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
rna0415@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 AI-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 AIhub. 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

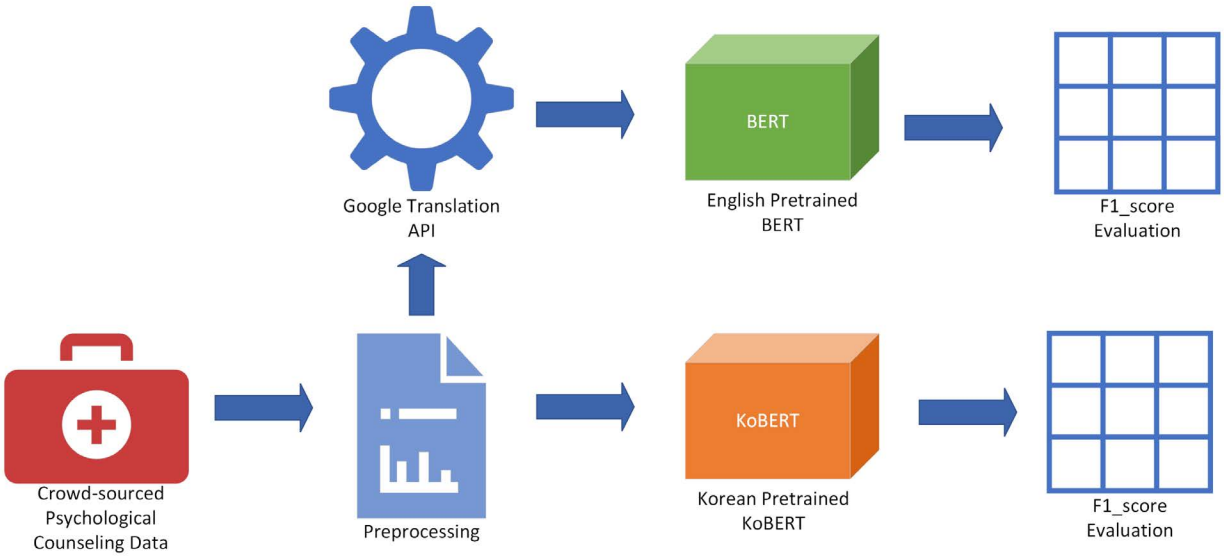


Fig. 1: An emotion classification comparison framework for BERT and KoBERT

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

1	감정_대분류	사람문장1
5087	상처	남편이 사귈 때랑 결혼하고 나서의 모습이 너무 달라.
5088	기쁨	나 직장에 합격했어!
5089	분노	난 당구에 관심도 없는데 남자 친구가 자꾸 같이 하자고 해서 짜증 나.
5090	분노	회사 화장실에서 우연히 내 험담을 듣게 되었어. 막상 들어보니 화가 치밀어 오르네.
5091	슬픔	오늘 남자친구 때문에 눈물이 났어.
5092	당황	자식들이 건강관리를 엉터리로 해.
5093	상처	이번 시험 결과 반에서 늘 상위권이었는데 성적이 떨어져 충격받았어.
5094	불안	나이가 들어서인지 조금이라도 찬 것을 먹으면 잇몸이 시려서 불안해.
5095	상처	오늘 저녁때 일찍 갈 건데 맛있는 반찬 있느냐고 물었더니 집사람이 짜증을 내는 거야.
5096	분노	친구에게 빌려준 돈을 받지 못하고 있어.
5097	분노	남편이 나이가 들면서 아픈 곳만 더 생겨서 속상해. 그런 남편을 보면 한심해서 툴툴거리게 돼.
5098	당황	나는 내 남자친구가 부끄러워. 외모도 직업도 다.
5099	불안	뭔가 새로운 일에 도전하는 건 너무 무서워.
5100	당황	형이랑 목욕탕에 갔는데 왜소한 내 체구랑 비교돼서 몸을 가렸어.
5101	슬픔	요즘 금리가 너무 낮아 슬퍼.
5102	슬픔	친구들과 싸움이 없는 날이 하루도 없네.
5103	분노	아파트에 불이 크게 나서 당분간 창고에서 지내게 됐어. 너무 화가 나.
5104	당황	취업이 안 되니 부모님에게 손만 벌리는 것 같아서 죄책감이 들어.
5105	불안	저번에 냄비를 태워서 엄마한테 혼나고 요즘 요리할 땐 최대한 조심하고 있어.
5106	기쁨	오늘 발표시간에 내 차례가 되니 눈앞이 캄캄해지는 거 있지.
5107	분노	직장에서 부당하게 해고를 당한 것 같아서 너무 화가 나.

