Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K): A Parallel Corpus for Explaining (Im)morality of Actions

Jasper Kyle Catapang
De La Salle-College of Saint Benilde
Manila City, Philippines
jasperkyle.catapang@benilde.edu.ph

Moses Visperas
University of the Philippines Diliman
Quezon City, Philippines
moses.visperas@eee.upd.edu.ph

Abstract

Grasping morality is vital in AI systems, particularly as they become more prevalent in human-focused applications. Yet, research is scarce on this topic. This study presents the Emotion-based Morality in Tagalog and English Scenarios (EMoTES-3K), a collection that shows commonsense morality in both Filipino and English. This dataset is instrumental for analyzing moral decisions in various situations and their justifications. Our tests show that EMoTES-3K is effective for moral text categorization, with the fine-tuned RoBERTa model scoring 94.95% accuracy in English and 88.53% in Filipino. The dataset also excels in text generation tasks, as shown by fine-tuning the FLAN-T5 model to produce clear moral explanations. However, the model faces challenges when dealing with actions that have mixed moral implications. This work not only bridges the gap in moral reasoning datasets for languages like Filipino but also sets the stage for future research in commonsense moral reasoning in artificial intelligence.

1 Introduction

Moral reasoning, a cornerstone of human cognition, allows individuals to discern right from wrong and make judgments grounded in ethical considerations. As the integration of artificial intelligence (AI) systems into our daily lives deepens, the imperative for these systems to comprehend and reason about moral dilemmas becomes increasingly pronounced. The challenge lies not just in teaching machines to mimic human moral judgments but in ensuring that these judgments are grounded in a robust understanding of ethical principles. This paper aims to bridge a significant gap in the field: the absence of moral reasoning datasets for low-resource languages such as Filipino. Specifically, we introduce a parallel corpus for commonsense morality—determined heavily by emotions—available in both Filipino and English and analyze its validity by using it in downstream tasks, namely text classification and text generation. We call this corpus the Emotion-based Morality in Tagalog and English Scenarios corpus or EMoTES-3K.

The following are the contributions of the researchers:

1. Introduce a commonsense morality dataset in Filipino and English.
2. Demonstrate the dataset's utility in moral text classification and text generation.
3. Demonstrate the (in)ability of large language models to generalize to tricky scenarios in explaining commonsense morality.

2 Background

2.1 Moral Reasoning Frameworks

Jiang et al. (2021) introduced Delphi, an AI system for commonsense moral reasoning. At its core is the Commonsense Norm Bank with 1.7M crowd-sourced ethical judgments. A notable subset is the ETHICS dataset (Hendrycks et al., 2021), covering diverse moral concepts. Building on this, Pyatkin et al. (2023) presented CLARIFYDELPHI, which emphasizes the importance of context in moral reasoning. In parallel, Zhou et al. (2023) suggested rethinking machine ethics with a top-down approach rooted in established moral theories, aiming for greater transparency in AI decision-making.

2.2 Challenges in Moral Enhancement of AI

Understanding the complexities involved in AI's moral reasoning leads us to explore the associated challenges. Serafimova (2020) discussed the differences between moral agency in humans and AI, highlighting the issue of replicating human moral autonomy in machines. The study also underlined the risks of biases in algorithm design, which can lead to computational and moral errors, affecting AI's moral outcomes and raising significant socio-political concerns.

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 1–6
December 1-3, 2023 ©2023 Association for Computational Linguistics

2.3 Human Values and AI Alignment

The relationship between human values and AI decision-making is further elucidated by Sorensen et al. (2023). They introduced VALUE PRISM, a dataset capturing human values in authored situations, and KALEIDO, a model adept at generating and assessing the relevance of these values. Additionally, Yao et al. (2023) focused on aligning Large Language Models (LLMs) with human values, highlighting the evolution of LLMs from basic capabilities to a deep value orientation. Complementing this, Schramowski et al. (2022) showed that LLMs can represent moral norms geometrically and learn moral biases, suggesting their potential in answering moral questions.

