User Guide for KOTE: Korean Online Comments Emotions Dataset

Duyoung Jeon and **Junho Lee** and **Cheongtag Kim**
Department of Psychology, Seoul National University
Seoul, Republic of Korea
wuju1201@gmail.com, {smbslt3, ctkim}@snu.ac.kr

Abstract

Sentiment analysis that classifies data into positive or negative has been dominantly used to recognize emotional aspects of texts, despite the deficit of thorough examination of emotional meanings. Recently, corpora labeled with more than just valence are built to exceed this limit. However, most Korean emotion corpora are small in the number of instances and cover a limited range of emotions. We introduce KOTE dataset[1]. KOTE contains 50k (250k cases) Korean online comments, each of which is manually labeled for 43 emotion labels or one special label (NO EMOTION) by crowdsourcing (Ps = 3,048). The emotion taxonomy of the 43 emotions is systematically established by cluster analysis of Korean emotion concepts expressed on word embedding space. After explaining how KOTE is developed, we also discuss the results of finetuning and analysis for social discrimination in the corpus.

1 Introduction

Sentiment analysis that classifies texts into positive or negative has been the most widely used method to analyze the emotional aspect of texts. Although sentiment analysis is simple, feasible, and useful in various situations, the need for more sophisticated emotions beyond just valence for text analysis is emerging. This is due to the advent of powerful language models that can accommodate complicatedly labeled data and the recent advancement in computing power.

The demand for an emotion analysis tool for the Korean language is high. However, most Korean emotion corpora are small in the number of instances and have coarse emotion taxonomies that cover only a limited range of emotions. As a result, GoEmotions (Demszky et al., 2020), an English dataset that is large (58k instances) and has

Text
you silly cat made a fuss just because you didn't want to take a bath?? LOL 😂
Labels
rater 1 *preposterous, attracted, care, happiness*
rater 2 *preposterous, attracted, embarrassment, realization*
rater 3 *preposterous, interest, embarrassment, irritation, dissatisfaction*
rater 4 *preposterous*
rater 5 *attracted, interest, excitement*

Table 1: A raw example in KOTE.

a fine-grained emotion taxonomy (27 emotions or neutral), is widely used for emotion analysis for Korean text using machine translation, despite the imperfect translation quality. The Korean language model trained with translated GoEmotions is downloaded up to hundreds of thousands of times a month in Hugging Face[2].

However, emotions are strongly related to culture since they are products of culture-specific schema. Accordingly, emotion taxonomies representing underlying emotion structures vary across cultures (Mesquita and Frijda, 1992) and the variation even holds for basic emotions (Gendron et al., 2014). This demonstrates the need to create a culturally relevant dataset that is labeled with a culturally relevant emotion taxonomy.

To create a culturally relevant database, we developed KOTE (Korean Online That-gul[3] Emotions), a large language dataset of 50k Korean online comments labeled for 43 emotions. The online comments in KOTE are collected from 12 different platforms of various domains (*news, online community, social media, e-commerce, video platform, movie*

---

[1] https://github.com/searle-j/KOTE

[2] https://huggingface.co/monologg/bert-base-cased-goemotions-original

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

review, microblog, and forum). 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.

The purpose of this study is twofold. The first purpose suggests a new emotion taxonomy that is suitable to the Korean language in general. The second purpose builds KOTE with the new taxonomy. We also finetuned the pretrained 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 versus 0.41). There is much room to improve since the results are not tuned. A diversity of strategies can possibly be applied on the raw data according to the individual purpose of an analyst because KOTE is fully open and contains rich information.

2 Related Work

2.1 Emotion Taxonomy

Constructing an emotion corpus requires an appropriate emotion taxonomy by which the texts are labeled. To find the appropriate emotion taxonomy, constructing an emotion words dataset must precede to obtain all available emotions each of which is treated as a candidate to be included 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 nonemotion 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 nonemotion 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 scrutinize approximately 18k psychological concepts and concluded that 717 words contained emotionality. For an example of a Korean study, 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. One classic way of the vectorization is by human rating, which is conducted by asking human annotators to rate each word in a few scales designed 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 via 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, p.75) 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). The notable points that the studies imply in common are: **i)** The fixed dimensionality of emotion may not exist, which varies depending on research setting, and **ii)** The emotion is a complex structure. More than six or seven basic emotions can stand alone. Accordingly, the emotion taxonomy of this study considers the 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 sub-

Dataset	Unit	# of instances	Label dimension
Korean Emotion Words Inventory (Park and Min, 2005)	word	434	4
Korean Emotion Vocabulary Taxonomy (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*).

ject of machine learning. Machine learning methods have introduced many useful tools to obtain rich information of words, which are competent when compared with the traditional approaches in a couple of ways. They are more efficient than the human annotation, and thus allow to handle big language data. They also impose more abundant information on language while the language annotated by a human contains information restricted in a specific research design.