Fig. 2: Preprocessed Korean dataset

1	data	label
40832	I want to get married, but I don't have enough money, so I'm going to ask my parents for help. I feel guilty that my	1
40833	I'm seriously thinking about my career path. embarrassed	1
40834	Yesterday my children came home and fought in front of me.	5
40835	I really hate my mom.	1
40836	I'm worried I'll have a miscarriage because of work.	3
40837	I guess it's something a little simpler. Work stress these days is no joke.	4
40838	My friend's father passed away. He went to the funeral yesterday. It's really sad.	5
40839	I am grateful to my parents these days.	0
40840	I think I have a problem with interpersonal relationships.	2
40841	When I see new employees coming in one by one these days, I feel a sense of wonder.	3
40842	I'm worried that the pay won't go up no matter how hard I work.	3
40843	It's good that the wages are normal these days in the Corona era.	3
40844	I haven't met the people of the fortress, so I feel lonely.	1
40845	Now that I am preparing for retirement, my heart aches with all kinds of thoughts.	5
40846	I told my wife not to make the food salty, but I get angry because I keep making it salty.	2
40847	It's disappointing that the father you trusted started gambling again without the knowledge of the family.	5
40848	I'm sad because even if I die, my family won't be surprised.	4
40849	It's annoying to meet my friends these days.	2
40850	I want to get out of my life as a non-regular worker, but I couldn't convert to a full-time job this time as well.	4
40851	Mom and Dad want to break up, but I'm upset that the reason they live is because of me.	4
40852	I had a fight with my girlfriend today and I'm so mad!	2
40853	It is disappointing that the severance pay is less than expected.	5

Fig. 3: Translated dataset

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.

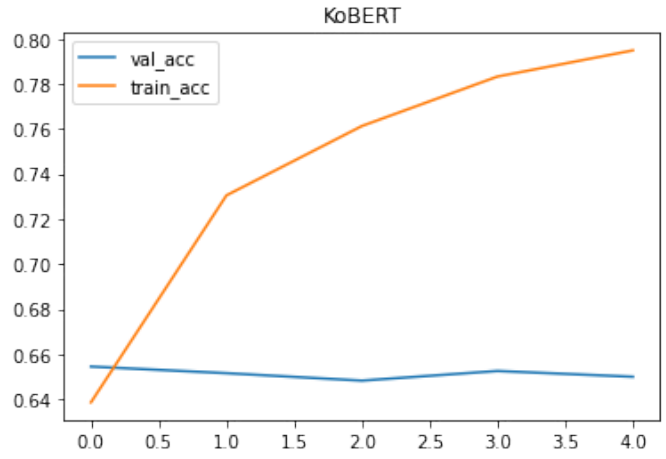
Psychological counseling data: The data collected through crowdsourcing is a dataset built through psychological counseling. We split the training set into 80% of the total dataset, and the validation and test datasets into 10% each.

Preprocessing: As shown in Fig. 2, the dataset created after preprocessing consists of more than 40000 contexts and corresponding emotion labels.

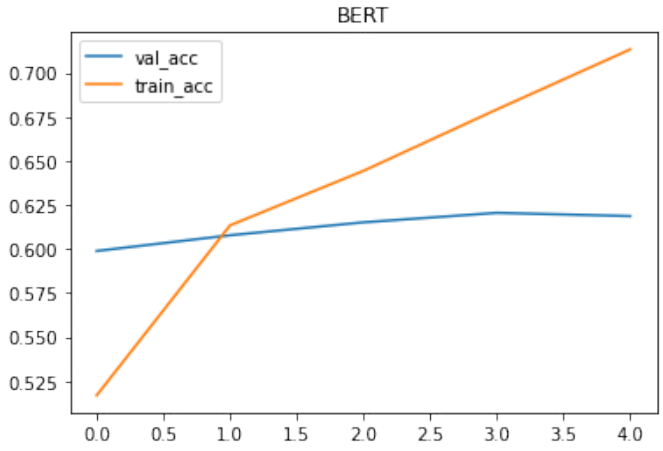
Google Translation API: Fig. 3 is a sample dataset reconstructed with the Google Translation API. The Korean context is translated into English and mapped with a labeled index.

