ACL 2024

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

Jaehong Kim*, Chaeyoon Jeong*, Seongchan Park*, Meeyoung Cha†, Wonjae Lee†

KAIST
ibs 데이터사이언스그룹 Data Science Group
MAX PLANCK INSTITUTE FOR SECURITY AND PRIVACY

01 Background

1) What is Moral Emotion?

Definition of Moral Emotion

- Unlike general emotions, **pro-social emotions** motivated by intentions for others and social justice (Haidt, 2003)

Type	Category	Definition
Moral Emotions	Other-Condemning	Emotions that condemn others, such as anger, contempt, or disgust
	Other-Praising	Emotions that praise others, such as admiration, gratitude, or awe
	Other-Suffering	Emotions of empathy for the suffering of others, such as compassion or sympathy
	Self-Conscious	Emotions that negatively evaluate oneself, such as shame, guilt, or embarrassment
Non-Moral Emotions	Neutral	A neutral category with few or no emotions
	Non-Moral Emotion	Emotions that are emotional but not one of the moral emotions, such as fear, surprise, joy, etc.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

01 Background

1) What is Moral Emotion?

Definition of Moral Emotion

- Unlike general emotions, **pro-social emotions** motivated by intentions for others and social justice (Haidt, 2003)

Type		Category	Definition
Moral Emotions		Other-Condemning	Emotions that condemn others, such as anger, contempt, or disgust
		Other-Praising	Emotions that praise others, such as admiration, gratitude, or awe
		Other-Suffering	Emotions of empathy for the suffering of others, such as compassion or sympathy
		Self-Conscious	Emotions that negatively evaluate oneself, such as shame, guilt, or embarrassment
Non-Moral Emotions		Neutral	A neutral category with few or no emotions
		Non-Moral Emotion	Emotions that are emotional but not one of the moral emotions, such as fear, surprise, joy, etc.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

01 Background

1) What is Moral Emotion?

Definition of Moral Emotion

- Unlike general emotions, **pro-social emotions** motivated by intentions for others and social justice (Haidt, 2003)

Type	Category	Definition
Moral Emotions	 Other-Condemning	Emotions that condemn others, such as anger, contempt, or disgust
	 Other-Praising	Emotions that praise others, such as admiration, gratitude, or awe
	 Other-Suffering	Emotions of empathy for the suffering of others, such as compassion or sympathy
	Self-Conscious	Emotions that negatively evaluate oneself, such as shame, guilt, or embarrassment
Non-Moral Emotions	Neutral	A neutral category with few or no emotions
	Non-Moral Emotion	Emotions that are emotional but not one of the moral emotions, such as fear, surprise, joy, etc.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

01 Background

1) What is Moral Emotion?

Definition of Moral Emotion

- Unlike general emotions, **pro-social emotions** motivated by intentions for others and social justice (Haidt, 2003)

Type		Category	Definition
Moral Emotions	😤	Other-Condemning	Emotions that condemn others, such as anger, contempt, or disgust
	🥳	Other-Praising	Emotions that praise others, such as admiration, gratitude, or awe
	🤗	Other-Suffering	Emotions of empathy for the suffering of others, such as compassion or sympathy
	😮‍💨	Self-Conscious	Emotions that negatively evaluate oneself, such as shame, guilt, or embarrassment
Non-Moral Emotions	😶	Neutral	A neutral category with few or no emotions
	😐	Non-Moral Emotion	Emotions that are emotional but not one of the moral emotions, such as fear, surprise, joy, etc.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

01 Background

1) What is Moral Emotion?

Definition of Moral Emotion

- Unlike general emotions, **pro-social emotions** motivated by intentions for others and social justice (Haidt, 2003)

Type	Category	Definition
Moral Emotions	😤 Other-Condemning	Emotions that condemn others, such as anger, contempt, or disgust
	🥳 Other-Praising	Emotions that praise others, such as admiration, gratitude, or awe
	🤗 Other-Suffering	Emotions of empathy for the suffering of others, such as compassion or sympathy
	😳 Self-Conscious	Emotions that negatively evaluate oneself, such as shame, guilt, or embarrassment
Non-Moral Emotions	😑 Neutral	A neutral category with few or no emotions
	😐 Non-Moral Emotion	Emotions that are emotional but not one of the moral emotions, such as fear, surprise, joy, etc.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

01 Background

1) What is Moral Emotion?

Why Does ‘Moral Emotion’ Matter?

- Moral emotional contents **capture human’s attention** (Boehm, 2012; Krebs, 2008, Brady et al., 2017)
  - Headlines criticizing celebrities for breaking social norms capture our attention

Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. Basic Books.

Krebs, D. L. (2008). Morality: An evolutionary account. *Perspectives on Psychological Science, 3*(3), 149–172.

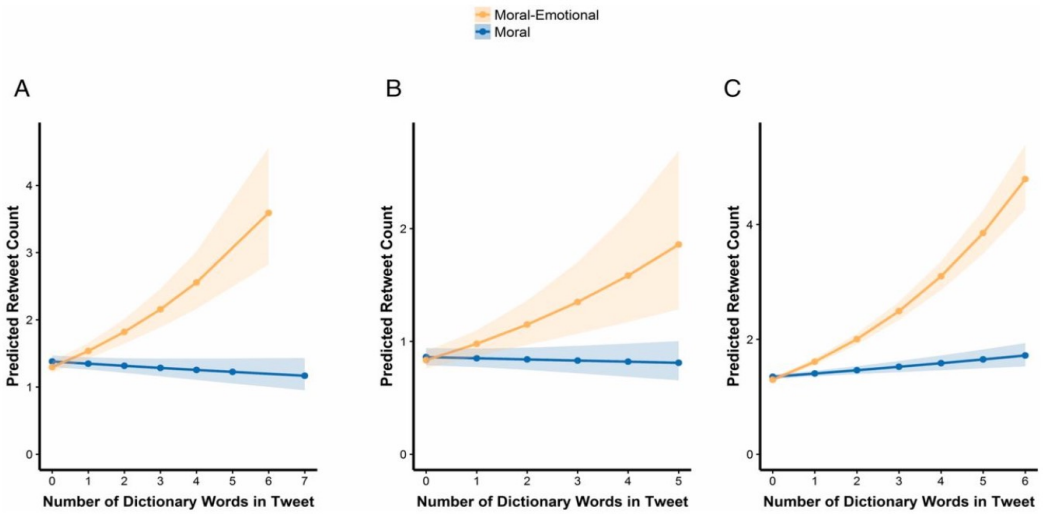
Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318.

01 Background

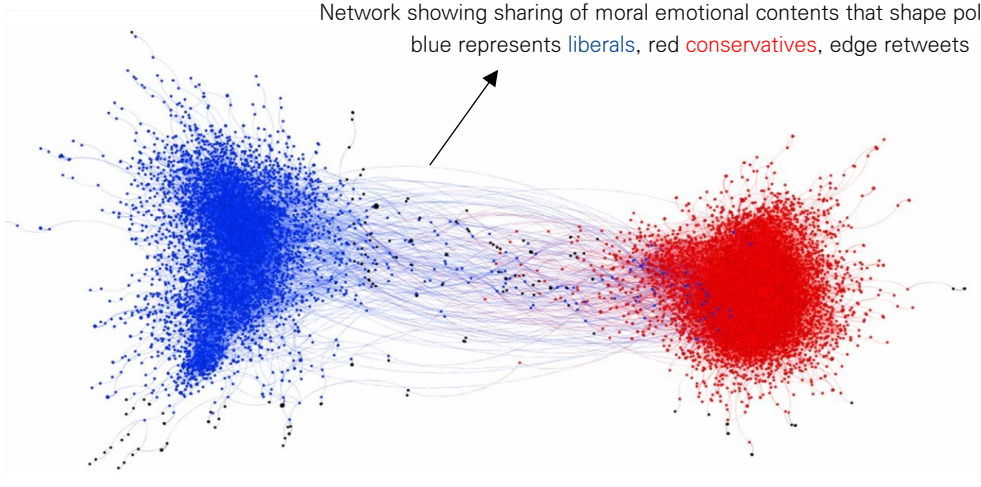
1) What is Moral Emotion?

