processing & Management, 57(4): 102263 (2020).
[8] Yoon Kim, "Convolutional Neural Networks for Sentence Classification", EMNLP 2014, Part number 1of1, pp. 1746-1751, Aug. 2014.
[9] Li. Hui 1, Chen. Ping Hua, "Improved backtracking-forward algorithm for maximum matching Chinese word segmentation", Applied Mechanics and Materials, v 536-537, p 403-406, 2014.
[10] Liyi. Zhang, Yazi. Li, Jian. Meng, "Design of Chinese word segmentation system based on improved Chinese converse dictionary and reverse maximum matching", Lecture Notes in Computer Science, v 4256 LNCS, p 171-181, 2006.
[11] Gai. Rong Li, Gao. Fei Duan, Li Ming, Sun. Xiao Hui, Li. Hong Zheng, "Bidirectional maximal matching word segmentation algorithm with rules", Advanced Materials Research, v 926-930, p 3368-3372, 2014.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[12] Young. Tom, Hazarika. Devamanyu, Poria. Soujanya, Cambria. ErikRecent, "Trends in Deep Learning Based Natural Language Processing", IEEE Computational Intelligence Magazine, v 13, n 3, p 55-75, August 2018.
[13] Pengfei Liu, Xipeng Qiu, Xuanjing Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning", IJCAI 2016, May 2016.
[14] Luong. Minh-Thang, Pham. Hieu, Manning. Christopher D, "Effective Approaches to Attention-based Neural Machine Translation", Conference on Empirical Methods in Natural Language Processing, Part number: 1of1, Pages: 1412-1421, September 21, 2015.
[15] Sutskever. Ilya, Vinyals. Oriol, Le. Quoc V, "Sequence to sequence Learning with Neural Networks", 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014, Pages: 3104-3112.
[16] Lai. Siwei, Xu. Liheng, Liu. Kang, Zhao. Jun, "Recurrent convolutional neural networks for text classification", Proceedings of the 29th AAAI Conference on Artificial Intelligence, Volume: 3, Part number: 3of6, Pages: 2267-2273, June 1, 2015.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)