

Large Language Models for Morality Classification in Texts

Nayeli Hernández-Ramírez, Grigori Sidorov, Obdulia Pichardo-Lagunas*, Ildar Batyrshin

Instituto Politécnico Nacional, Centro de Investigación en Computación,
Mexico

sidorov@cic.ipn.mx, opichardola@ipn.mx, batyr1@cic.ipn.mx

Abstract. Moral values have an important role in shaping the human behavior, influencing how we should act or react in any situation. However, we cannot identify someone's moral sentiment only by observing them; we have to infer from their behavior, and these inferences are often unreliable. Human beings use morally relevant language. In other words, we use words or phrases that can be linked to values and through which our conduct or way of acting can be conveyed. These words can be viewed as true or false signals, but, in any case, expressions of moral sentiment serve as or at least as indicators of the values one wishes to display. One way to compensate for the inaccuracy of this information is to express our moral values through language. Based on this observation, this work intends to illustrate how morality may be inferred from text by evaluating six large language models (LLMs) publicly available for moral classification and comparing accuracy using different machine learning methods (ML) as baselines. We introduce a novel corpus of text in Spanish language labeled with several moral concerns like Care, Fairness, Loyalty, Authority, and Fairness based on the Moral Foundation Theory (MFT). We collected opinions about death penalty in Mexico from native Spanish speakers along with their responses on the Moral Foundations Questionnaire (MFQ). We present evaluation of the corpus using Deep learning.

Keywords. Morality, moral foundation theory, natural language processing, text analysis.

1 Introduction

The number of users who use social media to express and/or debate situations through language is massive. Many of these conversations revolve around describing and judging the behaviors/beliefs of others, which we consider good or bad [1]. The tendency of people

to use social media to express moral values through language becomes a valuable research opportunity for scientists interested in analyzing feelings, emotions, and human behavior. Moral classification/analysis can also be a powerful tool for understanding social and cultural trends present in a country or region.

Human beings use morally relevant language; we use words or phrases that can be related to values that can convey behavior or ways of acting. Some texts published on social media describe moral situations in which personal traits, feelings, emotions, and even moral values can be identified. Thanks to these texts, we can define the stance taken on controversial social issues such as demographic characteristics, mental health, politics, and even climate change. For example, in the field of psychology, Schwartz et al. [2] present a tool that describes the similarities and differences between groups of people (belonging to the same region) through language. Their technique is based on what people write on social media to find words and phrases that determine certain information, such as gender, age, location, or psychological characteristics

In the field of medical opinion, Borghouts et al. [3] conducted a study on public sentiment among people from a specific state in the USA regarding the COVID-19 vaccine on social media. This research revealed a trend towards extreme ideologies among liberals and conservatives. Furthermore, it observed that the issue of vaccines is associated with political concepts, such as "science" and "death."

In politics, Atari et al. [4] using Moral Foundations Theory and the Moral Foundations Questionnaire as a basis for constructing their own moral classes focused on Iran. They employ

Table 1. Distribution of the MFRC

Subreddit	Care	Equality	Proportionality	Loyalty	Authority	Purity
AmltheAsshole	343	145	103	56	50	38
Conservative (French)	0	12	5	6	7	16
Conservative (US)	0	195	200	84	38	155
antiwork	304	132	202	46	86	29
confession	281	69	101	24	52	50
Europe	105	180	117	107	174	18
geopolitics	1	4	5	1	5	0
neoliberal	39	74	67	50	117	11
nostalgia	20	11	15	5	8	9
politics	148	119	72	61	161	34
relationship advise	0	418	137	84	98	31
worldnews	166	249	127	105	192	33
All	2032	1325	983	598	1047	323

the Moral Foundations Questionnaire and compare moral foundations between Iran and the United States.

They conduct interviews about moral values, and the findings suggest different moral classes for Iranian culture.

Mooijman et al. [5] conduct scientific research to understand the causes of violent clashes between protesters and police in the United States and study the increase in the number of violent protests occurring in the country.

Specifically, in this article we investigate the relevant language used in moral issues, which are present in social media, and the type of information that can be inferred from them by using a wide range of supervised machine learning and deep learning methods.

