Hybrid CNN-RNN Based Multi-Class Scientific Article Classification

# Introduction

There are currently at least 114 million English research articles online, according to a recent estimate . Studies show that every 10 to 15 years, the quantity of academic papers doubles (Larsen et al. 2010). The difficulty of precisely identifying relevant research papers rises with the expansion of scholarly publications, especially when articles from many subject categories (SCs) are mixed together in a search engine's database. The results of a web search for a certain subject can provide thousands of papers on that subject. For example, a Google Scholar search for "NLP for text classifications" returns 60,300 relevant findings(Google Scholar). As a result, it is quite difficult for readers to identify the necessary and targeted information they want.

Additionally, this frequently occurs when the same search phrases are used across other research fields. For instance, Google Scholar's "neuron" search returns articles from both computer science and neuroscience[2]. When the search phrases contain acronyms, the search results may also come from many domains. In the case of "SNA," which stands for "Social Network Analysis," and "Systems Network Architecture," respectively, a search for "SNA" produces results both in the social sciences and computer sciences. Therefore, it is crucial to choose the appropriate keywords for particular papers so that readers may quickly identify the content they're looking for. Finding more readers for a research paper requires choosing the right keywords. A mechanism that allows writers to submit the abstracts of their publications to locate appropriate and related keywords should be put into place. More readers will therefore discover their renowned articles. For this reason, the journal committee should adopt a deep learning model that will assist the authors in locating appropriate keywords for their work, which will draw readers and enhance the journal's recognition. Therefore, accurate text classification is necessary.

The goal of text classification, often referred to as text categorization, is to classify or tag textual units like sentences, queries, paragraphs, and documents. Text classification is a classic problem in natural language processing. Rule-based methods and machine learning (data-driven) methods are two categories into which text classification techniques can be divided(Lee et al. 2016). Rule-based systems use a set of predefined criteria to categorize texts into different groups, whereas machine learning-based methods learn to categorize texts based on data observation. Different approaches have been developed for short text classification, Such as Recursive Neural Network (Luong et al. 2013) and Convolutional Neural Network [5][6][7]–[11]

By using Natural Language Processing(NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content. The initial step in the procedure is to preprocess the text. This includes eliminating unnecessary words, punctuation, and terminology that don't add much to the text's context. This is crucial for improving analytical outcomes and ensuring the consistency of the dataset(Kalra et al. 2017). The initial stage in training a machine learning NLP classifier is feature extraction, which involves converting each text into a numerical representation in the form of a vector. Bag of words is a popular strategy in which a vector indicates the frequency of a word in a predetermined lexicon of terms. The machine learning algorithm is then fed training data consisting of pairs of feature sets (vectors for each text sample) and tags (e.g., sports, politics) to generate a classification model. It can begin to produce accurate predictions once it has been taught with enough training examples. The same feature extractor is used to convert unseen text into feature sets that can be put into the classification model to provide predictions on tags (for example, sports and politics). Machine learning text categorization is typically far more accurate than rule-based systems, especially on difficult NLP classification problems. Furthermore, machine learning classifiers are easier to maintain, usually give more accurate results.

A paper's abstract typically has 200 words. The majority of them contain a number of unusual and distinguishing terms. This makes choosing the ideal dimension for feature extraction crucial. Setting a small word dimension could result in the loss of some distinct words, but setting a large word dimension has a bad impact on the performance as a whole. Setting the precise dimension is therefore necessary, which is one of the major challenges. In this study, we compared three distinct word embedding approaches, namely Keras, word2vec, and the GloVe embedding model, to determine which of these approaches produces the most accurate results for classifying research topics. Although the classification of short texts using ML and DL has seen widespread use in various industries, this technique is still relatively novel when applied to the classification of abstract keywords.

# Related Works

[13] examined a Deep Learning (DL) strategy based on a Convolutional Neural Network (CNN) for the hierarchical XMTC issue. Using the PubMed scientific article as a test case, HLSE approach was utilized to classify the text of a MeSH collection including 27,775 different classes organized as a hierarchical graph. Several embedding models (NLP, BioASQ, POS, fastText, and Dependency Tree) were employed, with NLP preprocessed with HLSE proving to be the most effective. [14] developed a unified hierarchical model for high-dimensionality In the medical literature. They introduced a novel hierarchical neural network method that works at both the sentence and document levels. BIGRU was used to represent sentences and documents. Seven techniques were assessed (including CNN, LSTM, RCNN, HAN, SVM, Fasttext, Logistic Regression, and Ac-BiLSTM), and their approach outperformed other baselines. [15] developed Deep pyramid convolution neural networks for text categorization using datasets such as AG, Sogoru, Dbpedia, Yelp.p, Yelp.f, Ama.f, and Ama.p, and their model surpassed the previous best models on six benchmark datasets.(Fouzi et al. 2001)Used Artificial Neural Network (ANN) for test classification of hadiths and found that their SVD model is 52.4s faster than MPL NN, which gets the best performance of only 52%. [17] proposed the Recursive Data-Pruning Convolutional Neural Network (ReDP-CNN), which removes unnecessary words from a dataset without altering the network structure. 2.179 M samples were acquired from AGNews, Yahoo, Yelp Full, Yelp 10000, and trained, and 107k samples were tested. It was observed that the Recursive Data Pruning neural network outperformed the traditional neural network system.

[18]proposed a Multi-Stream Neural Network Method that works well in both English and Chinese corpora. The macro F1 score of Reuters 21578-R8 has increased to 95.02 percent, representing a 10.16 percent improvement, and the macro F1 score of Fudan University corpus has increased to 85.03 percent, representing an 8.75 percent improvement. The proposed method outperformed GRUs, Multinomial NB, SVM, Bayes Network, and KNN in Reuters R8 (97.67%) and Reuters R52 (94.35%) accuracy.(Nam et al. 2014) presented a neural network-based multi-label classification framework (which can be evaluated in two groups of measures: bipartition and ranking) that outperformed BP-MLL in predictive performance, computational complexity, and convergence speed. The experiment used nearly 1.2 million research documents from Reuters-21578, RCV1-v2, EUR-Lex, Delicious, Bookmarks, and the German Education Index. To avoid a computational bottleneck caused by many labels, the 1000 most common labels were chosen from 50,000.

# Methodology

## Data Collection

## Text Preprocessing

## Text Representation

## Model Construction

## Evaluation Metrics

# Result Analysis and Discussion

# Conclusion

# Reference