How Do Moral Emotions Shape Political Participation? A Cross-Cultural Analysis of Online Petitions Using Language Models

Anonymous ACL submission

Abstract

Understanding the interplay between emotions in language and user behaviors is critical. We study how moral emotions shape political participation of users based on cross-cultural online petition data. To quantify moral emotions, we employ a context-aware NLP model that is designed to capture the subtle nuances of emotions across cultures. For model training, we construct and share a moral emotion dataset comprising 50,000 petition sentences in Korean and English along with emotion labels annotated by a fine-tuned LLM. We examine two distinct types of user participation: general support (*i.e.*, registered signatures of petitions) and active support (*i.e.*, sharing petitions on social media). We discover that moral emotions like *other-suffering* increase both forms of participation and help petitions go viral, while *self-conscious* have the opposite effect. The most prominent moral emotion, *other-condemning*, led to polarizing responses among the audience. In contrast, *other-praising* was perceived differently by culture; it led to a rise in active support in Korea but a decline in the UK. Our findings suggest that both moral emotions embedded in language and cultural perceptions are critical in engaging the public in political discourse.

1 Introduction

Moral emotions influence group judgments and behaviors on social issues and further impact political participation (Van Bavel et al., 2023). They drive individuals to collectively act on issues, deter actions, polarize groups, and can lead to extremism (Inbar et al., 2012; Brady et al., 2021; Finkel et al., 2020). Their influence extends beyond the offline realm and affects political discussion on social media (Brady et al., 2020; Van Bavel et al., 2023). As online platforms connect vast networks of people, understanding how different types of moral emotions expressed through language affect user actions has become crucial.

Online petitions are excellent data forms to study the role of moral emotions and their influence on user response because they record the motivations, sentiments, and behaviors of individuals who engage in collective action. We use cross-cultural data to answer the following questions: (1) How can we measure moral emotions systematically? (2) Which moral emotions most effectively shape users' political participation? (3) Do moral emotions have the same effect across cultures? We gathered data from two petition websites of similar designs and social media functions: South Korea's Blue House National Petition and the United Kingdom's Government and Parliament Petitions. By utilizing these two datasets, we can test the cross-cultural generalizability of our findings on the role of moral emotions in political participation.

Our key contributions are proposing a 5-step framework to analyze moral emotions, as shown in Figure 1, and sharing a comprehensive moral emotion dataset in Korean and English. We consider broader emotion categories than previous work (Brady et al., 2017; Solovev and Pröllochs, 2022) and include: *other-condemning*, *other-praising*, *other-suffering*, *self-conscious*, *neutral*, and *non-moral emotion* (see Table 1). This dataset was labeled by large language models (LLMs) like GPT that are fine-tuned to inherit the knowledge of human annotators. As GPT models can become expensive for labeling large amounts of data, we trained in-house models like BERT and ELECTRA for cost-effective label predictions. We separately constructed language-specific versions of these models to reflect socio-cultural traits by language, as suggested in Havaldar et al. (2023). Our framework, using a combination of GPT and human annotation to label data and train light in-house models, can be reused for other low-resource languages and unseen tasks.

Two types of user actions are considered in this research. The first is the number of signatures on

Type	Category	Definition
Moral Emotions	Other-condemning	Emotions that condemn others (e.g., anger, contempt, disgust).
	Other-praising	Emotions that praise others (e.g., admiration, gratitude, awe).
	Other-suffering	Emotions of empathy for the suffering of others (e.g., compassion, sympathy).
	Self-conscious	Emotions that negatively evaluate oneself (e.g., shame, guilt, embarrassment).
Non-moral	Neutral	A neutral category with no or few emotions.
	Non-moral emotion	Emotional but not one of the moral emotions (e.g., fear, surprise, joy, etc).

Table 1: Definition of moral emotion categories. Examples of each category are introduced in Appendix Table 11.

petitions (called **general support** here), which is a direct political action that grants the signatory the right to receive a government response or even have the petition discussed in a legislative setting. The second is the number of "direct" sharings of petition information from the official government website via the share on social media button (called **active support**). The latter form represents a stronger commitment by making the sharer's public ID visible over the network (Kim and Yang, 2017; Proskurnia et al., 2017). By these definitions, our work seeks to understand the impact of moral emotions on substantial political participation, as opposed to studying discourse regarding politics on social media and the likes or retweets of such postings.

Our results point to an interplay between moral emotions and political participation. We discover that moral emotions like *other-suffering* that appeal to compassion and sympathy positively correlated with both political actions. However, emotions like *self-conscious* that emphasize feelings of public shame and guilt have the opposite effect, substantially reducing both forms of participation. *Other-condemning*, which blames others, polarized the audience. It negatively correlated with total signatures but positively correlated with social media sharing. In contrast, *other-praising* showed mixed patterns; while it negatively correlated with signatures in both countries, it led to a rise in social media sharing in Korea but a decline in the UK.

We discuss implications of our findings, which highlight the pronounced exposure of petitions with specific moral emotions, such as *other-condemning*, on social media. Our work also found cultural similarities and differences, which could lead to new research. We share the moral emotion dataset and the classifiers for wider use[1] for the political science and AI communities.

2 Related Work

2.1 Moral Emotion and Political Discourse

Moral emotions are key to spreading messages in political discourse on social media platforms (Brady et al., 2020; Van Bavel et al., 2023). Prior research identified their significance by using metrics like retweets. Brady et al. (2017) showed that including a single moral emotional word in a tweet on political topics can increase the retweet probability by 20%. Solovev and Pröllochs (2022) indicated the presence of emotions that "condemn" others amplified the spread of political rumors, regardless of whether they are true or false. While these studies offer fascinating insights, retweets alone cannot fully reflect the spectrum of political engagement, as they overlook the contributions of less vocal participants (Yang and Kim, 2017). Like shy supporters, such participants may engage in quiet, anonymous actions that are less visible. To address this gap, we study government-led online petitions and focus on visible and subtle forms of political participation.

2.2 Moral Emotion Detection

Unlike general emotions such as happiness, moral emotions are prosocial, driven by the intention to protect and support the interests of others over self-interest, often stemming from social injustice (Haidt et al., 2003; Van Bavel et al., 2023). The theoretical framework categorizes these emotions into four distinct types: other-condemning, other-praising, other-suffering, and self-conscious (Haidt et al., 2003). Moral emotion detection remains challenging due to the scarcity of datasets. Previous studies have used lexicon-based or word embedding approaches to identify moral emotions in texts. However, such efforts have been restricted to the English language and covered only a subset of the moral emotion categories, as detailed in Table 2.

[1] https://anonymous.4open.science/r/Moral-Emotion-Data-6E33/

Research	Detection Method	Target Moral Emotion (Independent Variables)	Political Metric (Dependent Variables)	Dataset	Language
Brady et al. (2017)	Lexicon	Moral emotional, non-moral categories	#Retweet	Twitter	Monolingual
Solovev and Pröllochs (2022)	Lexicon	Other-condemning, self-conscious	#Retweet	Twitter	Monolingual
Brady et al. (2021)	Word Embedding	Other-condemning or not	#Retweet, #Like	Twitter	Monolingual
Ours	Transformer	Complete categories, non-moral categories	#Signature, #Shares on Twitter	Online Petition, Twitter	Bilingual

Table 2: Comparison with previous studies. Our research employs more advanced methods in NLP and focuses on more comprehensive emotion categories and political variables. Examples of results from previous works and our final model's detection of moral emotion can be seen in Appendix Table 12.