We label opinions at 5 moral levels such as Care, Fairness, Loyalty, Authority and Purity, which are conceptualized by Moral Foundations Theory [6]. We also use the Moral Foundations Questionnaire as a measure to obtain the labels [7].

In this research work, we rely on Moral Foundations Theory (MFT) to deal with morality in texts.

This theory proposes that all existing moral values can be grouped into five basic positive and

negative classes, which can represent morally bad or good actions:

- Care/Harm: This foundation relates to emotional, the capacity to feel or hate pain in others.
- Fairness/Betrayal: This foundation relates to altruistic behavior.
- Loyalty/Betrayal: This foundation relates to the sense of permanence and constant support for a person, group, or ideal.
- Authority/Disobedience: This foundation was formed from hierarchical social.
- Purity/Degradation: This foundation was shaped by the emotion of disgust or the feeling of contamination, religious beliefs.

The work presented in [8] introduces the Moral Foundations Questionnaire (MFQ), a tool for estimating scores associated with each of an individual's five moral foundations.

The MFQ is composed of 32 multiple-choice questions/statements rated on a 6-point scale, this questionnaire has two types of questions: the individual's criteria for determining right and

wrong, and their level of agreement with specific moral statements.

Responses to the MFQ are combined as proposed in [9] to yield five numerical scores representing the individual's moral foundations.

Based on these observations, this work introduces several experiments using supervised machine learning to illustrate how morality may be inferred from text. Artificial Intelligence AI systems were prompted to evaluate the tasks of classifying moral values without any training, a new corpus created to test our experiments, and results that will be taken as a reference for future studies in the field.

The main contributions in this work are summarized as follows:

1. Experiments with different large language models applied to morality classification.
2. Novel labeled corpus with fine-grained features of participants and other demographic data, in addition to being aligned with moral foundation theory.
3. Comparative analysis of feature extraction methods.
4. Comparative analysis of deep learning and machine learning methods applied to the classification of moral values from text.

2 Data

Our studies are based on the Moral Foundations Reddit Corpus (MFRC) which is manually annotated for their moral content according to the MFT. It is a collection of 16,123 English Reddit comments recollected from 12 distinct subgroups and organized into three buckets: US politics, French politics, and everyday life (see Table 1). At least three trained annotators label the posts in eight categories of moral sentiment (e.g., Care, Proportionality, Equality, Purity, Authority, Loyalty).

The Reddit Corpus creators followed a methodology similar to Hoover et al. [10] for training annotators. We use only the five positive foundations listed on the MFT. Most existing studies based on moral value classification tend to use English text corpora labeled with information that serves as classes for supervised

machine learning methods, which require manual labeling. However, there are currently no tasks for classifying moral values in Spanish texts.

We introduce our labeled corpus in which we collected the opinion of 435 participants (mostly of them university students) based on only one topic. Participants who agreed with the terms of the research were requested to complete an on-line anonymous survey document about "Death penalty in México", the survey ask participants certain personal details like age, gender, education, etc. and moral scores representing their opinions. This was followed by the 30-item multiple choice MFQ questionnaire [11]. Sentiment scores were presented from questions as "Should Mexico introduce the death penalty?". For each question, participants were requested to choose from 0 (Strongly disagree) to 5 (Strongly agree) and also we ask participants to justify their moral selection by writing a short paragraph (at least 50 characters).

All participants were volunteers, native speakers of Spanish from Mexico City, Mexico, mostly male (65.2%) and, on average equals to 20 years-old (min=18, max=33). Mostly of them (98%) had studies on Information Technology (IT). Other parameters vary in education (from university studies to PhD), religiosity (from non-religious to religious) and political orientation (from left, neutral and right).

The MFQ questionnaire scores the moral foundation by averaging the item responses for each of the five moral subscales on a 0 to 5 scale. Each survey was averaged based on the MFQ score and the result were coded as the moral value label code from 1 to 5.

After that, each survey was examined by a psychologist. The doctor analyzes the 30 questions and the key words in the short paragraph with the opinions and diagnosed the presence of one of the five moral value presented in the MFT. The diagnosis results were coded as label moral value with an additional indication of the diagnosis.

3 Morality Classification

In this paper we propose a model, which consists of three phases working with the MFRC (Fig 1).

