AMORES: A Spanish Language Resource for an Extended Set of Moral Foundations

Oscar Araque[1], **Daniel Molina**[2], **Anny Álvarez-Nogales**[1], **Carlos Á. Iglesias**[1]
[1]Universidad Politécnica de Madrid, [2]Social Innolabs Foundation
[1]Avenida Complutense, 30, Madrid, Spain, [2]Calle de Santa Engracia 4, 3ºD , Madrid, Spain
dmolina@socialinnolabs.org
{o.araque, a.anogales, carlosangel.iglesias}@upm.es

Abstract

This work addresses the need for linguistic resources that enable language models to understand and adapt to subjective and abstract concepts in the domain of moral values within texts. In light of the growing interest in the study of moral values and its limited exploration in Spanish-speaking contexts, this work addresses this gap by developing a novel Spanish-language corpus. Furthermore, the corpus's development process ensures that the annotations capture a wide range of perspectives, resulting in a resource that reflects the diversity of moral interpretations in real-world contexts. Specifically, there are two main contributions. **1** The creation of the first large-scale Spanish corpus annotated according to Moral Foundations Theory. **2** We introduce an experimental framework that investigates how annotators' religious orientations could shape moral annotation patterns and propagate to model behavior. To do so, we employ a prompt-based alignment method that improves moral detection regardless of religious alignment for which the model was trained. In this scenario, we explore whether language models can align moral interpretations across divergent belief orientations.

Keywords: Corpus generation, Moral Values, Moral Foundation Theory, Language Models

1. Introduction

Understanding how people express moral values in everyday language is crucial for analyzing social interactions, persuasion, and the formation of collective identities. Moral language is prevalent in debates surrounding politics, health, and social issues, influencing how communities negotiate agreements and conflicts (Borghouts et al., 2023; Hurst and Stern, 2020; Troy et al., 2025; Day et al., 2014). Computational approaches that address morality depend on the quality of the resources employed. Most existing moral-centered language resources are oriented toward English, which limits cross-cultural comparison and prevents insight into how values are communicated across diverse contexts. Thus, although Spanish is one of the most widely spoken native languages in the world, there is currently no publicly available corpus that systematically captures moral expression in naturalistic Spanish texts.

We present **Annotated MORal ValuES in Spanish (AMORES)**, a large-scale Spanish moral corpus (Araque et al., 2025)[1] that addresses this gap by enabling researchers to examine how moral values are expressed in real discourse, from social media to political debates, and to identify cultural patterns that may remain hidden when focusing only on English. This resource also supports advancements in computational tasks such as morality detection, stance detection models, and propaganda and hate speech detection (Solovev and Pröllochs, 2023; Brugnoli et al., 2023; Martínez et al., 2023), tailored explicitly for Spanish, which are largely missing and often rely on translation or transfer methods from English datasets (Banea et al., 2013; Masad et al., 2023). In addition, we present a set of experiments to evaluate performance on different language models.

2. Related Work

Research on morality in language has produced a diverse range of resources, from psychometric questionnaires to large-scale annotated corpora. The study of morality in language has been strongly influenced by the Moral Foundation Theory (MFT) (Haidt et al., 2007). This social psychological framework provides a taxonomy of core moral dimensions and has motivated the development of both psychometric instruments and computational resources. The MFT suggests that human morality is grounded in several innate, evolutionarily derived principles, such as care, equality, loyalty, authority, property, purity, and liberty, which shape our moral intuitions across cultures and help explain differences in moral and political views (Haidt and Graham, 2007).

The Moral Foundations Questionnaire (MFQ) is the most widely used tool to measure the degree to

[1]The corpus is publicly available at https://zenodo.org/records/17045203 under a Creative Commons Attribution 4.0 International (CC-BY 4.0) license (DOI: 10.5281/zenodo.17045203).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7056–7068
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

which individuals endorse moral foundations (Graham et al., 2011). Later refinements, such as the MFQ-2, expanded the fairness foundation into two more specific constructs (Atari et al., 2023). Similarly, Alabèrnia-Segura et al. (2023) developed and validated a Spanish version of the Moral Identity Questionnaire (MIQ), thereby extending psychometric research into Spanish-speaking contexts. The first Moral Foundations Dictionary (MFD) provided a lexicon of 324 words mapped to the moral categories of MFT (Graham et al., 2009). This lexicon was later expanded into larger and more refined versions, including MFDv2 (Frimer et al., 2019), the extended eMFD (Hopp et al., 2021), MoralStrength, a lexicon of approximately 1,000 lemmas obtained as an extension of the MFD based on WordNet synsets, with crowdsourced numeric assessments of Moral Valence indicating the strength with which each lemma expresses a specific moral value (Araque et al., 2020), and LibertyMFD, developed to capture the dimension of liberty (Araque et al., 2022). Specifically Rezapour et al. (2019) demonstrated the benefits of human-in-the-loop lexicon expansion, enlarging the MFD to over 4,600 disambiguated entries. For Spanish, (Carvalho et al., 2022) produced a dictionary adapted to MFT categories, providing a crucial resource for analyzing moral language in Spanish.