2.4 Emphasis on Moral Judgment in Tagalog and Taglish

Transitioning to the linguistic aspect of moral reasoning, our research uniquely focuses on moral judgments in Tagalog and Taglish. These languages play a significant role in the global linguistic landscape, with Tagalog offering a distinctive cultural and moral framework and Taglish providing a blend of local and global moral perspectives. This dual focus allows us to analyze moral reasoning in two linguistically and culturally intertwined environments, highlighting the dynamic interplay of language and culture in moral reasoning.

2.5 Overview of Filipino Morality

To further understand the cultural context of our research, we delve into Filipino morality. Rooted in its cultural, historical, and sociological fabric, Filipino morality integrates indigenous values, colonial influences, and modern global perspectives (Jocano, 1997; Mercado, 1974). The concept of ’kapwa’ or shared identity is central to this ethos, emphasizing community and empathy (Enriquez, 2013). The influence of Catholicism and the Philippines’ colonial history have shaped a resilient and adaptable moral framework (Constantino, 2022; Doeppers, 2016), which is crucial for developing culturally sensitive AI systems.

2.6 Morphological Challenge in Natural Language Understanding of Filipino

Finally, we address a specific linguistic challenge in the Filipino language. The complex morphology of Filipino verbs (De Guzman, 1978) poses a unique challenge for AI-based moral judgment. Unlike English, with its simpler verb structure, Filipino verbs undergo significant morphological changes that affect moral implications. For example, the verb ’gawa’ (to do) in its root form is neutral, but when transformed into ’magagawa’ (can do), it implies capability or potential, introducing moral considerations like responsibility and choice. Conversely, ’nagawa’ (did) indicates completed action, shifting the focus to accountability for actions taken. Beyond verb forms, Filipino’s reliance on context sensitivity, non-verbal cues, and indirect communication style further complicates AI’s interpretation of moral nuances. The prevalent use of Taglish, blending Tagalog and English, adds another layer of complexity, reflecting cultural intermingling but posing challenges for AI models trained on monolingual datasets. Specialized algorithms are required to navigate these nuances, understanding the subtle shifts in meaning and cultural implications inherent in Filipino verb forms and communication styles.

3 Experimental Setup

3.1 Dataset Creation

In developing the EMoTES dataset, we meticulously followed an annotation process inspired by the methodology used in Hendrycks et al. (2021). Our team of annotators consisted of bilingual Filipino college graduates with specialized backgrounds in philosophy, psychology, and linguistics. This diverse academic expertise was crucial in ensuring a deep and accurate interpretation of the dataset.

Adopting a quality control approach parallel to that of Hendrycks et al. (2021), each entry in our dataset was subjected to multiple reviews. This rigorous process aimed to ensure consistency in annotations and minimize ambiguities. However, it is essential to acknowledge a potential limitation: all our annotators were from Metro Manila. This geographical concentration may introduce a cultural bias, potentially limiting the representation of diverse provincial values and norms in the Philippines. Therefore, we advise caution when applying our findings to broader, more culturally varied contexts.

Building upon the foundational work of Hendrycks et al. (2021), our research included the creation of approximately 2,400 original scenarios where commonsense moral judgments determine morality. Additionally, from the work of Hendrycks et al. (2021), we adapted and translated about 500 examples—providing not only classifications but also explanations and inferred personality traits from each scenario. In total, the EMoTES-3K dataset comprises 1,712 moral scenarios and 1,193 immoral scenarios.

To illustrate, consider this example from the dataset:

> `Filipino:` "Si Sofia ay nagbebenta ng pekeng COVID-19 test results upang makalusot sa mga travel restrictions."
> `English:` "Sofia is selling fake COVID-19 test results to bypass travel restrictions."
> `Annotation:` Immoral
> `Reason:` "Sofia's action of selling fake COVID-19 test results to evade travel restrictions is highly immoral as it compromises public safety and deceives authorities."
> `Personality Traits:` "deceptive"

