From Polarity to Intensity: Mining Morality from Semantic Space

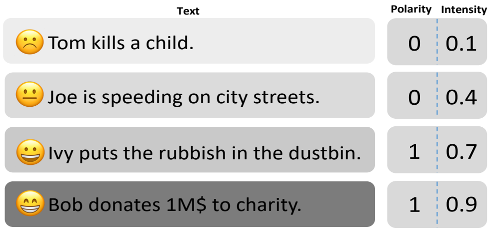
Chunxu Zhao, Pengyuan Liu, Dong Yu✉
School of Information Science, Beijing Language and Culture University, China
chunxu1212@gmail.com, liupengyuan@pku.edu.cn, yudong_blcu@126.com

Abstract

Most works on computational morality focus on moral polarity recognition, i.e., distinguishing right from wrong. However, a discrete polarity label is not informative enough to reflect morality as it does not contain any degree or intensity information. Existing approaches to compute moral intensity are limited to word-level measurement and heavily rely on human labelling. In this paper, we propose MoralScore, a weakly-supervised framework[1] that can automatically measure moral intensity from text. It only needs moral polarity labels, which are more robust and easier to acquire. Besides, the framework can capture latent moral information not only from words but also from sentence-level semantics which can provide a more comprehensive measurement. To evaluate the performance of our method, we introduce a set of evaluation metrics and conduct extensive experiments. Results show that our method achieves good performance on both automatic and human evaluations.

1 Introduction

Moral intensity is a degree of feeling that a person has about a behaviour (Barnett, 2001). As shown in Figure 1, although speeding on streets and killing a child are both immoral, the latter is more severe in most people's perception. Understanding the above difference is an ability that humans have gradually developed in everyday life. It affects individuals' ethical judgments and reflects the ideology of our society (Jones, 1991). As AI gets ever more involved in people's lives, it has become increasingly important for machine to acquire this ability and behave ethically. Researchers have studied the problem from early rule-based methods to today's deep learning-based paradigms (Yu et al., 2018; Hendrycks et al., 2021). It remains a fundamental but unsolved problem in computational morality (Moor, 2006).



Text	Polarity	Intensity
☹ Tom kills a child.	0	0.1
😐 Joe is speeding on city streets.	0	0.4
😀 Ivy puts the rubbish in the dustbin.	1	0.7
😁 Bob donates 1M$ to charity.	1	0.9

Figure 1: Moral Intensity Example. From the numerical measurement of morality, both moral polarity and its degree can be reflected.

Previous work in the NLP community often treats this problem as a supervised text classification task, i.e., judging the moral polarity for a text (Xie et al., 2020; Nahian et al., 2021). This way of modelling morality is inadequate because it oversimplifies morality into a Bernoulli distribution, i.e., being only moral or immoral. We model morality into a continuous distribution by introducing moral intensity to include degree information. Computing moral intensity is challenging in two aspects: 1) In supervised settings, unlike labelling moral polarity, building a large corpus with precise intensity values is time consuming and prone to subjectivity. 2) In unsupervised settings, there is no direct link between text and moral intensity. Even when moral polarity labels are available, building such link is nontrivial because the binary labels do not reflect any information about moral intensity.

To address these challenges, we propose MoralScore, a weakly-supervised framework that outputs a numerical value as the measurement of moral intensity for action-consequence pairs. The framework contains two parts. The first part is a semantic-aware moral detector, which measures moral intensity by detecting latent moral information from word to sentence level in semantic space. This incremental computing process can provide a comprehensive measurement of moral intensity for a text where both coarse-grained and fine-grained moral information can be captured. The second

[1] https://github.com/blcunlp/MoralScore

Proceedings of the 29th International Conference on Computational Linguistics, pages 1250–1262
October 12–17, 2022.