2.3 Data Annotation Using LLMs

High-quality labeled data is crucial for machine learning. However, creating large-scale, high-quality data requires extensive human labor, substantial cost, and time. As one way to assist the expensive task of data labeling, LLMs have been considered for its remarkable performance in various downstream tasks such as adaptation to unseen tasks (Brown et al., 2020). In particular, recent studies have investigated whether LLMs, such as GPT-3 or open-source LLMs, can reliably replace human annotation (Wang et al., 2022; Ding et al., 2023; Alizadeh et al., 2023). Some studies also propose innovative strategies to enhance the annotation quality using LLMs through data augmentation and developing effective prompt guidelines (Bansal and Sharma, 2023; He et al., 2023; Latif et al., 2023). In this paper, we contribute by constructing a new dataset labeled with LLMs, showcasing their ability to label and adapt to unseen tasks.

3 Moral Emotion Dataset

Dataset	Annotation	Korea	UK
Human annotation	Humans (§3.2)	640	640
Moral emotion	GPT-3.5 (§3.3)	49,930	49,896
Petition dataset	Classifier (§4.1)	4,705,292	210,304

Table 3: Overview of each dataset.

3.1 Data Preparation

Data Collection: We collected petition data from the Korean government archive[2] and the UK Government and Parliament Petition website.[3] The Korean archive recorded 459,447 petitions with 161,856,648 signatures from August 25, 2017, to May 9, 2022. The UK website logged 41,292 petitions with 47,554,399 signatures from March 2, 2020, to December 7, 2022. For comprehensive statistics, see Table 13 in the Appendix. The two platforms have common fields such as petition ID, URL, start date, end date, title, content, state (*e.g.*, open, closed, or rejected), and the signature count. The UK platform was launched in 2015 but only displays petitions created since March 2, 2020.

Both platforms offer a 'Share via Twitter' function that generates tweets with the petition's unique URL and the endorsement message (*e.g.*, https://petition.parliament.uk/petitions/xxxxxx). Following the syntax, we identified and collected all public tweets that are officially shared from the government websites, which lets us review which petitions were publicly shared on social media directly from the site. We collected 251,245 and 853,222 tweets in Korean and English.

Data Preprocessing: The petition titles and contents are cleansed using regular expressions to remove personal information such as email addresses, phone numbers, and special characters, including emojis. Subsequently, sentence tokenization is performed to segment text into sentences using the Kiwi (Lee, 2022) and PySBD library (Sadvilkar and Neumann, 2020). After removing short sentences with one or two words, we obtained 4,705,292 and 210,304 sentences from the Korean and UK petitions, respectively.

3.2 Human Annotation

We first collected human-annotated data to classify moral emotions following the method introduced in Field et al. (2022) to adapt human knowledge in the domain of moral emotions to language models. From both petition datasets, we selected approximately 700 sentences each. Annotations were received from five native speakers and citizens of each country who understand the politi-

[2] http://webarchives.pa.go.kr/19th/www.president.go.kr/petitions/

[3] https://petition.parliament.uk/

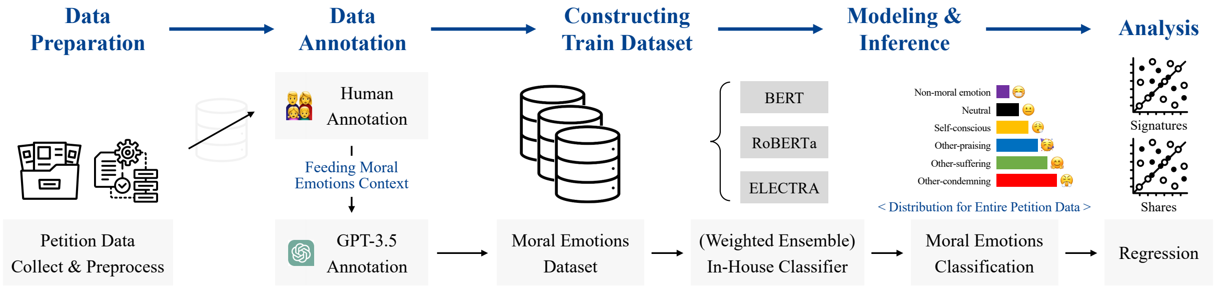


Figure 1: Method overview. We propose a full framework for constructing data, modeling classification, and conducting analysis on the theme of moral emotions.

cal context. Annotators were given guidelines on the moral emotion definitions and asked to choose multi-responses if they detected more than one emotion category. If annotators captured an emotion that did not fit the predefined categories, they were guided to select 'Emotional but not equivalent to the above'. Additionally, we included a 'Difficult to distinguish (Hard to tell)' checkbox for ambiguous cases. Please see Figure 4 in the Appendix for the example guideline. Only sentences that received a majority vote of at least three out of five annotators were considered ground truth labels, and sentences with no consensus were excluded.

The final human-annotated dataset consists of 640 sentences each for Korean and English, with the distribution presented in Appendix Figure 5. Table 4 shows the inter-annotator agreement score of the final human-annotated dataset. The observed discrepancy in agreement can be attributed to the differential propensity for multi-label responses between Korean (95.52% single-label) and English annotators (79.41% single-label).

Metric	Korean	English
Cohen's kappa	0.7218	0.4244
Fleiss' kappa	0.7253	0.4156
Krippendorff's alpha	0.7254	0.4158

Table 4: The inter-annotator agreement score of final human-annotated dataset. Annotator response scores for each emotion are in Appendix Table 8.

3.3 LLM-based Annotation

Two primary methods are used to perform unseen tasks. First is **in-context learning** or few-shot learning that enables a model to learn based on a few training samples within prompts. Second is **fine-tuning** which updates the model's weight parameters. Fine-tuning requires a large training dataset, whereas it can learn from more examples than can fit in the context window (Brown et al., 2020).

To determine a better-performing approach among the two options, we tested the in-context learning and fine-tuning prompts using the Chat Completions API (OpenAI, 2023). Each format includes 1) short definitions of moral emotion categories, 2) text instances from the human-annotated training dataset, and 3) labels of the corresponding sentences. We describe the detailed experimental setting in Section B.1 in the Appendix.

We split the human-annotated dataset into 300 and 340 samples for the training and testing sets. Table 5 summarizes the comparison of performance and costs. This table also shows the result of human annotation, comparing the accuracy of each human annotator's responses against the majority vote. The cost column indicates the estimated expense for labeling the entire 50,000 samples. We estimated the cost of human annotation from the Google Cloud Platform and used the pricing model based on the number of words.[4] Although Korean sentences typically contain fewer words, processing Korean texts with GPT approximately doubles the expense compared to English due to higher tokenization costs. In-context learning and fine-tuning methods cost substantially less than human annotation, which includes the combined cost of multiple annotators per task. The fine-tuning cost calculation includes both training and inference expenses.