Beyond lexicons, multiple corpora have been developed to apply MFT to user-generated texts.Trager et al. (2022) introduced the Moral Foundations Reddit Corpus, with over 16,000 comments annotated for moral sentiment, while Hoover et al. (2020) created the Moral Foundations Twitter Corpus of more than 35,000 tweets. In (Johnson and Goldwasser, 2018), political tweets were annotated for implicit foundations, and Roy and Goldwasser (2021) extended this work with tweets from U.S. Congress members. In the political and media domain, Simonsen and Widmann (2025) applied multilingual dictionaries to parliamentary speeches on immigration across eight Western democracies over six decades, enabling large-scale comparative analysis, while Weber et al. (2021) compiled a corpus of nearly 36,000 news articles. Several corpora have also targeted moral discourse in the context of COVID-19: Rojecki et al. (2021) collected 2,648 tweets, Beiró et al. (2023) annotated 4,498 Facebook comments with 12 moral values, including liberty and oppression, and Pacheco et al. (2022) compiled annotated tweets on vaccination. Argumentative datasets have also been developed and used in the study of argumentation (Kobbe et al., 2020; Alshomary et al., 2022).

Additional resources have been designed as controlled stimuli, such as YouTube clips crowdsourced to represent MFT dimensions (McCurrie et al., 2018) and 132 moral vignettes developed to elicit specific foundation violations (Clifford et al., 2015). Furthermore, the Social-Chem-101 dataset (Forbes et al., 2020) compiles 292,000 'rules of thumb' describing everyday moral and social norms, annotated across 12 dimensions. Similarly, the Moral Integrity Corpus (MIC) (Ziems et al., 2022) extends this approach with 38,000 question–answer pairs linked to moral rules. In addition, Araque et al. (2024) applied MFT-based moral lexicons to large-scale news texts to extract moral values, demonstrating the usefulness of lexicon-based pipelines for resource-rich domains. Other datasets have been developed to investigate moral reasoning: the ETHICS benchmark (Hendrycks et al., 2020) evaluates moral intuitions across justice, deontology, virtue, and utilitarian perspectives; scenarios of varying ambiguity were created to test Large Language Model (LLM) beliefs (Scherrer et al., 2023); and MoralExceptQA was developed to investigate exceptions to moral rules (Jin et al., 2022). The SCRUPLES dataset (Lourie et al., 2021) provides real-world anecdotes with distributions of community judgments, reflecting consensus and disagreement, while ProsocialDialog (Kim et al., 2022) provides dialogs, used to train dialog systems to respond prosocially to problematic content. Other weakly or automatically labeled resources include (Botzer et al., 2022), which used judgments from the Reddit r/AmItheAsshole community to derive large-scale moral valence annotations, and (Emelin et al., 2020), which created 'Moral Stories,' a dataset of structured narratives encoding moral decisions and their consequences.

Together, these works illustrate that moral annotation can be approached not only through the MFT taxonomy but also via norms, judgments, and prosocial principles. In parallel, several works have focused on moral phenomena beyond the scope of MFT. Yet, these are relevant for understanding moral reasoning in texts and for training LLMs. Although significant progress has been made in developing morally annotated corpora across languages and domains, most large annotated corpora are in English, and none are currently available in Spanish.

While existing efforts provide valuable groundwork, they do not deliver large-scale, systematically annotated datasets capable of capturing the richness and complexity of moral discourse in authentic user-generated content. To address this gap, this study introduces the first Spanish corpus annotated according to Moral Foundations Theory, using an extended set of foundations in comparison to previous corpora. This corpus advances both cross-linguistic research and the analysis of moral expression in online communities.

3. Corpus/Dataset Creation Methodology

3.1. Data Collection

The data of the presented corpus (AMORES) consists of Spanish comments from different Reddit[2] communities and online forums. This source has been used previously to capture morality in texts (Botzer et al., 2022), including language resources (Trager et al., 2022).

To construct the corpus, we implemented a scraping process that monitored Reddit communities (subreddits) for content that potentially expressed morality. To collect this data, we used the Python Reddit API Wrapper (PRAW)[3], which enables automated, scalable extraction of user comments. After analyzing different Spanish subreddits with potential moral foundation reflection, a list of 33 subreddits was selected as the most relevant data sources. These communities were chosen not only for the presence of moral themes in their discussions, but also for their diverse geographic and linguistic variations within the Spanish language and for reflecting a wide range of ideological and cultural perspectives.

The selected communities are the following: *cine*, *es*, *askspain*, *Espana*, *spain*, *futbol*, *Madrid*, *Desahogo*, *chile*, *SpainPolitics*, *Asi va España*, *videojuegos*, *agerntiina*, *catalunya*, *libros*, *esConversacion*, *chistes*, *askspain*, *espanol*, *lationamerica*, *mexico*, *preguntaareddit*, *putoscoches*, *valencia*, *venezuela*, *MujeresEnReddit*, *Negacionistas*, *OpinionesPolemicas*, *allinspanish*, *colapsoES*, *Antinatalismo esp*, *podemos*, *vayacurro*. For additional context, brief English descriptions of less self-evident communities are provided in Appendix A.

A total of approximately 60,000 texts were collected. After automatic language identification to remove non-Spanish comments (e.g., English or Catalan), the material was manually screened prior to annotation. This screening retained comments containing interpretable content and potential moral evaluations, and discarded texts that could not be understood in isolation, such as isolated reactions, short jokes, or fragments requiring missing conversational context.No automatic ambiguity or harmfulness detection was applied at this stage.

3.2. Annotation Methodology

The annotation process was carried out using Qualtrics[4] for the flexible creation of online surveys and Prolific[5] as the crowdsourcing platform. A total of 52 distinct annotation tasks (annotation batches) were designed, each comprising 60 texts. Five annotators were assigned to each task to ensure reliability and to incorporate diverse perspectives.

The annotators were recruited from several Spanish-speaking countries where Prolific offered availability (Spain, Chile, and Mexico) to provide a broad range of perspectives from different countries and various interpretations of expressions or slang used on the social media platform. In the case of Spain, the annotators were recruited from a mix of university students and Prolific participants.