Why Does 'Moral Emotion' Matter?

- Moral emotional contents **capture human's attention** (Boehm, 2012; Krebs, 2008, Brady et al., 2017)
  - Headlines criticizing celebrities for breaking social norms capture our attention
- Moral contagion: Moral emotions are key to **the spread of political discourse online** (Brady et al., 2017)
- Moral contagion tends to be **shaped by political ideology** → political polarization (Brady et al., 2017)



Network showing sharing of moral emotional contents that shape polarization
blue represents liberals, red conservatives, edge retweets



Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. Basic Books.

Krebs, D. L. (2008). Morality: An evolutionary account. *Perspectives on Psychological Science, 3*(3), 149–172.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318.

01 Background

2) Limitations of Previous Research

Theoretical Limitations

- **Narrow Range of Moral Emotions**: e.g.) Moral vs Non-Moral, Other-condemning vs Self-conscious
- **Challenges in Generalizing Social Media Metric to Entire Political Engagement**:

  Relying on retweet counts only may overrepresent a segment of the population that is vocal on social media
- **Single-Culture Focus in U.S Political Behavior Analysis**: Limiting its broader applicability

01 Background

2) Limitations of Previous Research

Theoretical Limitations

- **Narrow Range of Moral Emotions**: e.g.) Moral vs Non-Moral, Other-condemning vs Self-conscious
- **Challenges in Generalizing Social Media Metric to Entire Political Engagement**:

  Relying on retweet counts only may overrepresent a segment of the population that is vocal on social media
- **Single-Culture Focus in U.S Political Behavior Analysis**: Limiting its broader applicability

Technical Limitations

- **Dictionary-Based Measurement**:

  Less accurate emotion classification compared to recent deep learning-based methods

02 Research Framework

1) Technical Framework

Dictionary-Based Measurement — Moral Emotions Training Dataset — Transformer-Based Measurement

Lower Performance

- Cannot capture context
- Less accurate

Constructing Training Dataset Using LLMs

- Any Moral Emotions dataset does not exist covering all Moral Emotion categories
- By combining the knowledge of human annotators of Moral Emotions with the LLMs, we achieve a more cost-effective approach

High Performance

- Training with Transformer model
- Aiming for enhanced accuracy in Moral Emotion classification

Sentence	Dictionary-Based	Transformer-Based
Animal **cruelty** is taken **seriously** in the UK.	Moral Emotional (Cruel, Seriously)	Neutral
For example, a **judge** cannot also be a referee and a referee cannot also judge **fights**.	Moral Emotional (Fight, Judge)	Neutral
Set customer service KPI's that all utility and telecom companies need to hit to show they are providing a **good** service.	Moral Emotional (Good)	Neutral

02 Research Framework

1) Technical Framework

Dictionary-Based Measurement — Moral Emotions Training Dataset — Transformer-Based Measurement

Lower Performance

- Cannot capture context
- Less accurate

→

Constructing Training Dataset Using LLMs

- Any Moral Emotions dataset does not exist covering all Moral Emotion categories
- By combining the knowledge of human annotators of Moral Emotions with the LLMs, we achieve a more cost-effective approach

→

High Performance

- Training with Transformer model
- Aiming for enhanced accuracy in Moral Emotion classification

Sentence	Dictionary-Based	Transformer-Based
Animal **cruelty** is taken **seriously** in the UK.	Moral Emotional (Cruel, Seriously)	Neutral
For example, a **judge** cannot also be a referee and a referee cannot also judge **fights**.	Moral Emotional (Fight, Judge)	Neutral
Set customer service KPI's that all utility and telecom companies need to hit to show they are providing a **good** service.	Moral Emotional (Good)	Neutral

02 Research Framework

2) Theoretical Framework

Narrow range — Simple Metrics — Single-Culture

Moral Emotions

- Other-Condemning
- Other-Praising
- Other-Suffering
- Self-Conscious

Non-Moral Emotions

- Neutral
- Non-Moral Emotion

Detailed Individual Moral Emotions
(Independent variables)

Active Support (#Shared on tweets)

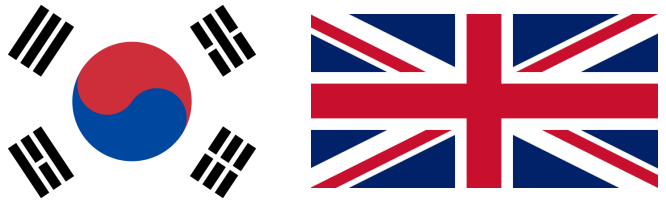
- Highly visible
- Vocal, but not a direct indicator of political outcomes unlike voting

General Support (#Signature)

- Anonymous voting
- Quiet, yet has a more direct effect

Political Participation Into Two Types
(Dependent variables)

South Korea & United Kingdom



Cross-Cultural analysis beyond U.S

Wide range — Various Metrics — Cross-Cultural

02 Research Framework

2) Theoretical Framework

Narrow range — Simple Metrics — Single-Culture

Moral Emotions

- Other-Condemning
- Other-Praising
- Other-Suffering
- Self-Conscious

Non-Moral Emotions

- Neutral
- Non-Moral Emotion

Detailed Individual Moral Emotions
(Independent variables)

Active Support (#Shared on tweets)

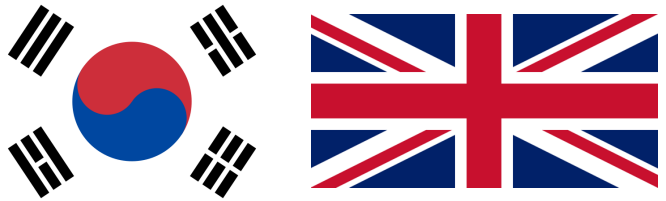
- Highly visible
- Vocal, but not a direct indicator of political outcomes unlike voting

General Support (#Signature)

- Anonymous voting
- Quiet, yet has a more direct effect

Political Participation Into Two Types
(Dependent variables)

South Korea & United Kingdom



Cross-Cultural analysis beyond U.S

Wide range — Various Metrics — Cross-Cultural

02 Research Framework

2) Theoretical Framework

Narrow range — Simple Metrics — Single-Culture

Moral Emotions

- Other-Condemning
- Other-Praising
- Other-Suffering
- Self-Conscious

Non-Moral Emotions

- Neutral
- Non-Moral Emotion

Detailed Individual Moral Emotions
(Independent variables)

Active Support (#Shared on tweets)

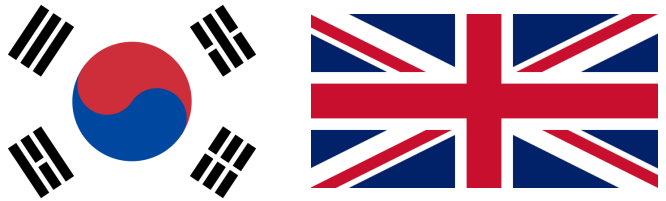
- Highly visible
- Vocal, but not a direct indicator of political outcomes unlike voting