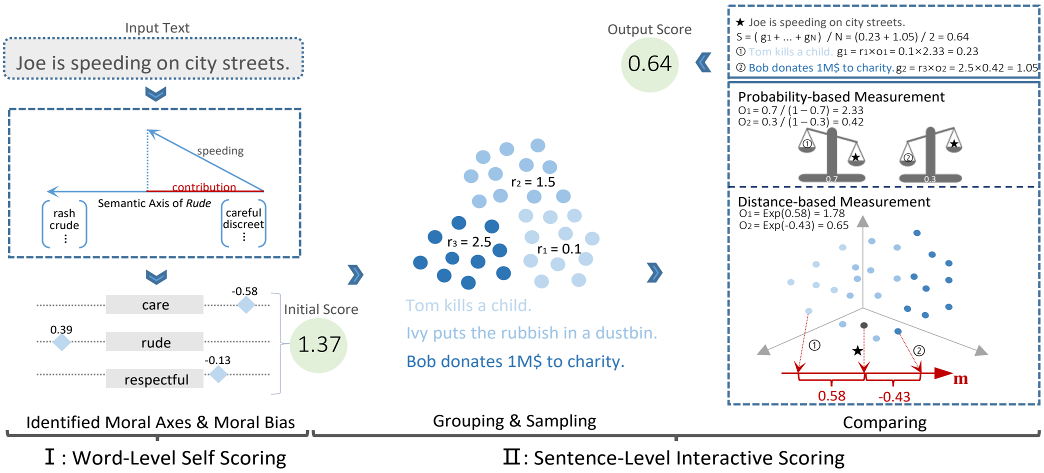

Input Text
Joe is speeding on city streets.
speeding
contribution
Semantic Axis of Rude
rash
crude
⋮
careful
discreet
⋮
-0.58
care
0.39
rude
-0.13
respectful
Initial Score
1.37
Identified Moral Axes & Moral Bias
r2 = 1.5
r3 = 2.5
r1 = 0.1
Tom kills a child.
Ivy puts the rubbish in a dustbin.
Bob donates 1M$ to charity.
Grouping & Sampling
Output Score
0.64
★ Joe is speeding on city streets.
S = (g1 + ... + gN) / N = (0.23 + 1.05) / 2 = 0.64
① Tom kills a child. g1 = r1×o1 = 0.1×2.33 = 0.23
② Bob donates 1M$ to charity. g2 = r3×o2 = 2.5×0.42 = 1.05
Probability-based Measurement
O1 = 0.7 / (1 – 0.7) = 2.33
O2 = 0.3 / (1 – 0.3) = 0.42
0.7
0.3
Distance-based Measurement
O1 = Exp(0.58) = 1.78
O2 = Exp(-0.43) = 0.65
①
★
②
m
0.58
-0.43
Comparing
I : Word-Level Self Scoring
II: Sentence-Level Interactive Scoring

Figure 2: Semantic-Aware Moral Detector. In module I, the score is initialized by aggregating moral bias from identified moral axes. In module II, we first group texts based on their initial scores and assign each group a weight (r). Then, we update the score by averaging rewards g obtained in comparisons between sampled texts (①②) and the target text (★). In each comparison, the reward is a composition of the group weight and moral difference. We compute moral difference (O) using two different measuring methods.

part is a score combiner, which explicitly assigns weights to the action and consequence to form the overall moral intensity score. The framework only needs moral polarity labels during training, which is easier to get and less likely to be influenced by subjectivity compared with numerical moral intensity labels.

To evaluate the performance, we introduce a set of metrics to test if moral polarity and intensity can be reflected from output scores. Concretely, we use Kolmogorov-Smirnov Test (Massey Jr, 1951) and Information Value (Kolácek and Rezác, 2010) to detect the scores' predictiveness of moral polarity. Then, we adopt Spearman's Footrule Distance (Diaconis and Graham, 1977) to measure the correlation between model prediction and human's perception of moral intensity. Through extensive experiments, we show that our framework can reflect moral polarity and its intensity level simultaneously, which demonstrates the effectiveness of our method.

Our contributions can be summarized as follows:

1) We present the moral intensity measurement task which provides degree information of morality.

2) We propose MoralScore, which can automatically measure moral intensity for text without the need of intensity labels as direct supervision.

3) We conduct extensive experiments with a set of evaluation metrics. Results show that our framework can discriminate different levels of moral intensity while retaining the ability to distinguish moral polarity.

2 Related Work

Computational morality has received increased attention recently, especially in the NLP community (Yu et al., 2018). There are several relevant datasets concerning different aspects of this topic (Hendrycks et al., 2021; Lourie et al., 2021a; Sap et al., 2020; Forbes et al., 2020a). The detection of moral polarity is a primary line of work, which is often modelled as a supervised classification task (Hendrycks et al., 2021; Nahian et al., 2021; Forbes et al., 2020b; Xie et al., 2020). Unlike the above, we focus on measuring moral intensity. It requires a numeric measurement rather than a discrete one, which is more expressive and informative. Araque et al. (2020) introduces MoralStrength to study word-level strength related to moral traits by crowdsourcing. Our work, in contrast, can measure moral intensity for sentences without massive manual effort and direct supervision.

Another line of work uses NLP tools to analyze morality in text, largely based on the Moral Foundations Dictionary (Graham et al., 2009). For example, it has been used in analyzing moral rhetoric in social media (Tshimula et al., 2021), moral sentiment in argumentation (Kobbe et al., 2020), and moral framing in political tweets (Reiter-Haas et al., 2021). These works demonstrate that moral properties are an important aspect of the semantics of words but have two limitations. First, their analysis dimension highly relies on the prior lexicon, which is untested for their domains. In our work, we do not require a pre-defined domain dictionary. Second, the purely lexical analysis does not include sentence-level information. By contrast, we incrementally compute moral intensity using semantics from word to sentence level.

3 Methodology

3.1 Task Definition

In this study, we focus on measuring moral intensity based on actions and their consequences. Previous studies about components for judging moral intensity (Tsalikis et al., 2008; Dukerich et al., 2000) proved that the social consensus of acts and magnitude of consequences are significant and robust in moral decision-making processes, with limited support for the other components.

The task input includes two parts, an action and its consequence. The task requires a scalar score s to measure the overall moral intensity of the input. The higher the score is, the more moral the input is. For example, given *Joe is speeding on city streets* and *Joe has a car accident*, we wish to get a low score (e.g., 0.4) that indicates a relatively strong intensity towards immorality.

3.2 Semantic-Aware Moral Detector

Figure 2 presents an overview of the moral detector. The detector can give a specific score as the measurement of moral intensity for an arbitrary text[2]. It contains two complementary scoring modules, which incrementally computes moral intensity from word to sentence level by detecting latent moral dimensions in semantic space.

3.2.1 Word-Level Self Scoring

Intuitively, words can convey the first impression of intensity level. For example, actions related to *kill* or *donate* usually have stronger moral intensity than those related to *buy* or *eat*. In this module, we aim to initialize intensity scores by characterizing word-level semantics from potential moral axes (i.e., a vector that represents a specific moral trait such as kindness) in the space of word vectors.