The results of the GPT-3.5 few-shot in-context experiment indicate that performance improves with the training samples. However, the fine-tuned model consistently outperformed the in-context learning models in all settings, most likely because fine-tuning allows training on more examples than what can be accommodated within via prompting. Fine-tuned GPT-3.5 achieved comparable performance to that of human annotators. GPT-4 few-shot experiment also showed high performance, but

[4] https://cloud.google.com/ai-platform/data-labeling/pricing#labeling_costs

	Train size	Korean (KOR)			English (UK)		
		F1	Acc.	Cost ($)	F1	Acc.	Cost ($)
In-Context Learning (GPT-3.5)	6	0.5810	0.6029	74.57	0.6111	0.5912	41.42
	12	0.5825	0.6118	104.87	0.6223	0.5824	57.98
	18	0.6050	0.6353	148.07	0.6501	0.5794	75.86
In-Context Learning (GPT-4)	6	0.8259	0.8206	499.23	0.7056	0.7176	278.46
	12	0.8642	0.8588	701.06	0.7054	0.7118	389.24
	18	0.8458	0.8382	989.29	0.7023	0.7088	508.42
Fine-Tuning (GPT-3.5)	150	0.8518	0.8471	336.20	0.7169	0.7029	163.10
	200	0.8530	0.8471	338.17	0.7348	0.7471	164.10
	250	0.8580	0.8471	340.17	**0.7436**	**0.7500**	165.03
	300	**0.8678**	**0.8618**	342.16	0.7426	0.7294	165.93
Human Annotation	–	0.8678	0.8360	1480.96	0.7091	0.5816	2021.96

Table 5: Comparison of performance and costs in USD ($) across various labeling methods for multi-label tasks. Performance are measured in macro F1 score (F1) and accuracy (Acc.).

the fine-tuned GPT-3.5 model yielded more cost-effective and better-performing results. For this reason, we chose the fine-tuned GPT-3.5 version as our annotator model. Additional experiments are included in the Appendix Section B.

3.4 Dataset Description

Data Selection: In preparation for labeling with our fine-tuned GPT-3.5, we curated petition sentences using methods inspired by GoEmotion (Demszky et al., 2020). Our initial dataset consisted of approximately 4.7 million Korean and 210K UK petition. To ensure these sentences reflected societal engagement, we selected those with at least one signature and share. We also sought a balanced distribution in sentence length, choosing sentences between 3 and 30 tokens with the aid of the NLTK word tokenizer (Bird et al., 2009). To aim for a balanced representation of emotions in our dataset, we focused on reducing bias by trying to even out the distribution of sentences across emotion categories. We employed a pilot model trained on human-annotated examples to estimate the emotional content of petition sentences. This approach helped us identify and initially extract 5,000 sentences for each emotion label, creating a set of 30,000 sentences. Then, 20,000 sentences were randomly selected to achieve 50,000 samples. After removing duplicates and samples that were incorrectly labeled by the GPT's inference process (*e.g.*, Overall Condemning), our dataset was finalized with 49,930 Korean and 49,896 English sentences.

Data Statistics: Figure 2 shows the distribution of moral emotion labels. We make four observations. First, *other-condemning* is the most prevalent moral emotion, taking up one-third in English and one-fourth in Korean. This moral emotion is more common than general emotions (*i.e.*, non-moral emotions). Second, *other-suffering* is the next frequently used moral emotion, with UK petitions exhibiting nearly twice the exposure (19.0%) compared to Korea (10.1%). Third, *other-praising* and *self-conscious* make up a small proportion of moral emotions. Fourth, petitions contain a substantial proportion of *neutral* sentences, which may offer factual statements to support the petition.

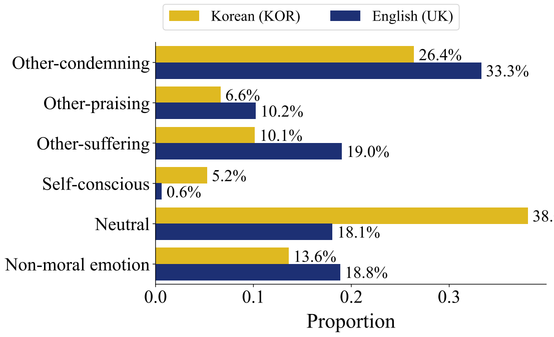


Figure 2: Distributions of moral emotion labels.

4 Political Participation Analysis

We built a classifier from our fine-grained moral emotion dataset to analyze the entire collection of petitions and performed regression analysis to assess the impact of these emotions on political participation.

4.1 Moral Emotion Measurement

Moral Emotion Classifier: Our analysis, which aims to measure moral emotion in online petition data using a Transformer-based model, utilizes the approach suggested by Wang et al. (2021). This work has established that compact, in-house lan-

	Korean (KOR)		English (UK)	
	F1	Acc.	F1	Acc.
Fine-tuned GPT-3.5	0.8678	**0.8618**	0.7436	**0.7500**
RoBERTa	0.8785	0.8471	0.7134	0.6971
BERT	0.8858	0.8500	0.6760	0.6588
ELECTRA	0.8914	0.8559	0.7523	0.6971
Ensemble	0.8950	0.8471	0.7367	0.5588
Weighted Ensemble	**0.8978**	0.8559	**0.7536**	0.6971

Table 6: Performance comparison of fine-tuned GPT-3.5 vs. in-house language models on human-annotated data.

	Korean (KOR)		English (UK)	
	General	Active	General	Active
Other-condemning	-0.056*	2.364**	-0.042	1.315**
Other-praising	-0.520**	2.100**	-1.025**	-0.768**
Other-suffering	1.383**	2.280**	0.217*	0.475**
Self-conscious	-4.009**	-5.187**	-5.326**	-2.463**
Neutral	-0.109**	0.595**	0.084	1.755**

Sign. levels: * $p < 0.05$, ** $p < 0.001$

Table 7: Result of regression in general and active support for emotions (positive , negative). For complete regression outcomes, refer to Appendix Table 14.

guage models, when trained with LLM-generated labels, can outperform the raw LLM while also reducing costs. Employing in-house models like BERT, RoBERTa, and ELECTRA, pre-trained on Korean (Park et al., 2021; Lee, 2021, 2020) and English corpora (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020), we analyzed petition data across these languages. Each model was added with a fully connected classification layer and fine-tuned on a GPT-labeled moral emotion dataset. For multi-label prediction, we applied binary cross-entropy loss, set the learning rate at $2e$-5, and trained the models for up to four epochs with early stopping.

The models were evaluated on the human-annotated dataset, and the results are presented in Table 6. Here we also report ensemble models. On the macro F1 score, we observe that small models like ELECTRA perform well compared to the fine-tuned GPT. We chose the weighted ensemble with the best F1 score as our final model.

Moral Emotion Score: We employed weighted ensemble models for both countries to compute petition's moral emotion score. For tokenized petition sentences, emotion scores were predicted as six-dimensional vectors with sigmoid outputs ranging from 0 to 1, and the average of these values determined the final score for each petition document.

4.2 Regression Model Specification

We separately estimated two count variables (*i.e.*, signatures and social media shares) using negative binomial regressions against the emotion variables and control variables. All overdispersion parameters were significant at the $\alpha = .01$ level. Here, only the four moral emotions and neutral were used as emotion categories; non-moral emotion was excluded due to its low F1 score for English (see Figure 9 in the Appendix). The control variables included information about the text length, URL usage, and time information. Temporal information was added in the regression to account for year-specific and seasonal variability.

4.3 Regression Result

Table 7 summarizes the regression results demonstrating the effects of moral emotion on two count variables after adjusting for control variables. The color background represents the prominent direction of correlation, which shows both positive and negative directions. The two dependent variables of signatures (*i.e.*, general support) and social media shares (*i.e.*, active support) themselves have a positive correlation (Korea: Pearson's $r = 0.49$, $p < 0.001$; UK: $r = 0.67$, $p < 0.001$). In both countries, other-suffering and self-conscious appear to drive these results. Other-suffering, positively correlated with both forms of support, effectively secured the number of signatures and shares. In contrast, self-conscious is negatively correlated, implying a reduction in political support for both types.

Although petitions with many signatures tend to be shared more frequently on social media, moral emotions could explain the subtle response patterns of users. For example, increase in the other-condemning emotion negatively correlates with general support and positively with active support in the two countries. Such an inverse trend may reflect nuanced divisions among the supporter groups, even for the same petition.