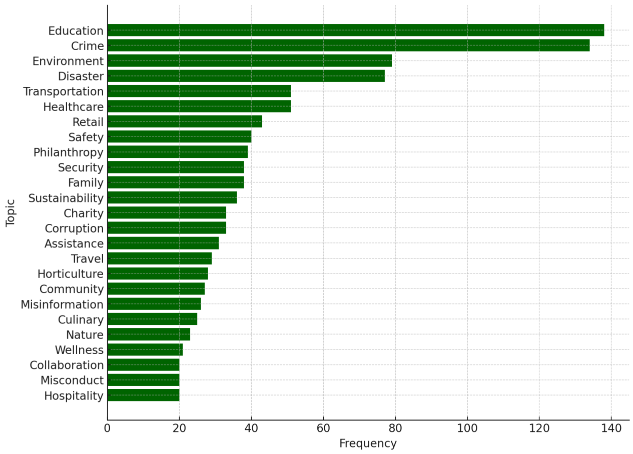

Figure 1: Top 25 topics in EMoTES-3K

The EMoTES-3K dataset explores a diverse array of topics, reflecting a wide spectrum of emotional and thematic elements in textual data. As shown in Figure 1, 'Education' and 'Crime' are the most prominent topics, each with over 130 instances, signifying a strong focus on these societal aspects. Other significant topics include 'Environment', 'Disaster', and 'Transportation', collectively addressing global challenges and daily life concerns. The inclusion of topics such as 'Healthcare', 'Retail', and 'Safety' highlights the dataset's relevance to public welfare and economic activities. Furthermore, the presence of subjects like 'Horticulture', 'Misinformation', and 'Hospitality' provides insights into specialized areas. This assortment of topics in the EMoTES-3K dataset ensures comprehensive coverage of societal themes and underscores its utility in analyzing the complex interplay between topic content and emotional expression. This variety is vital for the development of robust natural language processing tools capable of discerning the nuanced relationships between topics and emotional expressions.

Each scenario within the dataset has an average word count of 12.65 words for English and 15.41 words for Filipino. The distribution of the most common words for each language scenario is illustrated in Figure 2.

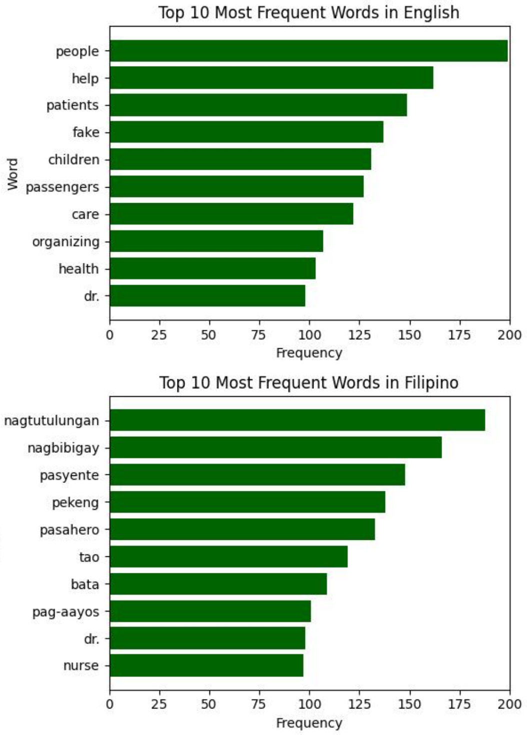

Figure 2: Common words in EMoTES-3K per language

To explain the structure clearly, the data fields of EMoTES-3K are described in Table 1.

Field	Description
entry_id	unique identifier
Filipino	scenario in Filipino
English	scenario in English
Annotation	Moral/Immoral
Explanation	why action is moral or immoral
Personality Traits	inferred traits from action

Table 1: Description of the proposed dataset fields

3.2 Moral Text Classification

One possible application of the EMoTES-3K dataset is the classification of a text's commonsense morality. To show how the corpus can be used for such a task, we use the RoBERTa architecture (Liu et al., 2019). Assuming Schramowski et al. (2022)'s findings hold, language models, like RoBERTa, should yield favorable results. For both language subsets of EMoTES-3K, we use Google Colab's free-tier GPU T4 runtime to fine-tune RoBERTa. No text preprocessing is made.

