**Text Classification Based on BERT and Convolutional Neural Networks**

Fatemeh Kamyabkalantari

Student number: 501087713

Ryerson University

CIND 820: Big Data Analytics Project

Dr. Tamer Abdou

Winter 2022

Contents

[Abstract 5](#_Toc95727826)

[Literature Review 6](#_Toc95727827)

[Proposed BERT-CNN Deep Learning Model 8](#_Toc95727828)

[Database 9](#_Toc95727829)

[Reference 11](#_Toc95727830)

**List of Figures**

Figure 1:Architecture of the BERT-CNN model 8

Figure 2:Architecture the Convolutional Neural Network 8

**List of Tables**

Table 1:Summary of Yahoo Answers Train Dataset 9

Table 2:Summary of Yahoo Answers Test Dataset 9

Table 3:Summary of Yahoo IMDB Dataset 10

Table 4:Summary of Yahoo AG’s Train news dataset 10

Table 5:Summary of Yahoo AG’s Train news dataset 10

# Abstract

Recently, deep learning has achieved impressive success in Natural Language Processing tasks. BERT is one of the remarkably rewarding deep learning models employed in various NLP classification tasks such as topic detection, question answering, and sentiment analysis. This study presents a hybrid technique of combining the BERT which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers and the **C**onvolutional **N**eural **N**etwork (CNN). This model embraces the BERT to train the word semantic representation language model. According to the word context, the semantic vector is dynamically generated and then placed into the CNN to predict the output.

In this study, we employ the most common BERT-Base model including ALBERT, RoBERTa, DistilBERT,BERT-Large in order to compare their performance to the hybrid model in which, the BERT embedding method uses to transform all the texts into numerical vectors. Then, the convolutional neural network will be applied to these numerical vectors to classify these texts into their appropriate classes. For the illustration, in this paper, we use the AG’s news, the IMDB movie reviews, and Yahoo! Answers datasets to perform our experiments, showing that the performance of the convolutional neural network model is better than the performances of the BERT-Base models.

**Keywords**: BERT, Bidirectional Encoder Representations from Transformers, Convolutional Neural Network, CNN, Text Classification.

# Literature Review

Text classification is a classical problem in machine learning and deep learning research areas. The goal of text categorization is to classify documents into predefined categories, which is one of the traditional tasks in natural language processing (NLP). This task has a wide range of applications, such as topic classification, sentiment analysis, and question answering in which the important intermediate step is text representation.

Over the last decade, word representation has revolutionized the field of NLP. The traditional method uses bag-of-words (BoW) to represent documents. Although the BoW model is effective, this method only regards documents as a collection of words, each word in the text is independent and ignores the variability of word meaning in different linguistic contexts (e.g., polysemy) as well as word order, grammar, and syntactic structure. For instance, for these two sentences: “He deposited his money in this *bank*.” and “His soldiers were arrayed along the river *bank*.”, the word bank has different meanings in different contexts.

In order to effectively use word order information for text classification, researchers generally use n-grams method. In n-gram models, (Miller&Selfridge,1950) the representation of a word is as a string of letters, or an index in a vocabulary list. Vector Semantic is the standard way to represent word meaning in which represents a target word as a vector with dimensions corresponding either to the documents in a large collection, the term-document matrix (Osgood, 1957), or to the counts of words in some neighboring window, the term-term matrix (Joos,1950; Harris,1954; Firth,1957). However, vectors are sparse and long since most values are zero and it is difficult to apply large n-grams in practice.

After that, the method of applying neural networks in text categorization has proved to exceed the traditional method. Recurrent neural networks (RNN) and convolutional neural networks (CNN) are two commonly used text representation networks. RNN can model the entire sequence and capture long-term dependencies. However, RNN sometimes ignores the key information in the text when modeling the whole sequence. By contrast, CNN can extract local feature information well and has positional invariance. Compared with traditional classifiers, neural networks provide greater expressive power and produce better performance. However, most previous methods apply shallow representations and use static word representations that represent the words by a single vector. Therefore, they fail to capture higher-level information like anaphora, long-term dependencies, agreement, negation. The shortcoming is addressed by the introduction of contextualizedword embeddings.