General Support (#Signature)

- Anonymous voting
- Quiet, yet has a more direct effect

Political Participation Into Two Types
(Dependent variables)

South Korea & United Kingdom

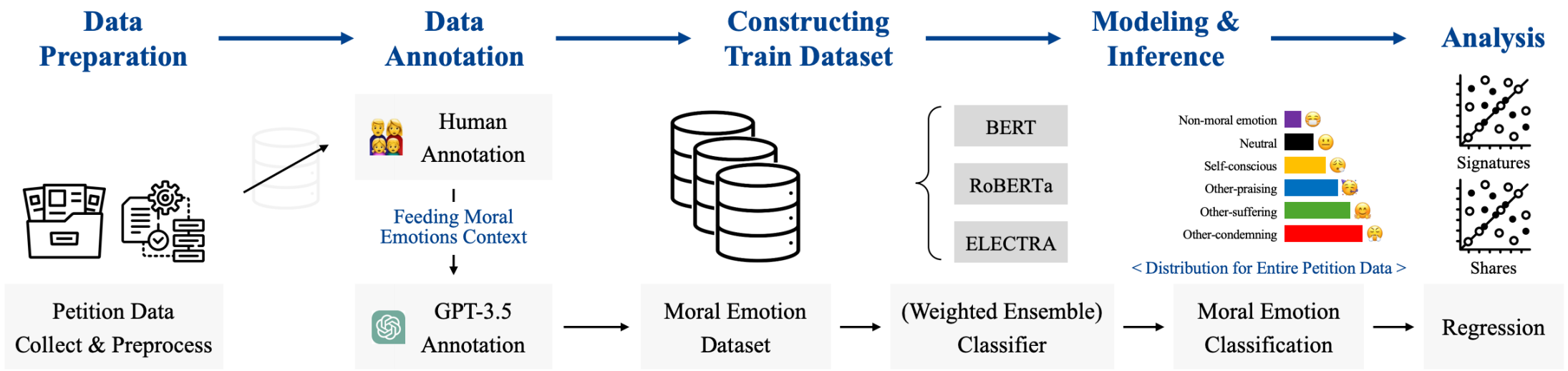


Cross-Cultural analysis beyond U.S

Wide range — Various Metrics — Cross-Cultural

03 Methodology

1) Schematic Diagram _Overview



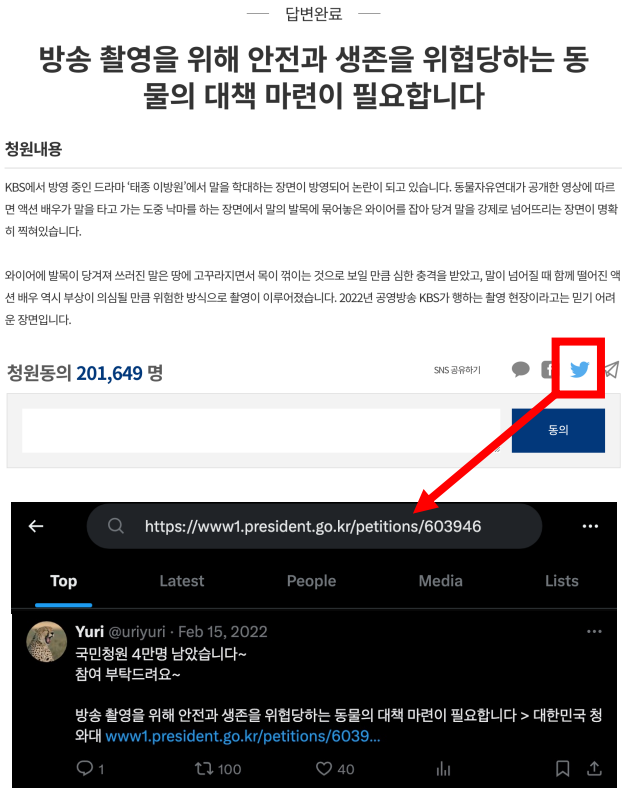
〈 Method Overview 〉

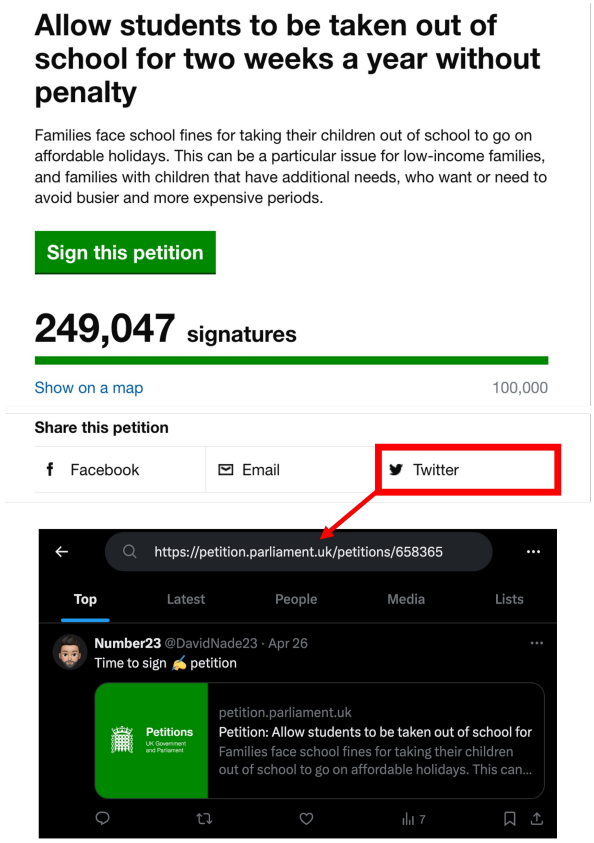
03 Methodology

1) Schematic Diagram _Overview

Data Preparation

Petition Data
Collect & Preprocess

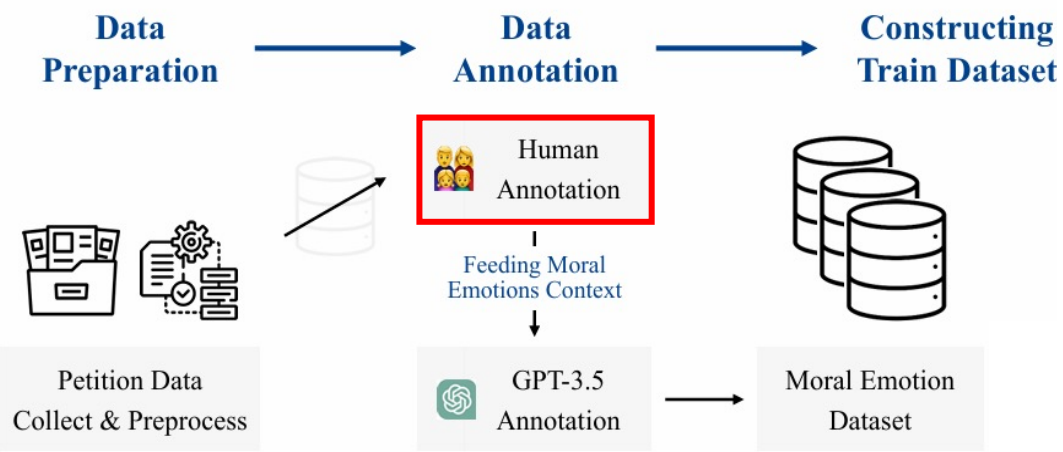




Government-led Petition Data Examples (Korea, UK)

03 Methodology

2) Creating Training Dataset



	Train size	Korean (KOR) F1	Acc.	Cost ($)	English (UK) F1	Acc.	Cost ($)
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.).

