**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

**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.

**Reference**

**Database:**

<https://www.kaggle.com/Cornell-University/arxiv>

<https://www.cse.msu.edu/~tangjili/trust.html>

<https://files.pushshift.io/reddit/comments>

<https://www.yelp.com/dataset/download>

**Book chapters**

Jurafsky and Martin. Speech and Language Processing (ed. 2017; draft).  
Chapters 3, 6, and 11, web.stanford.edu/~jurafsky/slp3/

**Paper:**

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019.BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT.*

Fano, R. M. 1961. Transmission of Information: A Statistical Theory of Communications. MIT Press

Firth, J. R. 1957. A synopsis of linguistic theory 1930– 1955. In *Studies in Linguistic Analysis*. Philological Society. Reprinted in Palmer, F. (ed.) 1968. Selected Papers of J. R. Firth. Longman, Harlow

Harris, Z. S. 1954. Distributional structure. *Word*, 10:146–162. Reprinted in J. Fodor and J. Katz, *The Structure of Language*, Prentice Hall, 1964 and in Z. S. Harris, *Papers in Structural and Transformational Linguistics*, Reidel, 1970, 775–794.

Joos, M. 1950. Description of language design. JASA, 22:701–708

Miller, G. A. and J. A. Selfridge. 1950. Verbal context and the recall of meaningful material. *American Journal of Psychology*, 63:176–185.

Mikolov, T., K. Chen, G. S. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. *ICLR 2013*.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

**Videos**

Manning on word2vec, youtube.com/watch?v=ERibwqs9p38

BERT and Other Pre-trained Language Models, youtube.com/watch?v=knTc-NQSjKA

Contextual Word Embeddings, youtube.com/watch?v=S-CspeZ8FHc

Introduction and Word Vectors, youtube.com/watch?v=8rXD5-xhemo