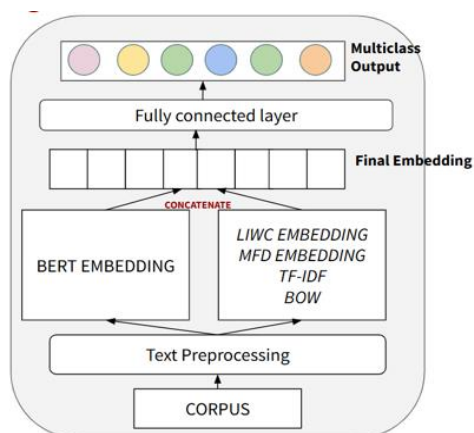


Fig. 1. Proposed model

The first phase is the preprocessing. Several operations were performed on the corpus, such as cleaning or replacement of special characters, numbers, and punctuation marks; removal of URLs; removal of stop words; tokenization, stemming, and lemmatization.

Then feature extraction phase goes. Several methods were used to obtain two final word vectors. For instance, in a multi-class classification task, labels can contain important information about the texts. However, this information is not captured by embeddings derived from a pre-trained model like BERT, as that is an unsupervised method. Therefore, we combined a model like BERT with a supervised method during the feature extraction phase. Examples of such methods include the use of dictionaries (LIWC [12] and MFD [13]) or TF-IDF. The feature extraction model involves two components: word vectors based on the pre-trained BERT model and feature generation via a supervised method, resulting in the fusion of two feature vectors for multi-label classification.

Finally, the sentiment classification phase goes. The resulting feature vector of the concatenation of the two selected feature extraction will be processed in a fully connected layer using a softmax activation function for multi-class classification. The metrics used to evaluate the proposed model are recall, accuracy, and F1 score.

Working with the proposed database, we select a training data set of 348 records that were randomly selected from the complete database,

maintaining the division of the database in the ratio of 80:20. We have to work with a limited size for the test sample (equals to 98 records), that was determined by the limit use of some of publicly available AI systems, which cannot accept queries that are too long, and when a query length exceeds 100 records, system failures occur. The training data set was prompted to several publicly available AI systems without any training. The verification was executed based on the comparison with the test set, which was obtained from the survey results. The diagnosis made by the psychologist was also encoded as a number in the training set and was absent from the test set. Based on the AI systems predictions or responses of the surveys, we were able to make comparisons with results of moral values obtained by the averaging of the MFT, allowing us to judge on what basis the AI systems has at the predicted value.

We select 6 free access AI systems: Grok-3 (reasoning) [14], Gemini 2.0 (reasoning) [15], Gemini 2.5 Pro (thinking) [16], Gemma 3 27B [17], DeepSeek-R1 (reasoning) [18] and GPT o3-mini (reasoning) [19]. We feed the AI systems with a specific prompt. The responses (scale 0 to 5) of the 30 moral foundation questionnaire were provided as a list of considerations and the label, for example:

```
Hi! I have a list
{{considerations={list},MoralValue} of
97
entries:{MoralData{4,3,2,3,5,2,3,2,4,3,
4,5,1,3,3,5,4,5,3,4,5,5,5,5,4,1,4,5,4,3
,1,4},{MoralValue=5}}, {MoralData{3,2,0,
1,1,0,3,1,2,0,3,4,3,0,1,0,5,5,0,4,5,5,5
,5,2,0,0,4,2,1,2,0},{MoralValue=1}}, {Mo
ralData{5,4,3,5,5,1,4,5,4,3,5,5,5,4,4,2
,2,5,4,5,1,2,5,5,3,0,2,2,0,4,4,0},{Mora
lValue=2}}, {MoralData{2,4,0,5,1,0,3,2,2
,5,5,4,3,1,5,0,4,3,0,4,5,5,3,3,2,0,4,3,
4,1,1,0},{MoralValue=4}} (97 additional
records placed here) predict the Moral
Values for these list of 30
entries:MoralData{2,4,0,5,1,0,3,2,2,5,5
,4,3,1,5,0,4,3,0,4,5,5,3,3,2,0,4,3,4,1,
1,0}.
```

Statistical methods were performed in the same database as baseline methods in order to