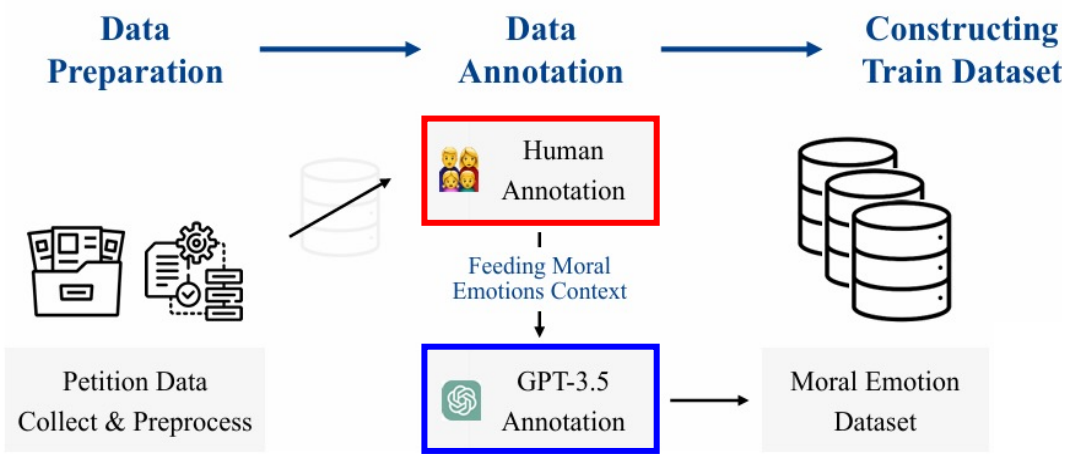
Moral Emotion Training Dataset (Labeled by Fine-tuned GPT-3.5)

1. Sample 700 sentences from petition texts (700 Korean / 700 English) & 5 annotators label moral emotions

   → Output: Human-Annotated Dataset

03 Methodology

2) Creating Training Dataset



	Train size	Korean (KOR) F1	Acc.	Cost ($)	English (UK) F1	Acc.	Cost ($)
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.).

Moral Emotion Training Dataset (Labeled by Fine-tuned GPT-3.5)

1. Sample 700 sentences from petition texts (700 Korean / 700 English) & 5 annotators label moral emotions

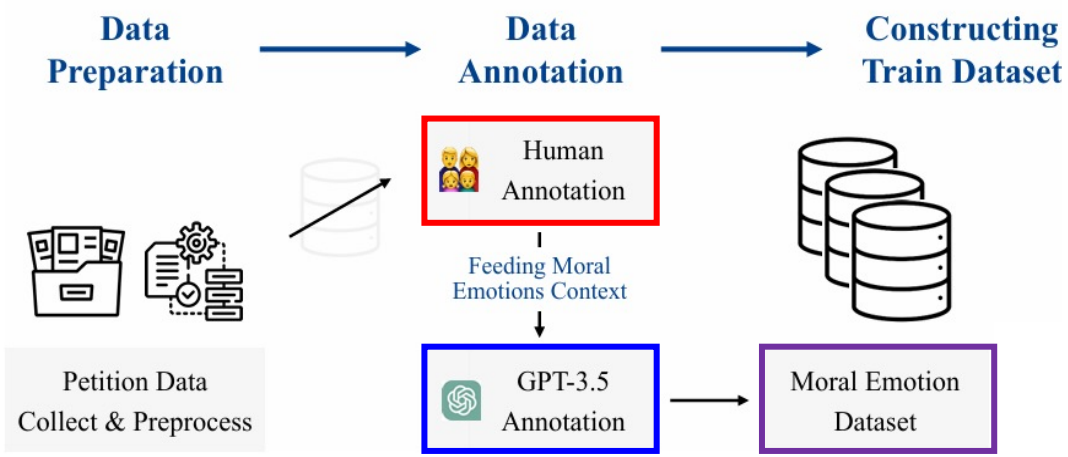
   → Output: Human-Annotated Dataset

2. Fine-tuning GPT-3.5 using the Human-Annotated Dataset allows GPT to inherit human knowledge

   → Output: Fine-tuned GPT-3.5

03 Methodology

2) Creating Training Dataset



	Train size	Korean (KOR) F1	Acc.	Cost ($)	English (UK) F1	Acc.	Cost ($)
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.).

Moral Emotion Training Dataset (Labeled by Fine-tuned GPT-3.5)

1. Sample 700 sentences from petition texts (700 Korean / 700 English) & 5 annotators label moral emotions

   → Output: Human-Annotated Dataset

2. Fine-tuning GPT-3.5 using the Human-Annotated Dataset allows GPT to inherit human knowledge

   → Output: Fine-tuned GPT-3.5

3. Fine-tuned GPT-3.5 labels 50,000 unseen sentences (50,000 Korean / 50,000 English)

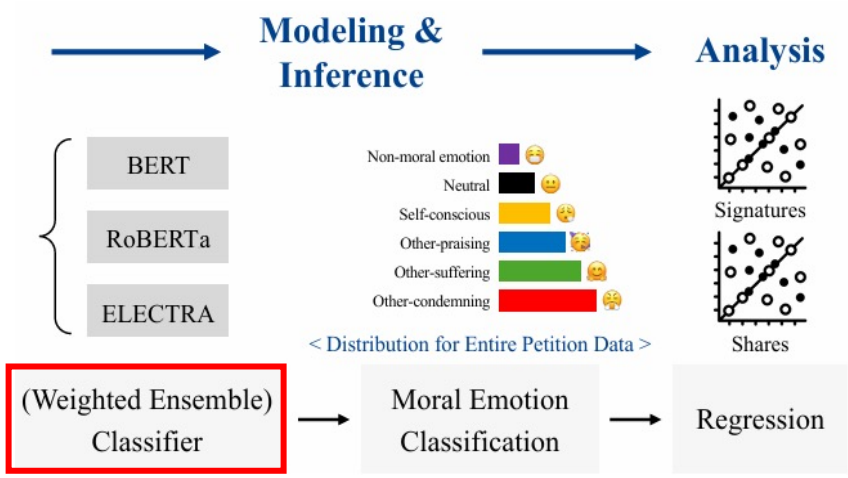
   → Output: Moral Emotion Training Dataset Labeled by Fine-tuned GPT-3.5

03 Methodology

3) Modeling

	Korean (KOR)		English (UK)	
	F1	Acc.	F1	Acc.
Fine-tuned GPT-3.5	0.8678	**0.8618**	0.7436	**0.7500**
BERT	0.8858	0.8500	0.6760	0.6588
RoBERTa	0.8785	0.8471	0.7134	0.6971
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. open-source models on human-annotated data.