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.

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)** Inputting the emotion words into a large pretrained word vector model to obtain a vector for every word; 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 Korean Emotion Words Inventory (Park and Min, 2005), Korean Emotion Vocabulary Taxonomy (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 emo-

tional 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 used these three datasets to categorize emotions.

However, the lexicons include some expressions that express emotions figuratively (e.g., *many*). These expressions are excluded because they are more often not used as emotional usage. Moreover, some expressions are missing, and thus we manually added extra expressions. Then, the expressions were tokenized by python package, KoNLPy (Park and Cho, 2014) and function words as well as stop words were deleted. We chose 3,017 expressions that we considered directly represent human emotions, which were inputted into the pretrained 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) pretrained with large language datasets such as the Korean Wikipedia[4]. 1,787 words were included in our candidate emotion words list and 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.

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 reckoned as noise. Clusters are considered valid, only if they survive long enough in this process. The `HDBSCAN` es-

---

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

timates 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. **Figure 1** (a) shows the histogram of the number of clusters. **Figure 1** (b) - (f) show histograms marginalized on each hyperparameter space. 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 noninterpretable 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
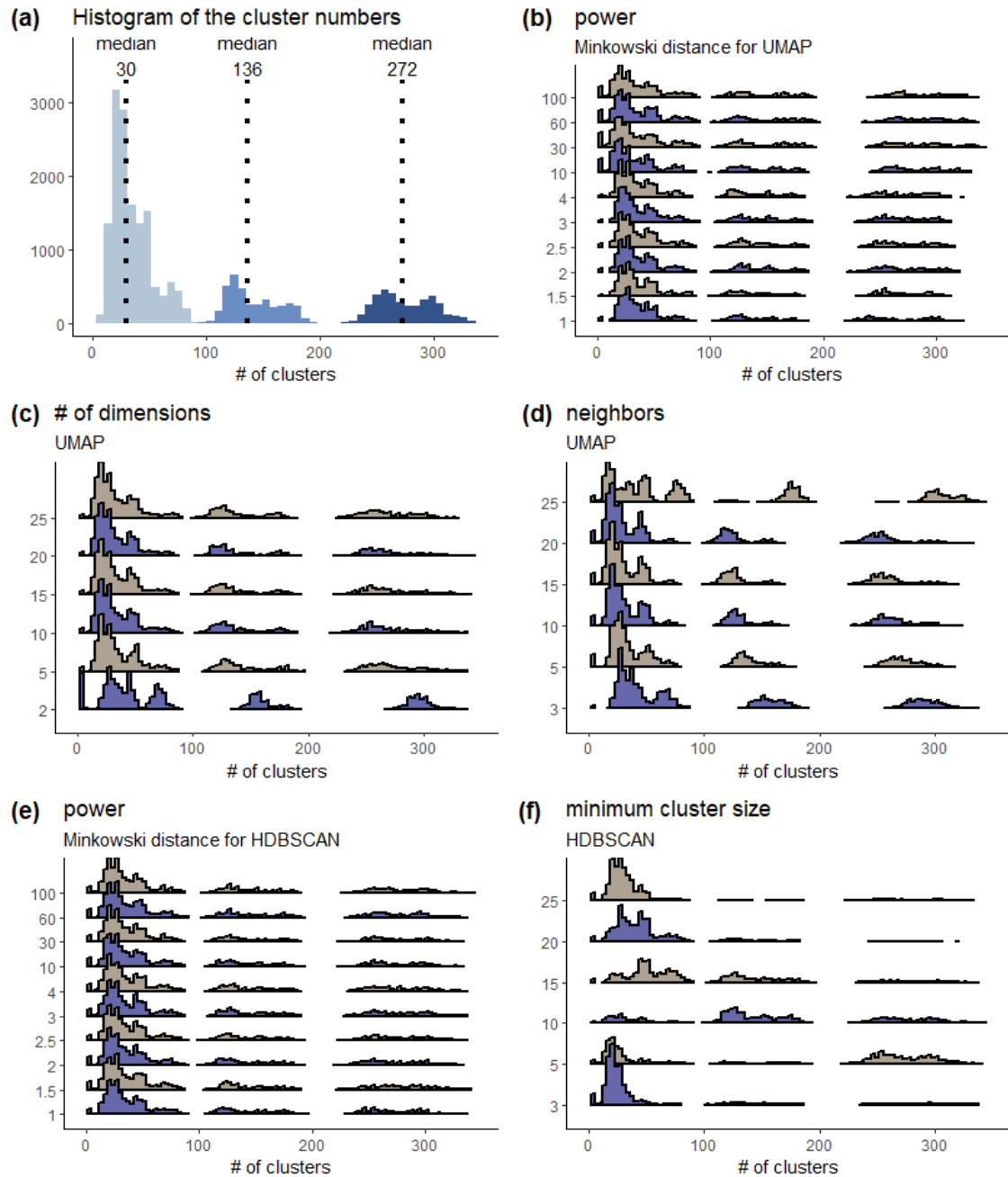
Figure 1: (a) is the histogram for the number of clusters in 21,562 partition sets. Three distributions are identified. (b) – (f) are histograms marginalized on each hyperparameter space. The y-axes represent the searched values of the hyperparameters. Three distributions are consistently identified. 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.