In the training phase for the RoBERTa model on the English dataset, several hyperparameters are meticulously chosen to ensure optimal performance. The batch size for both training and evaluation is set to 64. The model is trained for a total of 30 epochs. A learning rate of 1×10^{-6} is employed, accompanied by a weight decay of 0.005 to prevent overfitting. The evaluation strategy is configured to evaluate at regular step intervals, specifically every 100 steps, which is also the frequency at which the model's performance metrics were logged. The metrics in consideration are training loss, evaluation loss, accuracy, and F1 score.

3.3 Text Generation

Using EMoTES-3K, we fine-tune the FLAN-T5 large model (Chung et al., 2022) for text generation in both English and Filipino subsets. The prefix "*Explain the morality of this scenario*" was appended to each scenario in the dataset, following FLAN-T5's instruction style training. The model is trained over 30 epochs with a learning rate of 1×10^{-4}, using a linear learning rate scheduler. We set batch sizes at 2 for both training and evaluation and utilize a 32GB V100 GPU from DOST-ASTI's COARE for training[1]. This fine-tuning adapts FLAN-T5 to the EMoTES-3K dataset's specifics, optimizing its performance for subsequent tasks.

The evaluation metrics for this task are ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR. While ROUGE (Lin, 2004) and METEOR(Banerjee and Lavie, 2005) are primarily designed for assessing the quality of machine-generated text in the context of summarization and machine translation, we employ them as a proxy metric because our primary aim is to ensure that our model's outputs align with the desired standard of coherence and relevance to our gold explanations.

4 Results and Discussion

4.1 RoBERTa Sequence Classification

code	train loss	val loss	acc	f1
en	0.076	0.154	0.950	0.959
tl	0.240	0.251	0.885	0.907

Table 2: Training and validation results of RoBERTa on the English and Filipino scenarios at different steps.

RoBERTa model's training and validation results for both English and Filipino datasets are detailed in Table 2, respectively. Two separate models were fine-tuned [2]. For the English dataset, the model's peak performance was at step 800 with an accuracy of 94.95% and an F1 score of 0.9586. In contrast, the Filipino dataset saw its best results at step 900, achieving an accuracy of 88.53% and an F1 score of 0.9070.

The model's stronger performance in English can be attributed to its inherent design for the English language. Additionally, the Filipino language's complexity, marked by its diverse verb inflections, poses challenges in discerning moral implications, making it a more intricate task than in English.

4.2 FLAN-T5 Text Generation

4.2.1 Inherently (Im)moral Scenarios

The training and validation loss of two fine-tuned FLAN-T5 large models are shown in Table 3. After 30 epochs, the English model achieved a training loss of 0.0001 and a validation loss of 0.0546, while the Filipino model reached a training loss of 0.0001 and a validation loss of 0.0711.

Metric	English	Filipino
Training Loss	0.0001	0.0546
Validation Loss	0.0001	0.0711
ROUGE-1	0.7601	0.7263
ROUGE-2	0.7084	0.6562
ROUGE-L	0.7513	0.7146
ROUGE-Lsum	0.7515	0.7142
METEOR	0.8677	0.8249

Table 3: Performance metrics of fine-tuning FLAN-T5 on the text generation task via EMoTES-3K dataset.

Table 3 displays the ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and METEOR scores for both the English and Filipino datasets. The

[1] DOST-ASTI COARE's website: https://asti.dost.gov.ph/projects/coare/

[2] Data and fine-tuned models can be found here

ROUGE scores represent the count of matching n-grams, including unigrams, bigrams, and the longest common subsequence, between output and its corresponding reference. Meanwhile, the METEOR metric is based on the precision and recall of matching unigrams. It is evident that our fine-tuned models exhibit exceptional text generation capabilities, as most of our ROUGE metrics achieve scores of 0.70 or higher. The only exception is ROUGE-2 for Filipino. The METEOR scores were also higher for English than Filipino with a score of 0.8677 compared to 0.8249. In general, when we compare text generation capabilities between English and Filipino, we observe that Filipino falls short, likely due to the linguistic intricacies associated with this language, as mentioned earlier.