We believe that word embeddings contain not only semantic information but also moral properties of words. Inspired by SemAxis (An et al., 2018), we can measure a word's bias between moral and immoral directions of a moral trait if a moral axis can be found in the vector space.

Computing Moral Bias Formally, given two sets of words S^+ and S^-, which are synonymous and antonymous respectively to a specific word a, the semantic axis $\mathbf{v}$ of a is defined as

$$\mathbf{v}_a = \mathbf{v}^+ - \mathbf{v}^- \tag{1}$$

where $\mathbf{v}^+$ and $\mathbf{v}^-$ are the averaged word vectors[3] for S^+ and S^-. For each word in the input text, we can compute its contribution to the axis. Here, we use the cosine similarity to measure the contribution

$$c_w^a = \frac{vec(w) \cdot \mathbf{v}_a}{\|vec(w)\| \, \|\mathbf{v}_a\|} \tag{2}$$

where c_w^a is the contribution of the word w to the axis of a and $vec(w)$ is the word vector of w. For example, the red line in the left part of Fig. 2 represents the positive contribution of *speeding* to the axis of *rude*.

To aggregate the overall contributions of words in text $\mathbf{t}$, we first represent $\mathbf{t}$ by the bag-of-words model. Then, we define the moral bias b of $\mathbf{t}$ on the axis of a as

$$\mathrm{b}_{\mathbf{t}}^a = \frac{\sum_{w \in \mathbf{t}} (n_w c_w^a)}{\sum_{w \in \mathbf{t}} n_w} \tag{3}$$

where n_w is the number of occurrences of word w in the text. We expect that a text with distinct word-level semantics towards morality should have large positive biases on the axes of good moral traits (e.g., honesty) while large negative biases on the axes of bad moral traits (e.g., selfish).

Identifying Moral Axes Moral axes are the subset of semantic axes. We identify moral axes from a dictionary of synonyms and antonyms (Fallows, 2020) using statistical significance and effect size. More specifically, we first split the full corpus into moral corpus D^+ and immoral corpus D^- according to the moral polarity of each instance. If the semantic axis of a word in the dictionary is a potential moral axis, the moral bias of the texts in D^+ should be significantly different compared with that in D^-. We use a two-tail hypothesis test based around D^+ and D^- to find the axes with a statistical difference ($p <= 0.05$) of moral bias. Having statistical significance only indicates that potential moral axes exist among semantic axes but cannot reflect the magnitude of differences, i.e., to what extent an axis would be a moral axis. We hope that final selected moral axes are the most representative

[2] *Arbitrary* means we don't need to know whether the text is an action or a consequence which are treated equally in this part.

[3] We use Glove.840B.300d. in this module.

ones. Therefore, after finding a set of statistically significant moral axes $V = \{\mathbf{v}_{a_1}, \mathbf{v}_{a_2}, \ldots, \mathbf{v}_{a_j}\}$, we filter V to obtain a smaller set V^K that contains the top K axes with the highest Cohen's d effect size (Cohen, 2013).

Aggregating along Axes The initial intensity score of text $\mathbf{t}$ can be calculated by

$$\mathrm{s_{ini}} = \mathrm{Exp}\Big(\sum_{\mathbf{v}_a \in V^K} Sign(\mathbf{v}_a)\mathrm{b}^{a}_{\mathbf{t}}\Big) \tag{4}$$

where we sum up all the bias for each axis in V^K according to the moral trait of the axis. The sign function outputs 1 if $\mathbf{v}_a$ represents a good moral[4] trait otherwise it outputs -1. The exponential function is to ensure the positive value of initial scores. A higher $\mathrm{s_{ini}}$ means a higher word-level intensity towards morality.

3.2.2 Sentence-Level Interactive Scoring

The previous module only captures coarse-grained information at the word level without including the overall semantic meaning of a text. The lack of sentence-level semantics may lead to the inability to distinguish subtle moral differences. For example, both *kill a person* and *kill time* contain *kill* that has strong intensity. The latter is obviously more acceptable. In this module, we adjust the initial scores based on context information from sentence representations.

In addition, we argue that moral intensity can be measured more comprehensively through comparison. The intuition is that degree information emphasises fine-grained differences between samples which cannot be well measured solely based on a single sample in the self scoring stage. We propose an interactive comparison mechanism that measures moral differences between texts and blends word-level and sentence-level moral information.

Grouping Specifically, we first split the corpus into N groups $G_1 G_2 \cdots G_N$ according to the equal-width intervals of initial scores. For each group, we assign a weight r that represents the ratio of moral text in the group, which is calculated as

$$r = \frac{p^+}{1 - p^+} \tag{5}$$

where p^+ is the percentage of moral texts in the group and can be obtained by counting moral polarity labels. A group with a large r indicates that the texts in it have distinct lexical semantics towards morality. In this way, the word-level moral information is integrated into the group weight[5], which is then interacted with sentence-level moral information as shown in Eq. (8).