Training Dataset (Labeled by Fine-tuned GPT-3.5) -> Transformer Model for Moral Emotions Classification

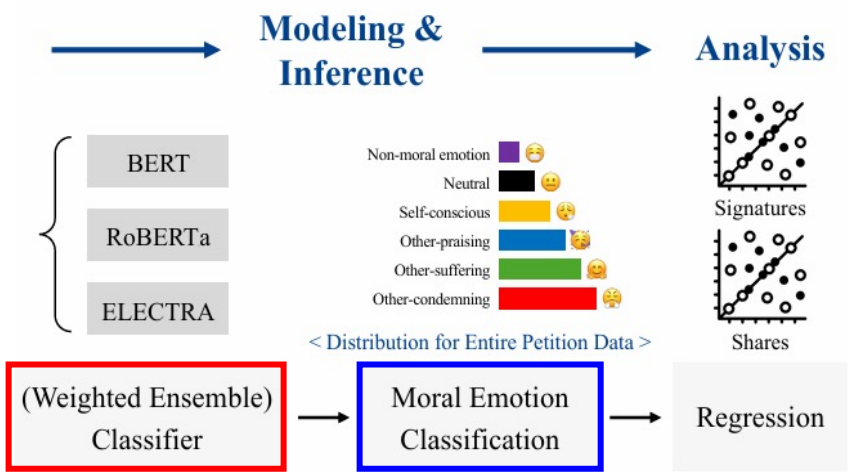
1. Training open-source language models with datasets labeled by Fine-tuned GPT-3.5

03 Methodology

3) Modeling

	Korean (KOR)		English (UK)	
	F1	Acc.	F1	Acc.
Fine-tuned GPT-3.5	0.8678	**0.8618**	0.7436	**0.7500**
BERT	0.8858	0.8500	0.6760	0.6588
RoBERTa	0.8785	0.8471	0.7134	0.6971
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. open-source models on human-annotated data.



Training Dataset (Labeled by Fine-tuned GPT-3.5) -> Transformer Model for Moral Emotions Classification

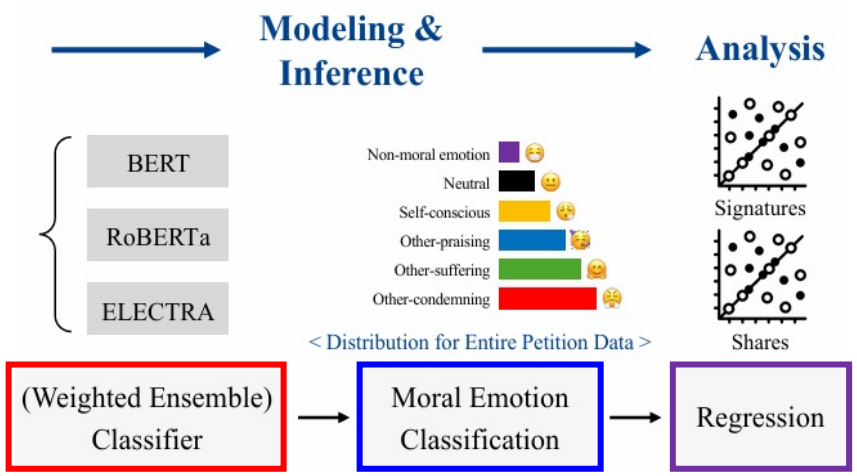
1. Training open-source language models with datasets labeled by Fine-tuned GPT-3.5
2. Inferring moral emotions from 4,705,292 Korean petition sentences / 210,304 the UK petition sentences

03 Methodology

3) Modeling

	Korean (KOR)		English (UK)	
	F1	Acc.	F1	Acc.
Fine-tuned GPT-3.5	0.8678	**0.8618**	0.7436	**0.7500**
BERT	0.8858	0.8500	0.6760	0.6588
RoBERTa	0.8785	0.8471	0.7134	0.6971
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. open-source models on human-annotated data.



Training Dataset (Labeled by Fine-tuned GPT-3.5) -> Transformer Model for Moral Emotions Classification

1. Training open-source language models with datasets labeled by Fine-tuned GPT-3.5
2. Inferring moral emotions from 4,705,292 Korean petition sentences / 210,304 the UK petition sentences
3. Performing regression analysis to check the relationship between
   - Moral emotions and # signatures (General Supports)
   - Moral emotions and # shares (Active Supports)

04 Result

1) Overview

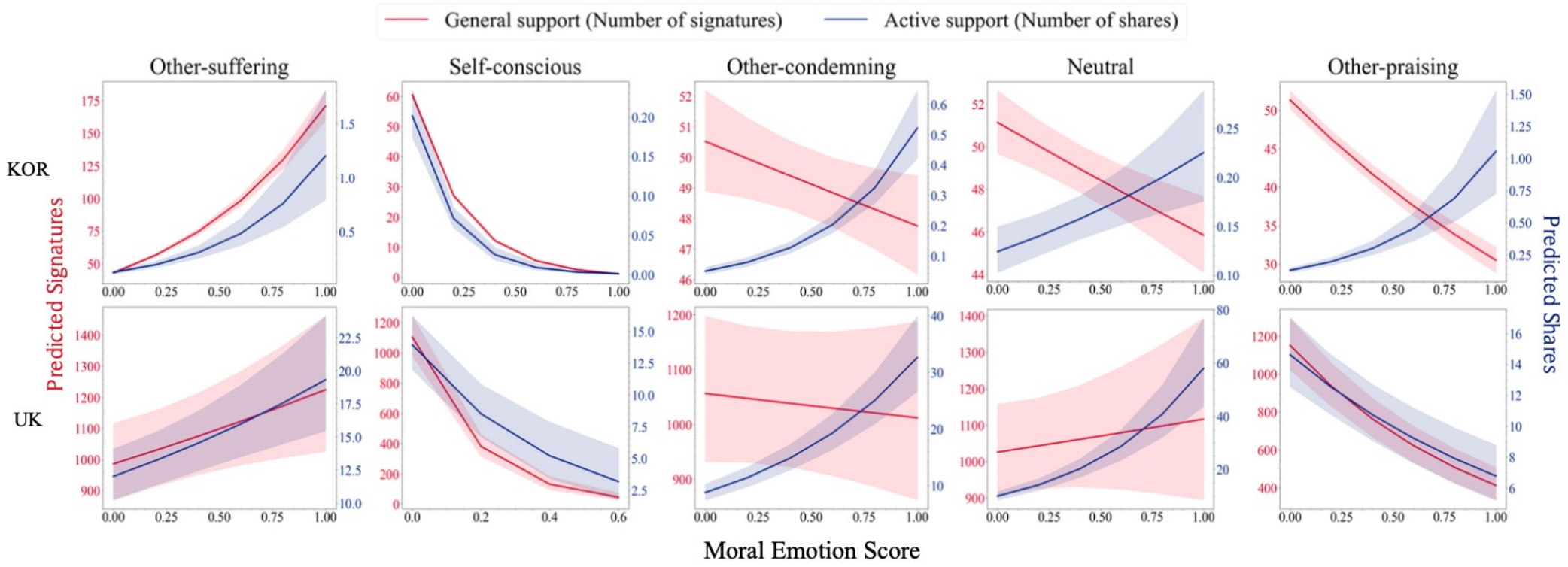
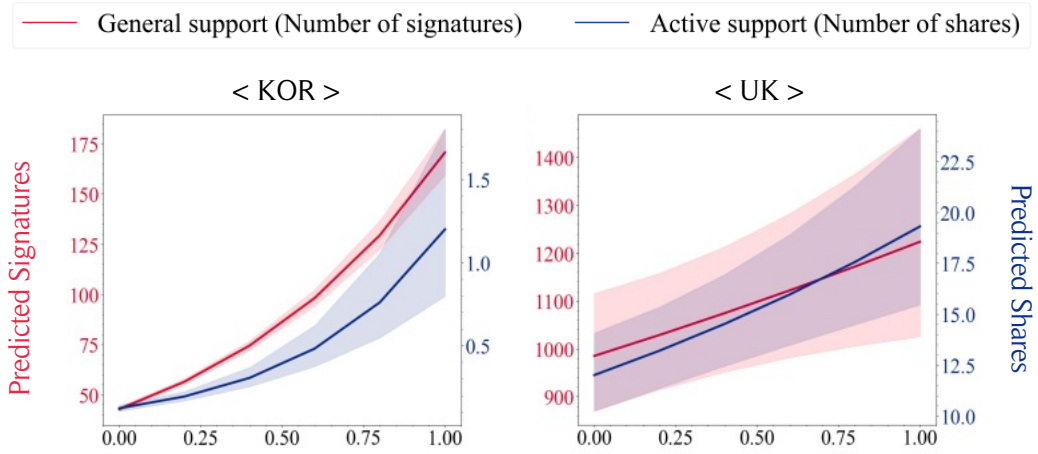