4.2.2 Scenarios with Mixed Moral Implications

To demonstrate how much the fine-tuned FLAN-T5 understands commonsense morality, we qualitatively analyze two scenarios wherein a moral action is done with ill intent (MB) and two other scenarios where an immoral action is done with good intent (IG). The examples are as follows:

1. `MB`: “Sarah helps her younger brother with his homework because she wants to show her parents that she’s the better sibling.”
2. `MB`: “Si John ay nagbigay ng malaking pera sa isang charity para matabunan ang masasamang balita sa kanya.”
   (TRANSLATED): “John gave a huge sum of money to a charity to get rid of the bad news about him.”
3. `IG`: “Mike doesn’t share his dad’s terminal condition with his mom because she will have a heart attack when she receives the news.”
4. `IG`: “Ninakaw ni Laura ang susi sa kotse ng kanyang lolang may malabong mga mata upang hindi magamit ng lola niya ang kotse.”
   (TRANSLATED): “Laura stole the car keys of her grandma with poor eyesight so that her grandma can’t use the car.”

The fine-tuned FLAN-T5 models gave the following responses:

1. “This action is considered moral because helping one’s younger sibling with their homework is supportive and educational.”
2. “This action is considered **moral** because donating money to a charity with the intention of embezzling money from it is a highly **immoral** and illegal act.”
3. “Mike’s action of not sharing his dad’s terminal condition to his mom is immoral as it puts her life in immediate danger and puts her health at risk.”
4. “Laura’s action of stealing her grandma’s car **seat** is immoral as it is disrespectful and inconveniences her family members.”

In Example 1, it can be inferred that the model only focused on the action itself and excluded the reason, so the model considered the person’s action as moral. In Example 2, it can be inferred that the model is confused since it was inconsistent in determining whether the action is moral or immoral. In Example 3, the model does not seem to know why the action is immoral, i.e., the judgment is correct but the reasoning is not. Lastly, in Example 4, the model changed the keys to a seat but chose to consider the act of stealing immoral as it is disrespectful.

5 Conclusion

This research introduced the corpus: Emotion-based Morality in Tagalog and English Scenarios or EMoTES-3K, providing a moral reasoning dataset for the Filipino language. The dataset was validated through text classification and text generation tasks. Notably, the RoBERTa model achieved an accuracy of 94.95% for English and 88.53% for Filipino in moral text classification. Furthermore, the fine-tuned FLAN-T5 models showcased impressive text generation capabilities, with most ROUGE metrics surpassing 0.70. However, the models exhibited challenges in discerning complex moral scenarios, especially those with conflicting intents. This study not only fills a gap in moral reasoning datasets for languages like Filipino but also highlights the intricacies and challenges of teaching AI systems to reason about morality.

Acknowledgements

We extend our gratitude to the Advanced Science and Technology Institute (ASTI) and the Computing and Archiving Research Environment (COARE) for providing our team with access to their High-Performance Computing (HPC) resources. We are also thankful to the manuscript reviewers and dataset annotators, whose insights and efforts were essential in elevating the quality of our research.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Renato Constantino. 2022. The philippines: A past revisited.

Videa P De Guzman. 1978. Syntactic derivation of tagalog verbs. *Oceanic linguistics special publications*, (16):i–414.

Daniel F Doeppers. 2016. *Feeding Manila in peace and war, 1850–1945*. University of Wisconsin Pres.

Virgilio Enriquez. 2013. From colonial to liberation psychology: The philippine experience. *Philosophy East and West*, 63(2).

Dan Hendrycks, Andrew Critch, Collin Burns, Jerry Li, Dawn Song, Steven Basart, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *OpenReview*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint*.

F Landa Jocano. 1997. Filipino value system: a cultureal definition. *(No Title)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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*.

Leonardo N Mercado. 1974. Elements of filipino philosophy.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *ACL Anthology*.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain humanlike biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Silviya Serafimova. 2020. Whose morality? which rationality? challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7:119.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *arXiv preprint*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values — a survey of alignment goals for big models. *arXiv preprint*.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories? *arXiv preprint*.