interpreted (see **Appendix A**).

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 finetuning on a pretrained 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 crawling unless no guideline was provided. If a website supports a search engine, randomly selected emotion words from KNU SentiLex were searched for crawling 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, the minimum length of the texts is set as 10, and the maximum as the 90th percentile of each platform. The grand maximum length is 404, the mean is 57.32, and the median is 42[5].

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 is 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 label but a special label, NO EMOTION. They are also instructed to select plausible emotions and

---

[5]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.

not NO EMOTION, 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 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 'expectancy' 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 comment, and thus 250k cases of 50k comments are created as a result. Five binary labels of a comment are summed to be the final label. Thus, the range of a label is 0–5. (see **Table 1**. Three out of the five raters agreed that the text contains *attracted*, so the value of *attracted* label is 3)

4.3 Data Description

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.

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.

No additory preprocessing is applied on the data

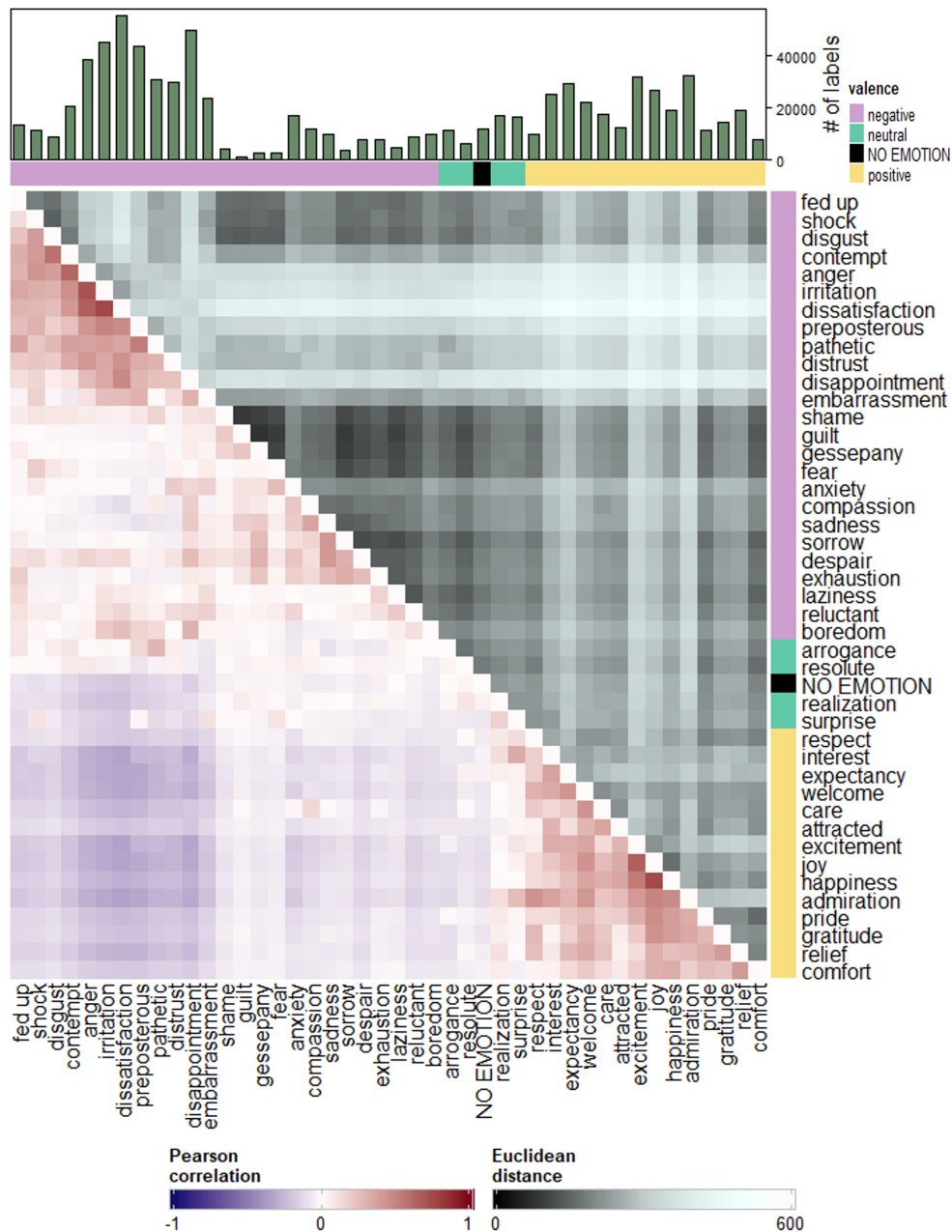| agreement					
at least one label of x or higher	x=1	x=2	x=3	x=4	x=5
# of texts	50,000	49,663	42,845	28,650	11,760
(% to total)	(100%)	(99%)	(86%)	(57%)	(24%)
texts labeled for NO EMOTION					
NO EMOTION	0	1	2	3	4	5
# of texts	42,156	5,243	1,592	644	264	101
(% to total)	(84%)	(10%)	(3%)	(1%)	(0.5%)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

to merge or exclude emotions even though some emotions are linearly related. This is not only because the emotion taxonomy is derived by a nonlinear method, but also the ELECTRA model, which would be finetuned, is nonlinear and potentially able to distinguish linearly similar emotions. In addition, significant emotions differ depending on the method and the criterion. There is no panacea to the best of our knowledge. Lastly, nonsignificant dimensions can additionally provide useful information, despite the risk of redundancy.

4.4 Finetuning

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

Training. We finetuned KcELECTRA, a language model pretrained 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 is 32, and the input token size is 512. If the number of tokens of an input is less than 512, it is padded with a special token, [PAD]. No input exceeds 512 in length. One linear layer is added on the [CLS] token of the last hidden layer for multi-label classification. The loss is binary cross entropy for each label. We use 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 switch 5% of tokens with a random token (except [CLS], [SEP], and [PAD]), and mask 5% of tokens with a special token, [MASK]. The maximum number of epochs is set as 15, but 9 epochs are enough to reach the optimum in almost all cases. The loss of the validation set is monitored during the learning. 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 is set as 0.3. We use 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) are 0.56, 0.88, and 0.59, respectively (see **Appendix B** for full description).

As mentioned in the Introduction section, these results are obtained with arbitrarily decided hyperparameters. Therefore, the performance can 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. If a good result is obtained, we hope it would be shared without hesitation.

5 Conclusions

The model finetuned with our dataset achieved a better performance than the existing model finetuned 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 emotions versus 27 emotions). 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 emotions. **ii)** Emotion is a dynamic structure, but we treat it as a static structure 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 C** 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

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, Jeong-woo 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.