Table 2. Results of the moral predictions using MFRC and the proposed model

Metrics	Models with 5-fold cross-validation					
	TF-IDF + SVM	TF-IDF + RF	LIWC + SVM	GLOVE + BiLSTM	BERT	LIWC + BERT
Accuracy, %	44.9	43.7	57.3	68.8	71.0	75.2
Precision, %	38.2	32.4	37.2	77.4	82.1	80.1
Recall, %	31.5	21.5	53.2	78.4	75.2	85.0
F1-score, %	41.7	40.0	43.8	77.3	78.1	82.2

Table 3. Results of the moral predictions with several AI systems using our corpus

Metrics	AI System					
	Gemini 2.0	Grok-3	Gemini 2.5	GPT o3	Gemma 3 27b	DeepSeek R1
Accuracy, %	78.7	63.8	75.1	68.8	71.0	66.2
Precision, %	68.2	44.3	58.8	51.3	55.0	37.2
Recall, %	50.0	50.8	66.7	46.6	73.3	53.3
F1-score, %	57.7	46.8	62.5	47.1	62.9	43.8
5-fold cross-validation						
Accuracy, %	73.0	57.3	70.1	36.0	67.0	57.0
Precision, %	81.1	37.2	53.4	30.2	50.3	37.2
Recall, %	83.0	53.3	52.3	40.5	45.6	53.3
F1-score, %	81.6	43.8	52.6	35.0	45.1	43.8

determine what maximum accuracy can be expected.

The baselines selected methods were: K-means clustering (KM), support vector machines (SV), decision tree (DT), random forest (RF), and gradient boosting (GB).

Training without cross-validation and with cross-validation (5-fold cross-validation) were considered for all the methods.

The selection of the optimum Hyper-parameters was obtained by doing an exhaustive experimentation.

For SVM the best hyper-parameters were $\{C: 100, 'gamma': 0.001\}$, for RF best hyper-parameters: $\{max_depth: 6, 'n_estimators': 20\}$, for KN best hyper-parameters are $\{metric: 'euclidean', 'n_neighbors': 7\}$, for GB best hyper-parameters are $\{learning_rate: 0.1, 'max_depth': 3\}$ and DT best parameters are $\{max_depth: 6, 'min_samples_leaf': 8, 'min_samples_split': 2\}$.

4 Results

In table 2 we present the results of several models. Specifically, we used dictionaries such as MFD and LIWC to generate initial predictions for each moral value category; these dictionaries provide moral valence scores and indicate lemma intensity.

We utilized MFT scores to categorize values discretely and align them with the proposed model's output.

Additionally, we employed machine learning techniques specifically Random Forest (RF) and SVM classifiers to predict each moral category.

We also used TF-IDF to generate word vectors; however, when combined with machine learning methods like RF and SVM, these yielded results inferior to those obtained via dictionaries.

Experiments were also conducted using GloVe, a widely used model in natural language

Table 4. Results of the moral predictions using several machine learning methods with our corpus

Metrics	Machine Learning Algorithm				
	Support Vector Machine	Random Forest	K Nearest Neighbor	Gradient Boosting	Decision Trees
Accuracy, %	68.8	64.1	64.1	71.0	63.8
Precision, %	51.3	46.1	46.1	67.9	50.0
Recall, %	46.6	46.1	46.1	63.3	63.3
F1-score, %	47.1	46.1	46.1	65.5	55.9
5-fold cross-validation					
Accuracy, %	63.0	61.2	66.0	66.0	52.0
Precision, %	55	50	60	64	47
Recall, %	50	67	75	65	47
F1-score, %	52	57	67	68	54

processing known for its word embedding capabilities. GloVe was designed to handle large datasets and demonstrates high performance in terms of both accuracy and computational efficiency. However, it is worth noting that GloVe embeddings lack context dependency, making them less suitable for analyzing sentences at a broader level.

We analyzed the results of the proposed model and compared its accuracy and efficiency. Initially, we observed that using dictionaries labeled by individuals trained in Moral Foundations Theory yielded slightly better results than frequency-based word vector methods; specifically, the LIWC + SVM approach outperformed TF-IDF + SVM by 21%, while results were similar for RF combined with the MFD dictionary.