Before starting the annotation phase, the participants completed a formation stage in which they were introduced to the Moral Foundations framework and the annotation guidelines. This formation emphasized the need to identify moral foundations intuitively, even when they were not explicitly stated in the text, and to consider sarcasm, which poses a valuable challenge for the corpus, given its difficulty in automatic detection. In addition, a set of profiling questionnaires was administered to annotators. These included general demographic information (gender, age, political orientation, income level, and religious sentiment), the Mood and Feelings Questionnaire (MFQ30) (Angold et al., 1987) in its Spanish translation, and an extended version of the Moral Foundations Questionnaire (MFQ-2) (Atari et al., 2023). The MFQ-2 is adapted to current issues, comprising 36 questions across six moral foundations (care, equality, proportionality, loyalty, authority, and purity), thereby enriching annotator profiles.

Regarding the annotation tasks, each text has been annotated by at least three and no more than five annotators. Annotators were required to indicate whether a text reflected any moral value and, if so, to specify the foundation, its leaning towards virtue or vice, and their confidence in the annotation (low, medium, or high). AMORES includes a broader range of moral foundations than prior corpora. These are defined in the updated MFT framework (Graham et al., 2013) (Atari et al., 2023), resulting in the following categories: Care, Equality, Proportionality, Loyalty, Honor, Purity, Property, Authority, and Liberty. If no moral dimension is identified, the option 'Non-Moral' is available.

Annotators are allowed to annotate more than one foundation for each text, which substantially broadens the scope of moral assessment. Following the paradigm of perspectivism (Federico et al., 2023), the annotation methodology has been designed to accommodate a variety of perspectives, especially given the subjectivity and scope of the morality phenomenon. The annotation phase took place between March and June 2025. The com-

[2] https://www.reddit.com/

[3] https://praw.readthedocs.io/en/stable/

[4] https://www.qualtrics.com/

[5] https://www.prolific.com/

plete annotation process resulted in 37,150 individual annotations. The final version of the corpus, after removing duplicates and correcting other errors, comprises 3,062 annotated instances.

All participants were compensated for their contributions. To maintain quality, responses were monitored to verify completion times, consistency, and the proportion of texts labeled by annotators without a moral foundation. After the annotation phase, annotations and texts were aggregated, and inter-annotator agreement was applied to ensure control and analysis of the corpus.

4. Corpus Analysis

This section covers the analysis of the presented corpus. Section 4.1 explores the characteristics of the annotators, reflecting on their impact on moral assessments. Section 4.2 presents a detailed study on the agreement among annotators.

4.1. Statistics of Annotators

In a first approach to the corpus, Table 1 summarizes the total counts for each moral value and the percentage breakdown of virtue and vice annotations within each moral class. Overall, the data indicate a predominance of vice polarity across most moral traits, suggesting that users in these subreddits tend to express themselves in a more negative or critical tone.

Moral	Count	Virtue (%)	Vice (%)
Authority	2345	22.56	77.44
Care	6462	55.96	44.04
Equality	5450	26.04	73.96
Liberty	3232	43.72	56.28
Honor	5631	23.23	76.77
Loyalty	2077	42.85	57.15
Property	1166	30.19	69.81
Proport	4537	28.52	71.48
Purity	5308	22.14	77.86
Non-Moral	942	-	-

Table 1: Total counts of each moral class with the percentage of instances annotated as virtue or vice.

Regarding the annotators' metadata, texts were annotated by 57 annotators from diverse demographic and ideological backgrounds, allowing us to capture a wide range of perspectives across different socioeconomic backgrounds and religious viewpoints. Figures 1 and 2 show the distribution of the annotators' characteristics. Regarding the gender report, the distribution is relatively balanced: 34 participants identified as male, 22 as female, and one as other. Political orientations were distributed as follows: 31.6% left-leaning, 22.8% center-left, 22.8% center-right, 17.5% centrist, and 5.3% did not disclose (NA). Regarding religion, 37 annotators identified as non-religious, 18 as religious, and two did not disclose (NA). The age and income distributions were also diverse (Fig. 2). The largest age group is 25–35 years old (50.9%), followed by 18–25 years old (21.1%), 35–45 years old (17.5%), 45–55 years old (8.8%), and 55–65 years old (1.8%). Income ranged from €0–4,999 (28.1%) to over €35,000 (5.3%), with the remainder distributed across intermediate brackets or left blank (NA).

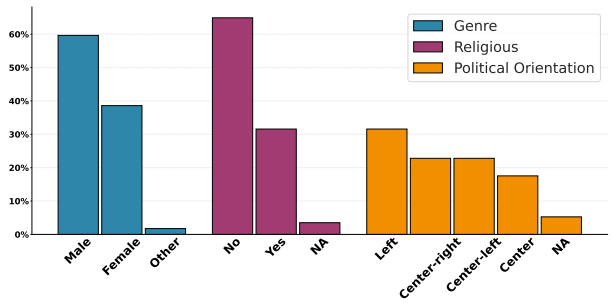

Figure 1: Annotator's gender, religious and political orientation.

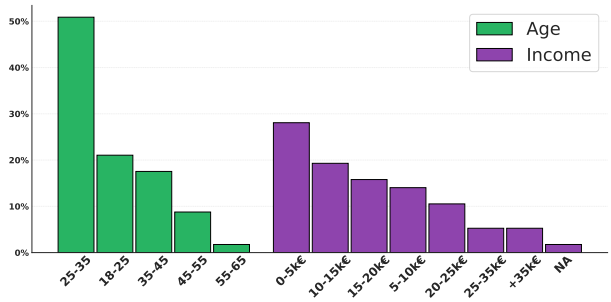

Figure 2: Annotator's age and income.