Figure 3: Predictive margins for general support (depicted by a red line) and active support (blue line) across five emotions in studied countries, with 95% confidence intervals.

05 Implications for Social Science

'Other-suffering' Shapes Political Consensus



Emotion Score for 'Other-suffering'

→

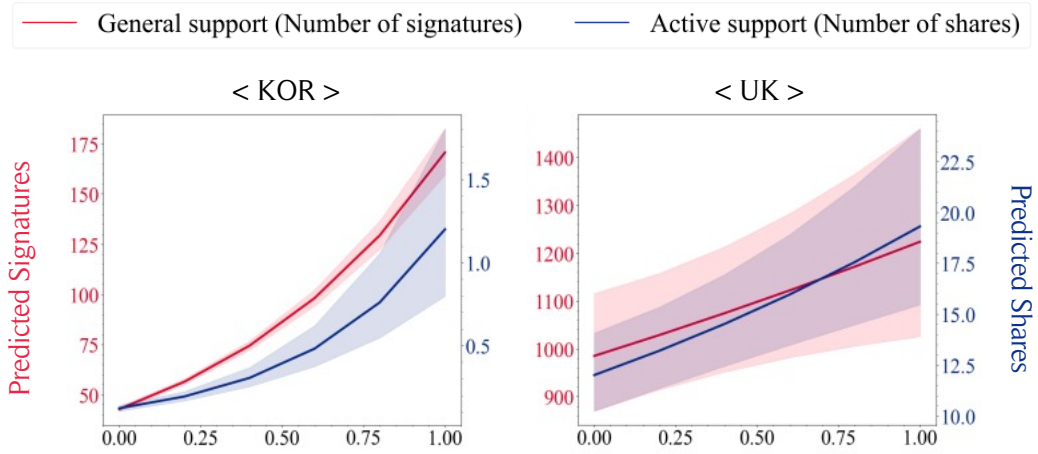
In both country, **positive** impact on

number of signatures (general support) and

number of shares (active support)

05 Implications for Social Science

'Other-suffering' Shapes Political Consensus



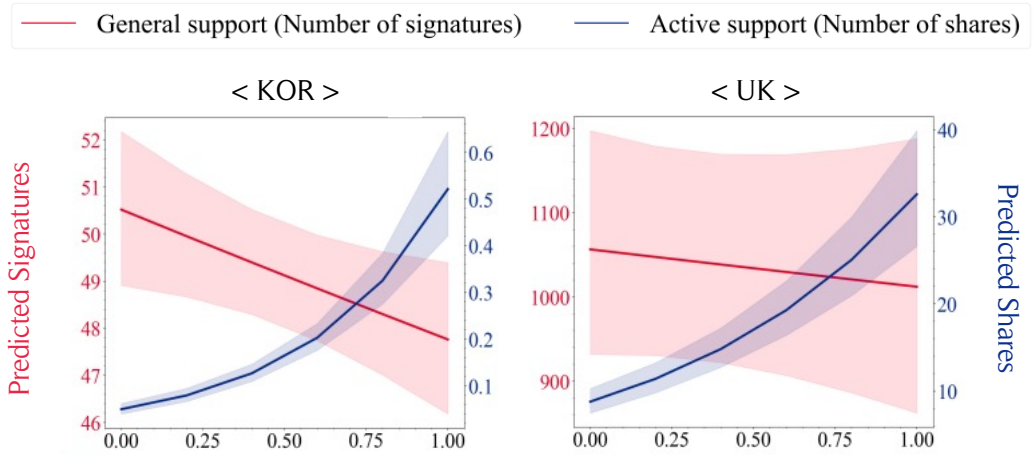
→ In both country, **positive** impact on number of signatures (general support) and number of shares (active support)

The expression of **other-suffering** harmonizes the perspectives of both general and active supporters, fostering widespread consensus on the petition (Sirin et al., 2016) => **Political Consensus**

Sirin, C. V., Villalobos, J. D., & Valentino, N. A. (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.

05 Implications for Social Science

'Other-condemning' Shapes Political Polarization



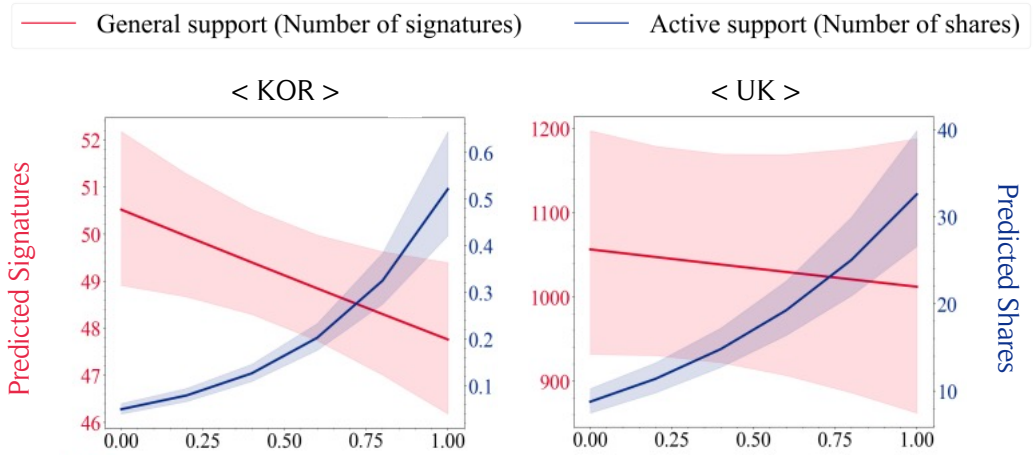
Emotion Score for 'Other-condemning'

→

In both country, **negative** impact on number of signatures (general support) and **positive** impact on number of shares (active support)

05 Implications for Social Science

'Other-condemning' Shapes Political Polarization



Emotion Score for 'Other-condemning'

→ In both country, **negative** impact on number of signatures (general support) and **positive** impact on number of shares (active support)

Other-condemning tends to lower the broader base of general supporters, sharply increase participation among a more dedicated segment (Crockett, 2017; Finkel et al., 2020; Brady et al., 2021) => **Polarization**

Other-condemning messages on social media are highly visible, but **do not reflect general public opinion**, as most supporters feel antipathy towards these messages

