Comparison of KoBERT and BERT for Emotion Classification of Healthcare Text Data

Mose Gu  
Department of Computer Science and Engineering  
Sungkyunkwan University  
Suwon, Republic of Korea  
mo0415@skku.edu

Jaehoon (Paul) Jeong  
Department of Computer Science and Engineering  
Sungkyunkwan University  
Suwon, Republic of Korea  
pauljeong@skku.edu

Abstract—In recent times, the rapid progress of digital technology has led to a substantial increase in the popularity of digital health. Identifying depression, which is a prevalent mental illness, is crucial in digital healthcare to prevent further harm and provide timely support. This study proposes an AI model that automates the identification of depressive patients. By leveraging Natural Language Processing (NLP) and pre-trained language models like BERT, we aim to classify emotions into six categories. Training the model requires a Korean emotional conversation corpus, which we obtain through crowd-sourcing and Al-Hub’s user case studies. To extend the applicability to English-speaking countries, we plan to translate the Korean corpus using the Google Translation API and fine-tune the BERT model with English data. The feasibility of the English model was evaluated by comparing the performance of KoBERT and BERT in emotion understanding. The findings will offer valuable insights into these models’ efficacy and contribute to the field of emotion classification.

Index Terms—AI, Deep Learning, NLP, Digital Health

I. INTRODUCTION

The rapid advancement in digital technology has led to the widespread application of digital health, encompassing various healthcare services such as prescription, disease prevention, and counseling. One significant area of interest in digital healthcare is the early identification of depression, which is a prevalent mental illness. Early identification can prevent further harm and enable prompt intervention for individuals suffering from depression. In this paper, we propose an artificial intelligence (AI) model that automates the identification of depressive patients.

Natural Language Processing (NLP) has made significant progress, primarily driven by transformer-based models such as the Transformer architecture [1]. These models have demonstrated remarkable capabilities in translation and decoding tasks. Deep learning-based language models have evolved to understand and interpret the context of complex and lengthy sentences. Pre-trained models, when fine-tuned with large amounts of data, can achieve high classification performance. Among these models, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a state-of-the-art model and serves as a baseline for numerous studies.

We utilize BERT, a renowned NLP model, to learn and classify human emotions into six categories. Although this study already has precedents [2], the importance of the data was noted nonetheless. Our model is trained on a corpus of 270,000 emotional conversation texts collected through crowd-sourcing, targeting 1,500 ordinary individuals. The emotional conversation data used for training is not readily available through web crawling, thus requiring direct production. To overcome this challenge, we leverage data from user case studies conducted by the Korean AI lab called Alhub. To process the Korean emotional conversation corpus, we employ KoBERT, which is a BERT model pre-trained specifically with Korean data. However, the current scenario is limited to classifying Korean input, while depression is a global issue. Therefore, it is crucial to extend the applicability of our approach to English-speaking countries.

To address this, we propose a methodology that involves translating a large-scale Korean emotional conversation corpus using the Google Translation API. Subsequently, we fine-tune the BERT model, which is pre-trained with a substantial English dataset, to classify emotions in English text and compare the result with the KoBERT model. This approach allows us to classify emotions with English input and from an evaluation perspective, this provides an opportunity to compare the performance with KoBERT and BERT.

The following is a summary of this paper’s significant contributions.

- (AI) Datasets for English learning depression validation models: We have collected a crowd-sourced dataset that can be trained for an emotion classification task of depression diagnosis.
- Evaluation of BERT and KoBERT: We examined and investigated the two state-of-the-art models. Through an analysis of their classification performance on specific tasks, we have contributed toward evaluating the models that excel in different emotional classifications within our framework.

The rest of this document is structured in a subsequent manner. Section II outlines the context and examination of our study. In section III, an outline of the suggested framework is presented, accompanied by an elucidation of its constituents and execution. Performance evaluation in detail, including
model experimentation results, is demonstrated in section IV. Section V offers concluding remarks for this paper, as well as a glimpse into future prospects.

II. RELATED WORK

A. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary (NLP) model introduced by Google in 2018. It is built upon the transformer architecture, specifically the "attention is all you need" concept [1], to bring about significant advancement in understanding the context of language. BERT’s significance lies in its ability to deeply comprehend the contextual relationships of words in a sentence, resulting in state-of-the-art performance across various NLP tasks.

The transformer architecture, introduced in the paper "Attention Is All You Need" [1] by Vaswani et al. in 2017, addressed the limitations of recurrent and convolutional neural networks in capturing long-range dependencies in sequences. The primary breakthrough of the transformer lies in its self-attention mechanism, enabling the model to assess the significance of distinct words within a sequence as it handles each individual word. This enables the model to capture relationships between distant words, resulting in a more comprehensive understanding of context [1].

BERT takes this self-attention mechanism a step further by processing a sentence in both directions (left-to-right direction and right-to-left direction), allowing it to consider the entire context for each word. Traditional models like Long Short Term Memory (LSTM) process sequences sequentially [3], which means they cannot access future words during prediction. BERT’s bidirectional approach enables it to leverage information from both preceding and succeeding words, resulting in a richer contextual understanding [4].