To examine how moral assessments vary within each moral trait, Fig. 3 shows how confidence and polarity levels are distributed within each moral foundation. The values have been normalized within each foundation, so the percentages represent the internal proportions of confidence and polarity for that moral trait. Overall, the Care, Equality, and Purity moral traits show the highest proportion of high confidence annotations, indicating that annotators generally felt more certain in their assessments for these traits. In contrast, Property, Authority, and Liberty exhibit greater variability, with

more annotations in lower confidence categories and fewer high confidence instances.

This pattern likely reflects the relative ease with which annotators can identify certain moral foundations. Care and equality relate directly to fairness, equal and just treatment among people, while purity can often be associated with intuitions about contamination, sanctity, or culturally prominent ideals (Graham et al., 2013). In contrast, Property, Authority, and Liberty involve judgments that are more context-dependent, influenced by possible political preferences and nuanced by language or expression (Atari et al., 2023). We argue that these traits make the latter set of foundations more challenging to assess consistently, leading to greater variability in annotators' confidence. In addition, it should be noted that most annotations reference the vice associated with the moral foundations, which is consistent with debates on social media, where topics are often controversial. In online debates, users also tend to express criticism, complaints, or moral violations rather than moral praise, leading to a greater prevalence of negative moral views.

To examine how moral interpretations vary within each political or religious group, we analyzed the internal distribution of moral foundations within each group. Table 2 illustrates the internal distribution of moral foundations within each ideological group. Each row represents the moral composition of each political orientation, expressed as percentages, showing which values are most prominent within that group. Columns are defined as C (center), C-L (center left), C-R (center-right), Left, and NA (not answered or prefer not to say). In general, the moral dimensions of care and equality prevail in all political groups, as shown in Figure 3. These three categories remain the most frequently detected foundations across the entire corpus.

Political	C	C-L	C-R	Left	NA
Authority	5.72	6.14	8.05	6.51	5.99
Care	14.11	17.54	21.07	19.20	19.28
Equality	15.39	14.60	14.56	14.90	17.34
Liberty	10.66	7.77	8.00	8.42	10.63
Honor	16.67	15.53	15.35	14.77	14.91
Loyalty	5.95	6.35	5.59	5.11	5.90
Property	2.89	3.53	4.34	2.52	30.29
Proport	14.77	11.79	8.28	13.71	13.15
Purity	13.86	16.76	14.76	14.87	9.50

Table 2: Percentages of moral foundations within political orientations normalized by political group.

However, different patterns emerge for specific groups. For instance, texts annotated by centrist participants show stronger associations with Proportionality. In contrast, center-right participants exhibit a higher prevalence of Authority-related annotations. This reflects the existence of distinct interpretations of moral evaluations across different political groups. We believe that this and other similar characteristics of the dataset will be of the most significant interest to the research community.

Following a similar approach to the one above, Table 3 presents the distribution of normalized moral foundations by annotators' religious orientations. We observe an interesting pattern: annotators who identify as religious show a more balanced distribution of moral categories, suggesting a broader moral interpretation across all values. In contrast, participants without a religious orientation have annotated higher proportions of the Purity foundation, a somewhat unexpected finding given that this foundation is often associated with religious or spiritual discourse. We view this as a starting point for future work, as the role of religion in moral assessment remains underexplored in large-scale studies.

Religious	No	Yes	NA
Authority	6.18	6.73	9.59
Care	18.40	16.74	19.28
Equality	13.92	17.10	14.66
Liberty	7.66	11.32	7.24
Honor	16.52	13.68	17.38
Loyalty	5.56	6.04	05.70
Property	3.36	2.90	4.07
Proport	12.19	13.44	9.14
Purity	16.22	12.06	12.94

Table 3: Percentages of moral foundations by religious orientation (normalized by religious group).

4.2. Agreement among Annotators

To perform a deeper analysis by subreddit and group, we define an agreement metric to explore potential patterns in annotations across topics or subreddits and among annotator groups categorized by gender, religion, and political orientation. During annotation, each annotator can assign multiple labels to a single text. Consequently, traditional agreement metrics, which rely on exact matching, are not suitable.

Given that annotators could assign multiple moral categories to each text, we treat annotations as sets and use Jaccard similarity to measure agreement (Verma and Aggarwal, 2020).

For each text t and each annotator i, we calculate an individual agreement g_i representing the average similarity between the annotations of the annotator i and those of all other annotators.

$$g_i = \frac{1}{N-1} \sum_{j \neq i} s(A_i, A_j) \quad (1)$$

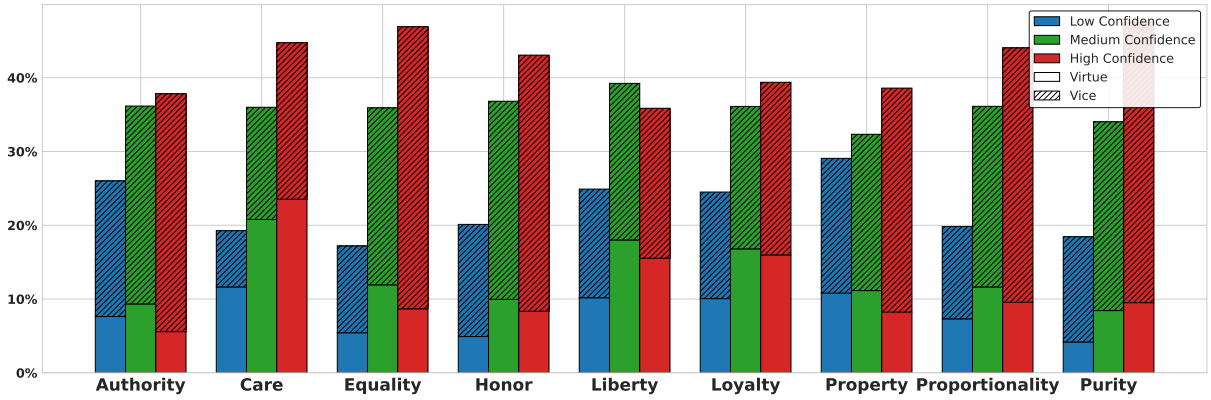

Figure 3: Moral distribution polarity by trust.