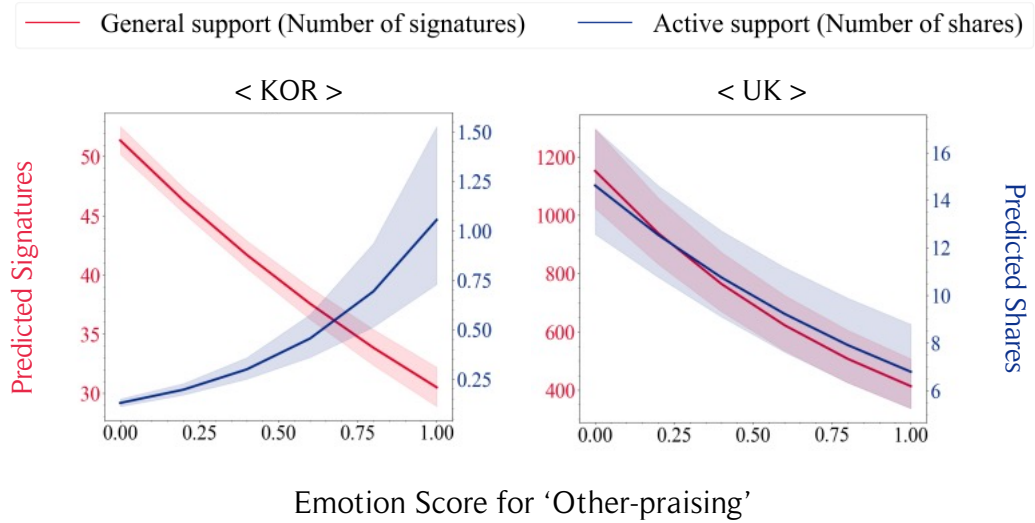
Crockett, M. J. (2017). Moral outrage in the digital age. *Nature human behaviour, 1*(11), 769-771.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., ... & Druckman, J. N. (2020). Political sectarianism in America. *Science, 370*(6516), 533-536.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances, 7*(33), eabe5641.

05 Implications for Social Science

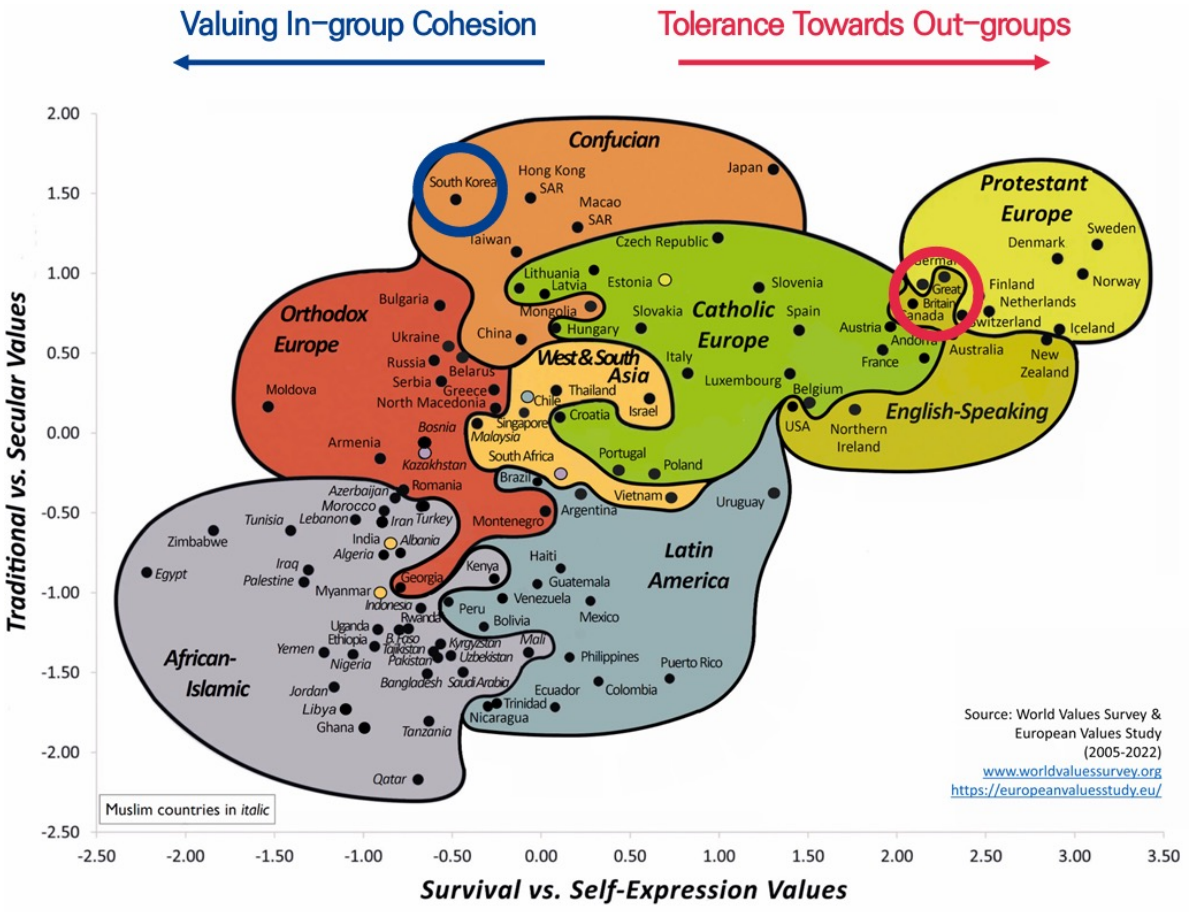
Other-praising's Impact Varies Across Cultures



- In Korea, **negative** impact on number of signatures, **positive** impact on number of shares
- In UK, **negative** impact on number of signatures and number of shares
- Distinct cultural impacts on the role of **other-praising** in political participation