Figure 3 shows the trends of the count variables (y-axis) across the studied emotions (x-axis) for Korea (top row) and the UK (bottom). Patterns with opposing regression trends have a crossing sign, such as other-condemning, neutral (Korea), and other-praising (Korea). In these cases, active support correlates positively with the corresponding emotion, whereas general support correlates negatively. This finding suggests that active support does not consistently translate into general support

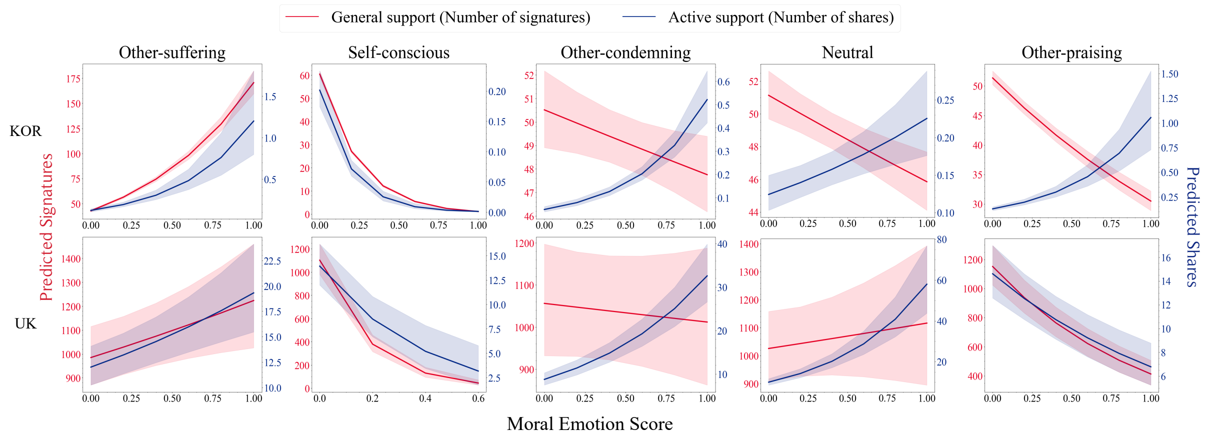


Figure 3: Predictive margins for general support (depicted by a red line) and active support (blue line) across five emotions in studied countries, with 95% confidence intervals.

in the presence of a specific emotion and culture.

Other-praising shows different results by culture. The number of shares in Korea increases with this emotion, whereas it decreases in the UK. Conversely, the number of signatures decreases in both countries with emotion, indicating that the enhancement of universal support is not consistent. The correlation between social media shares and other-praising in Korea mirrors the dynamics seen with other-condemning. In the UK, it aligns with self-conscious emotion, reducing support levels.

5 Discussion

5.1 Implications for Social Science

Political and moral psychology research has identified moral emotions as politically motivating through metrics like retweets (Brady et al., 2017; Solovev and Pröllochs, 2022). Our study expands the literature to encompass comprehensive moral emotion categories and two direct forms of political participation: general and active support seen in government-led online petitions. Additionally, our research extends cross-cultural insights by analyzing petitions from multiple countries, addressing the urgent need for broader cultural analysis in this domain (Van Bavel et al., 2023).

Prior studies have focused on the role of other-condemning emotions in social sharing (Brady et al., 2021; Solovev and Pröllochs, 2022). However, our findings highlight the significant role of other-suffering in amplifying both active and general support in two distinct countries. These results suggest that the expression of other-suffering harmonizes the perspectives of both dedicated and general supporters, fostering widespread consensus on petitions addressing social issues (Sirin et al., 2016, 2017). This implies that policymakers and activists could cultivate a more engaged and unified public discourse by leveraging this emotional dynamic.

Aligning with past research, our study indicates that other-condemning may boost active support through social media sharing. Additionally, our analysis shows that other-condemning will likely diminish petition signatures, illustrating a polarization effect. This effect lowers the broader base of general supporters' willingness to engage but also sharply increases participation among a more dedicated segment. This observation aligns with earlier research discussing other-condemning as a catalyst for political polarization (Crockett, 2017; Finkel et al., 2020; Brady et al., 2021). Thus, our study highlights the complex nature of online political participation: while other-condemning can enhance issue visibility on social media, it also poses deepening societal divisions. Interestingly, neutral, the absence of emotional engagement, also acts as a polarizing force, encouraging active support but not broadening general support. This indicates that political polarization can manifest in both emotional and rational forms (Singer et al., 2019).

Our cross-cultural data analysis reveals distinct cultural impacts on the role of other-praising in political participation between Korea and the UK. According to the WVS Inglehart-Welzel World Cultural Map 2023, Korea has lower self-expression values (-0.47) emphasizing in-group cohesion, while the UK with higher values (2.24) reflects a societal norm of more tolerance towards out-groups (Haerpfer et al., 2022). In political contexts, expressing other-praising often enhances the reputation of one's in-group and reinforces internal unity (Brady et al., 2020). Consequently, this emotion may create a clear division between in-groups and out-groups (Brady et al., 2020).

In cultures valuing in-group cohesion like Korea, promoting petitions with other-praising emotion on social media can emphasize the in-group's prestige and strengthen belonging, delineating a clear divide between in-groups and out-groups. Considering cultural backgrounds and emotional attributes provides a compelling explanation for the role of other-praising expressions in Korean online petitions in contributing to polarized support tendencies. Conversely, in cultures like the UK, where tolerance towards out-groups is emphasized, such expressions might fail to garner support and provoke antipathy. This finding reiterates the call for research that considers cultural variation in moral emotions (Haidt et al., 2003; Malti and Keller, 2010; Van Bavel et al., 2023).

5.2 Implications for AI Community

We measure moral emotions cost-effectively, leveraging fine-tuned GPT-3.5 to inherit human knowledge, significantly reducing annotation costs. Further cost reductions in inference are achieved through in-house modeling. Our strategy led to the development of a classifier proficient in identifying moral emotions within both Korean and English texts. This is a notable achievement, considering the complexity of moral emotion classification and its application to the niche area of online political petitions. Applying this method opens up possibilities for reuse in languages with fewer resources and in tasks that previously faced challenges due to the high costs associated with developing domain-specific data.

Using real-world data, we also analyzed how moral emotions in language influence political actions. Our insights offer valuable implications for the AI community, which is increasingly interested in understanding the mechanics of political persuasion through content (OpenAI, 2024; Bai et al., 2023). We confirmed through our experiment that LLMs can understand and classify moral emotions, even from a limited sample of sentences. These discoveries prompt future work into the potential of generative AI in crafting content that may influence public opinion (Bai et al., 2023). For instance, the ability of generative AI to quickly generate content expressing other-condemning emotions presents a risk of polarizing public discourse and deepening societal divisions (Coeckelbergh, 2022).

6 Conclusion and Limitations

This study proposed a 5-step framework for analyzing moral emotions and their effects on political participation, leveraging cross-cultural data from online petitions. Our framework addresses research gaps using a comprehensive Korean and English moral emotion dataset and is adaptable to low-resource languages and topic domains. The dataset, annotated with fine-tuned GPT-3.5, enables training language-specific transformer models, allowing for the precise quantification of moral emotions within the petitions. Our analysis reveals that other-suffering enhanced both general and active political participation. In contrast, other-condemning led to polarization in these cultural contexts. Patterns of other-praising by countries underscores the cultural difference of moral emotions' influence on political engagement. The discussion of these findings, particularly the pronounced effects of petitions on specific moral emotions, provides valuable insights into the fields of social science and AI.