where N is the total number of annotators for text t, A_i denotes the set of annotations from annotator i, and s is the Jaccard similarity coefficient. The similarity between two annotators i and j for text t is defined as:

$$s_t(i,j) = \frac{|A_t^{(i)} \cap A_t^{(j)}|}{|A_t^{(i)} \cup A_t^{(j)}|} \tag{2}$$

We use Jaccard similarity rather than exact matching because it handles multi-label annotations and captures partial overlap. Finally, for each subreddit with S texts, we define the aggregated agreement as follows.

$$\bar{g}_S = \frac{1}{|S|} \sum_{t \in S} \bar{g}_t \tag{3}$$

Figure 4 presents a heatmap summarizing the agreement and divergence scores across all subreddits (x-axis) and annotator groups (y-axis). Annotator groups were compared along several dimensions, including gender (Male, Female), religiosity (Religious, Non-Religious), and political orientation (Center, Center-Left, Center-Right, Left), as well as across all annotators combined. The mean column reflects the average agreement across all subreddits.

Generally, agreement values vary within a moderate range among all annotators. They are between 0.20 and 0.33 on average (see row 'All'), indicating a moderate degree of agreement. When examining gender differences, Female annotators tend to show slightly higher levels of agreement than Male annotators, with values exceeding 0.5 in some subreddits. In terms of religiousness, participants who identify as religious exhibit significantly higher agreement, whereas the Non-Religious group shows greater heterogeneity and lower agreement. Annotators who identify as politically center-right tend to exhibit the highest levels of internal consistency, with values exceeding 0.6 in some subreddits. By contrast, the center-left group has fewer agreements.

Interestingly, the agreement patterns vary substantially among the subreddits. Those communities focused on cultural or entertainment topics like cinema, video games, and books[6]. They show relatively high levels of agreement In contrast, subreddits dedicated to political and social issues[7], show the lowest agreement values.

To provide a more visual example of differences in text annotations, Figures 5 and 6 show two examples from different subreddits that exhibit annotation variability. Arrows indicate annotation confidence, categorized as low, medium, or high. Moral labels are color-coded by polarity: green denotes virtue and red denotes vice.

In Figure 5, a higher level of agreement among the annotators is shown. Most annotations show medium to high confidence, and almost all moral foundations are marked as virtues, except for the non-moral and honor labels. By comparison, Figure 6 shows an example with greater variability in confidence and polarity. The same moral foundations are indicated with both positive and negative polarities, and confidence levels also reflect almost the same number of annotations.

These examples reinforce our earlier point: texts, especially political ones, are more challenging to annotate consistently. In the first text (Fig. 5), although the comment appears ironic, the criticism is subtle and could be interpreted as positive by annotators. The second (Fig. 5) expresses a strong critique of politics, leading to greater variability in interpretations.

Finally, to quantify the overall level of consistency among annotators, we report the average inter-annotator agreement (IAA) scores, Fleiss' Kappa, PABAK, and Krippendorff's Alpha, calcu-

[6]Original names are: 'cine', 'videojuegos', and 'libros'

[7]E.g., 'SpainPolitics', 'Antinatalismo_esp', and 'Podemos', a Spanish political party

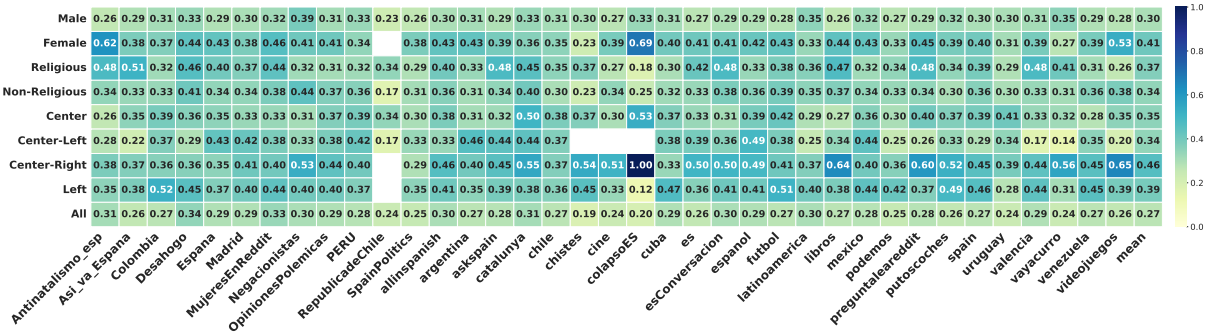

Figure 4: Mean agreement by subreddit and annotators group.

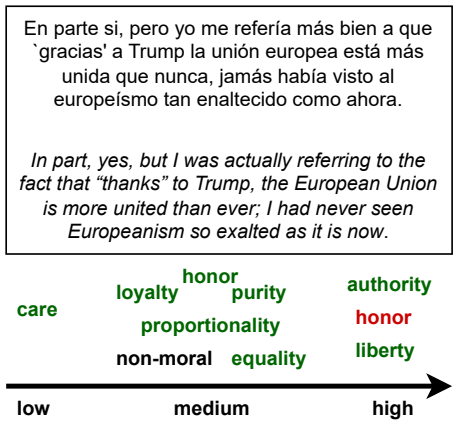

Figure 5: Annotation example of text from subreddit 'askspain', including confidence and polarity of each annotation.