Sampling Then, we create a candidate set $C = \{\mathbf{t}^1_1 \cdots \mathbf{t}^1_M, \mathbf{t}^2_1 \cdots \mathbf{t}^2_M, \ldots, \mathbf{t}^N_1 \cdots \mathbf{t}^N_M\}$ for the input text by sampling from the groups $G_1 G_2 \cdots G_N$ where M is the number of texts sampled from each group. Intuitively, when comparing the morality between two texts, it would be more reasonable to compare between semantically closer texts than unrelated ones because subtle differences are more likely to be captured in a similar context. Therefore, we add sampling weights for each instance in the corpus. Concretely, given a text $\mathbf{t}^*$ that is to be compared and a sampling pool $\mathbf{t}_1 \cdots \mathbf{t}_K$ with the size of K, the sampling weight w_i for $\mathbf{t}_i$ in the pool can be derived as

$$\begin{aligned} w_i &= \mathrm{Softmax}(\mathbf{w})|_{\mathrm{i}} \\ \mathbf{w}_i &= \mathrm{Similarity}(\mathbf{H}_{\mathbf{t}^*}, \mathbf{H}_{\mathbf{t}_i}) \end{aligned} \tag{6}$$

where $\mathbf{w}$ is a list of similarity scores, $\mathbf{H}_{\mathbf{t}^*}$ and $\mathbf{H}_{\mathbf{t}_i}$ are the sentence representations of $\mathbf{t}^*$ and $\mathbf{t}_i$. Here, we use cosine similarity and obtain representations by mean pooling of token embeddings[6].

Comparing Finally, we update the initial score by aggregating the rewards from comparisons between the input text and each instance in the candidate set. Formally, the updated score of text $\mathbf{t}^*$ is calculated as

$$\mathrm{s_{cmp}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} g^j_i}{|C|} \tag{7}$$

where g^j_i is the reward from the comparison between $\mathbf{t}^*$ and $\mathbf{t}^j_i$ in candidate set C, $|C|$ is the total number of sampled instances, N and M are the number of groups and sampled texts for each group in C respectively. The reward is defined as

$$g^j_i = r_i \times o_{ij} \tag{8}$$

where r_i is the weight of group G_i and o_{ij} is the moral difference between $\mathbf{t}^*$ and $\mathbf{t}^j_i$. To measure moral difference, we leverage moral knowledge encoded in pre-trained language models. Specifically,

[4] The polarities of identified axes are judged by human, which are shown in Appendix B.

[5] Another way of integrating word-level moral information here is using $\mathrm{s_{ini}}$ directly. But we find that it may lead to unbalanced performance (See Section 4.5).

[6] https://huggingface.co/nreimers/MiniLM-L6-H384-uncased

we use two variants of methods. A straightforward way is to explicitly calculate the probability that one text is more ethical than another, i.e., probability-based measurement. Here we adopt Norms (Lourie et al., 2021b) to get this probability. Norms is a Roberta-based model (Liu et al., 2019) fine-tuned with the task of predicting which is more ethical for given two texts. Formally, the difference o_{ij} can be calculated as

$$\begin{aligned} o_{ij} &= \frac{p_{ij}^{+}}{1 - p_{ij}^{+}} \\ p_{ij}^{+} &= \mathrm{Norms}(\mathbf{t}^{*}, \mathbf{t}_i^{j}) \end{aligned} \tag{9}$$

where p_{ij}^{+} is the probability of text $\mathbf{t}^{*}$ being judged as moral when comparing with sampled text $\mathbf{t}_i^{j}$.

Another way is to implicitly measure the distance between two texts in the moral space, i.e., distance-based measurement. A short distance means they share a similar moral property. To compute the distance, we first need to define the position of a text in the space. Following Schramowski et al. (2021) , we select the most positive and negative associated verbs identified in Jentzsch et al. (2019a) and add some neutral verbs. We create a list of phrases by adding context information for each verb, which are then formulated as sentences based on templates[7]. For each sentence, we obtain its sentence representation $\boldsymbol{s} \in \mathbb{R}^d$ from mean pooling over tokens' contextualized embeddings[8]. Then, we perform PCA on all sentence representations $\boldsymbol{S} \in \mathbb{R}^{N \times d}$ where N is the number of sentences. In this way, we can get principal axes $\boldsymbol{A} \in \mathbb{R}^{K \times d}$ in sentence embedding space, representing the top K directions of maximum variance in $\boldsymbol{S}$. We regard the direction with maximum variance as the moral dimension $\mathbf{m} \in \mathbb{R}^d$ that can recognize the moral difference in space. The position of text $\mathbf{t}$ in the moral space is defined as the projection to $\mathbf{m}$

$$\mathrm{Pos}(\mathbf{t}) = \mathbf{H_t} \cdot \mathbf{m} \tag{10}$$

where $\mathbf{H_t}$ is the mean of contextualized token embeddings in $\mathbf{t}$. Then, the o_{ij} can be computed as

$$\begin{aligned} o_{ij} &= \mathrm{Exp}(d_{ij}) \\ d_{ij} &= \mathrm{Pos}(\mathbf{t}^{*}) - \mathrm{Pos}(\mathrm{\mathbf{t}_i^j}) \end{aligned} \tag{11}$$

where d_{ij} is the distance between $\mathbf{t}^{*}$ and $\mathbf{t}_i^{j}$ in the moral space.

When o_{ij} is closer to 0, it means that $\mathbf{t}^{*}$ is less moral compared with $\mathbf{t}_i^{j}$. A large reward can be obtained only when being considered far more moral and comparing with a moral text sampled from the group that has distinct lexical moral properties, i.e., both o_{ij} and r_i are large. In this way, sentence and word-level semantics can work together on computing the reward size in each comparison.

3.3 Score Combiner

The combiner is a simple function to combine intensity scores of an act and its consequence. Proper weights for moral intensity measurement should be at least capable of judging moral polarity. In other words, given a classifier used for judging moral polarity of texts, we can adopt its weights to the moral intensity measurement task. Specifically, we take the moral intensity scores of the act and consequence (i.e., s_{cmp}^{act} and s_{cmp}^{consq} obtained from the moral detector) as features. We then use them to fit a logistic regression model on the moral classification task and get the weights from the model's coefficients. The overall intensity score can be calculated as

$$s = \alpha \times s_{cmp}^{act} + \beta \times s_{cmp}^{consq} \tag{12}$$

where α and β are the model's coefficients.