Our dataset, aimed at capturing cross-cultural nuances, may not fully represent moral emotions beyond Korea and the UK. Additionally, the data directed collected from government-led petitions could exhibit biases specific to this context. Such biases manifest in the expression of moral emotions and in shaping public opinion to prompt government action. Users of this dataset should be aware of these limitations. Additionally, our analysis of political participation is limited to the data's scope and cannot be generalized to other cultural contexts or media platforms. This observational study highlights associations without claiming causality, suggesting the need for future experimental research to explore the causal influence of moral emotions in language on political participation.

Potential Risks: Here, we designed prompts to train LLMs as annotators and constructed a dataset enriched with various moral emotion categories. However, we acknowledge the potential for modifying prompts to enable language models to generate texts infused with specific moral emotions. As discussed in Section 5.2, texts charged with moral emotions could be misused to sway public opinion for specific political agendas. Hence, we stipulate that the prompts used in our research and examples contained within the moral emotion dataset should be utilized solely for research purposes.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks. *arXiv preprint arXiv:2307.02179*.

Hui Bai, Jan Voelkel, Johannes Eichstaedt, and Robb Willer. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints*.

Parikshit Bansal and Amit Sharma. 2023. Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost. *arXiv preprint arXiv:2306.15766*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

William J Brady, Molly J Crockett, and Jay J Van Bavel. 2020. The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4):978–1010.

William J Brady, Killian McLoughlin, Tuan N Doan, and Molly J Crockett. 2021. How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33):eabe5641.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Mark Coeckelbergh. 2022. *The Political Philosophy of AI: An Introduction*. John Wiley & Sons.

Molly J Crockett. 2017. Moral outrage in the digital age. *Nature human behaviour*, 1(11):769–771.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a Good Data Annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11173–11195.

Anjalie Field, Chan Young Park, Antonio Theophilo, Jamelle Watson-Daniels, and Yulia Tsvetkov. 2022. An analysis of emotions and the prominence of positivity in #BlackLivesMatter tweets. *Proceedings of the National Academy of Sciences*, 119(35):e2205767119.

Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. 2020. Political sectarianism in America. *Science*, 370(6516):533–536.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2022. *World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0*. JD Systems Institute & WVSA Secretariat, Madrid, Spain & Vienna, Austria.

Jonathan Haidt, Richard J Davidson, Klaus R Scherer, and H Hill Goldsmith. 2003. The moral emotions. *Handbook of affective sciences*, 11(2003):852–870.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual Language Models are not Multicultural: A Case Study in Emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. *arXiv preprint arXiv:2303.16854*.

Yoel Inbar, David Pizarro, Ravi Iyer, and Jonathan Haidt. 2012. Disgust Sensitivity, Political Conservatism, and Voting. *Social Psychological and Personality Science*, 3(5):537–544.

Cheonsoo Kim and Sung-Un Yang. 2017. Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, 43(2):441–449.

Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. 2023. Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers. *arXiv preprint arXiv:2307.06090*.

Junbum Lee. 2020. KcBERT: Korean comments BERT. In *Proceedings of the 32nd Annual Conference on Human and Cognitive Language Technology*, pages 437–440.

Junbum Lee. 2021. KcELECTRA: Korean comments ELECTRA. `https://github.com/Beomi/KcELECTRA`.

Minchul Lee. 2022. Kiwi, Korean Intelligent Word Identifier. `https://github.com/bab2min/Kiwi`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Tina Malti and Monika Keller. 2010. The development of moral emotions in a cultural context. *Emotions, aggression, and morality in children: Bridging development and psychopathology*.

OpenAI. 2023. Chat Completions API.

OpenAI. 2024. How OpenAI is approaching 2024 worldwide elections.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337.

Julia Proskurnia, Przemyslaw Grabowicz, Ryota Kobayashi, Carlos Castillo, Philippe Cudré-Mauroux, and Karl Aberer. 2017. Predicting the Success of Online Petitions Leveraging Multidimensional Time-Series. In *Proceedings of the 26th International Conference on World Wide Web*, page 755–764.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic Sentence Boundary Disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software*, pages 110–114.

Daniel J Singer, Aaron Bramson, Patrick Grim, Bennett Holman, Jiin Jung, Karen Kovaka, Anika Ranginani, and William J Berger. 2019. Rational social and political polarization. *Philosophical Studies*, 176:2243–2267.

Cigdem V Sirin, Nicholas A Valentino, and José D Villalobos. 2017. The social causes and political consequences of group empathy. *Political Psychology*, 38(3):427–448.

Cigdem V Sirin, José D Villalobos, and Nicholas A Valentino. 2016. Group empathy theory: The effect of group empathy on US intergroup attitudes and behavior in the context of immigration threats. *The Journal of Politics*, 78(3):893–908.

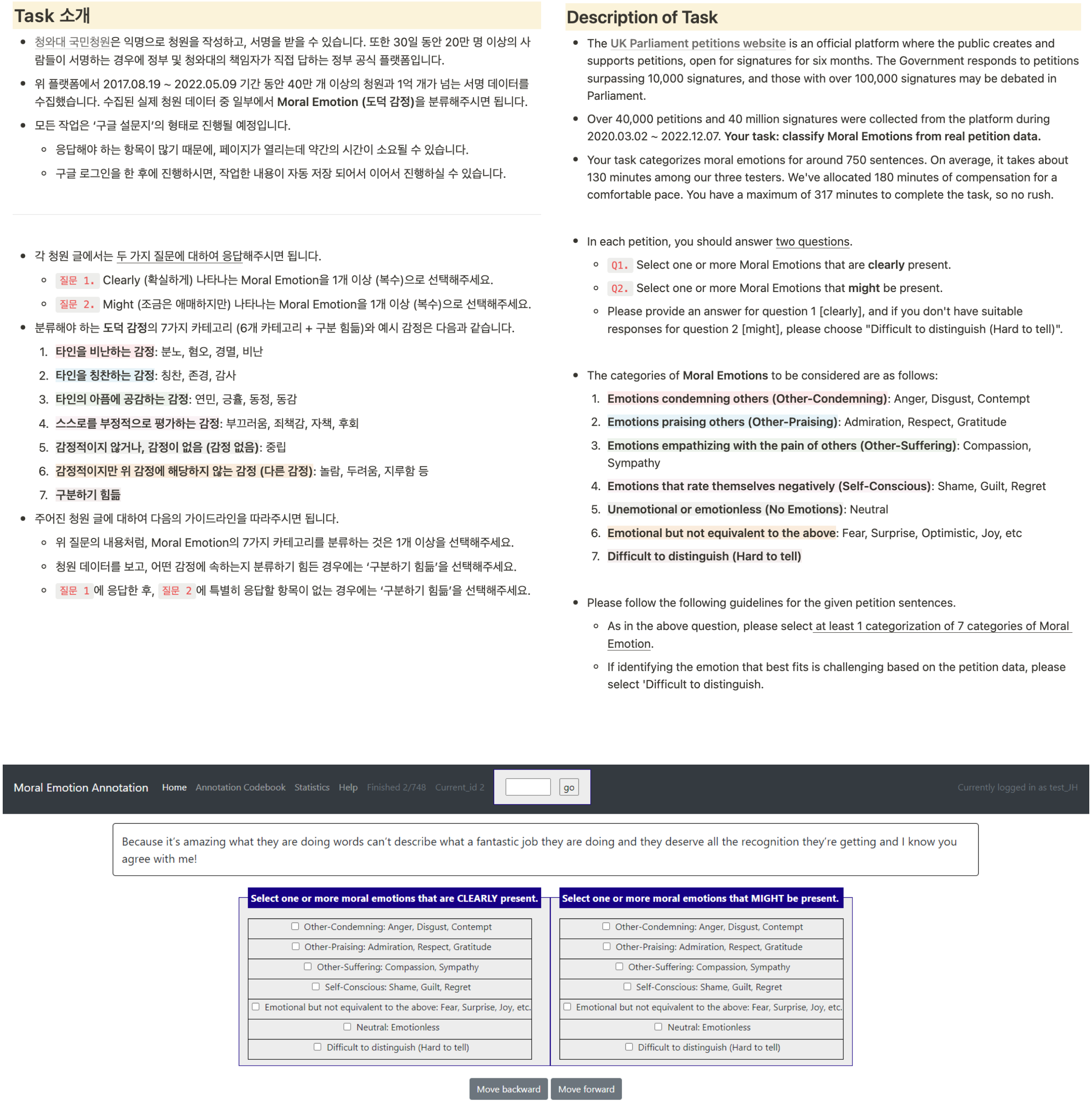
Kirill Solovev and Nicolas Pröllochs. 2022. Moral Emotions Shape the Virality of COVID-19 Misinformation on Social Media. In *Proceedings of the ACM web conference 2022*, pages 3706–3717.