A Appendix: Emotion Clusters

Valence	Interpretation	Example words in the cluster
Negative	*dissatisfaction*	dissatisfied, oppose, criticize, complaint
	embarrassment	embarrassed, disconcerted, awkward, untoward
	irritation	irritated, pissed off, ridiculous
	sadness	sad, miss, lonely, tear
	despair	frustrated, joys & sorrows, hurt, grief, letdown
	shame	ashamed, humiliated
	boredom	bored, tedium, trite, dull
	disappointment	disappointed, sorry, upset, deplorable, regretful
	disgust	disgusted, repulsive, dirty
	shock	shocked, flabbergasted, pass out, freaked out
	reluctant	unwilling, denial, pressure, cannot be bothered, give up
	fear	fear, anxious, tense, pressed
	contempt	contempt, hatred, scorn, vilifying
	guilt	guilt, blamed, repentance, remorse
	anxiety	apprehensive, worry, threatened
	distrust	suspicious, doubtful, lie
	anger	anger, rage, obsessed, fury
	gessepany	failure, miserably, extorted
	laziness	bothered, dawdling
	sorrow	sorrowful, mirthless, weary, sobbing, upset, complicated
	fed up	fed up, struggle, arduous, sick and tired
	preposterous	dumbfounded, stunned, sttufy, enervated, WTF
	compassion	pity, sadly, chocked up, heartrending
	pathetic	pathetic, belittled, stupid, impudence
	exhaustion	tired, peak, exhausted
Positive	*admiration*	admiring, great, praise, compliment
	happiness	happy, affection, valuable, hope, luck
	joy	delight, ecstasy, love
	gratitude	praiseworthy, commendable, favor, blessing, mercy
	excitement	excited, funny
	care	caring, adore, dear
	expectancy	new, achieve, together, harmonious, vitality
	comfort	comfortable, ease, cozy, cool, warm
	welcome	welcome, approval, kindness, enthusiastic
	interest	interested, curious
	relief	relief, trust, intimate, close
	respect	respect, loyal, veneration, follow, obedience
	attracted	handsome, pretty, sweet, thrilled, cute, aegyo
	pride	successful, victory, worthwhile, accomplish
Neutral	*arrogance*	arrogance, pompous, ignore, bragging, boast, gasconade
	surprise	astonished, startled
	realization	realize, enlightened, wakened, conviction, belief
	resolute	resolute, determination

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

B Appendix: Performance Metrics

			F1-score						
emotion	precision	recall	F1	#	emotion	precision	recall	F1	#
dissatisfaction	0.78	0.89	0.83	2113	*admiration*	0.67	0.86	0.75	1323
embarrassment	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	*excitement*	0.69	0.86	0.77	1321
shame	0.30	0.05	0.08	306	*care*	0.56	0.69	0.62	897
boredom	0.67	0.54	0.60	470	*expectancy*	0.58	0.81	0.67	1359
disappointment	0.68	0.88	0.77	2185	*comfort*	0.45	0.51	0.48	458
disgust	0.48	0.59	0.53	516	*welcome*	0.56	0.83	0.67	1109
shock	0.45	0.50	0.47	704	*interest*	0.57	0.77	0.66	1346
reluctant	0.43	0.33	0.37	606	*relief*	0.53	0.75	0.62	945
fear	0.36	0.26	0.30	164	*respect*	0.52	0.68	0.59	460
contempt	0.66	0.77	0.71	984	*attracted*	0.60	0.64	0.62	524
guilt	0.00	0.00	0.00	84	*pride*	0.42	0.56	0.48	602
anxiety	0.55	0.65	0.59	960	*arrogance*	0.44	0.50	0.47	743
distrust	0.61	0.78	0.69	1539	*surprise*	0.55	0.62	0.58	922
anger	0.73	0.86	0.79	1538	*realization*	0.52	0.58	0.54	1030
gessepany	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
exhaustion	0.53	0.46	0.49	473	**samples avg**	0.61	0.75	0.65	39651

				AUC					
dissatisfaction	0.94	*embarrassment*	0.84	*irritation*	0.92	*sadness*	0.90	*despair*	0.84
shame	0.74	*boredom*	0.88	*disappointment*	0.88	*disgust*	0.89	*shock*	0.84
reluctant	0.79	*fear*	0.89	*contempt*	0.93	*guilt*	0.86	*anxiety*	0.86
distrust	0.87	*anger*	0.94	*gessapany*	0.84	*laziness*	0.82	*sorrow*	0.85
fed up	0.83	*preposterous*	0.89	*compassion*	0.87	*pathetic*	0.88	*exhaustion*	0.85
admiration	0.93	*happiness*	0.92	*joy*	0.93	*gratitude*	0.92	*excitement*	0.93
care	0.89	*expectancy*	0.88	*comfort*	0.88	*welcome*	0.89	*interest*	0.87
relief	0.89	*respect*	0.92	*attracted*	0.92	*pride*	0.87	*arrogance*	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

C Appendix: 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.

