



## 저작자표시-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master's Thesis of Psychology

# Construction and Evaluation of Korean Online Comment Emotions Dataset KOTE

한국어 댓글 정서 데이터셋  
KOTE의 구축과 평가

August 2022

Graduate School of Social Science  
Seoul National University  
Psychology Major

Jeon Duyoung

# Construction and Evaluation of Korean Online Comment Emotions Dataset KOTE

Kim Cheongtag

Submitting a master's thesis of  
psychology

August 2022

Graduate School of Social Science  
Seoul National University  
Psychology Major

Jeon Duyoung

Confirming the master's thesis written by  
Jeon Duyoung  
August 2022

Chair	<u>Park Jooyong</u>	(Seal)
Vice Chair	<u>Hahn Sowon</u>	(Seal)
Examiner	<u>Kim Cheongtag</u>	(Seal)

# Abstract

Sentiment analysis that classifies texts into positive or negative has been dominantly used to recognize emotion of text in the field of natural language processing, despite the deficit of thorough examination of emotional meaning. Recently, corpora labeled with more than just valence are built to exceed the limit of the sentiment analysis. However, most Korean emotion corpora are small in the number of instances and cover a limited range of emotions. In addition, the labeling criteria are uncritically adopted from existing studies, or even decided by heuristics. We introduce the KOTE dataset. 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 ( $P_s = 3,048$ ). The emotion taxonomy of the 43 emotions is systematically established by clustering 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.

**Keyword :** emotion, sentiment, NLP, dataset

**Student Number :** 2020-29582

# Table of Contents

Chapter 1. Introduction .....	1
Chapter 2. Emotion Taxonomy .....	5
Chapter 3. KOTE.....	9
Chapter 4. Bias .....	18
Chapter 5. Conclusions .....	21
 Bibliography.....	 24
Appendix .....	28
Abstract in Korean .....	29

# Chapter 1. Introduction

Sentiment analysis aims to extract various information encompassing sentiments, emotions, opinions, appraisals, and attitudes contained in data (Zhang et al., 2018). It is applicable to various types of data, such as text and image (e.g., Yuan et al., 2013; Hutto & Gilbert, 2014; You et al., 2015; Mohammad, 2016; You et al., 2017). The text, in particular, is a major subject of the sentiment analysis because it contains emotions to some extent in most cases unless it is an explanatory document like Wikipedia.

The text sentiment analysis has been actively utilized in many disciplines, such as politics, communication, sociology, recommendation system, and psychology (e.g., Hu et al., 2013; Bakliwal et al., 2016; Haselmayer & Jenny, 2017; Zucco et al., 2017; Sailunaz & Alhajj, 2019; Provoost et al., 2019). It is evident that the sentiment analysis has gained lots of attention and shown its usefulness as well as feasibility, in that numerous studies are still being conducted.

To conduct a text sentiment analysis, a machine learning model trained with an emotion corpus is mainly used. The emotion corpus is a corpus in which the texts have emotion labels. For example, a review dataset can be considered as a simple emotion corpus with three-dimensional labels when the review texts are classified into positive, negative, or neutral according to the star ratings annotated on the texts. The label dimensions can be expanded as much as desired in order to recognize more detailed emotions.

However, there exist only few available Korean emotion corpora because the construction of an emotion corpus is costly; computer-based automatic labeling is inappropriate since emotions are too complex, rather each text should be labeled by human raters. Furthermore, the Korean emotion corpora are mostly small in the number of instances and cover only a limited range of emotions (see **Table 1**).

For this reason, GoEmotions (Demszky et al., 2020), an English dataset that is large (58k instances) and has a fine-grained emotion taxonomy (27 emotions or neutral), is widely used for emotion analysis for Korean text using machine translation. The Korean language model trained with machine-translated GoEmotions is downloaded about hundreds of thousands of times

Dataset	Unit	# of instances	Label dimension	Labeling method
Korean Emotion Words Inventory (Park & Min, 2005)	Word	434	4	Continuous
Korean Emotion Vocabulary Taxonomy (Sohn et al., 2012)	Word	504	11	Continuous
KOSAC (Jang et al., 2013)	Sentence	7.7k	2*	Continuous
NSMC (Naver, 2015)	Sentence	200k	1	Binary
KNU SentiLex (Park et al., 2018)	n-gram	14k	1	Continuous
Korean Continuous Dialogue Dataset with Emotion Information (KETI, 2020)	Dialogue	10k (55k sentences)	7	One-hot
Korean One-off Dialogue Dataset with Emotion Information (KETI, 2020)	Sentence	38k	7	One-hot
Emotional Dialogue Corpus (AI HUB, 2021)	Dialogue	15k (270k sentences)	60	One-hot

**Table 1:** Korean emotion corpora.

\* KOSAC contains far more plentiful information, but two dimensions are closely related to emotion (*polarity* and *intensity*).

a month in Hugging Face<sup>1</sup>.

The model trained with translated GoEmotions has several limitations: **i)** Since the texts in GoEmotions are Reddit comments that include lots of grammatically incorrect expressions and slangs, the performance of the machine translation is not satisfactory. **ii)** The emotions differ across cultures; thus, it is inadequate to uncritically adopt the emotion taxonomy of GoEmotions.

Especially the cultural difference is a major problem. Since the emotions are products of culture-specific schema, emotion taxonomies vary across

<sup>1</sup> <https://huggingface.co/monologg/bert-base-cased-goemotions-original>

cultures (Mesquita & Frijda, 1992) and the variation even holds for so called 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.

Constructing an emotion corpus requires an appropriate emotion taxonomy by which the texts are labeled. To find the appropriate emotion taxonomy, constructing an emotion word 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 emotions. Vocabularies representing emotion concepts can be used to this end. In traditional approaches, the distinction between emotion and nonemotion can be determined by human rating. Shields (1984) attempted to conceptualize *emotionality* by asking participants to categorize 60 feeling words (e.g., *happy*, *curious*, *hungry*) into emotion words or nonemotion words. Clore, Ortony, and Foss (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 decision of 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 mathematically analyzable form. This step is essential to create a semantic space spanned by the words and conduct further analysis about its dimensionality and the relations among the words. One popular way is vectorization, which imposes vector-shaped information on words by a certain measure. One classic way of the vectorization is by using human rating, which is performed 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*).



In case not information itself of each word, but the relation among the words is of interest, similarity (or distance) measure can be used. Storm and Storm (1987) utilized a sorting method to extract co-occurrence information from emotion words. More recently, 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.

