User Guide for KOTE: Korean Online That-gul Emotions Dataset

Duyoung Jeon[1], Junho Lee[2], Cheongtag Kim[1,2]
[1]Department of Psychology, Seoul National University
[2]Interdisciplinary Program in Cognitive Science, Seoul National University
wuju1201@gmail.com, {smbslt3, ctkim}@snu.ac.kr

Abstract

Despite the lack of comprehensive exploration of emotional connotations, sentiment analysis, which categorizes data as positive or negative, has been widely employed to identify emotional aspects in texts. Recently, corpora labeled with more than just valence or polarity have been built to surpass this limitation. However, most Korean emotion corpora are limited by their small size and narrow range of emotions covered. In this paper, we introduce the KOTE dataset. The KOTE dataset comprises 50,000 Korean online comments, totaling 250,000 cases, each manually labeled for 43 emotions and NO EMOTION through crowdsourcing. The taxonomy for the 43 emotions was systematically derived through cluster analysis of Korean emotion concepts within the word embedding space. After detailing the development of KOTE, we further discuss the results of fine-tuning, as well as analysis for social discrimination within the corpus.

Keywords: sentiment analysis, emotion, dataset, Korean

1. Introduction

Sentiment analysis, which classifies texts as positive or negative, is the most prevalent method for analyzing the emotional aspects of texts. While sentiment analysis is straightforward, practical, and applicable in many contexts, there is an emerging need for analyzing more complex emotions beyond mere polarity in texts. This is attributed to the advent of advanced language models capable of processing intricately labeled data, coupled with recent advancements in computing power.

There is a significant demand for emotion analysis tools tailored to the Korean language. However, the majority of Korean emotion text datasets are limited by their small size and coarse emotion taxonomies, which encompass only a narrow spectrum of emotions. Consequently, the GoEmotions (Demszky et al., 2020), an extensive English resource with 58k instances and a detailed taxonomy of 27 emotions or neutrality, is frequently employed for analyzing Korean texts through machine translation despite the imperfect translation quality.

However, emotions are deeply intertwined with culture since they are products of culture-specific schema. Accordingly, emotion taxonomies, which map out underlying emotional structures, vary across cultures (Mesquita and Frijda, 1992) and the variation even holds for basic emotions (Gendron et al., 2014). This underscores the necessity of developing datasets labeled with emotion taxonomies that are culturally pertinent.

In response to this need, we created KOTE (Korean Online That-gul[1] Emotions; pronounced as [kot]), a large language dataset of 50k Korean online comments labeled for 43 emotions. The online comments were sourced from 12 diverse platforms spanning various domains, including news, online communities, social media, e-commerce, video platforms, movie reviews, microblogs, and forums. The 43 emotions befitting to the Korean language are derived from the clustering results of Korean words that refer to emotion concepts. Table 1 shows a raw example in KOTE.

Text	
러브크래프트 소설 단편 에피같다ㄷㄷ 레알광기ㄷㄷㄷ It's like a short Lovecraftian episode 😟 True madness 😱	
Labels	
rater 1	공포/무서움, 놀람, 감동/감탄 *fear/scary, surprise, impressed/admiration*
rater 2	놀람, 감동/감탄, 신기함/관심 *surprise, impressed/admiration, curiosity/interest*
rater 3	부담/안_내킴, 공포/무서움, 놀람 *burden/unwillingness, fear/scary, surprise*
rater 4	놀람, 감동/감탄, 즐거움/신남, 기대감, 신기함/관심 *surprise, impressed/admiration, pleasure/excitement, anticipation, curiosity/interest*
rater 5	깨달음, 즐거움/신남, 신기함/관심 *realization, pleasure/excitement, curiosity/interest*

Table 1: A raw example in KOTE.

The purpose of this study is twofold. The first objective is to propose a new emotion taxonomy tailored to the Korean language in general. The second objective involves constructing the KOTE dataset utilizing this new taxonomy. We also fine-tuned the pre-trained KcELECTRA (Korean comment ELECTRA; Clark et al., 2020; Lee, 2021) model with KOTE. This achieves a better performance than the existing model trained with translated GoEmotions (F1-scores are 0.56 ver-

[1]'That-gul' or 'Daet-gul' is a Korean word that refers to 'online comment'.

LREC-COLING 2024, pages 17254–17270

sus 0.41). Significant improvement is possible, as the results have not been fully optimized. Given KOTE's open access and wealth of information, analysts can apply a variety of strategies to the raw data to meet specific objectives.

2. Related Work

2.1. Emotion Taxonomy

Constructing an emotion corpus requires an appropriate emotion taxonomy by which the texts are labeled. To identify the appropriate emotion taxonomy, it is essential first to compile a dataset of emotion words, gathering all possible emotions as candidates for inclusion in the taxonomy.

Thus, the very first question is how to identify the types of emotion. Vocabulary representing emotions can be used to this end. In traditional approaches, the distinction between emotion and non-emotion is determined by human rating. Shields (1984) attempted to conceptualize *emotionality* by asking participants to categorize 60 feeling words (*happy, curious, hungry, etc.*) into emotion or non-emotion words. Clore et al. (1987) measured the emotionality of 585 feeling words by asking participants to rate their confidence in a 4-point scale of how emotional each word is. Apart from the survey approaches, the emotionality can be determined by experts. Averill (1975) recruited graduate students to examine around 18k psychological concepts, including words and phrases, concluding that 717 exhibited emotionality. In the case of Korean, Sohn et al. (2012) collected 65k Korean words from a variety of text sources and manually checked their properties to confirm 504 emotional expressions.

The next question after identifying the emotion words is how to transform the words into a mathematically analyzable form. One popular way is vectorization, which imposes vector-valued information on words by a certain measure. A traditional method of vectorization involves human rating, where human annotators assess each word on several scales devised by researchers. For example, Block (1957) asked the participants to rate fifteen emotion words in twenty 7-point scales (e.g., *good-bad, active-passive, tense-relaxed*). Similarly, Sohn et al. (2012) vectorized 504 emotion words in eleven 10-point emotion scales (e.g., *joy, anger, sadness*). Park and Min (2005) rated emotion words in four scales (i.e., *prototypicality, familiarity, valence, and arousal*).

The vector of a word can be indirectly estimated by rating similarity (or distance) among words. Storm and Storm (1987) utilized a sorting method to extract co-occurrence information from emotion words. Cowen et al. (2019) suggested that a pseudorandom assignment for similarity rating is sufficient to embed the local similarity of 600 emotion words.

The last question is how to uncover an adequate structure of the emotion words using the information. 'How many emotions are there?' has always been one of the biggest and the most mesmerizing questions in the field of emotion research. Many emotion researchers have actively suggested *core emotions* or *emotion taxonomy* from their own disciplines, such as evolution, neural system, facial expression, physiology, culture (e.g., Osgood, 1966; Izard, 1977, 1992; Plutchik, 1980; Willcox, 1982; Mano and Oliver, 1993; Lee and Lim, 2002; Cowen and Keltner, 2017; Keltner et al., 2019), and language (e.g., Shaver et al., 1987; Storm and Storm, 1987; Hupka et al., 1999; Cowen et al., 2019). Common findings across these studies suggest: **i)** Emotion may not have a fixed dimensionality, as it varies with the research context; **ii)** Emotion forms a complex structure, indicating that beyond six or seven basic emotions exist as fundamentally distinct entities. Accordingly, the emotion taxonomy of this study considers these two implications.