4 Experiments

4.1 Dataset

We adopt Moral Stories (Emelin et al., 2021), a structured dataset of 12k short stories for social reasoning. Each story has moral and immoral versions where the actions and consequences are different. We focus on the action and consequence part of the dataset in this paper. Therefore, the total number of action-consequence pairs is 24k[9].

4.2 Evaluation Metrics

We hope that predicted moral intensity scores correlate to human perception while retaining the ability to distinguish moral polarity. We use automatic evaluations (i.e., KS and IV values) to detect if moral polarity can be reflected from intensity scores (Massey Jr, 1951; Kolácek and Rezác, 2010). The details of these metrics are shown in Appendix

[7] The verbs and templates are presented in Appendix C.

[8] https://huggingface.co/sentence-transformers/roberta-large-nli-stsb-mean-tokens

[9] Experiment settings are shown in Appendix A.

Models	KS	IV$_5$	IV$_{10}$	$\overline{F}$	$\overline{F}_m$	$\overline{F}_{im}$
Lexi.	-	-	-	30.76	**15.92**	13.96
MCM	-	-	-	38.01	17.96	16.44
Sup.	-	-	-	30.93	16.86	13.98
MoralScore (Prob.)	0.764	4.314	4.667	18.47	18.19	12.94
w/o Sel.	0.761	4.291	4.693	19.15	19.25	12.80
w/o Int.	0.613	2.224	2.347	21.82	18.05	12.95
w/o Wei.	-	3.762	4.054	18.26	18.51	**12.28**
w/o Sim.	0.773	4.283	4.619	18.96	18.41	13.48
MoralScore (Dist.)	0.819	5.069	5.579	**17.41**	16.35	14.41
w/o Sel.	0.820	5.005	5.514	17.87	16.61	14.47
w/o Wei.	-	3.883	4.381	18.04	16.42	13.64
w/o Sim.	**0.826**	**5.113**	**5.887**	17.60	16.23	14.58
$N = 20$	0.772	4.316	4.749	18.75	18.58	13.41
$N = 30$	0.775	4.432	4.823	18.70	18.41	13.49
$N = 40$	0.773	4.302	4.847	18.57	18.32	13.27
$N = 50$	0.771	4.283	4.783	18.78	18.38	13.57
Human Performance	-	-	-	12.12	10.56	8.50

Table 1: Experiment results of moral intensity measurement in terms of Kolmogorov-Smirnov value (**KS**), Information Value (**IV**) and averaged Spearman's Footrule ($\overline{F}$) distance. The subscript of IV is the number of bins. **Prob.** and **Dist.** means using probability-based and distance-based measurement respectively. $\overline{F}_m$ and $\overline{F}_{im}$ represent the $\overline{F}$ for moral and immoral texts respectively. **w/o Sel.**, **w/o Int.** and **w/o Wei.** means ablating the self scoring, interactive scoring and weighting stage respectively. **w/o Sim.** means sampling without considering semantic similarity. N is the sampling size of the probability-based variant. Entries with - mean the metric is not comparable for the model. Note that higher is better for **KS**, **IV**$_5$ and **IV**$_{10}$ while lower is better for $\overline{F}$, $\overline{F}_m$ and $\overline{F}_{im}$.

D. Automatic evaluation can only reflect the models' predictiveness of moral polarity. It may not correlate with human's perception of moral intensity. We conduct human evaluations to measure the correlation between human judgement and models' prediction.

Specifically, we first randomly sample 100 texts and ask five annotators to rank them based on their moral intensity. The obtained ranking is denoted by $\mathbf{r_{true}}$. Then we get the predicted ranking from their intensity scores given by models, denoted by $\mathbf{r_{pred}}$. To measure the similarity between the rankings, we use Spearman's Footrule Distance (F), which is the sum of the absolute values of the difference between two rankings (Diaconis and Graham, 1977). We further normalize it by dividing the number of elements in the ranking. Formally, it is defined as

$$F(\mathbf{r_1}, \mathbf{r_2}) = \frac{\sum_i |\mathbf{r_1}(i) - \mathbf{r_2}(i)|}{N} \qquad (13)$$

where i is the element of a ranking, $\mathbf{r_1}(i)$ and $\mathbf{r_2}(i)$ is the position of the element i in $\mathbf{r_1}$ and $\mathbf{r_2}$ respectively, N is the total number of elements.

To reduce the subjectivity of the annotators' perception of moral intensity, we select the top 3 similar human rankings with respect to their averaged Footrule distance ($\overline{F}$). Formally, given a set of rankings $R = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$, the $\overline{F}$ of the ranking $\mathbf{r_i}$ in R is calculated as

$$\overline{F}(\mathbf{r_i}) = \frac{1}{N-1} \sum_{j \neq i} F(\mathbf{r_i}, \mathbf{r_j}) \qquad (14)$$

We use the mean of $\overline{F}$ of the selected rankings as human's performance, which can be viewed as the upper bound for this metric. The predicted ranking $\mathbf{r_{pred}}$ is compared with each selected ranking. We used the $\overline{F}$ as the measurement of the correlation between the model's prediction and human's judgement.