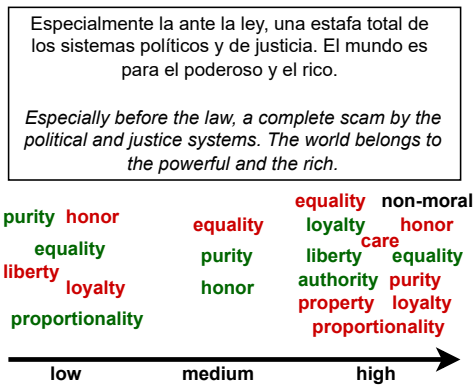

Figure 6: Annotation example of text from subreddit 'chile', including confidence and polarity of each annotation.

lated across all annotation tasks. These metrics provide a general measure of agreement among annotators across the entire set of texts.

The metrics show moderate but positive consistency among annotators, with no negative agreement values. Fleiss' Kappa (0.20) and Krippendorff's Alpha (0.22) indicate slight to fair agreement, which is acceptable given the interpretive nature of the task. PABAK (0.40) is noticeably higher, reflecting its adjustment for chance agreement and the distribution of categories. Overall, these findings confirm that, while moral interpretation varies among annotators, the observed agreement is meaningful and exceeds expectations by chance alone.

These values are consistent with those reported for comparable English-language moral corpora. For instance, (Hoover et al., 2020) report an overall Fleiss' Kappa of 0.27 and PABAK of 0.29. Although several foundation-specific Kappa values are relatively low (e.g., .16 for Purity and .18 for Betrayal). Similarly, (Trager et al., 2022) report generally low Fleiss' Kappa values, often ranging between 0.0 and 0.5, but agreement increases substantially after prevalence adjustment.

5. Experiments

To evaluate the developed linguistic resource, we conducted a series of experiments. The first one involved establishing baselines using all moral classes and an aggregated label for a single moral prediction task.

The gold label for each text was computed using a weighted voting scheme across annotators, based on each annotator's confidence in their annotations. Each annotator's vote was weighted by a confidence score (high, medium, low), and votes were counted across the combined moral categories and polarity (e.g., care-virtue). In the event of a tie, we apply a hierarchical tie-breaking procedure: first, we aggregate annotations with confidence weights; second, if a tie remains, we prefer the category with the higher raw count; and finally, if a tie persists, we choose the most frequent polarity. This ensures that the resulting label reflects the usage of annotator consensus and incorporates all confidence levels. We use different models to evaluate classification ability: models pre-trained mainly on English and/or multilingual data, and models pre-

	macro	Care	Equality	Propor	Honor	Purity	Liberty	Authority	NM	Loyalty	Property
roberta	17.06	23.62	31.29	22.41	17.54	8.51	16.39	25.45	25.40	0.0	0.0
bertin	13.05	22.86	16.07	18.65	23.62	11.51	13.68	8.70	15.38	0.0	0.0
llama-1b	16.62	30.22	10.08	**20.87**	16.00	16.39	**24.78**	14.08	**26.67**	0.0	7.14
llama-3b	21.03	**37.14**	22.73	20.17	28.57	15.25	18.35	**32.43**	20.69	9.52	5.41
salamandra-2b	19.25	33.08	25.42	10.29	27.94	16.84	22.61	18.92	32.26	5.13	0.0
salamandra-7b	**21.84**	31.65	24.00	19.86	**30.53**	**20.17**	21.78	17.65	18.18	**22.86**	**11.76**

Table 4: Macro F1 scores for different models in moral classification task. Column 'macro' reports the overall Macro F1 score, while the remaining columns show F1 scores for each specific moral value.

trained on Spanish data. Specifically, Roberta (Liu et al., 2019) was pre-trained primarily on English data; LLaMA (Grattafiori et al., 2024) and Salamandra (Gonzalez-Agirre et al., 2025) are multilingual models; and BERTin (de la Rosa et al., 2022) was pre-trained on Spanish data. This distinction is essential for capturing the nuanced interpretations of Spanish texts, which may be more difficult for one model than for another. There are also differences between smaller and larger models. RoBERTa and BERTin are based on the RoBERTa base architecture, whereas LLaMA and Salamandra are based on the large language model LLaMA architecture.

Building upon these baselines, we also designed an experiment to investigate whether moral interpretations vary between annotator groups with different religious orientations. Since a text may reflect multiple moral foundations, this experiment adopts a multilabel classification approach. For this class, four classification models were trained in two scenarios: with labels provided by religious annotators and by non-religious annotators.

After the initial training, a prompting process was introduced in which each model received input augmented through a textual prompt indicating the moral annotation that the other group would have produced. For instance, the religious model was exposed to prompts describing non-religious annotations, and vice versa. This prompt-based method could enable the model to integrate contextual cues about alternative moral framings and allow us to examine whether exposure to opposing perspectives can lead to greater predictive alignment across groups.

Table 4 shows the results for all the annotated moral categories, which include Care, Equality, Proportionality (Proport), Honor, Purity, Liberty, Authority, Loyalty, Property, and Non-Moral. This scenario is particularly challenging due to the ambiguity of specific categories and the limited number of existing instances. F1 scores for categories such as Loyalty and Property are near zero in some models, reflecting the difficulty of predicting these moralities.

In general, larger models such as LLaMA-3B and Salamandra-7B achieve higher overall macro-F1 scores (21.03 and 21.84, respectively), suggesting that greater model size and representational capacity have advantages in capturing moral semantics.