We briefly looked at how emotion researchers have constructed the concepts of emotion via emotion vocabulary. One can see that most studies relied on human participants. However, due to the recent advancement of machine learning in natural language processing, words, including emotion words of course, are becoming a full-fledged subject of machine learning. Machine learning methods have unveiled a plethora of tools for deriving detailed information from words, offering advantages over traditional approaches in several key aspects. These methods surpass human annotation in efficiency, enabling the processing of large text datasets with ease. Moreover, machine learning can encode texts with richer information than is possible through human annotation, which is often limited by specific research designs.

Therefore, in this study, we actively utilize machine learning techniques to follow the fundamental procedure above; identifying and vectorizing emotion words to propose a new emotion structure for the Korean language.

2.2. Emotion Text Datasets

In the past few years, many emotion text datasets have been developed, driven by a great interest in emotion analysis. **Table 2** lists currently available Korean emotion text datasets by chronological order of the publication dates.

The datasets are mostly small in size and have rough emotion taxonomies. The lack of a proper emotion corpus is the major motivation of this study.

Dataset	Unit	# of instances	Label dimension
List of Korean Emotion Terms (Park and Min, 2005)	word	434	4
Korean Emotion Vocabulary (Sohn et al., 2012)	word	504	11
KOSAC (Jang et al., 2013)	sentence	7.7k	2*
NSMC (Naver, 2015)	sentence	200k	1
KNU SentiLex (Park et al., 2018b)	n-gram	14k	1
Korean Continuous Dialogue Dataset with Emotion Information (KETI, 2020a)	dialogue	10k	7
Korean One-off Dialogue Dataset with Emotion Information (KETI, 2020b)	sentence	38k	7
Emotional Dialogue Corpus (AIHUB, 2021)	dialogue	15k	60

Table 2: Korean emotion text datasets.
* KOSAC contains far more plentiful information, but two dimensions are closely related to emotion (*polarity* and *intensity*).

3. Korean Emotion Taxonomy

In this study, we construct a new Korean emotion taxonomy with which our dataset is labeled. The taxonomy is constructed by finding and interpreting the meaning of clusters of emotion concepts. The basic process is as follows: **i)** Identifying emotion words out of all existing words; **ii)** Vectorizing the emotion words with a pre-trained word vector model; and **iii)** Clustering the words and interpreting the meaning of the clusters. One interpretable cluster is considered as one emotion in the emotion taxonomy.

3.1. Emotion Words

There are a few available emotion words datasets such as List of Korean Emotion Terms (Park and Min, 2005), Korean Emotion Vocabulary (Sohn et al., 2012), and KNU SentiLex (Park et al., 2018b). KNU SentiLex contains the greatest number of emotion expressions. The researchers preliminarily filtered emotion expressions out of the whole contents of the Korean dictionary by reading the glosses using `Bi-LSTM` (Bidirectional Long-Short Term Memory; Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Graves and Schmidhuber, 2005), and manually added emotional slangs and emoticons. Subsequently, they confirmed the emotionality of the expressions by the scrutiny of human raters. As a result, 14k emotion expressions were confirmed and suggested. This study utilizes these three datasets for emotion categorization.

However, the lexicons include some expressions that express emotions figuratively (e.g., 많다*many*). These expressions were excluded due to their infrequent use in emotional contexts. Furthermore, to address the absence of certain expressions, we manually supplemented the dataset with additional terms. Subsequently, expressions were tokenized using the Python package KoNLPy (Park and Cho, 2014), and both function words and stop words were removed. We chose 3,017 expressions that we considered directly represent human emotions, which were inputted into the pre-trained word vector model in the next step.

3.2. Word Vectorization

The 3,017 emotion words were inputted into a `fastText` model (Bojanowski et al., 2017) pre-trained with large language datasets such as the Korean Wikipedia[2]. From the list of candidate emotion words, 1,787 words were included in the model. Hence, the vectors of 1,787 emotion words were used for clustering.

3.3. Exploring Dimensionality of Emotion

Base Clustering. The purpose of the *base clustering* is to find the most likely number of clusters of the Korean emotion concepts. In other words, we attempt to answer the question, 'How many emotions are there, especially in Korean?' in this stage.

[2] https://github.com/ratsgo/embedding/releases

The base clustering is conducted in two steps: **i)** dimension reduction with UMAP (Uniform Manifold Approximation and Projection; McInnes et al., 2018) is performed and **ii)** the reduced vectors are clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Application with Noise; McInnes et al., 2017). The HDBSCAN determines the number of clusters by a survival algorithm. Clusters in a model diminish as its criteria, by which a data point is considered to belong to a cluster, gradually becomes strict and an increasing number of data points are regarded as noise. Clusters are considered valid, only if they survive long enough in this process. The HDBSCAN estimates the likely number of clusters by this algorithm. Consequently, the number of clusters is given as the final output after the two-step procedure.

The major goal of the two-step strategy is to explore the dimensionality of the emotions as exhaustively as possible. Thus, a grid search was applied on the hyperparameters of each step. The hyperparameters to be searched and the searched values are presented in **Figure 1**. 21,600 points in the hyperparameter space were searched in total. 21,562 partition sets remained, after partition sets with less than three clusters were eliminated. Three distributions are robustly identified regardless of the hyperparameters, and the cluster numbers are not correlated to the hyperparameters except for the minimum cluster size. The most likely number of clusters is 30 as in **Figure 1** (a), the median of the largest distribution. This result is consistent with many previous studies. However, we believe that the emotion is so complicated that just 30 categories are insufficient to represent the structure effectively. In addition, recently developed language models are powerful enough to handle complicatedly labeled data. Hence, we decided to proceed for the next most likely number, 136.

Clustering Ensemble to Build a New Emotion Taxonomy. It is not necessary to implement a cluster analysis from scratch to extract 136 clusters, because 21,562 partition sets are already acquired in the base clustering. A cluster ensemble is employed to utilize the partition sets.

The cluster ensemble, literally, is a method that aggregates multiple clustering results to derive one single agreed outcome. We use HBGF (Hybrid Bipartite Graph Formulation; Fern and Brodley, 2004), which exploits both instance- and cluster-based graph formulation (See also Vega-Pons and Ruiz-Shulcloper, 2011; Karypis and Kumar, 1998). In other words, the 21,562 partitions sets were fitted by a HBGF model to reach a consensus of how to split 1,787 emotion words into 136 groups.

The meaning of each cluster is interpreted. Some clusters are considered non-interpretable and dropped because seemingly unrelated words are entangled together. If antonyms are in the same cluster, they are regarded as two separate emotions (i.e., *sadness* and *joy*). 43 emotions were clearly interpreted (see **Appendix 8**).

4. KOTE

We developed KOTE (Korean Online That-gul Emotions), a Korean language dataset containing 50k online comments labeled for the 43 emotions in the new taxonomy. In this chapter, we explain how KOTE is compiled and provide the results of fine-tuning on a pre-trained language model.