Jay J Van Bavel, Claire E Robertson, Kareena Del Rosario, Jesper Rasmussen, and Steve Rathje. 2023. Social media and morality. *Annual Review of Psychology*, 75(1):311–340.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Xun Wang, Tao Ge, Allen Mao, Yuki Li, Furu Wei, and Si-Qing Chen. 2022. Pay Attention to Your Tone: Introducing a New Dataset for Polite Language Rewrite. *arXiv preprint arXiv:2212.10190*.

JungHwan Yang and Young Mie Kim. 2017. Equalization or normalization? Voter-candidate engagement on Twitter in the 2010 U.S. midterm elections. *Journal of Information Technology & Politics*, 14(3):232–247.



Task 소개

- 청와대 국민청원은 익명으로 청원을 작성하고, 서명을 받을 수 있습니다. 또한 30일 동안 20만 명 이상의 사람들이 서명하는 경우에 정부 및 청와대의 책임자가 직접 답하는 정부 공식 플랫폼입니다.
- 위 플랫폼에서 2017.08.19 ~ 2022.05.09 기간 동안 40만 개 이상의 청원과 1억 개가 넘는 서명 데이터를 수집했습니다. 수집된 실제 청원 데이터 중 일부에서 **Moral Emotion (도덕 감정)**을 분류해주시면 됩니다.
- 모든 작업은 '구글 설문지'의 형태로 진행될 예정입니다.
  - 응답해야 하는 항목이 많기 때문에, 페이지가 열리는데 약간의 시간이 소요될 수 있습니다.
  - 구글 로그인을 한 후에 진행하시면, 작업한 내용이 자동 저장 되어서 이어서 진행하실 수 있습니다.

- 각 청원 글에서는 두 가지 질문에 대하여 응답해주시면 됩니다.
  - 질문 1. Clearly (확실하게) 나타나는 Moral Emotion을 1개 이상 (복수)으로 선택해주세요.
  - 질문 2. Might (조금은 애매하지만) 나타나는 Moral Emotion을 1개 이상 (복수)으로 선택해주세요.
- 분류해야 하는 **도덕 감정**의 7가지 카테고리 (6개 카테고리 + 구분 힘듦)와 예시 감정은 다음과 같습니다.
  1. **타인을 비난하는 감정**: 분노, 혐오, 경멸, 비난
  2. **타인을 칭찬하는 감정**: 칭찬, 존경, 감사
  3. **타인의 아픔에 공감하는 감정**: 연민, 긍휼, 동정, 동감
  4. **스스로를 부정적으로 평가하는 감정**: 부끄러움, 죄책감, 자책, 후회
  5. **감정적이지 않거나, 감정이 없음 (감정 없음)**: 중립
  6. **감정적이지만 위 감정에 해당하지 않는 감정 (다른 감정)**: 놀람, 두려움, 지루함 등
  7. **구분하기 힘듦**
- 주어진 청원 글에 대하여 다음의 가이드라인을 따라주시면 됩니다.
  - 위 질문의 내용처럼, Moral Emotion의 7가지 카테고리를 분류하는 것은 1개 이상을 선택해주세요.
  - 청원 데이터를 보고, 어떤 감정에 속하는지 분류하기 힘든 경우에는 '구분하기 힘듦'을 선택해주세요.
  - 질문 1에 응답한 후, 질문 2에 특별히 응답할 항목이 없는 경우에는 '구분하기 힘듦'을 선택해주세요.

Description of Task

- The UK Parliament petitions website is an official platform where the public creates and supports petitions, open for signatures for six months. The Government responds to petitions surpassing 10,000 signatures, and those with over 100,000 signatures may be debated in Parliament.
- Over 40,000 petitions and 40 million signatures were collected from the platform during 2020.03.02 ~ 2022.12.07. **Your task: classify Moral Emotions from real petition data.**
- Your task categorizes moral emotions for around 750 sentences. On average, it takes about 130 minutes among our three testers. We've allocated 180 minutes of compensation for a comfortable pace. You have a maximum of 317 minutes to complete the task, so no rush.

- In each petition, you should answer two questions.
  - Q1. Select one or more Moral Emotions that are **clearly** present.
  - Q2. Select one or more Moral Emotions that **might** be present.
  - Please provide an answer for question 1 [clearly], and if you don't have suitable responses for question 2 [might], please choose "Difficult to distinguish (Hard to tell)".

- The categories of **Moral Emotions** to be considered are as follows:
  1. **Emotions condemning others (Other-Condemning)**: Anger, Disgust, Contempt
  2. **Emotions praising others (Other-Praising)**: Admiration, Respect, Gratitude
  3. **Emotions empathizing with the pain of others (Other-Suffering)**: Compassion, Sympathy
  4. **Emotions that rate themselves negatively (Self-Conscious)**: Shame, Guilt, Regret
  5. **Unemotional or emotionless (No Emotions)**: Neutral
  6. **Emotional but not equivalent to the above**: Fear, Surprise, Optimistic, Joy, etc
  7. **Difficult to distinguish (Hard to tell)**

- Please follow the following guidelines for the given petition sentences.
  - As in the above question, please select at least 1 categorization of 7 categories of Moral Emotion.
  - If identifying the emotion that best fits is challenging based on the petition data, please select 'Difficult to distinguish.

Because it's amazing what they are doing words can't describe what a fantastic job they are doing and they deserve all the recognition they're getting and I know you agree with me!

Select one or more moral emotions that are CLEARLY present.

☐ Other-Condemning: Anger, Disgust, Contempt
☐ Other-Praising: Admiration, Respect, Gratitude
☐ Other-Suffering: Compassion, Sympathy
☐ Self-Conscious: Shame, Guilt, Regret
☐ Emotional but not equivalent to the above: Fear, Surprise, Joy, etc.
☐ Neutral: Emotionless
☐ Difficult to distinguish (Hard to tell)

Select one or more moral emotions that MIGHT be present.

☐ Other-Condemning: Anger, Disgust, Contempt
☐ Other-Praising: Admiration, Respect, Gratitude
☐ Other-Suffering: Compassion, Sympathy
☐ Self-Conscious: Shame, Guilt, Regret
☐ Emotional but not equivalent to the above: Fear, Surprise, Joy, etc.
☐ Neutral: Emotionless
☐ Difficult to distinguish (Hard to tell)

Figure 4: Labeling guidelines and annotation tools provided to Korean and British annotators (Potato).