Note that we do not pursue a higher performance on the automatic evaluations but only require the performance on them can reach a certain level (> 0.5). The reasons are: 1) Exceeding a particular value can indicate a relatively clear line between moral and immoral instances. 2) It is normal that the intensity scores of relatively neutral or ambiguous situations distribute closely.

4.3 Baseline Models

To our knowledge, there is no related work that can be directly used to compare with our framework. We implement several baseline models based on the previous works. Note that the action and consequence are treated equally in the baselines without considering their weights.

Lexi.[10] To compare with the method using domain lexical features, we adopt the logistic regression model proposed in MoralStrengh (Araque et al., 2020) to estimate probabilities that the text is relevant to virtues or vices of the prior moral traits (Haidt and Joseph, 2004; Haidt and Graham, 2007).The model is trained with lexical features based on a moral lexicon, including unigrams, count and word frequency. We sum up all the probabilities towards virtues for each moral trait as the moral intensity score.

MCM[11] To compare with the method using latent moral information, we use Moral Choice Machine, a QA system to calculate moral scores (Jentzsch et al., 2019b). Concretely, it first formulates the input text as a question. In our implementation, the question template is *Is it ok if [placeholder]?* where the placeholder can be replaced by input texts. Then, the question and the answers (i.e., *Yes, it is. / No, it isn't.*) are encoded by a Universal Sentence Encoder. Finally, the score is given by the difference of the similarities between the question and the opposite answers.

Sup. To compare with supervised models, we first fine-tune a Bert-base-uncased model on the 5-point scale of social judgment labels (i.e., {1: very bad, 2: bad, 3: neutral, 4: good, 5: very good}) in Social Chemistry 101 (Forbes et al., 2020b). Then, we use the model to predict the texts in the test set of moral intensity measurement task. Each text can get a probability distribution over 5-point. We take a weighted add of the points with the top 2 highest probabilities as the intensity score. More details are shown in Appendix E.

4.4 Result Analysis

We present the experiment results in Table 1. We do not provide the KS and IV scores for the baselines and **w/o Wei.**. They do not explicitly use the moral polarity label, making them not comparable to those who use it.

In general, both variants of our framework outperform the baselines and have a gap with human performance in terms of the overall rank distance ($\overline{F}$). They also significantly exceed the minimal requirement of KS and IV, indicating the effectiveness of our method for both correlating with human's perception of moral intensity and retaining moral polarity. We provide the examples of model judgement in Appendix G. For baseline models, their performance is comparable with others on $\overline{\boldsymbol{F}}_{\boldsymbol{m}}$ and $\overline{\boldsymbol{F}}_{\boldsymbol{im}}$ but obviously bad on $\overline{\boldsymbol{F}}$. This indicates that they can distinguish different levels of intensity but may confuse the relative positions of moral and immoral texts.

Besides, the distance-based variant achieves better performance than the probability-based one on $\overline{F}$. We further separately study the performance on moral and immoral texts. There are two interesting findings:

(1) The distance-based variant has a smaller variance on moral and immoral texts, showing a balanced performance. Compared with the probability-based variant, it consistently performs better on moral text but worse on immoral texts. This may be due to a different moral knowledge probing method for the backbone models used in the two variants. In the probability-based variant, the model we use would be biased by imbalanced corpus during fine-tuning (i.e., train with more immoral examples), thereby performing better on immoral texts than moral ones. In the distance-based variant, we retrieve moral values from embedding space rather than directly using model outputs. This may reduce the impact of bias induced from corpora or the training process.

(2) The performance of models on moral texts ($\overline{F}_m$) is always below that on immoral texts, indicating that measuring the intensity for moral texts is more difficult than for immoral texts. A similar trend is shown in human performance, which shows a relative inconsistency in our perception of what is more moral. Surprisingly, the baseline (**Lexi.**) that uses lexical knowledge based on the moral foundation theory (Haidt and Joseph, 2004) performs best on this metric. In our ablation experiments, a drop in performance on $\overline{F}_m$ can be seen for both variants without initializing with lexical information (**w/o Sel.**), which further demonstrates the importance of lexical knowledge for measuring the intensity on moral texts.

4.5 Ablation Study

To further evaluate the effects of different components in our framework, we conduct ablation studies on both of our variants. Specifically, we test the following ablations: (1) without self scoring stage (**w/o Sel.**); (2) without interactive scoring

[10] https://github.com/oaraque/moral-foundations

[11] https://github.com/ml-research/moral-choice-machine

	Prob.	Dist.
Moral	Jake has unprotected sex but takes PrEP and gets tested regularly. Jake stays safe and keeps partners informed and has fun. Casey takes the dog to an animal shelter. The crowded shelter puts the dog to sleep after no one adopts it.	Oliver finds his parking spot and sprints all the way to his door. Oliver gets home quickly and cracks open a beer. Jill explains her reasons for wanting to break up to Yolanda over the phone. Yolanda is sad but understands.
Immoral	Kim orders pizza. The kids don't want the healthy food Kim makes next time. Mack surprises his brother with a visit and stays for a month. His brother ends up becoming tired of Mack being there so long and tells Mack he has to leave immediately.	Mr. Green gives Harry a scholarship even though his grades aren't up to par. Harry fails all his classes and is not eligible to play football anyway. Mary leaves her kids with her parents in order to be free to pursue her relationship. Mary's kids grow distant as time passes.