4.1. Text

50k online comments in KOTE are collected from 12 different platforms of various domains (*news, online community, social media, e-commerce, video platform, movie review, microblog, and forum*) to cover general online environments. The robots.txt guideline of every website was obeyed during the scraping unless no guideline was provided. If a website supports a search engine, randomly selected emotion words from KNU SentiLex were searched for scraping to maximize the emotionality of the collected texts. 3.2 million comments were collected in total, and 50k were sampled being balanced in the number of comments of each website. In the sampling process, the minimum length of the texts was set as 10, and the maximum as the 90th percentile of each platform. The grand maximum length was 404, the mean was 57.32, and the median is 42[3].

In all texts, personal information, such as user ID, was deleted without leaving the original. The comments were also supervised for a privacy check by a credible third-party institution designated by the Korea Data Agency, the supporter of this study. They confirmed that no comment contains inappropriate personal information.

4.2. Label

The 50k comments were labeled by crowdsourcing in which 3,084 raters whose mother tongue is Korean participated with monetary reward. The labeling procedure was as follows: 50 randomly selected comments are given to a rater. The rater chooses all emotions that the speaker of each comment intends to express. If they identify no emotion, they choose no emotion label but a special label, NO EMOTION. They are also instructed to select plausible emotions and not NO EMOTION,

[3]The unit of length is a syllable. In the Korean system, 2-3 letters are combined to create one character, which basically corresponds to one syllable. Therefore, the length is 2-3 times longer if the unit is a letter.

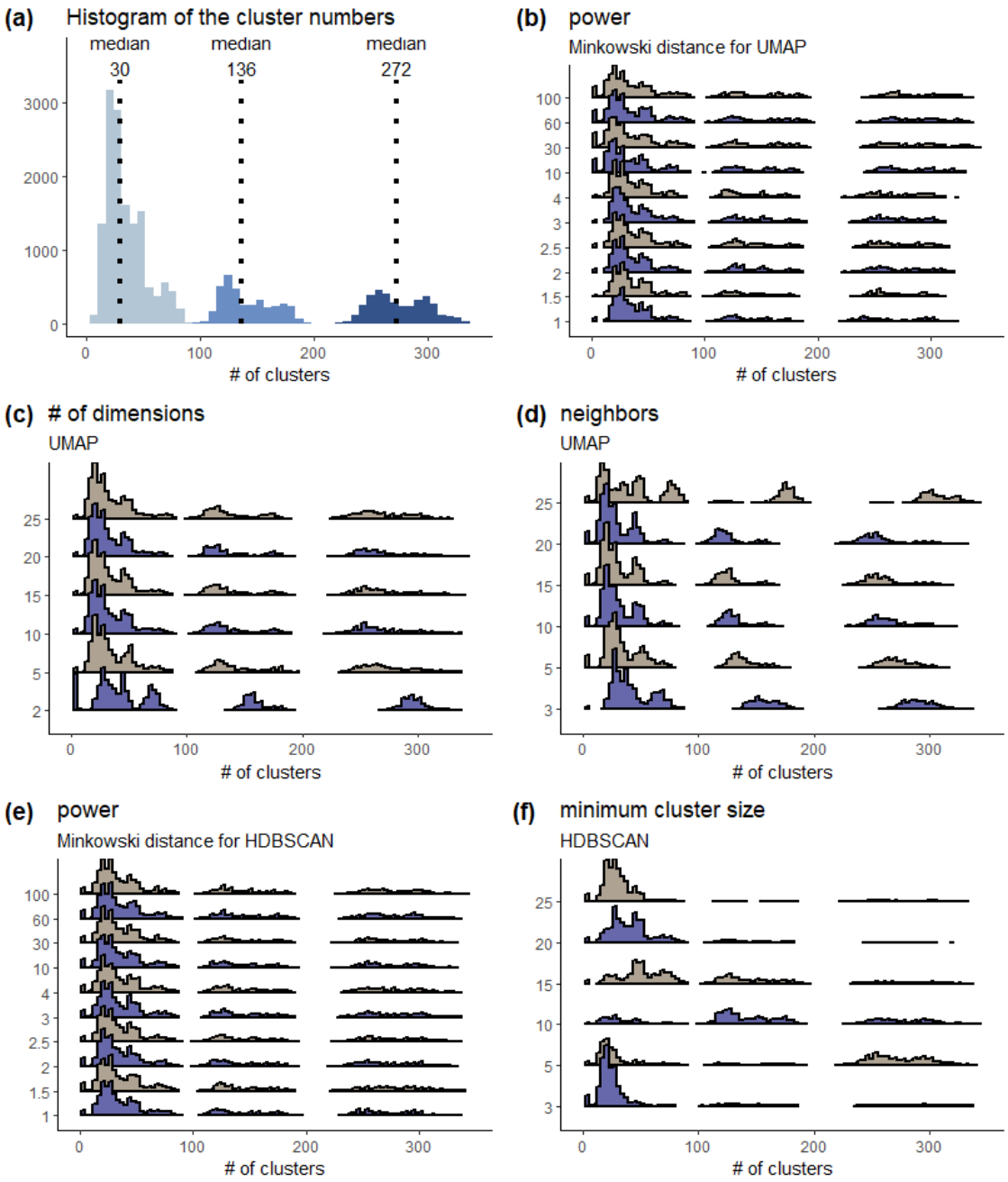

Figure 1: (a) is the histogram for the number of clusters in 21,562 partition sets. Three distributions are identified. (b) – (f) are histograms marginalized on each hyperparameter space. The y-axes represent the searched values of the hyperparameters. Three distributions are consistently identified. The hyperparameters and the number of clusters are not correlated, except for the minimum cluster size (r = -0.2). (plot packages; ggplot2 (Wickham, 2011), ggpubr (Kassambara and Kassambara, 2020) and ggridges (Wilke, 2021).) **Hyperparameters:** (b): the power in Minkowski distance used to compute the distance matrix for `UMAP`. (c): the number of dimensions after the reduction by `UMAP`. (d): the number of neighbors of each data point in `UMAP`. (e): the power in Minkowski distance used to compute the distance matrix for `HDBSCAN`. (f): the minimum size of a group of data points that would be considered as a cluster in `HDBSCAN`.

if they think a comment obviously contains some emotion but the exact emotion is not in the given category. Lastly, they are instructed to choose all possibly relevant emotions if the text could have different emotions according to context. The minimum and the maximum number of labels they can choose for one comment are 1 and 10, respectively. The rater can request one more set of 50

agreement						
at least one label of x or higher	x=1	x=2	x=3	x=4	x=5	
# of texts	50,000	49,663	42,845	28,650	11,760	
(% to total)	(100%)	(99%)	(86%)	(57%)	(24%)	
texts labeled for NO EMOTION						
# of NO EMOTION	0	1	2	3	4	5
# of texts	42,156	5,243	1,592	644	264	101
(% to total)	(84%)	(10%)	(3%)	(1%)	(0.5%)	(0.2%)

Table 3: Description of the labels.