Pre-trained Model: Pre-trained models have previously been trained, and the pre-trained dataset is used to establish the model's parameters. The advantage of the pre-trained model is obvious from the following. A method to employ a portion of a neural network trained in a specific area for training a neural network used in a new field is proposed to overcome this constraint. Acquiring a large number of data is necessary to develop a model with good performance, but because it is expensive, this limitation must be overcome [6]. Since the pre-trained model has undergone enough training, there is no need to have many training epochs in transfer learning. We conducted the experiment by setting epoch to 5 as a learning hyperparameter and setting the learning rate to $5 * e^{-}6$. In addition, since the length of the preprocessed context data was not long enough to apply the default vector size of 512, we set the batch size to 16 and reduced the length to 256. The reduced length size of the batch is slightly larger or similar to the average context length. Note that the batch size is set in proportion to the max length..

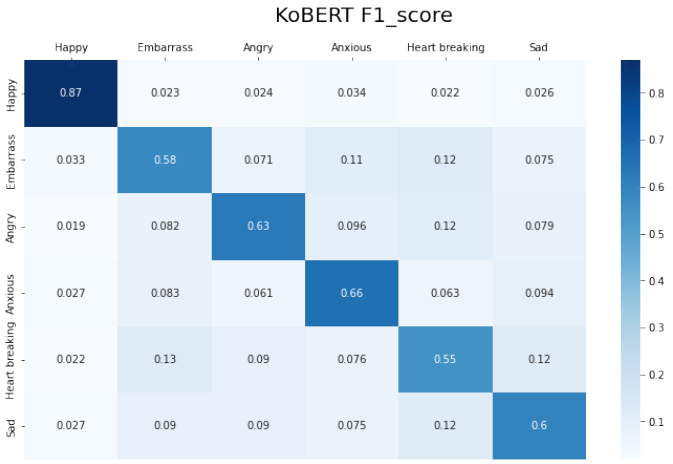
Evaluation: We used two metrics for evaluation, such as an accuracy plot and an F1 score matrix. The equation shown in equation (4) is the evaluation metric's choice for the F1 score. For the metric depiction of the F1 results, we utilized a confusion matrix. The confusion matrix serves as an evaluation metric to see how accurately the model predicted each class. We have F1 scores for absolute evaluation in any situation in our confusion matrix. Performance evaluation in //Section IV Section IV includes specific confusion matrix output values. The F1 score was chosen as an evaluation metric because, by representing both precision and recall, it can demonstrate that it is a good model in every context using an evaluation index that takes objectives and different conditions into account.



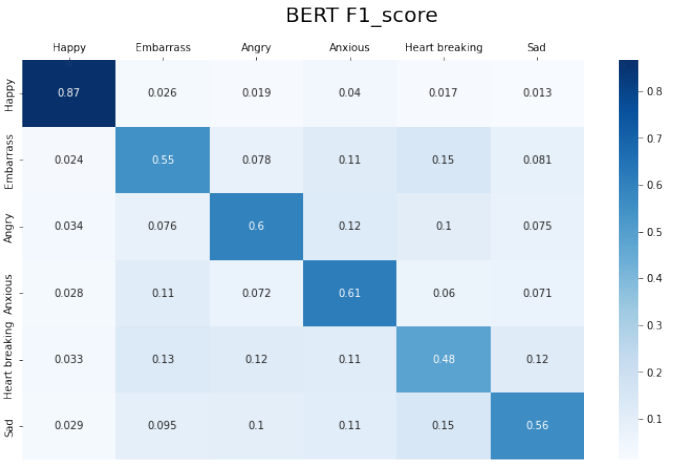
(a) Classification accuracy in training and validation for KoBERT



(b) Classification accuracy in training and validation for BERT



(c) F1 score matrix of KoBERT



(d) F1 score matrix of BERT

Fig. 4: Evaluation metric KoBERT vs BERT

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%. \quad (1)$$

$$Precision = \frac{TP}{TP + FP} * 100\%. \quad (2)$$

$$Recall = \frac{TP}{TP + FN} * 100\%. \quad (3)$$

$$F1\_score = 2 * \frac{Precision * Recall}{Precision + Recall} * 100\%. \quad (4)$$

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 KoBERRT, 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

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] M. Gu, J. Kwon, J. P. Jeong, and S. Kwon, "An emotion classification scheme for english text using natural language processing," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2022, pp. 1941–1946.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[5] J. Lee, "Kcbert: Korean comments bert," in *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, 2020, pp. 437–440.

[6] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, 2017.