A Human Annotation

A.1 Human Annotation Guideline

Korean petition annotations were conducted via Google Surveys, while the UK petition utilized a Potato data annotation tool web interface (Pei et al., 2022). Figure 4 displays our Korean and English guidelines given to the users. Regarding compensation for the annotation task, each of five Korean annotators was paid ₩70,000, leading to a total of ₩350,000. On the other hand, each of the five UK annotators received £37.5, along with an additional Prolific service fee of £62.5, resulting in an overall expenditure of £250.

A.2 Human Annotation Dataset Result

We obtained 640 human-annotated data for the Korean petition, comprising 619 single-label and 21 multi-label instances. The distribution of emotion labels was as follows: other-condemning (136), other-praising (113), other-suffering (110), self-conscious (87), neutral (97), and non-moral emotion (118). Similarly, we acquired 640 human-annotated examples for the UK petition, consisting of 590 single-label and 50 multi-label instances. The distribution of the labels is as follows: other-condemning (226), other-praising (111), other-suffering (141), self-conscious (25), neutral (68), and non-moral emotion (119).

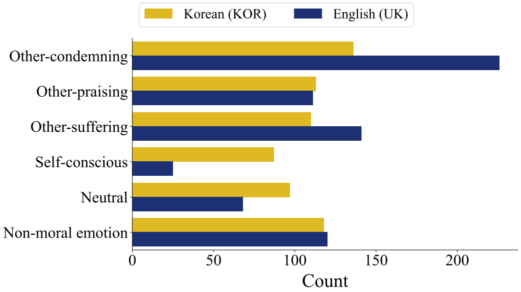


Figure 5: Distribution of moral emotion labels in human-annotated datasets, with 640 instances each in Korean and English.

A.3 Inter-annotator Agreement Score

Table 8 presents the inter-annotator agreement score for each class in a human-annotated dataset, measured by Cohen's kappa (C), Fleiss' kappa (F), and Krippendorff's alpha (K). The *other-praising* shows the highest score in both Korean and English, in contrast to *non-moral emotion*, which achieves lower score. The average scores across all emotion categories provide a comprehensive view of the annotation reliability, with the Korean dataset exhibiting a higher average agreement score than the English dataset.

	Korean			**English**		
	C	F	K	C	F	K
Other-condemning	0.5890	0.5812	0.5813	0.5018	0.4980	0.4982
Other-praising	0.9554	0.9554	0.9554	0.8005	0.8010	0.8010
Other-suffering	0.8107	0.8112	0.8112	0.2454	0.2064	0.2067
Self-conscious	0.7970	0.7993	0.7993	0.3738	0.3566	0.3568
Neutral	0.6043	0.6305	0.6306	0.4438	0.4765	0.4767
Non-moral emotion	0.5745	0.5743	0.5744	0.1813	0.1552	0.1555
Average	0.7218	0.7253	0.7254	0.4244	0.4156	0.4158

Table 8: Inter-annotator agreement score calculated using three different metrics for a dataset of 640 final human annotation responses.

B LLM Annotation

B.1 Prompt Format

Figure 6 and Figure 7 show the example few-shot prompt input and fine-tuning data formats that we gave to GPT-3.5 and GPT-4 for the labeling task. Each prompt and fine-tuning data is constructed in a chat format, consisting list of messages from the "system", "user", and "assistant". Both formats start with short definitions of moral emotion categories. While the few-shot learning prompt contains N number of training samples and human-annotated labels, fine-tuning training examples include a single training sample and corresponding labels per chat object. Finally, the few-shot learning prompt provides a task to label an unlabeled sentence in a multi-label manner.

System: You are an AI expert in moral emotion. You are asked to classify unlabeled sentences for 6 moral emotion labels. A text is classified into a moral emotion category if the author is expressing the emotion in the text. 'Moral Emotions' is divided into six categories, as follows:

1. Other-condemning: Emotions that condemn others, such as anger, contempt, or disgust.
2. Other-praising: Emotions that praise others, such as admiration, gratitude, or awe.
⋮
6. Neutral: A neutral category with no or few emotions.

User: I will give you a sentence and you will classify it into one or more of the moral emotion categories.
Make sure you answer the label among the predefined 6 moral emotions.
Assistant: Yes, I understand. I am ready to choose what kind of moral emotion is the writer expressing in the following petition sentences.

User: It's shameful and it make me appalled to be British. We have seen the impact that plastic is having on all our lives.
Assistant: Other-condemning, Self-conscious

User: · · ·
Assistant: · · ·

User: They were put in a financially adverse situation having to close during the Covid lockdown.
Assistant: Neutral

Figure 6: Prompt format composed of system, user, and assistant message for in-context learning.

System: You are an AI expert in moral emotion. You are asked to classify unlabeled sentences for 6 moral emotion labels. A text is classified into a moral emotion category if the author is expressing the emotion in the text. 'Moral Emotions' is divided into six categories, as follows:

1. Other-condemning: Emotions that condemn others, such as anger, contempt, or disgust.
2. Other-praising: Emotions that praise others, such as admiration, gratitude, or awe.
⋮
6. Neutral: A neutral category with no or few emotions.

I will give you a sentence and you will classify it into one or more of the moral emotion categories.
Make sure you answer the label among the predefined 6 moral emotions.

User: It's shameful and it make me appalled to be British. We have seen the impact that plastic is having on all our lives.
Assistant: Other-condemning, Self-conscious

System: You are an AI expert in moral emotion · · ·
User: · · ·
Assistant: · · ·

Figure 7: Train data format composed of system, user, and assistant message for fine-tuning.

B.2 Monolingual vs Bilingual Fine-tuning

To compare the performance of monolingual and bilingual training approaches, we first fine-tuned GPT-3.5 models on separate Korean and English single-language datasets, creating two monolingual models. Subsequently, we fine-tuned another model on a combined bilingual dataset. Training on monolingual corpora may provide a deeper understanding of each language's unique nuances and socio-cultural contexts. Conversely, joint training could enhance multilingual understanding and performance by facilitating knowledge transfer across languages. Table 9 presents the training dataset size for each experiment along with the models' macro F1 scores and accuracy. The models achieved the best performance in Korean and English datasets under the monolingual condition when the training size was $N = 300$.

	Train size	Korean (KOR)		English (UK)	
		F1	**Acc.**	**F1**	**Acc.**
Bilingual	150	0.8094	0.8029	0.7138	0.6618
	300	0.8636	0.8588	0.7171	0.6912
Monolingual	150	0.8518	0.8471	0.7169	0.7029
	300	**0.8678**	**0.8618**	**0.7426**	**0.7294**

Table 9: Comparison of performance between fine-tuning using monolingual or bilingual train dataset. The bilingual dataset size is set as N = 150, $2N$ = 300 while the monolingual dataset size is N = 150 for each Korean and English.

B.3 Fine-tuning Train Dataset Size

We conducted experiments with different training set sizes ranging from 50, 100, 150, $\cdots$, to 300 to optimize the train data size and improve fine-tuned model performance. Figure 8 depicts the model macro F1 score and accuracy as a function of fine-tuning dataset size. The performance tends to improve as the train dataset size grows. The Korean model shows the highest F1 score and accuracy when the training dataset size N = 300, while the English model performs best when N = 250.