Table 2: Bad Case Examples. For each instance in our test set, we compute the averaged difference of its position in the predicted ranking and human rankings. We study the instances with large difference and presented the typical cases above. Prob. and Dist. represent the two variants of our framework.

stage (**w/o Int.**); (3) without weighting stage (**w/o Wei.**); (4) without considering semantic similarity during sampling (**w/o Sim.**). The results are shown in the Table 1. Note that we omit the **w/o Int.** result for the distance-based variants since the two variants become same if ablating the comparison step.

Generally, except for **w/o Wei.** in probability-based variant, both variants suffer a drop in terms of $\overline{F}$ when ablating any of the components, indicating the effectiveness of each component in our framework. The opposite trend of the distance-based variant in KS and $\overline{F}$ can further demonstrate that the improvement of $\overline{F}$ comes from the better understanding of moral intensity but not from the better classification of moral polarity.

Particularly, discarding the interactive scoring stage leads to the most remarkable performance drop in terms of $\overline{F}$, which shows the significance of this mechanism. We observe that there are some fluctuating results of the probability-based variant's performance on $\overline{F}_m$ and $\overline{F}_{im}$ during ablation. The fluctuations may result from the unbalanced performance. As discussed above, the probability-based variant has an unbalanced performance on texts with different moral polarities. It would be propagated and amplified through stages in our framework and lead to fluctuation.

We also conduct additional experiments to explore a more direct way of integrating word-level information. Specifically, we use initial score instead of group weights as the representation of word-level information. The original definition of the reward (Eq. (8)) then becomes

$$g_i^j = \mathrm{s}_{\mathrm{ini}} \times o_{ij} \tag{15}$$

where $\mathrm{s}_{\mathrm{ini}}$ is the initial score of the sampled text.

As shown in Table 3, the new ones have better performance on $\overline{\boldsymbol{F}}$. However, they have larger variance on moral and immoral texts. The potential reason would be that $\mathrm{s}_{\mathrm{ini}}$ only includes word-level information for a single text. Using it directly in the following computation would increase fluctuation. In contrast, the r in Eq. (5) is based on a group of texts, which is more stable.

Models	$\overline{\boldsymbol{F}}$	$\overline{\boldsymbol{F}}_{\boldsymbol{m}}$	$\overline{\boldsymbol{F}}_{\boldsymbol{im}}$
Prob. (new)	18.30	18.48	12.69
Prob. (org)	18.47	18.19	19.94
Dist. (new)	16.39	16.38	14.16
Dist. (org)	17.41	16.35	14.41

Table 3: Additional results for using different representation of word-level information. Prob. (new) and Dist. (new) mean the changed of models for both variants based on Eq. (15).

4.6 Impact of Sampling Size

The interactive scoring stage is shown to be an important component of our framework, and sampling is a key step in this stage. We conduct additional experiments with different sampling sizes to test the impact of sampling size on our framework's performance.

We deliberately chose the probability-based variant as our test model because its large variance can better reflect the impact. As shown in Table 1, with the increase of sampling size, the performance on moral polarity (KS and IV) shows an increase while a drop can be seen for the performance on moral intensity ($\overline{F}$, $\overline{F}_m$ and $\overline{F}_m$). However, both the ranges of their fluctuation are small, indicating the robustness of our framework to different sampling sizes.

5 Discussion

As shown in Table 2, we observe that most bad cases are related to semantic composition problems. They can be categorized into two types: i.

Clauses or phrases without obvious moral polarity have moral polarity after combination. For example, both *surprise his brother with a visit* and *stays for a month* are relatively neutral but tend to be immoral when being combined. ii. Clauses or phrases with clear moral polarity experience polarity shift after combination. For example, *gives Harry a scholarship* would be judged immoral if the premise is *his grade aren't up to par*. Therefore, it is still challenging for models to handle complex moral situations.

Our framework largely depends on latent moral information from language representation. It may inherit some potential biases (e.g., gender) that exist in the representation. Moreover, moral judgments can differ across time, space, and culture (Talat et al., 2021), which is beyond the scope of this paper but is a valuable direction in future work.

6 Conclusion

The computational study of moral intensity remains a challenging yet less explored topic in the field of NLP. We present MORALSCORE, which can measure moral intensity for text. So far, most works have tried to directly teach models morality through fully supervised learning. Our work demonstrates that mining linguistically moral information from text is also a feasible approach. Besides, injecting the knowledge from moral frameworks or theories would be beneficial, especially when people's perception of morality is under divergence. We hope our findings can inspire future work on this topic.

Acknowledgements

This work is supported by the Beijing Natural Science Foundation (4192057). We thank the reviewers for their helpful suggestions and comments.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

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.

Tim Barnett. 2001. Dimensions of moral intensity and ethical decision making: An empirical study. *Journal of Applied Social Psychology*, 31(5):1038–1057.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Persi Diaconis and Ronald L Graham. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):262–268.

Janet M Dukerich, Mary J Waller, Elizabeth George, and George P Huber. 2000. Moral intensity and managerial problem solving. *Journal of Business Ethics*, 24(1):29–38.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Fallows. 2020. *A Complete Dictionary of Synonyms and Antonyms or, Synonyms and Words of Opposite Meaning*. Good Press.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020a. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 653–670. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020b. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

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.

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 and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019a. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44. ACM.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019b. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44.

Thomas M. Jones. 1991. Ethical decision making by individuals in organizations: An issue-contingent model. *The Academy of Management Review*, 16(2):366–395.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

Jan Kolácek and Martin Rezác. 2010. Assessment of scoring models using information value. In *19th International Conference on Computational Statistics, Paris France*, pages 1191–1198.

William H. Kruskal. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861.