05 Implications for Social Science

Other-praising's Impact Varies Across Cultures

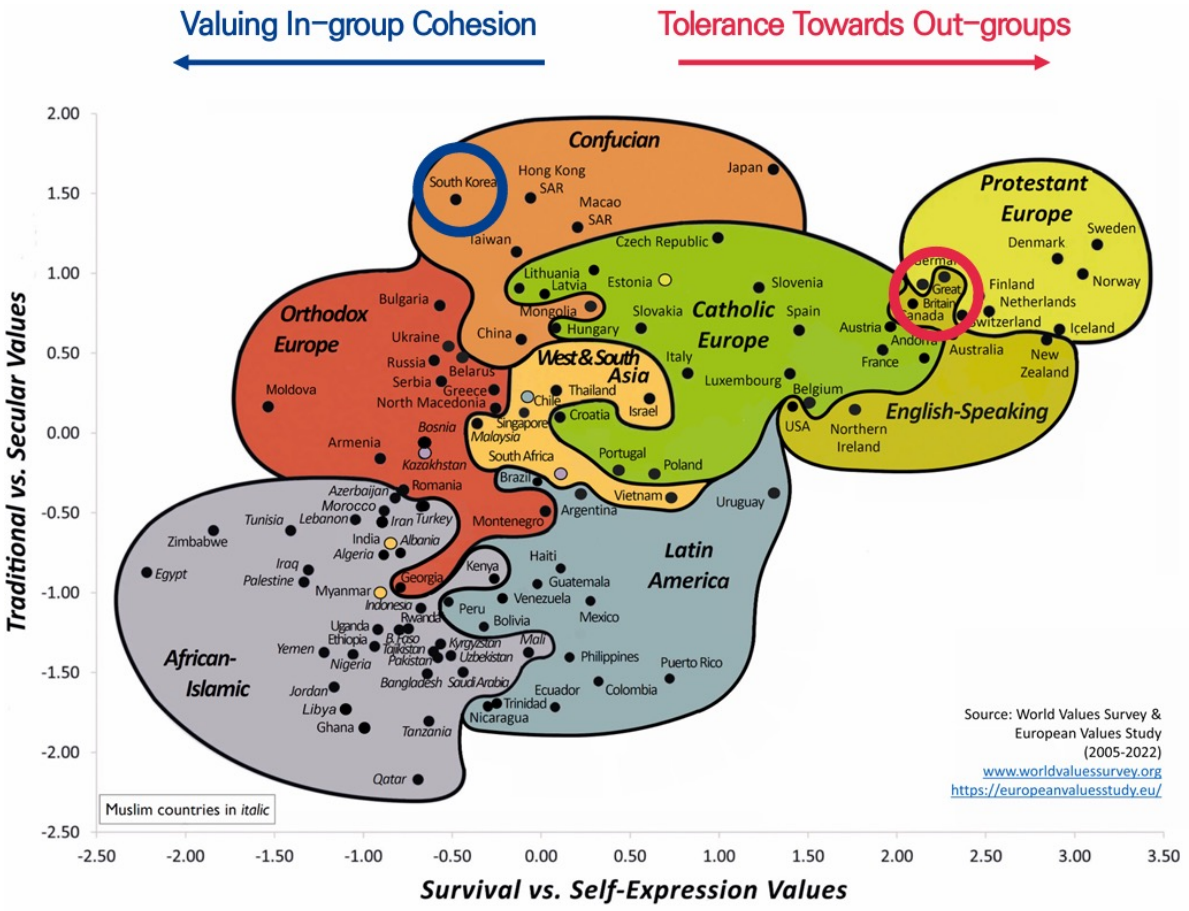


- Korea values in-group cohesion, while UK emphasizes tolerance towards out-groups (Haerpfer et al., 2022)

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... & Puranen, B. (2022). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat, 12*(10), 8.

05 Implications for Social Science

Other-praising's Impact Varies Across Cultures



- **Korea** values in-group cohesion, while **UK** emphasizes 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)

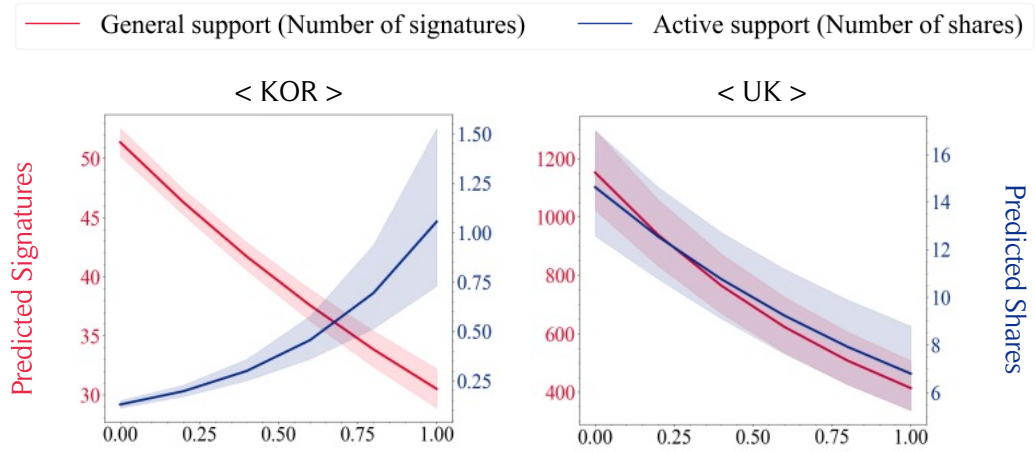
=〉 This theory works well in **Korea**, but not in the **UK**

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (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.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... & Puranen, B. (2022). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat, 12*(10), 8.

05 Implications for Social Science

Other-praising's Impact Varies Across Cultures



- In cultures valuing in-group cohesion like Korea, **other-praising** emotions in petitions can enhance in-group reputation and contributing to polarization
- In cultures emphasizing out-group tolerance like UK, such expressions fail to garner support and provoke antipathy
- These findings reiterate the call for research that considers cultural variation in moral emotions (Haidt, 2003; Van Bavel et al., 2024)

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences, 11*(2003), 852-870.

Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J., & Rathje, S. (2024). Social media and morality. *Annual review of psychology, 75*(1), 311-340.

06 Implications for AI Community

Ethical and Political Implications of Moral Emotions on LLMs

Cost-Effective Analysis of Moral Emotions

- Fine-tuning GPT-3.5: Leveraging AI to inherit human knowledge, reducing annotation costs
- Open-source language modeling: Achieving further cost reductions in inference

06 Implications for AI Community

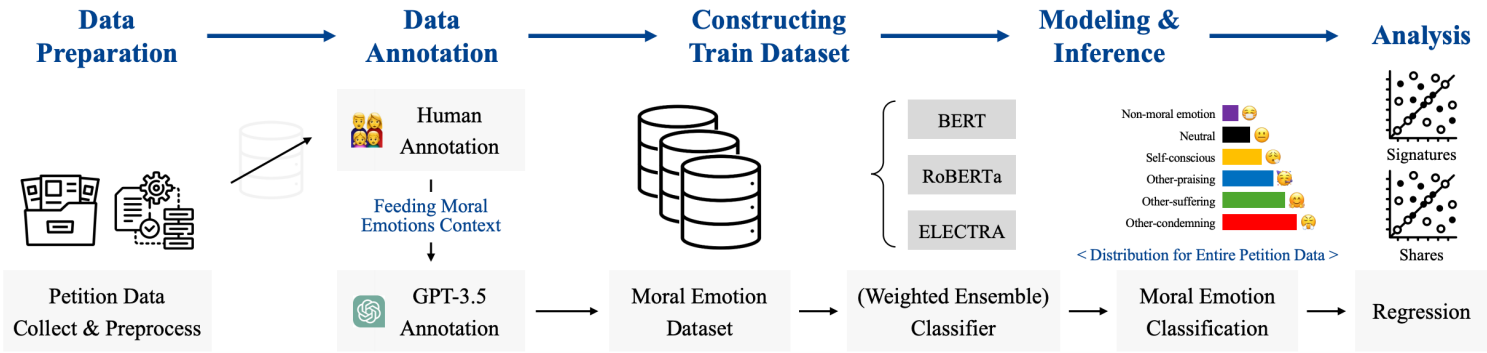
Ethical and Political Implications of Moral Emotions on LLMs

Cost-Effective Analysis of Moral Emotions

- Fine-tuning GPT-3.5: Leveraging AI to inherit human knowledge, reducing annotation costs
- Open-source language modeling: Achieving further cost reductions in inference

Moral Emotion Classifier for Korean and English

- Korean and English texts: Proficient classifier for complex moral emotion identification (especially, Korean)
- Application: Extending methods to less-resourced languages and cost-intensive domains



06 Implications for AI Community

Ethical and Political Implications of Moral Emotions on LLMs

Generative AI Risks (Political Persuasiveness)

- 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

〈 ChatGPT instantly adapted the petition to an **'Other-condemning'** context 〉

Abolish the juvenile law and create a new law that ensures that
young offenders receive proper punishment and realize the wrongs they have committed.

ChatGPT

The juvenile law is far **too lenient on irresponsible youths**. This law must be abolished,
and they should be justly punished to understand the gravity of their offenses.

Thank you ☺

Moral Emotion Training Dataset & Models