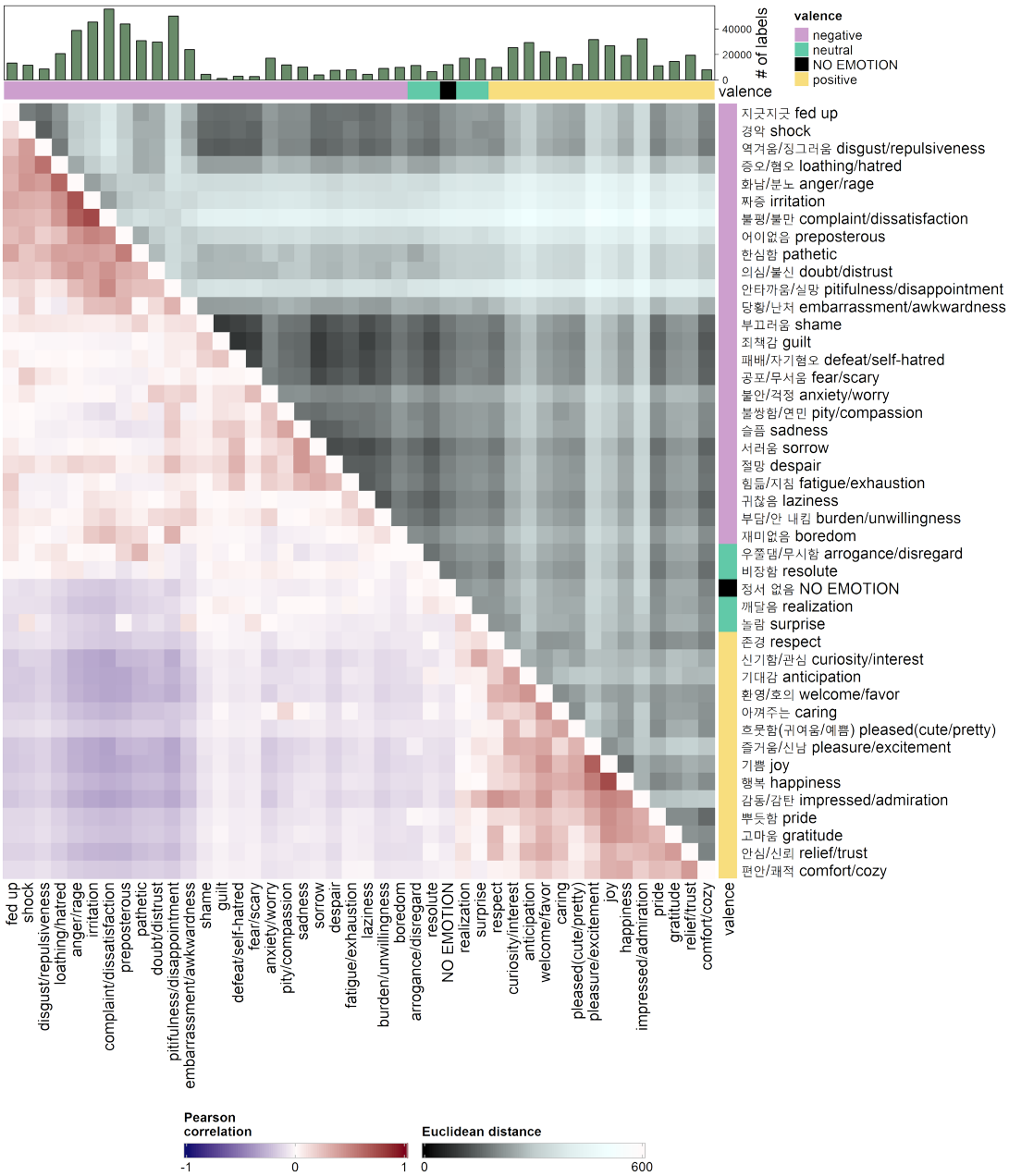

Figure 2: Heatmap of Pearson correlation and Euclidean distance among the labels. The lower and upper triangle represent the correlation coefficients and the Euclidean distances, respectively. The bars indicate the number of labels in 250k cases. The order of the labels follows Ward clustering with squared Euclidean distance (Ward Jr, 1963). (plot package; ComplexHeatmap (Gu et al., 2016).)

comments, and one rater can answer a maximum of two sets. After the labeling, the annotated texts are sent to other crowdworkers who examine the validity of the labels. If the examiner finds labels that they do not agree upon, the disagreed texts are sent back to the original labelers for relabeling. This back-and-forth examination can be repeated three times at maximum.

Two types of catch trials are given in the middle of the labeling. The raters were informed about the catch trials before answering and agreed that the labeling procedure would end with no reward if they did not answer the catch trials correctly. Type-1 catch trial directly instructs the raters to select a certain label, for example *"Please choose only 'joy' and no other labels for this question"*. Type-2 catch trial asks a question that has a correct answer, for example *"I finally realize what happened. Now I know… I understand everything"*. The selected labels must include *'realization'*, or the answer is regarded wrong. The correct answer label word is always in the presented text itself.

Five randomly selected raters are assigned to one text, and thus 250k cases of 50k texts are created as a result. Five binary labels of a text are summed to be the final label. Thus, the range of a label is 0–5. (see **Table 1**. Four out of the five raters agreed that the text contains *surprise*, so the value of *surprise* label is 4)

4.3. Data Description

The relations among the labels are presented in the heatmap in **Figure 2**. It shows Pearson correlation and Euclidean distance among the labels, each of which is a 50k-dimensional vector.

Table 3 describes the labels. 99% of the texts have at least one label of 2 or higher, which means that 99% have at least one label that two or more raters choose in common. It is evident that the raters did not have much difficulty to reach a consensus. Also, a moderate number of texts are labeled for `NO EMOTION`.

No additory preprocessing was applied on the data to merge or exclude emotions even though some emotions are linearly related. This was not only because the emotion taxonomy was derived by a nonlinear method, but also the ELECTRA model, which would be fine-tuned, was nonlinear and potentially able to distinguish linearly similar emotions.

4.4. Fine-tuning

Preparation. The labels ranging from 0 to 5 were dichotomized into 0 or 1. Minmax scaling was applied on the labels for each text. The purpose of the text-wise minmax scaling was to have the fine-tuned machine return several possible emotions when no emotion is confidently recognized. The labels exceeding 0.2 after the scaling were converted into 1, and 0 otherwise. One text has 7.91 labels in average as a result. The dataset was randomly split into train (80%), test (10%), and validation (10%) sets.

Training. We fine-tuned `KcELECTRA`, a language model pre-trained with Korean online comments, with three packages: `pytorch` (Paszke et al., 2019), `pytorch-lightning` (Falcon and Cho, 2020), and `transformers` (Wolf et al., 2019). The batch size was 32, and the input token size was 512. If the number of tokens of an input was less than 512, it was padded with a special token, `[PAD]`. No input exceeds 512 in length. One linear layer was added on the `[CLS]` token of the last hidden layer for multi-label classification. The loss was binary cross entropy for each label. We used a linear optimization scheduler, in which the initial learning rate is 2e-5 and the number of warmup steps and total steps are 2,500 and 12,500, respectively. We also switched 5% of tokens with a random token (except `[CLS]`, `[SEP]`, and `[PAD]`), and masked 5% of tokens with a special token, `[MASK]`. The maximum number of epochs was set as 15, but 9 epochs were enough to reach the optimum in almost all cases. We tried label smoothing (Szegedy et al., 2016), but the results are not reported since the performance rather declined.

Results. The decision threshold for predicted labels was set as 0.3. We used `scikit-learn` (Pedregosa et al., 2011) to compute the performance metrics. The average F1-score, AUC (Area Under Curve; Hanley and McNeil, 1982), and MCC (Mathews Correlation Coefficient; Matthews, 1975; Baldi et al., 2000; Chicco and Jurman, 2020) were 0.56, 0.88, and 0.59, respectively (see **Appendix 9** for full description).

