# 전이 학습 및 SHAP 분석을 활용한 트랜스포머 기반 감정 분류 모델

임수빈 [1], 이병천 [2], 전인수 [1], 문지훈 [1,2]
[1] 순천향대학교 의료과학과
[2] 순천향대학교 AI·빅데이터학과
{qlsl0519, qudcjs0208, jis601, jmoon22}@sch.ac.kr,

# A Transformer-Based Emotion Classification Model Using Transfer Learning and SHAP Analysis

Subeen Leem[1], Byeongcheon Lee[2], Insu Jeon[1], and Jihoon Moon[1,2]
[1] Department of Medical Science, Soonchunhyang University
[2] Department of AI and Big Data, Soonchunhyang University

**요 약**

In this study, we embark on a journey to uncover the essence of emotions by exploring the depths of transfer learning on three pre-trained transformer models. Our quest to classify five emotions culminates in discovering the KLUE (Korean Language Understanding Evaluation)-BERT (Bidirectional Encoder Representations from Transformers) model, which is the most exceptional among its peers. Our analysis of F1 scores attests to its superior learning and generalization abilities on the experimental data. To delve deeper into the mystery behind its success, we employ the powerful SHAP (Shapley Additive Explanations) method to unravel the intricacies of the KLUE-BERT model. The findings of our investigation are presented with a mesmerizing text plot visualization, which serves as a window into the model's soul. This approach enables us to grasp the impact of individual tokens on emotion classification and provides irrefutable, visually appealing evidence to support the predictions of the KLUE-BERT model.

## 1. Introduction

Conveying emotions during conversations is crucial, as how we respond varies depending on the emotions of the person we are talking to. For this reason, various studies [1, 2] are being conducted to recognize and predict emotions from audio and text information automatically. However, most research focuses solely on adjusting the model's hyperparameters and improving performance, leaving the model's final predictions needing more explanatory power.

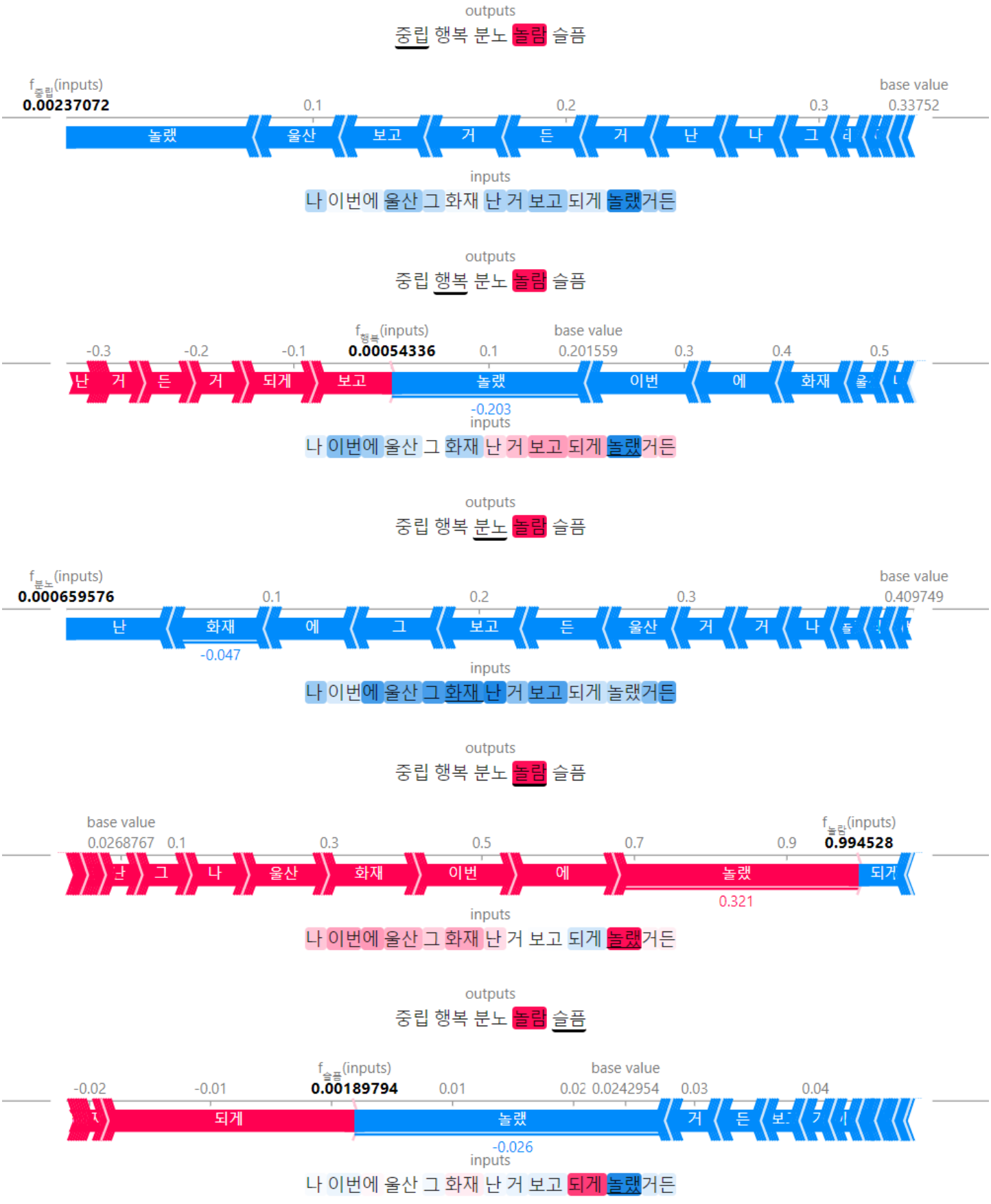
In such cases, explainable artificial intelligence (XAI) techniques can be utilized to derive more explanatory and trustworthy results from the model's predictions. There is research [3] that has applied the Shapley additive explanations (SHAP) technique to a model that classifies English text, providing explanations for its predictions. However, there have been no reported cases of using SHAP for Korean text data.