One can attempt to uncover an adequate structure of the emotion words with the mathematically analyzable 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 & Oliver, 1993; Lee & Lim, 2002; Cowen & Keltner, 2017; Keltner et al., 2019), and language (Shaver et al., 1987; Hupka, Lenton, & Hutchison, 1999; Cowen et al., 2019). The notable points that the studies imply in common are: **i)** The fixed dimensionality of emotion may not exist. It 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 and analyzed 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 the 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. Moreover, they impose more abundant information on language while the language annotated by humans 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 taxonomy for the Korean language.

To develop a culturally relevant and more accurate database with the new

taxonomy, we developed KOTE (Korean Online That-gul<sup>2</sup> Emotions), 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 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 2** shows a sample instance in KOTE.

The purpose of this study is twofold. The first is suggesting a new emotion taxonomy that is suitable to Korean language in general. The second is building KOTE with the new taxonomy. We also finetuned the pretrained KcELECTRA (Korean comment ELECTRA; Clark et al., 2020; Lee, 2021) model with KOTE and checked the performance. This achieves 0.56, 0.88, and 0.59 for average F1-score, AUC and MCC, respectively. 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 the data is fully open and contains rich information<sup>3</sup>.

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

**Table 2:** A sample instance in KOTE.

## Chapter 2. Emotion Taxonomy

In the present study, we construct a new 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

<sup>2</sup> ‘That-gul’ or ‘Daet-gul’ is a Korean word that refers to ‘online comment’.

<sup>3</sup> <https://github.com/searle-j/KOTE>

<sup>4</sup> The text is translated into English for readers’ understanding. The original texts barely include English.

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.

## 2.1. Emotion Words

There are a few Korean emotion lexicons such as Korean Emotion Words Inventory (Park & Min, 2005), Korean Emotion Vocabulary Taxonomy (Sohn et al., 2012), and KNU SentiLex (Park et al., 2018). In particular, 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 glosses using Bi-LSTM (Bidirectional Long-Short Term Memory; Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997; Graves & 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 used the three Korean emotion lexicons.

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 & Jo, 2014) and function words as well as stop words were deleted. We chose 3,017 expressions that we consider directly represent human emotions, which were inputted into the pretrained word vector model in the next step.

## 2.2. Word Vectorization

The 3,017 emotion words were inputted into a `fastText` model (Bojanowski et al., 2017) pretrained with a large language data such as the Korean Wikipedia<sup>5</sup>. 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.

---

<sup>5</sup> <https://github.com/ratsgo/embedding/releases>

## 2.3. Exploring Dimensionality of Emotion

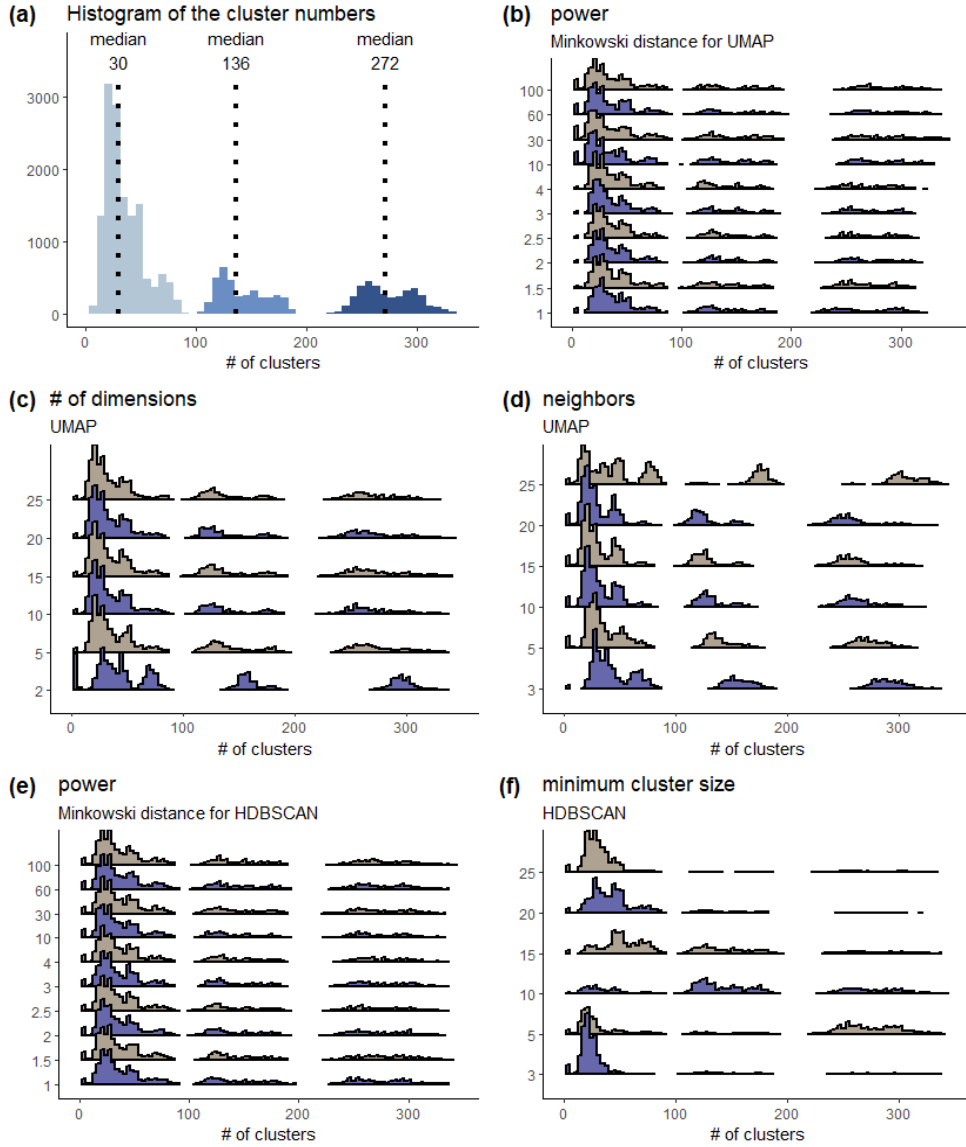
### Base Clustering

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 *survival algorithm*. Clusters in a HDBSCAN model diminish as its criteria, by which a data point is considered to belong to a cluster, gradually becoming 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 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. Let us consider a space where values represent the number of clusters, and the axes indicate the hyperparameters. By analogy, the approximate shape of the function in the space is grasped by fumbling it using the grid search, since the precise shape cannot be analytically calculated. The hyperparameters to be searched and the searched values are presented in **Fig 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. **Fig 1 (a)** shows the histogram of the number of clusters. **Fig 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 minimum cluster size of HDBSCAN ( $r = -0.2$ ). The most likely number of clusters is 30 as in **Fig 1 (a)**, the median of the largest distribution. This result is consistent with many previous studies. Ho-



**Fig 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 for HDNSCAN ( $r = -0.2$ ). (Drawn with R packages, ggpilot2 (Wickham, 2011), ggpubr (Kassambara & 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 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.

wever, 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, the median of the second largest distribution.