The GloVe + BiLSTM combination yielded a result 12% lower than LIWC + RF; this is because GloVe relies heavily on training across large, diverse datasets that may not be focused on moral values. Finally, we employed a pre-trained BERT model combined with the MFD dictionary, achieving a 12% improvement over the best-performing baseline model, LIWC + RF.

The predictions result of the six surveyed AI systems are presented in Table 3. The table 3 shows the corresponding results achieved by a specialist using five different methods of machine learning.

With the exception of GPT-o3, all evaluated AI systems achieved acceptable diagnostic accuracy on professional questionnaires, ranging between 60% and 70%. When assessed through cross-validation, however, accuracy estimates dropped to 50–60%.

Following the 5 cross-validation, Gemma 3 27B (achieving 71%), Gemini 2.0 (achieving 78%), and Gemini 2.5 (achieving 75%) achieved the highest performance metrics among the evaluated systems, whereas GPT-o3 recorded the lowest (achieving 68.8%).

When subject to cross-validated estimates, top performers included Gemma 3 (achieving 67%), Gemini 2.0 (achieving 73%), and Gemini 2.5 (achieving 70%), while GPT-o3 (achieving 36.0%), DeepSeek R1 (57%), and Grok-3 (achieving 57.3%) registered the lowest scores.

These performance scores varied significantly between multiple test runs and it depends at run time. For example, Grok-3 dipped to 62.5% without cross-validation in some instances, despite achieving a much higher top score overall.

While none of the AI models matched the benchmark set by a human specialist using machine learning, their top scores fell only slightly short of the human's best results in Table 4.

Among these techniques, Support Vector Machines achieved the highest accuracy at 63.3%, while the remaining methods scored

between 50% and 60%—matching or trailing the performance of several AI models.

Notably, multi-moral classification remains a challenging task for traditional machine learning approaches.

Under cross-validation, the k-Nearest Neighbors and Gradient Boosting methods achieve the strongest performance, each reaching 66% accuracy.

Other techniques maintained comparable results, with even the lowest-performing models (Decision Trees and Random Forests) slightly outperforming the lowest-tier AI systems.

Gemma 3 sought combinations of similar moral values across each list and, based on these similarities, calculated the most probable moral value outcome.

Gemini 2.5 Pro selected a moral value by identifying combinations of MFQ responses similar to those in the examples; Gemini 2.0 analyzed the MFQ opinion vector and, based on patterns found in the provided data, determined the most probable moral value.

This model could not calculate the diagnosis for all records simultaneously, so the test data were loaded in batches. DeepSeek-R1 calculated the diagnosis through pattern analysis of the data, including: high sums frequently pointing to the predicted value; key positions or indices within the list.

GPT o3-mini compared each set of symptoms against available examples, relied on a heuristic assessment, evaluating symptom structures and relative magnitudes against reference cases rather than applying a formal model.

Grok-3 inferred diagnoses by analyzing structural patterns within vector distributions and cross-referencing input lists with reference samples from its knowledge base. Its decision-making relied on several key heuristics:

- High values in the central region (positions 3–15) typically indicated moral values 2–3, whereas high values near the terminus (positions 20–29) correlated with values 4–5.
- A high terminating value (position 30 exceeding 5) generally signaled moral values 1–2.
- Regarding overall magnitude, early-vector peaks strongly associated with moral value 4.

- In sparse vectors with few non-zero entries, the precise spatial placement and combinations of active positions provided critical diagnostic signals.

5 Conclusion

This work proposes a multi-label classification model based on fusing the pre-trained BERT model with supervised embedding methods, aiming to capture rich feature information and enhance multi-class classification capabilities.

Moral foundations are linked to individual's judgments, opinions, and attitudes regarding specific situations; consequently, they can be subjective, shaped by culture and beliefs.

Therefore, it is important to acknowledge that the moral annotations used in this study may contain potential biases stemming from the cultural diversity of the annotators.

While fine-tuning models to predict moral values offers a useful foundation for studying human ethics, analyzing morality in text presents both significant opportunities and clear risks:

- **The potential:** Identifying moral themes in written content can improve communication strategies and strengthen social advocacy campaigns.
- **The risks:** These tools can be weaponized for manipulation or lead to automated mischaracterizations of an individual's actual ethical beliefs.
- **The disadvantages:** The publicly available AI systems are the inability to access them with large amounts of data.
- **The accuracy:** Prediction of publicly available AI systems varies significantly from one call to another.