As mentioned in the Introduction section, these results were obtained with arbitrarily decided hyperparameters. Therefore, the performance could be improved with additional methods, such as hyperparameter tuning. Otherwise, it would be a good attempt to employ different approaches for the preprocessing, such as label merging, dichotomization, or label balancing. Since the dataset is fully open, one can try anything necessary.

5. Conclusions

The model fine-tuned with our dataset achieved a better performance than the existing model fine-tuned with the translated GoEmotions dataset (F1-scores are 0.56 versus 0.41). Although direct comparison is difficult because of different emotion taxonomies, it is meaningful to achieve a comparable performance with a wider range of emotions (43

versus 27). The reasons for good performance can be summarized as follows. **i)** We derived emotion taxonomy by introducing machine learning to repeatedly validated psychological theories and methodologies. **ii)** The emotion taxonomy is befitting to Korean culture, which is beneficial in two respects; the human raters can easily understand the emotions in the taxonomy, and the Korean language model can infer the emotions of the texts efficiently. **iii)** We viewed the emotion as a complex structure according to the existing psychology literature, which motivated us to impose complex information on the texts in the labeling and to maintain the complexity in the preprocessing.

6. Limitations

However, there are limitations that the users should keep in mind: **i)** Emotion is a complex structure, which is impossible to perfectly capture with just tens of categories. **ii)** Although emotion is a dynamic structure, it is treated as a static one in this study. The emotions must interact complicatedly. For example, an emotion may be combined with other emotions to create a new one, or one single emotion can have different meanings according to the degree of emotionality and contextuality. **iii)** KOTE is large, but not large enough to cover different domains inside and outside the internet. KOTE may have limitations when one tries to apply the trained model to a different type of texts other than online comments. *Fear*, for example, is one of the core emotions but rarely appears in our dataset. Accordingly, linguistic expressions associated with *fear* might be scarce as well. **iv)** The discriminatory evaluation against protected groups is carried within our dataset, since it reflects the discrimination of the texts and the human raters. We highly recommend **Appendix 10** for ethical consideration.

Although future works are required to answer those questions, KOTE is still a new useful tool that helps to overstep the limit of mere sentiment analysis. We hope this user guide provides the users with useful information to utilize the dataset.

Acknowledgements

This study is supported by the 2021 Data Voucher Support Project organized by the Korea Data Agency under the Ministry of Science and ICT of the Government of Republic of Korea. We also thank Crowdworks, Inc. and all crowdworkers who sincerely helped us to annotate the data.

7. Bibliographical References

AIHUB. 2021. Emotional dialogue corpus. https://aihub.or.kr/aidata/7978.

James R Averill. 1975. *A semantic atlas of emotional concepts*. American Psycholog. Ass., Journal Suppl. Abstract Service.

Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Jack Block. 1957. Studies in the phenomenology of emotions. *The Journal of Abnormal and Social Psychology*, 54(3):358.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Gerald L Clore, Andrew Ortony, and Mark A Foss. 1987. The psychological foundations of the affective lexicon. *Journal of personality and social psychology*, 53(4):751.

Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. 2019. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, pages 4040–4054, Online. Association for Computational Linguistics.

William Falcon and Kyunghyun Cho. 2020. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*.

Xiaoli Zhang Fern and Carla E Brodley. 2004. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning*, page 36.

Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

Zuguang Gu, Roland Eils, and Matthias Schlesner. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ralph B Hupka, Alison P Lenton, and Keith A Hutchison. 1999. Universal development of emotion categories in natural language. *Journal of personality and social psychology*, 77(2):247.

Carroll E Izard. 1977. Differential emotions theory. In *Human emotions*, pages 43–66. Springer.

Carroll E Izard. 1992. Basic emotions, relations among emotions, and emotion-cognition relations.

Hayeon Jang, Munhyong Kim, and Hyopil Shin. 2013. KOSAC: A full-fledged Korean sentiment analysis corpus. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 366–373, Taipei, Taiwan. Department of English, National Chengchi University.

George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1):359–392.

Alboukadel Kassambara and Maintainer Alboukadel Kassambara. 2020. Package 'ggpubr'. *R package version 0.1*, 6.

Dacher Keltner, Disa Sauter, Jessica Tracy, and Alan Cowen. 2019. Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, 43(2):133–160.

KETI. 2020a. Korean continuous dialogue dataset with emotion information. https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-010.

KETI. 2020b. Korean one-off dialogue dataset with emotion information. https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-009.

Hak-Sik Lee and Ji Hoon Lim. 2002. Measuring the consumption-related emotion construct. *Korea Marketing Review*, 17(3):55–91.

Junbum Lee. 2021. Kcelectra: Korean comments electra. https://github.com/Beomi/KcELECTRA.

Haim Mano and Richard L Oliver. 1993. Assessing the dimensionality and structure of the consumption experience: evaluation, feeling, and satisfaction. *Journal of Consumer research*, 20(3):451–466.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Batja Mesquita and Nico H Frijda. 1992. Cultural variations in emotions: a review. *Psychological bulletin*, 112(2):179.

Naver. 2015. Nsmc: Naver sentiment movie corpus. https://github.com/e9t/nsmc.

Charles E Osgood. 1966. Dimensionality of the semantic space for communication via facial expressions. *Scandinavian journal of psychology*, 7(1):1–30.

E Park and S Cho. 2014. Konlpy: easy and concise korean information processing python package. In *Proceedings of the 26th Korean and Korean Information Processing Conference*, pages 1–4.

In-Jo Park and Kyung-Hwan Min. 2005. Making a list of korean emotion terms and exploring dimensions underlying them. *Korean Journal of Social and Personality Psychology*, 19(1):109–129.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018a. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Sang-Min Park, Chul-Won Na, Min-Seong Choi, Da-Hee Lee, and Byung-Won On. 2018b. Knu korean sentiment lexicon: Bi-lstm-based method for building a korean sentiment lexicon. *Journal of Intelligence and Information Systems*, 24(4):219–240.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061.

Stephanie A Shields. 1984. Distinguishing between emotion and nonemotion: Judgments about experience. *Motivation and Emotion*, 8(4):355–369.

Sun-Ju Sohn, Mi-Sook Park, Ji-Eun Park, and Jin-Hun Sohn. 2012. Korean emotion vocabulary: extraction and categorization of feeling words. *Science of Emotion and Sensibility*, 15(1):105–120.

Christine Storm and Tom Storm. 1987. A taxonomic study of the vocabulary of emotions. *Journal of personality and social psychology*, 53(4):805.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Sandro Vega-Pons and José Ruiz-Shulcloper. 2011. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Hadley Wickham. 2011. ggplot2. *Wiley interdisciplinary reviews: computational statistics*, 3(2):180–185.

Claus O Wilke. 2021. Ridgeline plots in 'ggplot2'[r package ggridges version 0.5. 3]. *January. https://cran. r-project. org/web/packages/ggridges/index. html*.