Contextualized word embeddings map type-level representations to token-level representations as a function of the linguistic context (McCann et al., 2017). They are widely used in NLP, constituting the semantic backbone of pre-trained language models (PLMs) such as ELMo (Peters et al., 2018a), BERT (Devlinet al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), and T5 (Raffel et al., 2020). Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder,2018).

BERT (Devlin et al., 2019), as a contextualized word embedding and pre-trained language model, is a conceptually simple and empirically powerful representation for learning dynamic contextual embedding. It is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. Also, it can be used for a variety of language tasks by adding a small layer to the core model. In the other words, the variants of the BERT model have been developed to cater to different types of NLP-based systems such as classification, question answering.

Moreover, BERT alleviates unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor,1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse the left and the right context, which allows us to pre-train a deep bidirectional Transformer.

In addition to the masked language model, “Next Sentence Prediction” (NSP) (Jernite et al. 2017; Logeswaran and Lee,2018) task is used that jointly pre trains text-pair representations which applied more specifically to Question Answering (QA) that are based on understanding the relationshipbetween two sentences, which is not directly captured by language modeling. In order to train a model that understands sentence relationships, the pre-train for a binarized *next sentence prediction* task is generated from any monolingual corpus. Specifically, when choosing the sentences, A and B for each pretraining example, 50% of the time B is the actual next sentence that follows A (labeled as IsNext), and 50% of the time it is a random sentence from the corpus (labeled as NotNext). Despite its simplicity, very beneficial to QA.

Although BERT has achieved amazing results in many natural language understanding (NLU) tasks, the existing text classification methods have some limitations in text feature extraction. Because these methods cannot model the local and global structure features in the text. In this paper, we investigate how to maximize the utilization of BERT for the text classification task by incorporating contextualized language models (BERT) into a shallow convolutional neural network for the document classification task. In the other words, we employ BERT to extract semantic features then we use CNN to extract the high-level features finally, the max-pooling layer retains the most important features for text classification.

In this study, we have two sections. First, in terms of baselines methods, we extensively inspect various BERT-Based fine-tuning models including ALBERT, RoBERTa, DistilBERT,BERT-Large on different text classification tasks such as Question Classification, Sentiment Analysis, Topic Classification. As a brief introduction to the BERT-Based model, AlBERTA is mainly involved in increasing the training speed of BERT and lower memory consumption. it performs better in multiple classification tasks and uses a very low number of parameters while doing sentiment analysis. Also, RoBERTa (Robustly optimized BERT) has slight advancements in BERT including training their model with more data and larger batch size, eliminating the next sentence prediction factor, having larger sequences, and making changes in masking pattern. Furthermore, DistilBERT is a lighter, fast, smaller, and cheap version of BERT with a size reduction of 40% with 60% more speed and 97%understanding of language capabilities. Lastly, BERT-large works similarly as BERT-base does, but it has a larger size than BERT-base. It is more expensive than BERT-base as it takes more time for computation and is applicable to large datasets.

In the second part, we will inspect finetuning models by adding simple fully connected Convolutional Neural Network (CNN) layers. Our objective is ending up with a general outperforming model as well as identify the most useful output from BERT’s layers. More specifically, for classification tasks, BERT takes the final hidden state of the first token [CLS] as the representation of the whole sequence. The most significant difference between our proposed method and the BERT-Base models is that we use the hidden state of all the final outputs of the BERT as the contextualized word vector. Then we use CNN to extract the high-level features, finally, the max-pooling layer retains the most important features for text classification.

## Proposed BERT-CNN Deep Learning Model

The proposed model, BERT-CNN deep learning model, has been shown. The BERT-CNN primarily consists of three main components: 1) Preprocessing the data. 2) BERT base model, in which the text was passed through 12 layers of self-attention to obtain the contextual vector representation. 3) CNN, which is used as a classifier.

Diagram

Description automatically generated with medium confidence

Figure 1:Architecture of the BERT-CNN model

Also, The CNN’s overall structure is formed of a Convolutional layer, Max pooling and fully connected layer are shown in Figure 2. the pre-training BERT model is used to learn the word and sentence embedding. The embedding vector is the CNN input subsequently extracted from BERT.