C.1 Bias Detection

The very first question is whether the texts in the source data are biased. We collected 3.2m comments for the source data 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, *women*, *men*, *female*, and *male*, are collected. Texts that have 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 nonoverlapping 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., *women* and *female*) or the male words (i.e., *men* and *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., *contempt, anger, disgust, pathetic*, and *irritation*).

C.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 (*female* to *male*, *women* to *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 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 attention, 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.
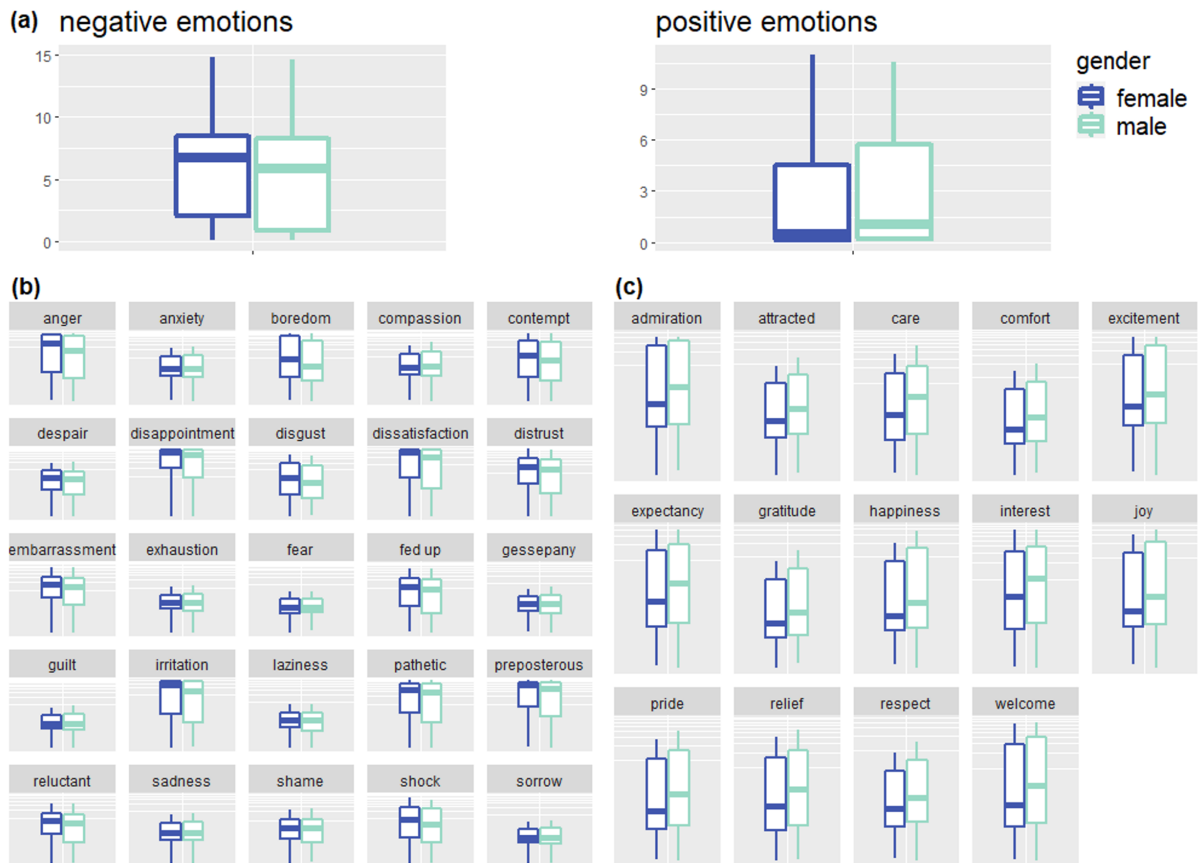
Figure 3: A comparison of emotions between female and male texts in which the gender tokens are masked. The first plot in (a) compares the sum of negative emotions of each comment in the gender text sets. The second plot in (a) compares the sum of positive emotions of each comment in the gender text 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)
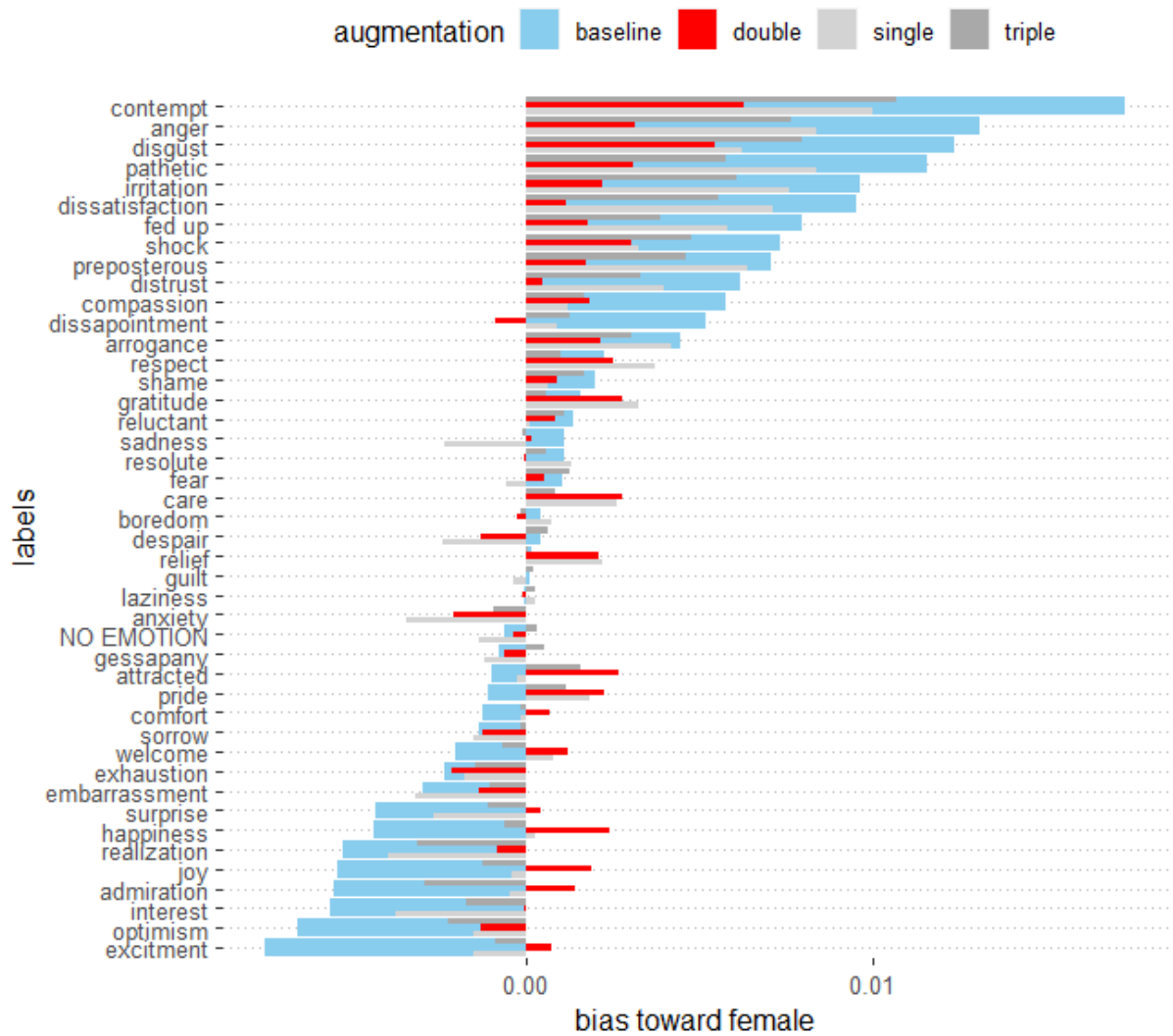
Figure 4: The bars indicate the mean difference of each label's probabilities between the texts in which two random tokens are replaced with the female words and the texts in which two random tokens are replaced with the male words. The texts with female words are evaluated more negative. The bias is 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.)