Gloria Willcox. 1982. The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

8. Appendix A: Emotion Clusters

Valence 극성	**Interpretation** 해석	**Example words in the cluster** 군집 안의 정서 단어 예시
Negative 부정	*complaint /dissatisfaction* 불평/불만	dissatisfied, oppose, criticize, complaint 불만, 반발, 비판, 항의
	embarrassment /awkwardness 당황/난처	embarrassed, disconcerted, awkward, untoward 당황, 당혹, 곤혹, 난처
	irritation 짜증	irritated, pissed off, ridiculous 짜증, 열 받다, 어이없다
	sadness 슬픔	sad, miss, lonely, tear 슬픔, 그리운, 외로운, 눈물
	despair 절망	frustrated, joys & sorrows, hurt, grief, letdown 절망, 애환, 아픔, 비탄, 허무감
	shame 부끄러움	ashamed, humiliated 부끄러움, 부끄럽다
	boredom 재미없음	bored, tedium, trite, dull 지루함, 재미없음, 식상, 답답함
	pitifulness /disappointment 안타까움/실망	disappointed, sorry, upset, deplorable, regretful 실망, 안타까움, 속상, 애석, 아쉬움
	disgust/repulsiveness 역겨움/징그러움	disgusted, repulsive, dirty 역겨움, 징그러움, 지저분
	shock 경악	shocked, flabbergasted, pass out, freaked out 경악, 기절초풍, 실신, 까무러치다
	burden/unwillingness 부담/안 내킴	unwilling, denial, pressure, cannot be bothered, give up 마지못해, 거부, 재촉, 고깝다, 단념
	fear/scary 공포/무서움	fear, anxious, tense, pressed 공포, 불안, 긴장, 압박감
	loathing/hatred 증오/혐오	loathing, hatred, scorn, vilifying 증오, 혐오, 죄악시, 경멸, 모멸, 멸시
	guilt 죄책감	guilt, blamed, repentance, remorse 죄책감, 죄의식, 가책, 참회, 속죄, 뉘우침
	anxiety/worry 불안/걱정	apprehensive, worry, threatened 우려, 염려, 위험
	doubt/distrust 의심/불신	suspicious, doubtful, lie 의심쩍다, 반신반의, 거짓
	anger/rage 화남/분노	anger, rage, obsessed, fury 증오, 분노, 사로잡힌, 분개, 격분, 격노
	defeat/self-hatred 패배/자기혐오	failure, miserably, extorted 실패, 처참히, 빼앗기다
	laziness 귀찮음	bothered, dawdling 귀찮음, 빈둥빈둥
	sorrow 서러움	sorrowful, mirthless, weary, sobbing, upset, complicated 서러움, 서글픔, 고달프다, 흐느낌, 속상, 착잡
	fed up 지긋지긋	fed up, struggle, arduous, sick and tired 지긋지긋, 애쓰다, 고되다, 질리다

Valence 극성	**Interpretation** 해석	**Example words in the cluster** 군집 안의 정서 단어 예시
Negative 부정	*preposterous* 어이없음	dumbfounded, stunned, sttufy, enervated, WTF 어처구니, 싱겁, 갑갑함, 맥빠지다, 이뭐병
	pity/compassion 불쌍함/연민	pity, sadly, chocked up, heartrending 짠하다, 슬프다, 울컥, 먹먹하다
	pathetic 한심함	pathetic, belittled, stupid, impudence 한심, 우스운, 멍청, 뻔뻔
	fatigue/exhaustion 힘듦/지침	tired, peak, exhausted 피로, 야위다, 수척
Positive 긍정	*impressed/admiration* 감동/감탄	admiring, great, praise, compliment 감탄, 대단하다, 칭찬, 찬사
	happiness 행복	happy, affection, valuable, hope, luck 행복, 친애, 소중, 희망, 행운
	joy 기쁨	delight, ecstasy, love 환희, 황홀, 사랑
	gratitude 고마움	praiseworthy, commendable, favor, blessing, mercy 기특함, 은혜, 은총, 베풀다
	pleasure/excitement 즐거움/신남	excited, funny 즐거운, 재밌는
	caring 아껴주는	caring, adore, dear 아낌, 흠모, 경애
	anticipation 기대감	new, achieve, together, harmonious, vitality 새로운, 이루다, 함께, 원활, 활력
	comfort/cozy 편안/쾌적	comfortable, ease, cozy, cool, warm 편안, 포근함, 안락, 시원, 따듯
	welcome/favor 환영/호의	welcome, approval, kindness, enthusiastic 환영, 우호, 호의, 열렬히
	curiosity/interest 신기함/관심	interested, curious 호기심, 관심
	relief/trust 안심/신뢰	relief, trust, intimate, close 신뢰, 안심, 친밀, 각별
	respect 존경	respect, loyal, veneration, follow, obedience 존중, 충성, 숭상, 본받다, 복종
	pleased(cute/pretty) 흐뭇함(귀여움/예쁨)	handsome, pretty, sweet, thrilled, cute, aegyo 멋있다, 예쁘다, 달달, 짜릿, 귀엽다, 깜찍, 애교
	pride 뿌듯함	successful, victory, worthwhile, accomplish 성공, 승리, 달성, 보람, 희열
Neutral 중립	*arrogance/disregard* 우쭐댐/무시함	arrogance, pompous, ignore, bragging, boast, gasconade 우쭐댐, 얕잡아보다, 무시, 업신여기다, 거만, 교만
	surprise 놀람	astonished, startled 질겁, 소스라치다
	realization 깨달음	realize, enlightened, wakened, conviction, belief 깨달음, 깨우침, 일깨워, 확신, 믿음
	resolute 비장함	resolute, determination 비장함, 결단, 결심

Table 4: Interpretation of each interpretable cluster and emotion words in it.

9. Appendix B: Performance Metrics

F1-score									
emotion	**precision**	**recall**	**F1**	**#**	**emotion**	**precision**	**recall**	**F1**	**#**
complaint /dissatisfaction	0.78	0.89	0.83	2113	*impressed /admiration*	0.67	0.86	0.75	1323
embarrassment /awkwardness	0.57	0.70	0.63	1319	*happiness*	0.57	0.80	0.67	906
irritation	0.74	0.86	0.80	1909	*joy*	0.65	0.85	0.73	1205
sadness	0.62	0.61	0.62	545	*gratitude*	0.54	0.70	0.61	637
despair	0.46	0.41	0.43	472	*pleasure /excitement*	0.69	0.86	0.77	1321
shame	0.30	0.05	0.08	306	*caring*	0.56	0.69	0.62	897
boredom	0.67	0.54	0.60	470	*anticipation*	0.58	0.81	0.67	1359
pitifulness /disappointment	0.68	0.88	0.77	2185	*comfort/cozy*	0.45	0.51	0.48	458
disgust	0.48	0.59	0.53	516	*welcome/favor*	0.56	0.83	0.67	1109
shock	0.45	0.50	0.47	704	*curiosity/interest*	0.57	0.77	0.66	1346
burden /unwillingness	0.43	0.33	0.37	606	*relief/trust*	0.53	0.75	0.62	945
fear/scary	0.36	0.26	0.30	164	*respect*	0.52	0.68	0.59	460
loathing/hatred	0.66	0.77	0.71	984	*pleased /(cute/pretty)*	0.60	0.64	0.62	524
guilt	0.00	0.00	0.00	84	*pride*	0.42	0.56	0.48	602
anxiety/worry	0.55	0.65	0.59	960	*arrogance /disregard*	0.44	0.50	0.47	743
doubt/distrust	0.61	0.78	0.69	1539	*surprise*	0.55	0.62	0.58	922
anger/rage	0.73	0.86	0.79	1538	*realization*	0.52	0.58	0.54	1030
defeat/self-hatred	0.39	0.21	0.27	208	*resolute*	0.47	0.43	0.45	416
laziness	0.39	0.20	0.26	290	*NO EMOTION*	0.54	0.59	0.56	725
sorrow	0.41	0.33	0.36	263					
preposterous	0.70	0.88	0.78	2055					
fed up	0.46	0.56	0.51	816	**micro avg**	0.60	0.72	0.66	39651
compassion	0.52	0.57	0.54	685	**macro avg**	0.54	0.61	0.56	39651
pathetic	0.64	0.80	0.71	1519	**weighted avg**	0.60	0.72	0.65	39651
fatigue/exhaustion	0.53	0.46	0.49	473	**samples avg**	0.61	0.75	0.65	39651