### **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 results from one or various clustering methods to derive one single agreed outcome. We use HBGF (Hybrid Bipartite Graph Formulation; Fern & Brodley, 2004), which utilizes both instance- and cluster-based graph formulation (See also Vega-Pons & Ruiz-Shulcloper, 2011; Karypis & Kumar, 1998). The 21,562 partitions sets were used in a HBGF model to reach consensus for how to split 1,787 emotion words into 136 groups.

The meaning of each cluster is interpreted, while some clusters are uninterpretable because seemingly unrelated words are entangled together. Theses clusters were dopped. If antonyms are in a same cluster, they are regarded as two separate emotions (i.e., *sadness* and *joy*). 43 emotions were clearly interpreted (see **Table 3**).

## **Chapter 3. KOTE**

We propose 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.

### **3.1. Text**

50k online comments in KOTE are collected from 12 different platforms (*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

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

**Table 3:** Interpretation of each interpretable cluster and emotion words in it. Of course, all the words are translated into English in this table. See the appendix for original Korean words.

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 90<sup>th</sup> percentile of each platform. The grand maximum length is 404, the mean is 57.32, and the median is 42<sup>6</sup>.

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.

### 3.2. Label

The 50k comments were labeled by crowdsourcing which 3,084 raters whose mother tongue is Korean participated in with monetary reward. The labeling process is as follow: 50 random comments are given to a rater. The raters choose 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 not NO EMOTION, if they think the 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 the context. The minimum and the maximum number of labels they can choose for one comment are 1 and 10, respectively. The raters 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 process can be repeated three times at maximum.

Two types of catch trials are given in the middle of the labeling. The

---

<sup>6</sup> 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.



raters were informed about the catch trials before the answering and agreed that the labeling process can 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** for a sample. Three out of five raters agreed that the text contains *attracted*, so the value of *attracted* label is 3)

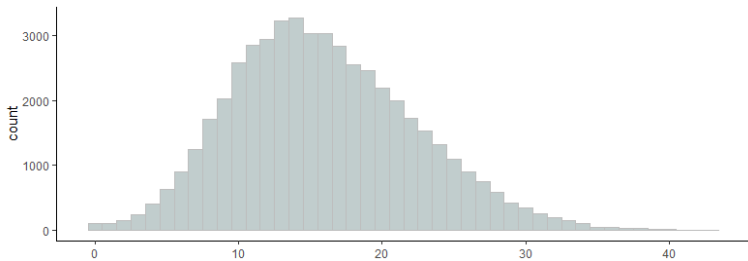
### 3.3. Data Description

**Table 4** describes the data. 99% texts have at least one label of 2 or higher, which means that 99% texts have at least one label that two or more labelers choose in common. It is evident that the raters did not have much difficulty to reach a consensus. A moderate number of texts are labeled for NO EMOTION. It is because online comments mostly contain emotion to some extent and the comments are collected in the search results of various emotion expressions.

The relations among the labels are presented in the heatmap in **Fig 2**, which shows Pearson correlation and Euclidean distance among the labels, each of which is a 50k-dimensional vector. In the perspective of Euclidean distance, the emotions of low intensity seem to be located close. **Fig 3** shows the phylogram generated by the Euclidean distances. In the correlation heatmap, on the other hand, negative-intense, negative-weak, and positive emotions seem related separately. **Fig 4** shows the 2-dimensional projection by UMAP with correlation distance. The valence is primary, but each valence has many ramifications.

No additory preprocessing is applied on the data 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

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



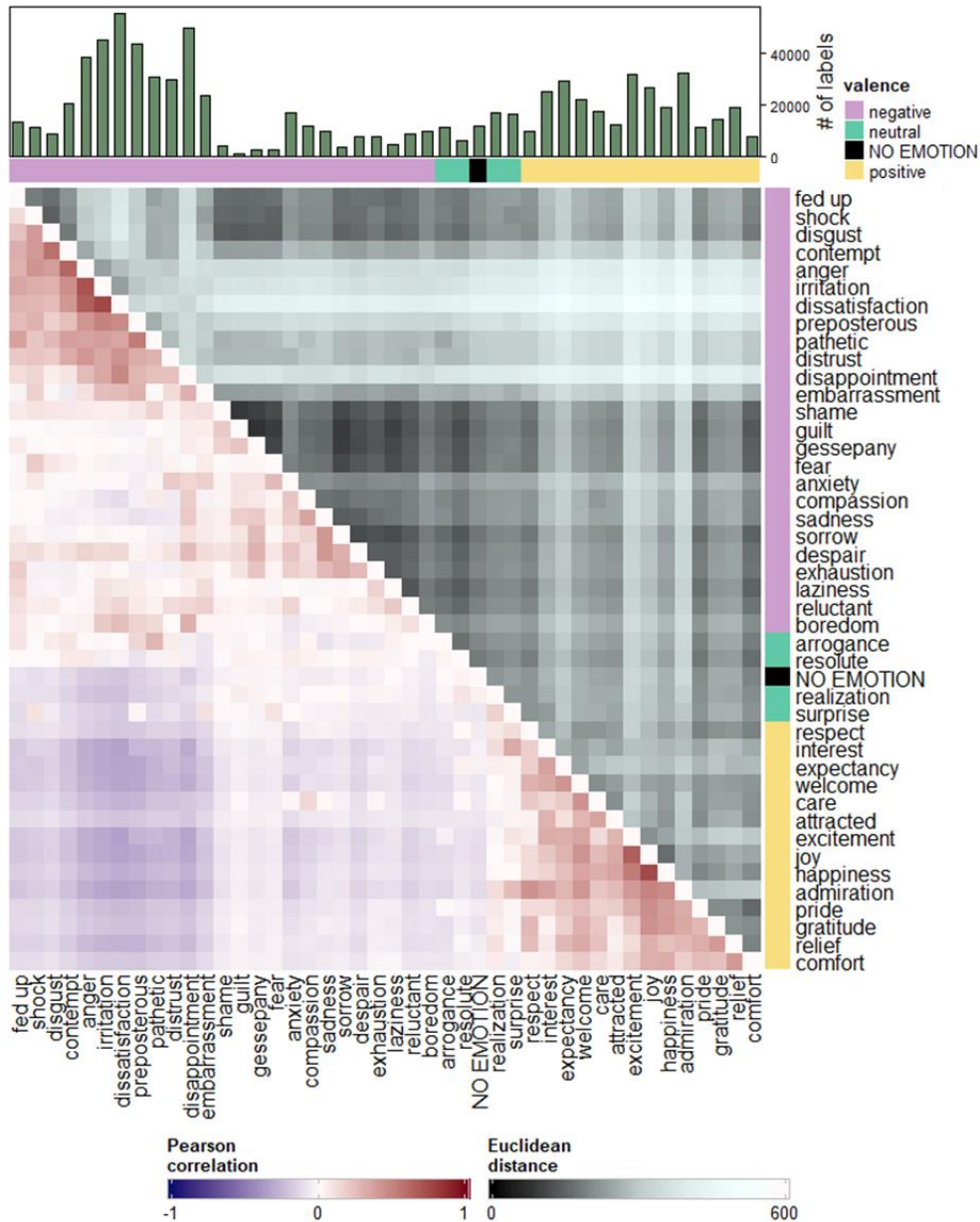
**Table 4:** Description of the labeled data.