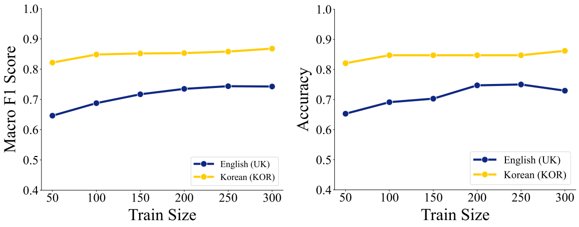


Figure 8: Macro F1 score and accuracy for fine-tuning train dataset size.

C Moral Emotion Dataset

C.1 Moral Emotion Dataset Statistics

The basic statistics and features are represented in Table 10. The moral emotion dataset comprises 49,930 Korean and 49,896 UK petition samples. Both datasets contain six classes: four pertaining to moral emotions (other-condemning, other-praising, other-suffering, self-conscious) and two to non-moral emotions (neutral, non-moral emotion). Most of our data samples are single-labeled, accounting for 99.93% in the Korean and 99.99% in the UK.

Properties	**Korean**	**English**
Number of instances	49,930	49,896
Number of classes	6	
Number of instances with single-label	49,894	49,892
Number of instances with multi-label	36	4

Table 10: Statistics of moral emotion dataset.

C.2 Moral Emotion Classifier Performance

Figure 9 presents the per-class F1 scores for Korean and English weighted ensemble classifier models, noting that the F1 score for the *non-moral emotion* class in the English model is below 0.7.

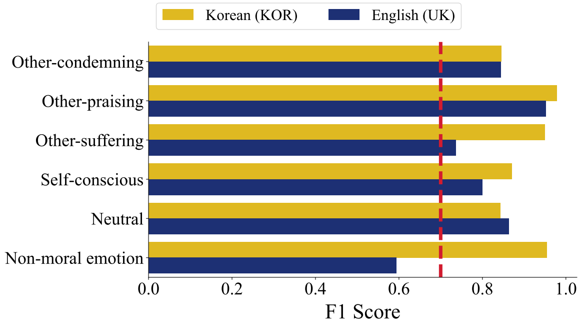


Figure 9: F1 scores for each class in both the Korean and English models.

Type	Category	Sample Text
Moral Emotions	Other-condemning	His actions and decisions have been chaotic & contradictory at best throughout the pandemic. Retailers are putting their staff and families in danger so they can make a profit.
		자신의 정치 권력을 위해 치열하게 투쟁하는 정치꾼일 뿐입니다. (They are merely self-serving politicians, fighting only for their own gain.) 언론과 정부가 제 역할을 못하고 있다고 생각합니다. (I believe that the media and government are failing to fulfill their roles.)
	Other-praising	This gentleman is an inspiration to us all and should be commended for his efforts. Let us not forget what these key workers are doing for this country.
		코로나 대응에 총력을 다하는 공무원, 의료진 분들을 응원합니다. (We support officials and medical staff who are working hard to respond to COVID-19.) 경찰관은 목숨을 걸고 달려갈 것이며, 당신을 살리기 위해 최선을 다할 것입니다. (The police officer will run for his life, and he will do his best to save you.)
	Other-suffering	Something needs doing NOW the government have to now take action to protect the vulnerable. The Government must fund this to help protect most vulnerable during pandemic.
		현재 우리나라도 살기 어렵고, 힘들게 살아가는 사람들이 많이 있습니다. (Currently, there are many people who are difficult to live in our country and have a hard time living.) 더 이상 아이를 잃는 아픔을 겪지 않게 법을 강화해 주시길 바랍니다. (I hope that laws will be strengthened to prevent any more pain of losing children.)
	Self-conscious	As a British citizen I am ashamed of our pitiful response towards those fleeing war zones. It will be embarrassing on the global stage to not have any government organized fireworks.
		이렇게 저희는 어머니의 임종도 지켜드리지 못하고, 갑작스럽게 어머니를 보내드려야 했습니다. (We couldn’t be there for our mother’s final moments and had to say goodbye to her abruptly.) 이 정권에 힘을 실어줬던 과거의 제 결정이 정말 후회스럽습니다. (I really regret my decisions in the past that gave this regime a boost.)
Non-moral	Neutral	I understand the reasons and the carbon foot print is very much in the fore front of our minds.
		해외연수 후 이행 내역을 임기후 5년까지 국민이 볼 수 있도록 해주세요. (Make the implementation details of overseas training to the public for up to five years after the term.)
	Non-moral emotion	I, and others, have serious concerns about the accuracy of the daily COVID-19 statistics.
		혹여나 아직 감염자 없는 지역에서 내가 가해자가 될까 두렵기도 합니다. (I fear becoming the perpetrator in areas where there are still no infected individuals.)

Table 11: Example sentences of both moral and non-moral emotion categories from the Korean and UK datasets.

Sentence	Lexicon-based	Ours
Animal **cruelty** is taken **seriously** in the UK Moral.	Emotional	Neutral
For example a judge cannot also be a referee and a referee cannot also **judge fights**.	Moral emotional	Neutral
Set customer service KPIs that utility and telecom companies must meet to show **good** service.	Moral emotional	Neutral
The process does not put my child’s needs at the forefront.	Neutral	Other-Suffering

Table 12: Examples of moral emotion classification results using lexicon-based and Transformer-based approaches on the same sentences, with bold indicating terms identified by the lexicon-based method.

Properties	Korean (KOR)	English (UK)
Number of Petitions	459,447	41,292
Number of Sentences	4,705,292	210,304
Number of Signatures	161,856,648	47,554,399
Number of Shares on Twitter	251,245	853,222
Date	2017.08.25 - 2022.05.09	2020.03.02 - 2022.12.07

Table 13: Statistics of collected petition dataset.

Variables	Korean (KOR)		English (UK)	
	General (Signatures)	Active (Shares)	General (Signatures)	Active (Shares)
Other-Condemning	-0.0561*	2.3649***	-0.0429	1.3158***
	(0.0249)	(0.174)	(0.0777)	(0.0979)
Other-Praising	-0.5209***	2.1001***	-1.0258***	-0.7685***
	(0.0276)	(0.1953)	(0.0904)	(0.1156)
Other-Suffering	1.3833***	2.28***	0.217*	0.4752***
	(0.036)	(0.2213)	(0.0873)	(0.1097)
Self-Conscious	-4.0095***	-5.1869***	-5.3267***	-2.4638***
	(0.055)	(0.4114)	(0.3994)	(0.5025)
Neutral	-0.109***	0.5957***	0.0847	1.7557***
	(0.0256)	(0.1672)	(0.111)	(0.1404)
URL Included	0.5963***	1.2724***	0.0178	0.757***
	(0.018)	(0.0935)	(0.0926)	(0.1164)
The Number of Sentences	-0.0623***	0.0163	-0.0838***	-0.0851***
	(0.001)	(0.0108)	(0.0082)	(0.0096)
The Number of Characters	0.0038***	0.001***	0.003***	0.0029***
	(0.00001)	(0.0001)	(0.00008)	(0.0001)
Before Apr. 2019	-1.9169***	-4.0316***	-	-
	(0.0136)	(0.1362)	-	-
(Intercept)	4.8415***	-1.9557***	4.2896***	-0.6444***
	(0.0351)	(0.287)	(0.1052)	(0.1299)
Year Fixed Effects	Yes		Yes	
Month Fixed Effects	Yes		Yes	
Weekday Fixed Effects	Yes		Yes	
Observations	438,871		40,130	

Sign. levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; *Standard errors in parentheses*

Table 14: Result of negative binomial regression on the number of signatures and shares. For Korea, we added an extra dummy variable. Before April 2019, to account for the fact that petitions with more than 100 signatures were listed on the board.