AUC									
complaint /dissatisfaction	0.94	*embarrassment /awkwardness*	0.84	*irritation*	0.92	*sadness*	0.90	*despair*	0.84
shame	0.74	*boredom*	0.88	*pitifulness /disappointment*	0.88	*disgust /repulsiveness*	0.89	*shock*	0.84
burden/unwillingness	0.79	*fear/scary*	0.89	*loathing /hatred*	0.93	*guilt*	0.86	*anxiety /worry*	0.86
doubt/distrust	0.87	*anger/rage*	0.94	*defeat /self-hatred*	0.84	*laziness*	0.82	*sorrow*	0.85
fed up	0.83	*preposterous*	0.89	*pity /compassion*	0.87	*pathetic*	0.88	*fatigue /exhaustion*	0.85
impressed/admiration	0.93	*happiness*	0.92	*joy*	0.93	*gratitude*	0.92	*pleasure /excitement*	0.93
care	0.89	*anticipation*	0.88	*comfort/cozy*	0.88	*welcome/favor*	0.89	*curiosity /interest*	0.87
relief/trust	0.89	*respect*	0.92	*pleased /(cute/pretty)*	0.92	*pride*	0.87	*arrogance /disregard*	0.83
surprise	0.85	*realization*	0.83	*resolute*	0.86	*NO EMOTION*	0.87	**macro avg**	0.88
MCC: 0.588									

Table 5: Performance metrics

10. Appendix C: Ethical Consideration

It is well known that a large dataset inevitably has discrimination against protected groups, and the demand of a fair model is not negligible. Our dataset is not an exception. In this section, we point out such problem and instantiate that a simple method helps to alleviate the discrimination. Here, we focus on gender discrimination as an example.

10.1. Bias Detection

The very first question is whether the texts in the source data are biased. We collected 3.2m comments for the our source dataset and sampled 50k for KOTE. To detect discrimination, we use comments not used for the learning. The comments that include words referring to protected groups and their counterparts are collected. Since we focus on gender discrimination, the texts containing one of the gender words, 여자*women*, 남자*men*, 여성*female*, and 남성*male*, are collected. Texts that mention both genders are removed. 53k and 38k texts are identified to have female words or male words, respectively. 30k texts are randomly sampled from each gender text set for emotion analysis.

The texts in both sets are analyzed by the `KcELECTRA` trained with KOTE, while the gender words are masked with the special token, [MASK]. As in **Figure 3**, the texts containing female words are generally evaluated more negatively, and the texts containing male words are generally evaluated more positively. In conclusion, the source data is biased in the first place, and thus the model could only be biased regardless of the potential discrimination of the raters.

The second question is whether and how much the trained model is biased. To answer this question, we borrow the basic idea of explainable machine learning via token switching. From the source data, we input 320k texts (10% of the total source data) into the model and select 500 non-overlapping texts that have the highest probabilities for each label (22k in total). Then, two randomly selected tokens (except `[PAD]`, `[CLS]`, and `[SEP]`) of each text are replaced with the female words (i.e., 여자*women* and 여성*female*) or the male words (i.e., 남자*men* and 남성*male*). As a result, 22k random-to-female switched texts and 22k random-to-male switched texts are produced. The model would evaluate the two text sets equally if it is fair.

The results are presented in **Figure 4**. The bars show the mean difference of each label's predicted probabilities between the two text sets. The light blue bars indicate the baseline model without a manipulation for fairness. The positive direction indicates the bias toward female. The baseline model evaluates the texts more negative on average when some tokens are replaced with the female words. In contrast, the same texts with the male words are evaluated more positive on average. In particular, the texts with the female words are evaluated discriminatorily for negative-intense emotions (e.g., 증오/혐오*loathing/hatred*, 화남/분노*anger/rage*, 역겨움/징그러움*disgust/repulsiveness*, 한심함*pathetic*, and 짜증*irritation*).

10.2. Unbiasing

One of the simplest but powerful methods to mitigate discrimination in a language dataset is data augmentation with token switching (Zhao et al., 2018; Park et al., 2018a). We swap the gender tokens to generate additional texts, and then add the generated texts on the train set.

940 texts in our train set are identified to have at least one gender word. The gender tokens in the texts are replaced with their antonym (여성*female* to 남성*male*, 여자*women* to 남자*men*, and vice versa) and these gender-swapped texts are added on the original train set to create 40,940 instances in total. Also, we trained a double and triple augmented model, in which the original texts and the gender-swapped texts are augmented one and two more times respectively, in order to accentuate the texts containing the gender tokens.

Figure 4 shows the results. The augmented models are less biased than the baseline model, and the double augmented model is the least biased. Furthermore, the augmented models cause no critical change in the performance metrics. In the case of double augmented model, the average F1-score increases by 0.002, the average AUC decreases by 0.0002, and the MCC hardly changes. Of course, there exist a variety of more thorough methods that help to mitigate biases (For survey and review, see Sun et al., 2019; Caton and Haas, 2020; Mehrabi et al., 2021). However, we would like to emphasize that bias can be alleviated with little effort, and the model performance may not be impaired much. Hence, it is recommended to use a fairer model. Especially, when the dataset is used for a machine designed for direct interaction with humans or other sensitive situations, a strong recommendation is to proceed with caution and go through the process of mitigating discrimination.