ELECTRA model, which would be finetuned, is nonlinear and potentially able to distinguish linearly similar emotions. In addition, significant dimensions as well as emotions related to them 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.

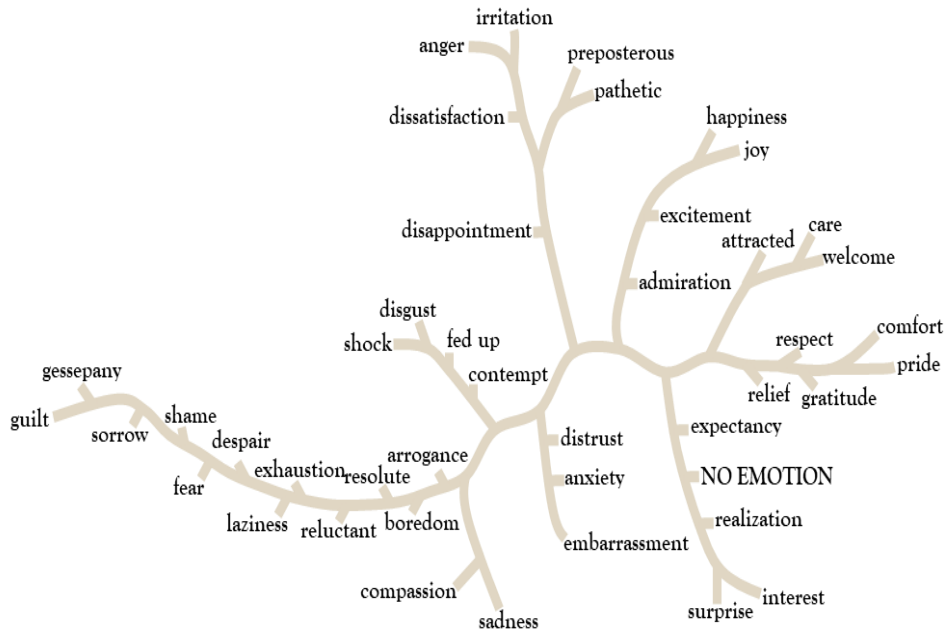
### 3.4. Finetuning

#### Preparation

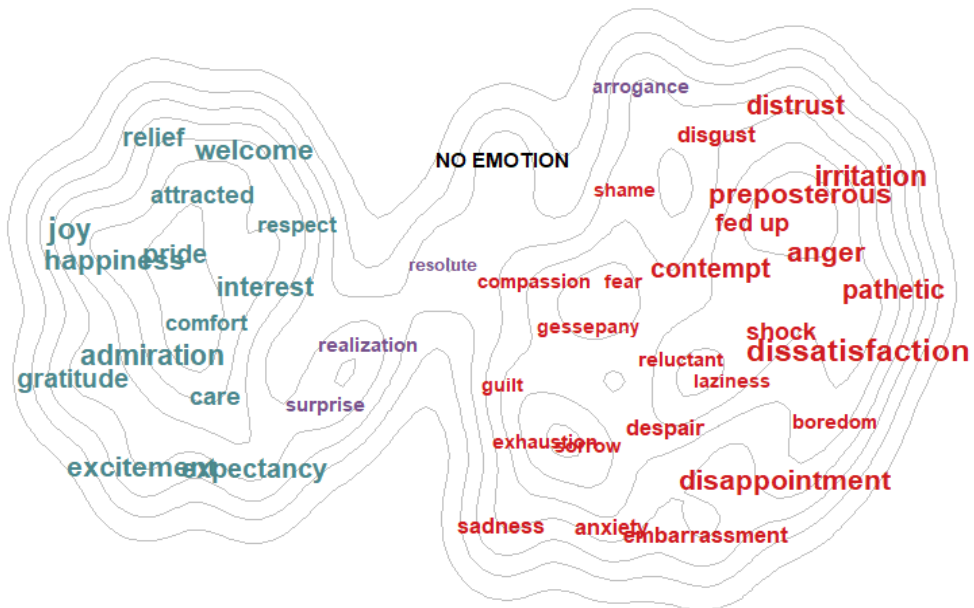
The labels ranging from 0 to 5 are dichotomized into 0 or 1. Minmax scaling is applied on the labels for each comment. The purpose of the comment-wise minmax scaling is to have the finetuned machine return several possible emotions when no emotion is confidently recognized. The labels exceeding 0.2 after the scaling are converted into 1, and 0 otherwise. One comment has 7.91 labels in average as a result. The data is randomly split into train (80%) and test (20%) sets. The test set is used as validation set in training.



**Fig 2:** Heatmap of Pearson correlation and Euclidean distance among labels. The lower and upper triangle represents the correlation coefficients and the Euclidean distances, respectively. The bars indicate the number of the labels in 250k cases. The order of the labels follows Ward clustering with squared Euclidean distance (Ward Jr, 1963) (Drawn with R package, `ComplexHeatmap` (Gu, Eils & Schlesner, 2016).)



**Fig 3:** Phylogram of the labels. The labels are hierarchically clustered by Ward's method (Ward Jr, 1963) with squared Euclidean distance. (Skeleton drawn with R package, ape (Paradis & Schliep, 2019) and redesigned.)



**Fig 4:** UMAP by correlation. The size of each text represents the sum of absolute values of its correlations with all other labels. Thus, the labels with larger text are correlated to the other labels more. The contours are added to help to judge the distances among the labels. (Drawn with ggplot2.)