To train BERT, a technique called the "masked language model" [4] is employed. During training, a certain number of words in a sentence are substituted randomly with masked tokens. The model is then tasked to forecast these masked tokens based on the neighboring tokens. Additionally, it trains to predict whether consequent sentences are consecutive or not. This dual training objective fine-tunes BERT to comprehend both individual-word meanings and sentence-level relationships [4].

The significance of BERT lies in its ability to perform a variety of NLP tasks with a single pre-trained model, eliminating the need for task-specific feature engineering and separate models. This is achieved by fine-tuning BERT on specific tasks, like text classification, named entity recognition, question-answering, and so on. The model’s pre-trained contextual understanding of language significantly boosts its performance on these tasks.

B. KoBERT

KoBERT was developed by SKT, which is a Korean telecommunications company, to overcome the Korean language performance limitation of the existing BERT. Composed of the same Transformer Encoder as BERT, KoBERT was trained with a large-scale corpus of millions of Korean sentences collected from Wikipedia and Korean news. KoBERT applied a data-based tokenization technique to reflect the irregular language change characteristics of Korean and achieved a performance improvement of more than 2.6% using only 27% of the tokens compared to the existing BERT [5].

III. DESIGN AND IMPLEMENTATION

A. Architecture for Emotion classification diagnosis

The architecture of our framework is depicted in Fig. 1. This architecture is made up of six components: A Korean
emotional conversation corpus consists of 6 classes for fine-tuning the architecture, a preprocessing task that omits unnecessary data and separates the corpus to be used for learning by separately tying the mapped labels together, the translator that translates the Korean corpus into English using Google Translation API, two SOTA models (i.e., BERT and KoBERT). An evaluation metric which shows training performance, is the F1 score for the test dataset that reveals the model’s potential.

There are two steps in this scheme: Fine-tuning and evaluation. Both the original data and the translated corpus data are divided into various tokens by a tokenizer and entered into the model throughout the fine-tuning process. Models whose parameters have been initialized after training is already complete are the ones getting input at this moment. Following each token operation through the model and entry with a ground truth label in the classification layer, the input example is written as a specific logit. The classifier layer outputs the input example as a six-dimensional linear vector, and the linear transformation layer’s final value with the highest logit is then subjected to supervised learning by computing the cost with its label. The model adjusts a number of previously taught parameters on the input data during training when error backpropagation is carried out using supervised learning. This approach advances by using training data and validation data, and it is adaptable by looking at validation accuracy.

In the evaluation procedure, the model completes fine-tuning with training data and validation data and receives test data to produce final output values. We calculate these output values as F1 scores and output them as a table in the form of a matrix to observe the results as shown in Figs. 4c and 4d in Section IV.

B. Components for Implementation

This subsection describes the components of the proposed architecture and its implementation. **Psychological counseling data**: The data collected through crowdsourcing is a dataset built through psychological counseling. We split the training set into 80% of the total dataset, and the validation and test datasets into 10% each.
**IV. Performance Evaluation**

A summary of the performance of emotion classification can be found in this section. The experiment continued by producing the model accuracy and F1 score matrix using identically sized data samples. Keep in mind that all of the hyperparameters are the same, including learning rate, optimizer, and loss function.

The displayed results shown in Figs. 4a and 4b is the classification accuracy’s fine-tuning during the model’s training and validation phases utilizing 6 classes of emotions. As a result of observation, it was confirmed that the acuity of KoBERT was higher than that of BERT. Our concern is validation accuracy, and we can see that KoBERT is up about 5% from BERT. However, in the case of BERT, it can be seen that the performance does not improve from 3 epochs. On the other hand, in the case of KoBERT, it was seen that the performance graph of KoBERT was lower at a glance, showing an unstable upward line compared to BERT. It is judged that the parameter of KoBERT is more likely to be overfitted than the parameter of BERT.

Figs. 4c and 4d show through the F1 score matrix that KoBERT is a model that learns the context representation better than BERT. Although both models recorded high F1 scores, it was found that KoBERT was more accurate than BERT for the emotion classification task. This suggests that KoBERT has higher fine-tuning performance and token understanding than BERT. However, since the performance of BERT is not so bad...
compared with KoBERT, BERT can be used to the emotion classification for English text in depression diagnosis.

V. CONCLUSION

We evaluated and contrasted the model power of the two types of NLP models (i.e., BERT and KoBERT) understood the emotional context. Through experiments, we demonstrated that the framework-translated English fine-tuning task of BERT in emotion classification was slightly lower than KoBERT, which is a Korean NLP Model, but sufficiently reasonable. In summary, crowd-sourced Korean data proved to be effective in fine-tuning activities via a translation framework. To discuss future work, we could observe that the Happy class has an F1 score of over 80%, while the other classes have relatively low percentages due to the similarity in negative sentiment contexts. The performance improvement through advanced research in clustering in dimensions of negative sentiment context seems promising in future work.

ACKNOWLEDGMENTS

This work was supported by the IITP grant funded by the Korea government, Ministry of Science and ICT (MSIT) (No. 2022-0-01199, Regional strategic industry convergence security core talent training business). This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-01015, Development of Candidate Element Technology for Intelligent 6G Mobile Core Network). Note that Jaehoon (Paul) Jeong is the corresponding author.

REFERENCES