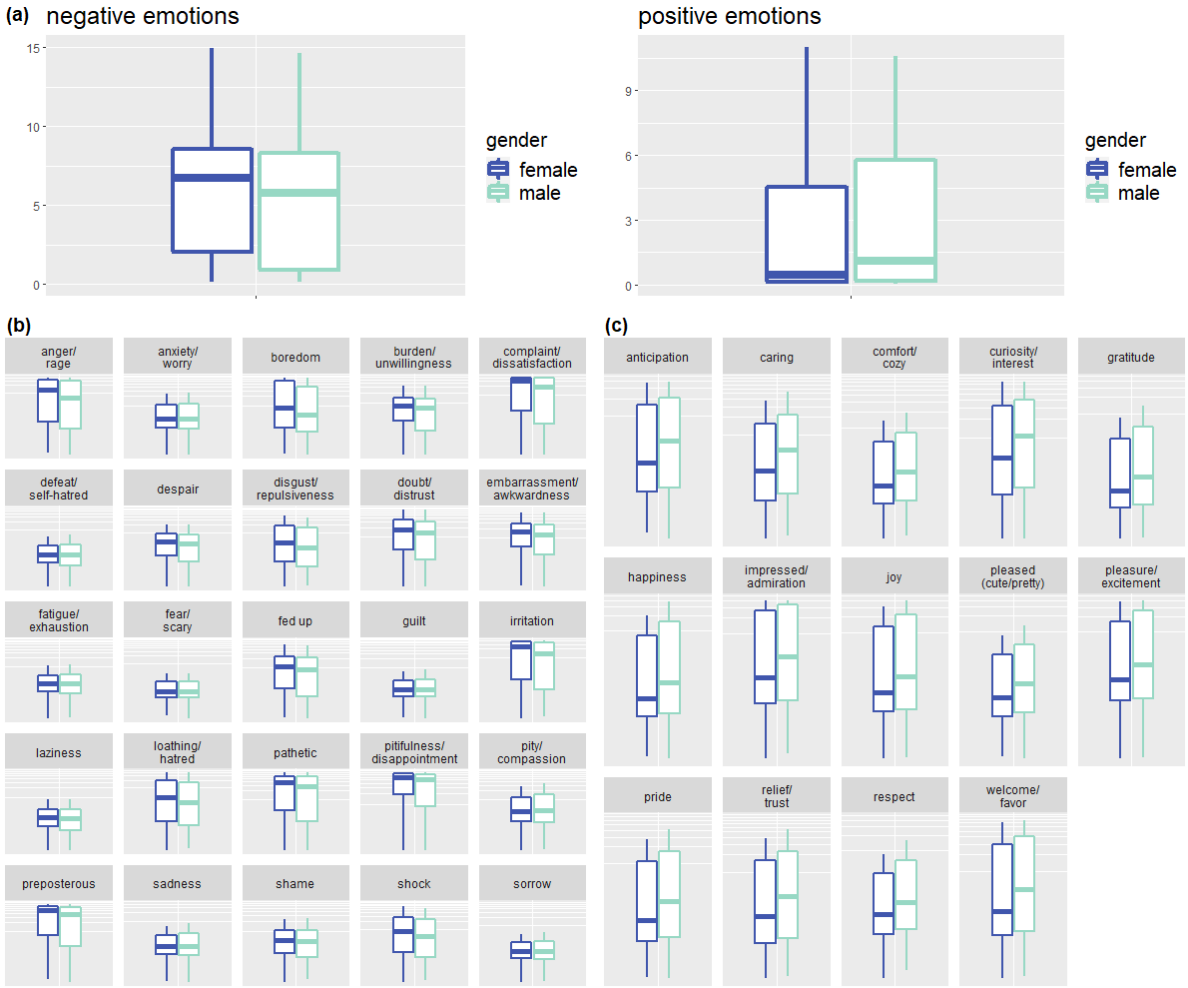

Figure 3: A comparison of emotions between female and male texts in which the gender tokens are masked. The first plot in (a) compares the sum of negative emotions of each comment in the gender text sets. The second plot in (a) compares the sum of positive emotions of each comment in the gender text sets. In (b) and (c), each box of each plot represents an emotion recognized in the 30k texts. (b) shows how different each negative emotion is by gender, and (c) shows how different each positive emotion is by gender. (b) and (c) are log transformed to illustrate the differences visually. (plot package; ggplot2)

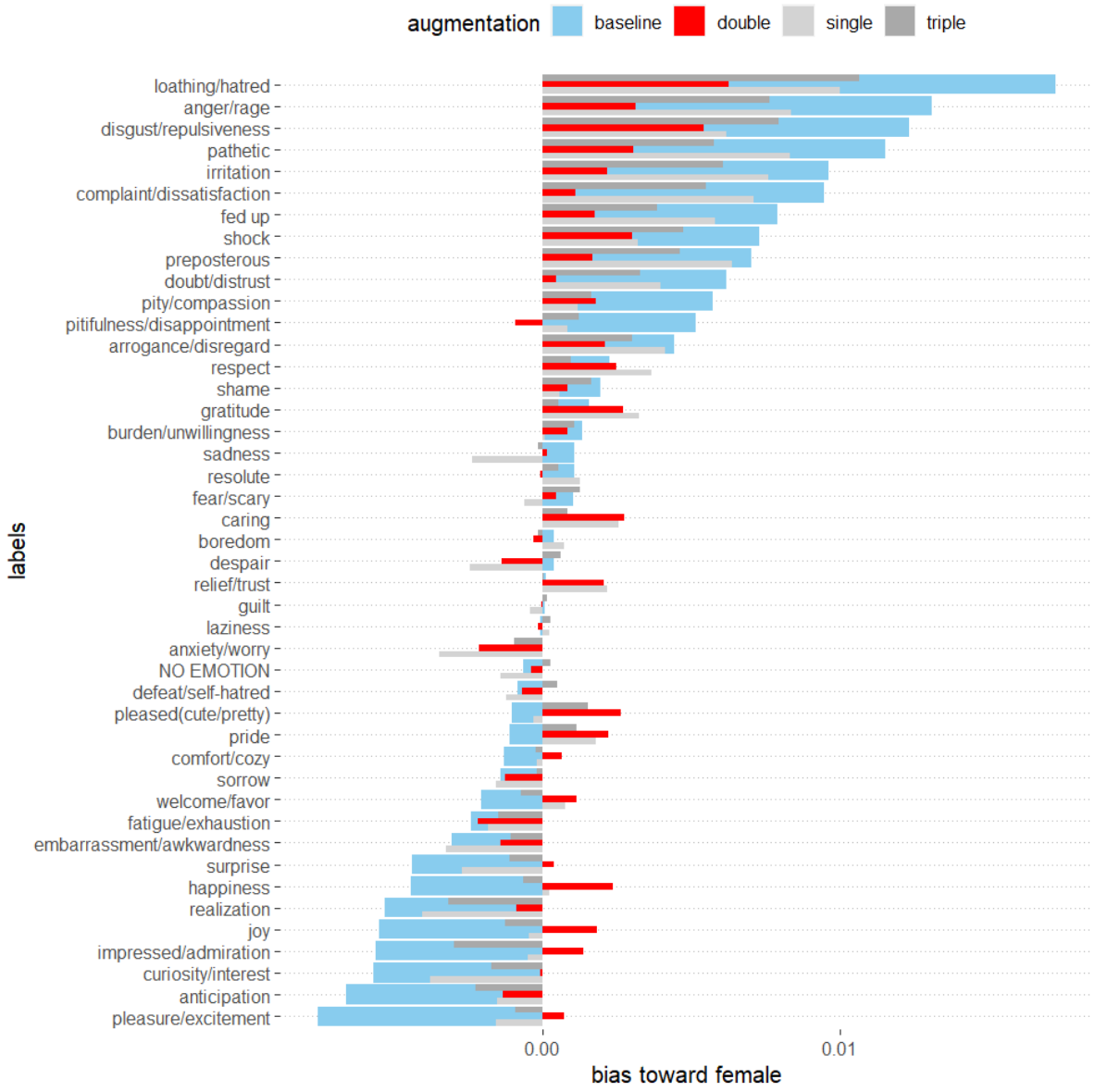

Figure 4: The bars indicate the mean difference of each label's probabilities between the texts in which two random tokens are replaced with the female words and the texts in which two random tokens are replaced with the male words. The texts with female words are evaluated more negative. The bias is most serious in the baseline model (the light blue bars). On the other hand, models trained with additional gender-swapped texts are relatively less biased, and the decrease of the bias is largest when the gender-swapped texts as well as the original texts containing gender words are augmented twice (the red bars). (plot packages; ggplot2 and ggpubr.)