Model		Religious	No Religious
Roberta	Baseline	16.55	29.63
	Prompt	**19.32**	**41.86**
Bertin	Baseline	08.61	**39.39**
	Prompt	**18.06**	34.49
LLaMA	Baseline	**22.98**	38.09
	Prompt	17.02	**39.78**
Salamandra	Baseline	25.16	37.64
	Prompt	**22.54**	**39.52**

Table 5: Macro F1 scores for models trained on religious and non-religious annotations, comparing baseline and prompt performance.

Evaluating the prompt and religious experiment, as seen in Table 5, there are noticeable differences in macro F1 scores between models; those trained on religious annotations seem to struggle more than those trained on non-religious annotations, capturing the distributions of moral values. Although prompt-based augmentation seems to encourage moral interpretation, it appears to benefit non-religious models more. Of these, the non-religious Roberta model benefits the most, achieving an increase in macro F1 of almost 13. However, larger language models do not appear to benefit as consistently, as seen in the case of Llama trained on religious annotations, where prompting seems to reduce performance. In this case, the prompt appears to interfere with the moral learning of larger models. However, smaller models can effectively leverage the additional guidance.

6. Conclusions

This research introduces **AMORES**, the first Spanish corpus based on Moral Foundations Theory. Alongside the corpus, we present an extensive analysis of trends and annotation patterns, examining how annotator backgrounds and orientations influence moral evaluations. Furthermore, we demonstrate the utility of the corpus by assessing and prompting language models in moral classification tasks.

Although subjectivity can introduce bias or variability in annotations, this work sought to mitigate these effects by implementing rigorous guidelines and systematic controls during the annotation process. However, it is essential to acknowledge that moral reasoning is inherently context sensitive. Although the results of the automatic classification are modest, the AMORES corpus represents a significant step forward in the computational modeling of moral discourse in Spanish. It provides a valuable resource for future research into moral reasoning, value alignment, and the interpretability of language models.

We have observed in our experiments that modern language models, despite their remarkable abilities in Natural Language Processing (NLP), struggle with abstract and subjective tasks, such as moral judgment. Based on this corpus, future research will seek to improve the reasoning and transparency of models when dealing with subjective moral content. The corpus will be made publicly available to support continued research within the scientific community.

7. Limitations

This work adopts the Moral Foundations Theory as its fundamental framework. Alternative moral taxonomies may capture different nuances, and our findings may not transfer across frameworks. Although large for Spanish and the moral domain, the corpus presented may not fully represent the diversity of the Spanish-speaking world. Choices such as gender, topic, and annotation platform could introduce biases, and some moral foundations remain underrepresented.

Moral assessments are inherently subjective. Despite guidelines and extensive quality control, inter-annotator variability persists, which represents a path for future work that this paper does not address. In this regard, the analyses identify associations between annotator features and model behavior but do not establish causal mechanisms, which is the focus of further work.

8. Ethics Statements

All annotators participating in corpus creation were aware of the project's objectives and annotations, and they gave their consent to the publication of their annotations. The metadata collected was correctly anonymized during the study. All annotators were adults who provided informed consent after reviewing the task description, data use, and their rights, including the right to withdraw from the study. Participation was voluntary and paid, compensated at fair local rates. We provided clear guidelines, training examples, and support.

For the experimental analysis, annotators could optionally self-report coarse-grained metadata. These attributes were collected solely to study the aggregate effects on annotation patterns. We explicitly avoid normative claims about any group. Results describe annotation behavior in this dataset and should not be used to stereotype individuals or communities.

The study involves analyzing publicly available texts and aggregating annotation behavior without collecting personally identifiable information. It constitutes minimal-risk research under standard institutional guidelines. All procedures complied with our institution's ethics policies for research involving human participants.

9. Acknowledgements

This work has been partially funded with the AMOR project (TSI-063100-2022-0002), supported by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union - NextGeneration EU, within the UNICO I+D Cloud Programme. This work has been partially supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement 2023-2026 with Universidad Politécnica de Madrid in the Line A, Emerging PhD researchers (project MORA, DOCTORES-EMERGENTES-24-9UMLXZ-37-IIGW); Spanish version of this acknowledgement: "Financiado por la Comunidad de Madrid a través del convenio-subvención para el fomento y la promoción de la investigación y la transferencia de tecnología en la Universidad Politécnica de Madrid, en la Línea A, Doctores Emergentes".

10. Bibliographical References

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. *arXiv preprint arXiv:2203.14563*.

Oscar Araque, Luca Barbaglia, Francesco Berlingieri, Marco Colagrossi, Sergio Consoli, Lorenzo Gatti, Caterina Mauri, and Kyriaki Kalimeri. 2024. Beyond the headlines: Understanding sentiments and morals impacting female employment in spain. *arXiv preprint arXiv:2402.07339*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2013. Porting multilingual subjectivity resources across languages. *IEEE Transactions on Affective Computing*, 4(2):211–225.

Judith Borghouts, Yicong Huang, Sydney Gibbs, Suellen Hopfer, Chen Li, and Gloria Mark. 2023. Understanding underlying moral values and language use of covid-19 vaccine attitudes on twitter. *PNAS nexus*, 2(3):pgad013.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. Analysis of moral judgment on reddit. *IEEE Transactions on Computational Social Systems*, 10(3):947–957.

Emanuele Brugnoli, Pietro Gravino, and Giulio Prevedello. 2023. Moral values in social media for disinformation and hate speech analysis. In *International Workshop on Value Engineering in AI*, pages 67–82. Springer.

Eugene Y Chan. 2021. Moral foundations underlying behavioral compliance during the covid-19 pandemic. *Personality and individual differences*, 171:110463.

Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198.