## Training

We finetune KcELECTRA, a language model pretrained with Korean online comments, with three packages; `pytorch` (Paszke et al., 2019), `pytorch-lightning` (Falcon & Cho, 2020), and `huggingface’s 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 dense 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 also tried BP-MLL (Back Propagation Multi-Label Loss; Zhang & Zhou, 2006), but the results are not reported since the performance was not satisfactory. We use a linear optimization scheduler, in which the initial learning rate is  $2e-5$ , and the number of warmup steps and the total steps are 2,500 and 12,500, respectively. At every epoch, we switch 5% of tokens with a random token (except [CLS], [SEP], and [PAD]), and mask 5% of tokens with a special token, [MASK]. We tried label smoothing (Szegedy et al., 2016), but the results are not reported since the performance rather declined. The number of epochs is set 15 at maximum, but 9 to 10 epochs were enough to reach the optimum in almost all cases.

## 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 (**Table 5**). The average F1-score, AUC (Area Under Curve; Hanley & McNeil, 1982), and MCC (Mathews Correlation Coefficient; Mathews, 1975; Baldi et al., 2000; See also Chicco & Jurman, 2020) are 0.56, 0.88, and 0.59, respectively. This result is comparable to the existing model trained with translated GoEmotions (The F1-scores are 0.56 versus 0.41). Although a direct comparison is difficult because of different emotion taxonomies in KOTE and GoEmotions, it is meaningful to achieve comparable performance with wider range of emotions (43 versus 27 emotions).

As mentioned in the Introduction section, these results are obtained with arbitrarily decided hyperparameter. Therefore, the performance can be improved with additional methods, such as hyperparameter tuning. Otherwise, it would be a good try to employ a different approach for 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

F1-score									
emotion	precision	recall	F1	#	emotion	precision	recall	F1	#
<i>dissatisfac-</i> <i>tion</i>	0.78	0.89	0.83	4257	<i>admiration</i>	0.67	0.86	0.75	2616
<i>embarr-</i> <i>assment</i>	0.57	0.71	0.63	2605	<i>happiness</i>	0.57	0.80	0.67	1802
<i>irritation</i>	0.74	0.87	0.80	3781	<i>joy</i>	0.65	0.86	0.74	2387
<i>sadness</i>	0.57	0.60	0.58	1047	<i>gratitude</i>	0.55	0.72	0.62	1293
<i>despair</i>	0.45	0.40	0.43	927	<i>excitement</i>	0.68	0.87	0.76	2594
<i>shame</i>	0.36	0.05	0.09	592	<i>care</i>	0.56	0.72	0.63	1781
<i>boredom</i>	0.67	0.54	0.60	937	<i>expectancy</i>	0.58	0.81	0.68	2746
<i>disappo-</i> <i>intment</i>	0.69	0.88	0.77	4423	<i>comfort</i>	0.45	0.53	0.49	909
<i>disgust</i>	0.47	0.59	0.52	998	<i>welcome</i>	0.55	0.82	0.66	2230
<i>shock</i>	0.46	0.50	0.48	1399	<i>interest</i>	0.56	0.77	0.65	2634
<i>reluctant</i>	0.41	0.32	0.36	1184	<i>relief</i>	0.54	0.77	0.63	1932
<i>fear</i>	0.39	0.27	0.32	298	<i>respect</i>	0.53	0.68	0.59	945
<i>contempt</i>	0.67	0.78	0.72	1958	<i>attracted</i>	0.60	0.65	0.62	1042
<i>guilt</i>	1.00	0.01	0.01	173	<i>pride</i>	0.44	0.58	0.50	1258
<i>anxiety</i>	0.52	0.61	0.56	1888	<i>arrogance</i>	0.46	0.50	0.48	1515
<i>distrust</i>	0.61	0.79	0.69	3018	<i>surprise</i>	0.53	0.60	0.57	1821
<i>anger</i>	0.73	0.86	0.79	3107	<i>realization</i>	0.51	0.57	0.54	2059
<i>gessepanty</i>	0.41	0.22	0.28	391	<i>resolute</i>	0.48	0.46	0.47	815
<i>laziness</i>	0.39	0.20	0.26	602	<i>NO EMOTION</i>	0.55	0.57	0.56	1494
<i>sorrow</i>	0.38	0.29	0.33	497					
<i>preposte-</i> <i>rous</i>	0.70	0.87	0.78	4116					
<i>fed up</i>	0.46	0.58	0.51	1621	<b>micro avg</b>	<b>0.60</b>	<b>0.72</b>	<b>0.66</b>	<b>79011</b>
<i>compassion</i>	0.52	0.58	0.55	1353	<b>macro avg</b>	<b>0.56</b>	<b>0.61</b>	<b>0.56</b>	<b>79011</b>
<i>pathetic</i>	0.63	0.81	0.71	3016	<b>weighted avg</b>	<b>0.60</b>	<b>0.72</b>	<b>0.65</b>	<b>79011</b>
<i>exhaustion</i>	0.50	0.43	0.46	950	<b>samples avg</b>	<b>0.61</b>	<b>0.75</b>	<b>0.65</b>	<b>79011</b>

#### AUC

<i>dissatisfac-</i> <i>tion</i>	0.94	<i>embarrass-</i> <i>ment</i>	0.85	<i>irritation</i>	0.93	<i>sadness</i>	0.90	<i>despair</i>	0.84
<i>shame</i>	0.76	<i>boredom</i>	0.87	<i>disapp-</i> <i>ointment</i>	0.88	<i>disgust</i>	0.90	<i>shock</i>	0.85
<i>reluctant</i>	0.78	<i>fear</i>	0.87	<i>contempt</i>	0.93	<i>guilt</i>	0.86	<i>anxiety</i>	0.85
<i>distrust</i>	0.87	<i>anger</i>	0.94	<i>gessapany</i>	0.84	<i>laziness</i>	0.80	<i>sorrow</i>	0.85
<i>fed up</i>	0.83	<i>preposte-</i> <i>rous</i>	0.89	<i>compassion</i>	0.87	<i>pathetic</i>	0.88	<i>exhaustion</i>	0.85
<i>admiration</i>	0.93	<i>happiness</i>	0.92	<i>joy</i>	0.93	<i>gratitude</i>	0.92	<i>excitement</i>	0.93
<i>care</i>	0.89	<i>expectancy</i>	0.88	<i>comfort</i>	0.88	<i>welcome</i>	0.89	<i>interest</i>	0.87
<i>relief</i>	0.90	<i>respect</i>	0.92	<i>attracted</i>	0.92	<i>pride</i>	0.88	<i>arrogance</i>	0.83
<i>surprise</i>	0.84	<i>realization</i>	0.82	<i>resolute</i>	0.86	<i>NO EMOTION</i>	0.87	<b>macro avg</b>	<b>0.88</b>

MCC: 0.59

**Table 5:** The performance metrics.

hope it would be shared without hesitation.