In this study, we employ Korean text data and use a transformer-based pre-trained model to perform transfer learning, evaluating the model's generalization ability. Through this process, we select the model that demonstrates the best performance. We aim to enhance the explanatory power of the selected model using SHAP, ultimately deriving more accurate and interpretable results.

## 2. Proposed Method

### 2.1. Data and Pre-trained Model

In this study, we utilize the KEMDy20 [4] dataset. This dataset includes text, audio, and biometric signal data for individual utterances from 80 Korean adult participants. Ten evaluators set the emotion labels based on the most frequently chosen emotion. To balance the labels, we undersample 200 instances for each of the five emotion labels (neutral, happy, angry, surprised, and sad). The training and testing data are split in an 8:2 ratio.




**(Figure 1)** Text plot for each label to understand the prediction influence of tokens (in the order of neutral, happiness, anger, surprise, and sadness).

Transformer-based pre-trained models have recently shown exceptional performance among natural language processing models. Among these, we conduct transfer learning for the five emotion classification tasks using KoELECTRA [5], KLUE (Korean Language Understanding Evaluation)-BERT (Bidirectional Encoder Representations from Transformers) [6], and KLUE-RoBERTa [6] models, which are used for Korean natural language processing. This allows us to compare the versatility and generalizability of each model.

### 2.2. Emotion Classification Performance and SHAP Analysis

As shown in Table 1, the performance of the three models is evaluated based on the F1-score. The KLUE-BERT model has higher overall performance than the other models, KoELECTRA and KLUE-RoBERTa. This could be interpreted as the KLUE-BERT model having better learning capabilities for new data. Furthermore, all three models consistently performed exceptionally well, classifying the 'surprised' label. This suggests that the tokens for predicting 'surprise' were more distinct in the training data. We perform SHAP analysis and visualize the results to find the basis for this.

**<Table 1>** Emotion classification performance evaluation.

| Data | KoELECTRA | KLUE-BERT | KLUE-RoBERTa |
|---|---|---|---|
| Neutral | 0.19 | 0.45 | 0.39 |
| Happy | 0.26 | 0.36 | 0.11 |
| Angry | 0.40 | 0.61 | 0.22 |
| Surprised | 0.54 | 0.63 | 0.56 |
| Sad | 0.53 | 0.67 | 0.00 |
| Average | 0.38 | 0.54 | 0.25 |

As shown in Figure 1, the SHAP text plot visually represents the influence of each token on the model's predictions. Red tokens have a positive impact on the model's predicted value, blue tokens have a negative impact, and white tokens have a neutral impact. Additionally, larger tokens indicate a greater influence on the model's prediction, and the baseline shows how much a specific token can change the predicted value. The length of the upper bar represents the impact of the entire text, while the lower bars represent the impact of individual tokens on the model's predictions. Thus, the influence of each token on emotion prediction can be confirmed through the length and color of the tokens.

Figure 1 shows the SHAP analysis results for the specific text "나 이번에 울산 그 화재 난 거 보고 되게 놀랬거든" (I was really surprised when I saw the fire in Ulsan this time) for each label. The correct label for this text is 'surprised,' and looking at the fourth plot that predicted 'surprised,' the token '놀랬' (surprised) is dark red with a long bar. This means that the keyword '놀랬' (surprised) significantly influences predicting 'surprise.' Through this visualization, we can quickly grasp how the tokens in the text contribute to the model's predictions.

## 3. Conclusions

Motivated by the need for effective emotion classification, this study conducted transfer learning on the KEMDy20 dataset using three transformer-based pre-trained models. Based on F1 scores, the KLUE-BERT model demonstrated superior performance compared to the other models. Subsequently, we performed a SHAP analysis on this model, utilizing text plots to assess the influence of individual tokens on class prediction. By visually verifying the rationale behind the model's predictions, we were able to facilitate a better understanding of the outcomes.

**Acknowledgements**

본 연구는 한국연구재단 4 단계 두뇌한국 21 사업(4 단계 BK21 사업)의 지원을 받아 작성되었음(과제번호: 5199990514663). 또한, 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구 결과로 수행되었음(2021-0-01399).

**References**

[1] S. Hong, T. Kim, S. Lee, J. Kim, and M. Lee, “MATE : the Multimodal model using Audio and Text for Emotion recognition,” in *Proc. of the KCC 2022*, 2022, pp. 2297–2299.

[2] J.-W. Kim, D.-H. Kim, J.-S. Do, and H.-Y. Jung, “Strategies of utilizing pre-trained text and speech model-based feature representation for multi-modal emotion recognition,” in *Proc. of the KCC 2022*, 2022, pp. 2282–2284.

[3] E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, and M. Robnik-Šikonja, “BERT meets shapley: Extending SHAP explanations to transformer-based classifiers,” In *Proc. of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 2021, pp. 16–21.

[4] K. J. Noh and H. Jeong, “KEMDy20,” https://nanum.etri.re.kr/share/kjnoh/KEMDy20/update?lang=ko_KR.

[5] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, and K. Cho, “KLUE: Korean language understanding evaluation.” *arXiv preprint,* arXiv:2105.09680, 2021.

[6] J. W. Park. “KoELECTRA: Pretrained ELECTRA Model for Korean.” Github Repository. https://github.com/monologg/KoELECTRA.