While AI systems are adept at detecting moral values within a text, they currently lack the precision and consistency required to deliver dependable insights for an expert analysis.

Acknowledgments

The work was done with partial support from grants 20260367 by Secretary of Research and Postgraduate Studies (SIP) of Instituto Politécnico Nacional, México.

References

1. Hoover, J., Johnson-Grey, K.M., Dehghani, M. (2017). Moral Values Coding Guide.
2. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, Vol. 8, No. 9, pp. 73791. DOI: 10.1371/journal.pone.0073791.
3. Borghouts, J., Huang, Y., Gibbs, S. (2023). Understanding Underlying Moral Values and Language Use of COVID-19 Vaccine Attitudes on Twitter. *PNAS Nexus*, Vol. 2, No. 3, pp. 13. DOI: 10.1093/pnasnexus/pgad013.
4. Atari, M., Graham, J., Dehghani, M. (2020). Foundations of Morality in Iran. *Evolution and Human Behavior*, Vol. 41, No. 5, pp. 367–384. DOI: 10.1016/j.evolhumbehav.2020.07.014.
5. Mooijman, M., Hoover, J., Lin, Y. (2018). Moralization in Social Networks and the Emergence of Violence during Protests. *Nature Human Behaviour*, Vol. 2, pp. 389–396. DOI: 10.1038/s41562-018-0353-0.
6. Koenraads, S.P.C., Aarts, M.C.J., van der Veen, E.L. (2016). Quality of Life Questionnaires in Otorhinolaryngology: A Systematic Overview. *Clinical Otolaryngology*, Vol. 41, pp. 681–688. DOI: 10.1111/coa.12586.
7. Meghaz, H.M., Mostefaoui, S.A., Maaskri, M. (2024). Deep Learning-Based Text Classification to Improve Web Service Discovery. *Computación y Sistemas*, Vol. 28, No. 2, pp. 529–542. DOI: 10.13053/cys-28-2-4556.
8. Graham, J., Haidt, J., Koleva, S. (2013). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism, *Advances in Experimental Social Psychology*. Elsevier, Vol. 47, pp. 55–130. DOI: 10.1016/B978-0-12-407236-7.00002-4.
9. Graham, J., Haidt, J., Koleva, S. (2011). Moral Foundations Theory: The Validity and Reliability of the Moral Foundations Questionnaire. *Advances in Research and Theory*, pp. 1–37. DOI: 10.31234/osf.io/p5fj8.
10. Trager, J., Ziabari, A.S., Davani, A.M. (2022). The Moral Foundations Reddit Corpus. *arXiv*. DOI: 10.48550/arXiv.2208.05545.
11. De Buck, A., Pauwels, L.J.R. (2023). Moral Foundations Questionnaire and Moral Foundations Sacredness Scale: Assessing the Factorial Structure of the Dutch Translations. *Psychologica Belgica*. DOI: 10.5334/pb.1188.
12. Boyd, R.L., Ashokkumar, A., Seraj, S. (2022). The Development and Psychometric Properties of LIWC-22. University of Texas at Austin. DOI: 10.13140/RG.2.2.23890.43205.
13. Graham, J., Haidt, J., Nosek, B.A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, Vol. 96, No. 5, pp. 1029–1046. DOI: 10.1037/a0015141.
14. xAI (2025). Grok 3 Announcement. x.ai/news/grok-3.
15. Google (2024). Gemini API Models Documentation. <https://ai.google.dev/gemini-api/docs/models>.
16. Google DeepMind (2024). Gemini Pro Technical Overview. deepmind.google/technologies/gemini/pro/.
17. Google DeepMind (2024). Gemini Technology Overview. deepmind.google/technologies/gemini/.
18. DeepSeek-AI (2025). DeepSeek-R1 Repository, GitHub. github.com/deepseek-ai/DeepSeek-R1.
19. OpenAI (2025). Introducing o3 and o4-mini. openai.com/index/introducing-o3-and-o4-mini/.

*Article received on 07/05/2026; accepted on 17/08/2026.

*Corresponding author is Obdulia Pichardo-Lagunas.