Martin V Day, Susan T Fiske, Emily L Downing, and Thomas E Trail. 2014. Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, 40(12):1559–1573.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling.

Cabitza Federico, Campagner Andrea, and Basile Valerio. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. Salamandra technical report.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. 2024. The llama 3 herd of models.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.

Jonathan Haidt, Craig Joseph, et al. 2007. The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. *The innate mind*, 3(3):367–391.

Kristin Hurst and Marc J. Stern. 2020. Messaging for environmental action: The role of moral framing and message source. *Journal of Environmental Psychology*, 68:101394.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining,*. Association for Computational Linguistics, ACL.

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.

Rubén Yáñez Martínez, Guillermo Blanco, and Anália Lourenço. 2023. Spanish corpora of tweets about covid-19 vaccination for automatic stance detection. *Information Processing and Management*, 60(3):103294.

Ofri Masad, Kfir Bar, and Amir Cohen. 2023. Automatic translation of span-prediction datasets. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–173.

Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 35–45.

Kirill Solovev and Nicolas Pröllochs. 2023. Moralized language predicts hate speech on social media. *PNAS nexus*, 2(1):pgac281.

Cassandra LC Troy, Nicholas Eng, and Chris Skurka. 2025. Green and good? examining intended and unintended effects of morally framed climate messages. *Environmental Communication*, 19(2):238–258.

Vijay Verma and Rajesh Kumar Aggarwal. 2020. A comparative analysis of similarity measures akin to the jaccard index in collaborative recommendations: empirical and theoretical perspective. *Social Network Analysis and Mining*, 10(1):43.

11. Language Resource References

Miquel Alabèrnia-Segura, Guillem Feixas, and David Gallardo-Pujol. 2023. Moral identity questionnaire (MIQ): Adaptation and psychometric properties in Spanish population. *Acción Psicológica*, 20(1):121–130.

Adrian Angold, Elizabeth J Costello, Andrew Pickles, F Winder, and D Silver. 1987. Mood and feelings questionnaire. *Unpublished document. Duke University Developmental Program*.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM conference on information technology for social good*, pages 154–160.

Araque, Oscar and Molina Cabrera, Daniel and Álvarez Nogales, Anny and Iglesias Fernandez, Carlos Angel. 2025. *Spanish Morality Corpus*. Zenodo.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157.

Mariano Gastón Beiró, Jacopo D'Ignazi, Victoria Perez Bustos, María Florencia Prado, and Kyriaki Kalimeri. 2023. Moral narratives around the vaccination debate on facebook. In *Proceedings of the ACM Web Conference 2023*, pages 4134–4141.

Flavio Carvalho, Gustavo Guedes, M Nussbaum, C Infante, and J Sánchez. 2022. Dicionário de fundamentos morais em espanhol. *Nuevas Ideas En Inform á tica Educativa*, 16:287–291.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Jeremy A Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehgani. 2019. Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, 53(1):232–246.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational*

linguistics (volume 1: long papers), pages 720–730.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Caitlin H McCurrie, Damien L Crone, Felicity Bigelow, and Simon M Laham. 2018. Moral and affective film set (maafs): A normed moral video database. *PloS one*, 13(11):e0206604.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the covid-19 vaccine debate. *arXiv preprint arXiv:2205.01817*.

Andrew Rojecki, Elena Zheleva, and Lauren Levine. 2021. The moral imperatives of self-quarantining. In *Annual meeting of the American Political Science Association*.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of us politicians through the lens of moral foundation theory. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 1–13.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Kristina Bakkær Simonsen and Tobias Widmann. 2025. When do political parties moralize?: A cross-national study of the use of moral language in political communication on immigration. *British Journal of Political Science*, 55:e33.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations Reddit corpus. *arXiv preprint arXiv:2208.05545*.

René Weber, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2021. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. In *Computational Methods for Communication Science*, pages 39–59. Routledge.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A. Description of Reddit Communities

Table 6 provides the full list of Reddit communities (subreddits) used as data sources for the AMORES corpus, along with a brief description of each community's primary topic and geographic focus.

es	General Spanish-language community
spain	General discussion about Spain
Espana	News and current affairs in Spain
askspain	Questions about life and culture in Spain
SpainPolitics	Political debate focused on Spain
Asi_va_España	Critical commentary on Spain's political situation
Madrid	Discussion about the city and region of Madrid
catalunya	Topics related to Catalonia, politics and culture
podemos	Community around the Spanish left-wing party Podemos
colapsoES	Debates on societal and environmental collapse in Spain
Antinatalismo_esp	Antinatalism discussion in Spanish
Negacionistas	Debates around denialism and conspiracy theories
OpinionesPolemicas	Controversial opinions on various topics
Desahogo	Venting and emotional release community
MujeresEnReddit	Community for women, gender issues and feminism
esConversacion	Casual conversation in Spanish
espanol	Spanish language learning and discussion
allinspanish	General Spanish-language community
chistes	Jokes and humor in Spanish
futbol	Football (soccer) discussion
videojuegos	Video games discussion
cine	Cinema and film discussion
libros	Books and literature
putoscoches	Cars and driving, often humorous
vayacurro	Work-related humor and complaints
preguntaleareddit	Spanish-language equivalent of AskReddit
chile	General discussion about Chile
argentina	General discussion about Argentina
mexico	General discussion about Mexico
latinoamerica	Pan-Latin American topics
venezuela	Discussion about Venezuela
valencia	Topics related to the Valencia region of Spain
cuba	Discussion about Cuba

Table 6: Reddit communities included in the AMORES corpus, with brief descriptions of their primary focus and geographic scope. Communities were selected to ensure diversity across topics, ideological orientations, and Spanish-speaking regions.