## Chapter 4. Bias

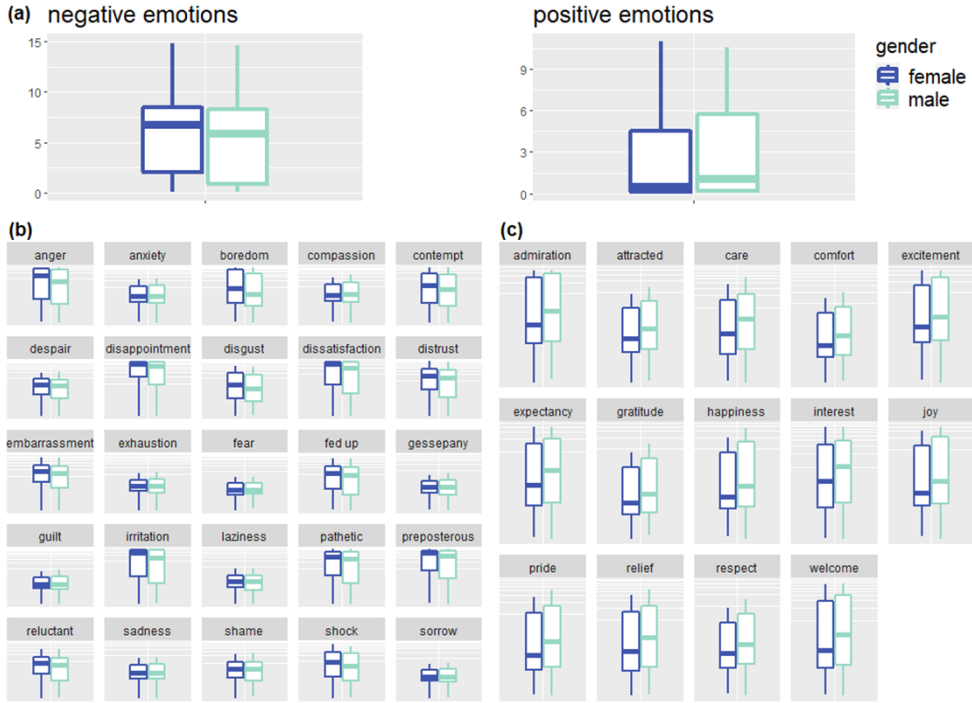
It is well known that a large dataset inevitably has discrimination against protected groups, and the demand of a fair model is not negligible. Our dataset is not an exception. In this section, we point out such problem and instantiate that a simple method helps to alleviate the discrimination. Here,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

### 4.1. 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 the 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 a KcELECTRA model trained with KOTE, while the gender words are masked with the special token, [MASK]. As in **Fig 5**, the texts containing female words are generally evaluated more negative, and texts containing male words are evaluated relatively more positive across all the positive emotions. 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 labelers.

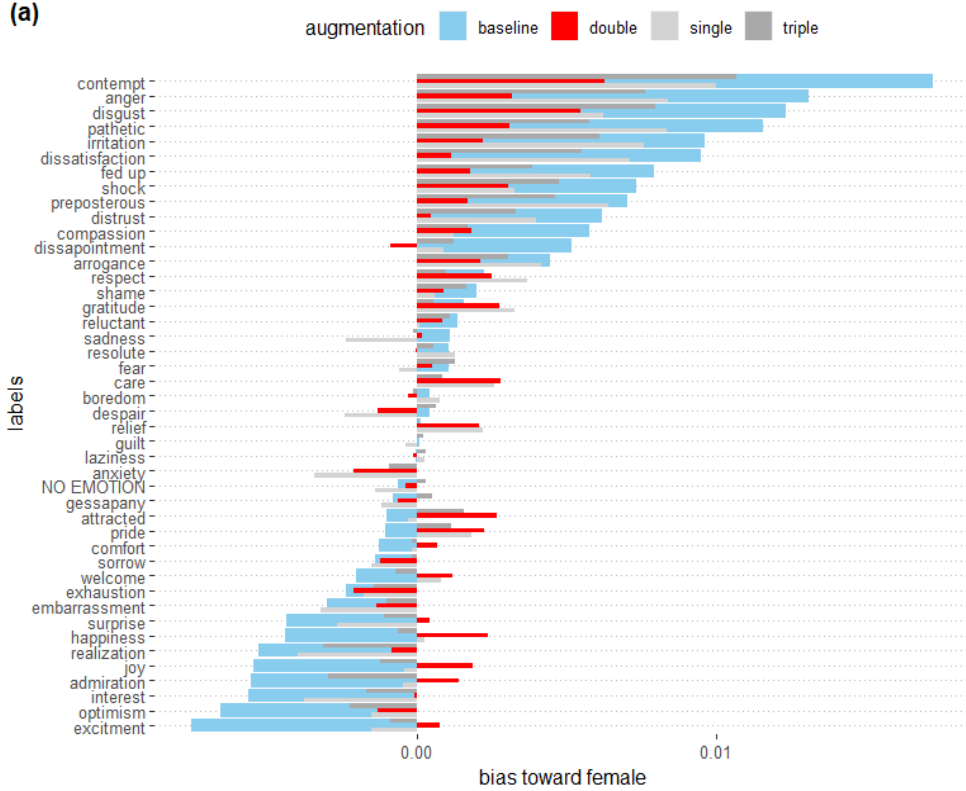
The second question is whether and how much the trained model is biased. To answer this question, we borrow the basic idea of explainable machine learning via token switching. From the source data, we input 320k texts (10% of the total source data) into the model and select 500 non-overlapping texts that have the highest probabilities for each label. As a result, 500 texts best representing each of the 44 labels (22k in total) are sampled for further analysis. Then, two randomly selected tokens (except [PAD], [CLS], and [SEP]) of each text are replaced with either female words (i.e., *women* and



**Fig 5:** 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. (Drawn with ggplot2.)

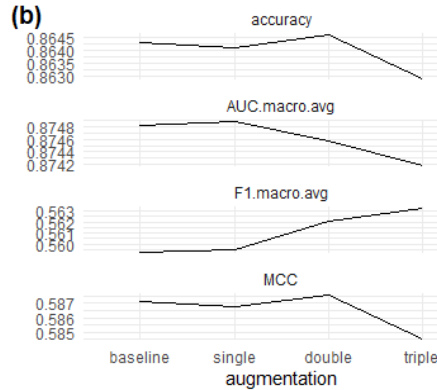
*female*) or male words (i.e., *men* and *male*). As a result, 22k random-to-female switched texts and 22k random-to-male switched texts are produced. The basic idea is the model would evaluate the two text sets equally if it is fair. The results are presented in **Fig 6 (a)**. The bars show the mean difference of each label’s predicted probabilities between the random-to-female switched texts and the random-to-male switched texts. The light blue bars indicate the baseline model without 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 male words are evaluated more positive on average. In particular, the texts with female words are evaluated discriminatorily for negative-intense emotions (e.g., *contempt*, *anger*, *disgust*, *pathetic*, and *irritation*).