Diagram

Description automatically generated

Figure 2:Architecture the Convolutional Neural Network

## Database

We evaluate our approach on three widely studied datasets. These datasets have varying numbers of documents and varying document lengths, covering common text classification tasks including sentiment analysis, question classification, and topic classification. Following is a brief description of three types of text classification and databases which are applied in this experiment.

Question classification

Question answering is the task of answering user questions employing NLP and information retrieval techniques. To evaluate our model for this task, we used Yahoo! Answers (Zhang et al.,2015) dataset including almost1.5M questions labeled with 10 largest main categories as classes. Each class has 140k training samples and 6k test samples. There are 4 columns, corresponding to class index (1 to 10), question title, question content, and best answer. The list of classes includes Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government. Yahoo! Answers dataset is a big dataset with train and test samples.

Text

Description automatically generated

Table 1: Summary of Yahoo Answers Train Dataset

Text

Description automatically generated

Table 2: Summary of Yahoo Answers Test Dataset

Sentiment Analysis

Sentiment analysis is the classification of the user’s text into emotions (positive, negative) using text analysis and NLP techniques. For sentiment analysis, we used the IMDB dataset (Maaset al., 2011) including 50K movie reviews with two labels. These 50k short texts belong to 2 classes which are 1 (positive/good review) and 0 (negative/bad review).

Text, letter

Description automatically generated

Table 3:Summary of Yahoo IMDB Dataset

Topic Classification

Topic classification is another NLP application that aims to find the topic of a given  
document or text. To target this task in our experiments, we used AG’s news dataset (Zhang et al., 2015) with more than 120K samples assigned to 4 labels. Each class has 25% train and test samples.

Text

Description automatically generated

Table 4:Summary of Yahoo AG’s Train news dataset

Text

Description automatically generated

Table 5:Summary of Yahoo AG’s Train news dataset

Up to know, all the above-mentioned steps are in the following GitHub address.

GitHub website: <https://github.com/Fatemeh1717/Ryerson-University/tree/main/CIND%20820>

I would like to say thank you to Dr. Tamer. The suggested websites Mendeley, and Overleaf helped me to get more insights into preparing the reference and the style of document.

# Reference

* Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1
* Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11856 LNAI. <https://doi.org/10.1007/978-3-030-32381-3_16>
* Lehečka, J., Švec, J., Ircing, P., & Šmídl, L. (2020). Adjusting bert’s pooling layer for large-scale multi-label text classification. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12284 LNAI. https://doi.org/10.1007/978-3-030-58323-1\_23
* Tran, L., Pham, L., Tran, T., & Mai, A. (2021). Text classification problems via BERT embedding method and graph convolutional neural network. International Conference on Advanced Technologies for Communications, 2021-October.

<https://doi.org/10.1109/ATC52653.2021.9598337>

* Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., & Dahou, A. (2021). Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. Information (Switzerland), 12(2). https://doi.org/10.3390/info12020052
* Mohammadi, S., & Chapon, M. (2020). Investigating the Performance of Fine-Tuned Text Classification Models Based-on Bert. Proceedings - 2020 IEEE 22nd International Conference on High Performance Computing and Communications, IEEE 18th International Conference on Smart City and IEEE 6th International Conference on Data Science and Systems, HPCC-SmartCity-DSS 2020. https://doi.org/10.1109/HPCC-SmartCity-DSS50907.2020.00162
* Yang, Y., & Cui, X. (2021). Bert-enhanced text graph neural network for classification. Entropy, 23(11). https://doi.org/10.3390/e23111536
* Lopez-Garcia, G., Jerez, J. M., Ribelles, N., Alba, E., & Veredas, F. J. (2021). Transformers for Clinical Coding in Spanish. IEEE Access, 9.

<https://doi.org/10.1109/ACCESS.2021.3080085>

* Daniel, J., & Martin, J. H. (2020). Speech and Language Processing A.1 Markov Chains. In Speech and Language Processing.