**Fig 6: (a)** shows 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 additory 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). **(b)** shows the performance metrics of the models. No critical change is identified.

(Drawn with ggplot2 and ggpubr.)



## 4.2. Unbiasing

One of the simplest but powerful methods to mitigate discrimination in language dataset is data augmentation with token switching (Zhao et al., 2018; Park, Shin & Fung, 2018). 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 make 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. The double and triple augmented train set has 42,820 and 44,700 instances, respectively.

**Fig 6 (a)** shows the results. The augmented models are less biased than the baseline model, and the double augmented model is the least biased. As in **Fig 6 (b)**, furthermore, the augmented models cause no critical change in the performance metrics.

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 & 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. In some cases, excessive bias can rather degrade the performance. 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.

## Chapter 5. Conclusions

In this paper, we introduced KOTE, and its user guide about establishment of the emotion taxonomy, text collecting, labeling via crowdsourcing, brief description of the data, results of finetuning, and detection as well as mitigation of gender bias. Since the information in the dataset is rich, lots of attempts to improve the dataset are possible.

The model finetuned with our dataset achieved better performance than the existing model finetuned with the translated GoEmotions dataset (F1-scores are 0.56 and 0.41 respectively). 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 reason for good performance can be summarized as follows. **i)** We derived emotion taxonomy by introducing machine learning into repeatedly validated psychological theories and methodologies. **ii)** The emotion taxonomy is befitting to Korean culture, which is beneficial in two

respects: the human labelers can easily understand the emotions in the taxonomy, and the Korean language model can infer the emotions of the texts more 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 labeling and to maintain the complexity in preprocessing.

However, there are limitations that the KOTE users should keep in mind:

- (1) Emotion is a complex structure, which is impossible to perfectly capture with just tens of emotions.
- (2) Emotion is a dynamic structure, but we treated it as a static structure. 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.
- (3) 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.
- (4) The discriminatory evaluation toward protected groups can be mitigated even by a simple method, but the extent of the mitigation as well as the severity of the discrimination is still opaque. Gender discrimination can be carried within more complicated expressions other than the four gender words we used. In addition, it is challenging to address all kinds of discriminations other than the gender discrimination.

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 paper provides the users with useful information to utilize the dataset.

## **Acknowledgement**

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 Korean Government. We also thank all crowdworkers who sincerely helped us to annotate the data.

# Bibliography

- AI HUB (2021). Emotional Dialogue Corpus. Retrieved December 30, 2021, from Korea AI HUB open data: <https://aihub.or.kr/aidata/7978>
- Averill, J. R. (1975). A sementic atlas of emotional concepts. *JSAS Catalog of Selected Documents in Psychology*, 5, 330.
- Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013, June). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, 49-58.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), 412-424.
- Block, J. (1957). Studies in the phenomenology of emotions. *The Journal of Abnormal and Social Psychology*, 54(3), 358.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146.
- Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1), 1-13.
- Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Clore, G. L., Ortony, A., & Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of personality and social psychology*, 53(4), 751.
- Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38), E7900-E7909.
- Cowen, A., Sauter, D., Tracy, J. L., & Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1), 69-90.
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- Falcon, W., & Cho, K. (2020). A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*.
- Fern, X. Z., & Brodley, C. E. (2004, July). Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on Machine learning* (p. 36).
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2), 251.
- Graves, A., & Schmidhuber, J. (2005, July). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. (Vol. 4, pp. 2047-2052). IEEE.
- Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18), 2847-2849.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.
- Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. *Quality & quantity*, 51(6), 2623-2646.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

- Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 537-546.
- Hupka, R. B., Lenton, A. P., & Hutchison, K. A. (1999). Universal development of emotion categories in natural language. *Journal of personality and social psychology*, 77(2), 247.
- Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, 8(1) 216-225.
- Izard, C. E. (1977). Differential emotions theory. In *Human emotions* (pp. 43-66). Springer, Boston, MA.
- Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, 99(3), 561-565.
- Jang, H., Kim, M., & Shin, H. (2013, November). KOSAC: A full-fledged Korean sentiment analysis corpus. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 366-373).
- Kassambara, A., & Kassambara, M. A. (2020). Package ‘ggpubr’.
- Karypis, G., & Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20(1), 359-392.
- Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, 1-28.
- KETI, Korea Electronics Technology Institute. (2020). Korean Continuous Dialogue Dataset with Emotion Information. Retrieved December 30, 2021, from AI HUB: <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-010>
- KETI, Korea Electronics Technology Institute. (2020). Korean One-off Dialogue Dataset with Emotion Information. Retrieved December 30, 2021, from AI HUB: <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-009>
- Lee, Haksik., & Lim, Jihoon. (2002). Measuring the consumption - related emotion constructs. *Korean Journal of Marketing*, 17(3). 55-91.
- Lee, Junbum. (2021). KcELECTRA: Korean comments ELECTRA. Retrieved December 29, 2021, from Github, Github repository: <https://github.com/Beomi/KcELECTRA>
- Mano, H., & Oliver, R. L. (1993). Assessing the dimensionality and structure of the consumption experience: evaluation, feeling, and satisfaction. *Journal of Consumer research*, 20(3), 451-466.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442-451.
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), 205.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1-35.
- Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: a review. *Psychological bulletin*, 112(2), 179.
- Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, 201-237.
- Naver. (2015). NSMC: Naver Sentiment Movie Corpus. Retrieved December 30<sup>th</sup>, 2021, from Github, Github repository: <https://github.com/e9t/nsmc>
- Osgood, C. E. (1966). Dimensionality of the semantic space for communication via facial expressions. *Scandinavian journal of psychology*, 7(1), 1-30.

- Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526-528.
- Park, Eunjung., & Jo, Sungjun. (2014). KoNLPy: Easy and Concise Python Package for Korean Information Processing. In *Proceedings of the 26<sup>th</sup> Annual Conference on Human and Cognitive Language Technology* (pp. 1-4).
- Park, In-Jo., & Min, Kyung-Hwan. (2005). Making a List of Korean Emotion Terms and Exploring Dimensions Underlying Them. *Korean Journal of Social and Personality Psychology*, 19(1), 109-129.
- Park, Ji Ho., Shin, Jamin., & Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.
- Park, S. M., Na, C. W., Choi, M. S., Lee, D. H., & On, B. W. (2018). KNU Korean sentiment lexicon: Bi-LSTM-based method for building a Korean sentiment lexicon. *Journal of Intelligence and Information Systems*, 24(4), 219-240.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 8026-8037.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Academic press.
- Provoost, S., Ruwaard, J., Van Breda, W., Riper, H., & Bosse, T. (2019). Validating automated sentiment analysis of online cognitive behavioral therapy patient texts: an exploratory study. *Frontiers in psychology*, 10, 1065.
- Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, 36, 101003.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673-2681.
- Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6), 1061.
- Shields, S. A. (1984). Distinguishing between emotion and nonemotion: Judgments about experience. *Motivation and Emotion*, 8(4), 355-369.
- Sohn, S. J., Park, M. S., Park, J. E., & Sohn, J. H. (2012). Korean emotion vocabulary: extraction and categorization of feeling words. *Science of Emotion and Sensibility*, 15(1), 105-120.
- Storm, C., & Storm, T. (1987). A taxonomic study of the vocabulary of emotions. *Journal of personality and social psychology*, 53(4), 805.
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
- Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03), 337-372.
- Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.
- Wickham, H. (2011). ggplot2. Wiley Interdisciplinary Reviews: *Computational Statistics*, 3(2), 180-185.
- Wilke, C. O. (2021). Ridgeline Plots in 'ggplot2'[R Package ggridges Version 0.5.3]. *January*.

<https://cran.r-project.org/web/packages/ggribes/index.html>.

- Willcox, G. (1982). The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4), 274-276.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- You, Q., Luo, J., Jin, H., & Yang, J. (2015, February). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.
- You, Q., Jin, H., & Luo, J. (2017, February). Visual sentiment analysis by attending on local image regions. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Yuan, J., McDonough, S., You, Q., & Luo, J. (2013, August). Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 1-8.
- Zhang, M. L., & Zhou, Z. H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10), 1338-1351.
- Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.
- Zucco, C., Calabrese, B., & Cannataro, M. (2017, November). Sentiment analysis and affective computing for depression monitoring. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 1988-1995.



# Appendix

극성	해석	군집 안의 정서 단어 예시
부정	불만	불만, 반발, 비판, 항의
	당황	당황, 당혹, 곤혹, 난처
	짜증	짜증, 열 받다, 어이없다
	슬픔	슬픔, 그리운, 외로운, 눈물
	절망	절망, 애환, 아픔, 비탄, 허무감
	부끄러움	부끄러움, 부끄럽다
	지루함	지루함, 재미없음, 식상, 답답함
	실망	실망, 안타까움, 속상, 애석, 아쉬움
	억겨움	억겨움, 징그러움, 지저분
	경악	경악, 기절초풍, 실신, 까무러치다
	부담감	마지못해, 거부, 재촉, 고깝다, 단념
	공포	공포, 불안, 긴장, 압박감
	혐오	증오, 혐오, 죄악시, 경멸, 모멸, 멸시
	죄책감	죄책감, 죄의식, 가책, 참회, 속죄, 뉘우침
	불안	우려, 염려, 위협
	불신	의심쩍다, 반신반의, 거짓
	분노	증오, 분노, 사로잡힌, 분개, 격분, 격노
	패배감	실패, 처참히, 빼앗기다
	귀찮음	귀찮음, 빈둥빈둥
	서러움	서러움, 서글픔, 고달프다, 흐느낌, 속상, 착잡
	지긋지긋	지긋지긋, 애쓰다, 고되다, 질리다
	어이없음	어처구니, 싱겁, 갑갑함, 맥빠지다, 이뉘병
	연민	째하다, 슬프다, 울컥, 먹먹하다
	한심함	한심, 우스운, 명칭, 뻔뻔
	지침	피로, 야위다, 수척
긍정	감탄	감탄, 대단하다, 칭찬, 찬사
	행복	행복, 친애, 소중, 희망, 행운
	기쁨	환희, 황홀, 사랑
	고마움	기특함, 은혜, 은총, 베풀다
	즐거움	즐거움, 재밌는
	아끼는	아낌, 흠모, 경애
	기대감	새로운, 이루다, 함께, 원할, 활력
	편안	편안, 포근함, 안락, 시원, 따듯
	환영	환영, 우호, 호의, 열렬히
	흥미	호기심, 관심
	안심	신뢰, 안심, 친밀, 각별
	존경	존중, 충성, 숭상, 본받다, 복종
	흐뭇함	멋있다, 예쁘다, 달달, 짜릿, 귀엽다, 깜찍, 애교
	뿌듯함	성공, 승리, 달성, 보람, 희열
중립	우쭐댐	우쭐댐, 앞잡아보다, 무시, 업신여기다, 거만, 교만
	놀람	질겁, 소스라치다
	깨달음	깨달음, 깨우침, 일깨워, 확신, 믿음
	비장함	비장함, 결단, 결심

## Abstract in Korean

텍스트를 단순히 긍부정으로 분류하는 감성 분석은 텍스트의 정서적 측면을 철저하게 조사하기 힘든 방법임에도 자연어 처리 분야에서 널리 쓰여왔다. 최근 이러한 한계를 뛰어넘기 위해 긍부정보다 더욱 풍부한 기준을 가진 정서 말뭉치들이 구축되고 있다. 그러나 대부분의 한국어 정서 말뭉치는 규모가 작고 제한적인 정서만 다룬다는 점에서 여전히 한계를 가진다. 또한, 말뭉치의 정서 분류 기준은 기존의 연구에서 무비판적으로 차용되거나 심지어 직관에 따라 정해지기도 한다. 우리는 KOTE(한국어 온라인 댓글 정서 말뭉치)를 제안한다. KOTE는 클라우드 소싱(참가자 수 3,084)을 통해 43개의 정서 혹은 ‘정서 없음’으로 레이블링된 5만 개의 댓글로 이루어져 있다. 43개 정서 분류 기준은 단어 임베딩 공간에 표상된 정서 단어를 군집화하는 방식으로 체계적으로 수립되었다. KOTE를 구축한 방법을 설명한 이후에는, 미세조정 결과와 말뭉치가 지니는 사회적 차별에 대해서도 논한다.