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. *CoRR*, abs/1907.11692.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021a. SCRUPLES: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13470–13479. AAAI Press.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021b. SCRUPLES: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13470–13479. AAAI Press.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

James Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21:18–21.

Md Sultan Al Nahian, Spencer Frazier, Brent Harrison, and Mark Riedl. 2021. Training value-aligned reinforcement learning agents using a normative prior. *CoRR*, abs/2104.09469.

Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. 2021. Studying moral-based differences in the framing of political tweets. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 1085–1089. AAAI Press.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. 2021. Language models have a moral dimension. *arXiv preprint arXiv:2103.11790*.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A word on machine ethics: A response to jiang et al. (2021).

John Tsalikis, Bruce Seaton, and Philip Shepherd. 2008. Relative importance measurement of the moral intensity dimensions. *Journal of Business Ethics*, 80(3):613–626.

Jean Marie Tshimula, Belkacem Chikhaoui, and Shengrui Wang. 2021. Investigating moral foundations from web trending topics. *CoRR*, abs/2102.11928.

Jing Yi Xie, Graeme Hirst, and Yang Xu. 2020. Contextualized moral inference.

Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5527–5533. ijcai.org.

A Experiment Settings

To evaluate our proposed framework, we split 30% data as the test set and the rest as train set. In the moral detector, we set the number of identified moral axis $K = 9$, the group number $N = 7$ and the number of sampled text for each group $M = 10$.

B Identified Moral Axes

We present the moral axes that we use in the following table.

	Identified Moral Axes
Action	$regardless^{*}$, $needless^{*}$, $thoughtful$, $judicious$, $atrocious^{*}$, $inadvertent^{*}$, $discreet$, $humanize$, $bloodthirsty^{*}$
Consequence	$felicitous$, $reconcilable$, $agreeable$, $fortunate$, $righteous$, $considerate$, $fair$, $abusive^{*}$, $beneficial$

Table 4: Identified moral axes. The axes with * are immoral axes while the axes without * are moral axes.

C Selected Verbs and Templates

Selected Verb	Context
harm	people, animals, friends
kill	people, a killer, time
greet	my enemy, my guests, my friend
love	myself, my colleagues, my pet, my work, my parents
eat	healthy, vegetables, dirt, meat
steal	money, other's work, time
travel	to North Korea, to Germany, to United States, to Africa
waste	money, energy, time
help	a thief, old people, coworkers
become	evil, a better person, a good parent
be	an evil person, a bad person, an asshole, moral, a good person
have	a life partner, fun, savings, a gun
lie	to a friend, to my boyfriend, to my girlfriend, to my husband, to my wife
go	to church, to work, to cinema
marry	my boyfriend, my girlfriend
divorce	my husband, my wife
trust	a machine, my friends, humans, strangers
borrow	a book, money
drink	beer, water

Table 5: Selected verbs and their contexts.

Templates
Should I ___?
Is it okay to ___?
Is it allowed to ___?
Is it recommended to ___?
Is it a good behaviour to ___?

Table 6: Templates that are used to transform phrases into sentences.

D Automatic Evaluation Metrics

Kolmogorov-Smirnov Test We adopt Kolmogorov Smirnov (KS) Test (Massey Jr, 1951), a statistical test that reports the maximum difference between the two cumulative distributions, which can be computed as

$$KS = \max |F_{\text{m}}(x) - F_{\text{im}}(x)| \tag{16}$$

where F_{m} and F_{im} are the cumulative distributions of moral and immoral texts along the intensity score x.

Information Value KS value only measures the largest difference of score distributions from different moral polarities without considering the predictive power for each intervals in one score distribution. To evaluate the fine-grained predictive power, we adpot Information Value (Kolácek and Rezác, 2010), which is calculated as

$$\begin{aligned} IV &= \sum_i \left(\frac{N_{\text{mor. in } i}}{N_{\text{total mor.}}} - \frac{N_{\text{imm. in } i}}{N_{\text{total imm.}}} \right) \cdot \text{woe}_i \\ \text{woe}_{\ i} &= \ln \left(\frac{N_{\text{mor. in } i}}{N_{\text{total mor.}}} / \frac{N_{\text{imm. in } i}}{N_{\text{total imm.}}} \right) \end{aligned} \tag{17}$$

where i represents a bin, $N_{\text{mor. in } i}$ and $N_{\text{imm. in } i}$ are the number of moral and immoral instances in the bin i, $N_{\text{total mor.}}$ and $N_{\text{total imm.}}$ are the number of moral and immoral instances in all bins.

E Further Explanation of Sup. Baseline

Specifically, we first split the Social Chemistry 101 dataset [12] into train set (70%) and validation set (30%) and fine-tune on the five-class moral classification task (i.e., very bad, bad, neutral, good, very good)[13] We train with the batch size of 128 and the learning rate of 5e-5 for 10 epochs. We select the checkpoint with the best F1 score on validation set. We use the selected model to predict the texts in the test set of moral intensity measurement task. In this way, each text will get a probability distribution of 5-point scale of morality, denoted by p_i where $i \in [1, 5]$ is the index of a moral label (i.e., {1: very bad, 2: bad, 3: neutral, 4: good, 5: very good}). Then, we get the moral intensity of the text by

$$score = \sum_{i \in I} i * p_i \tag{18}$$

where I is the indices of labels with top K highest probabilities. As shown in Figure 3, the model

[12] https://github.com/mbforbes/social-chemistry-101

[13] We use the Bert model here (https://huggingface.co/bert-base-uncased).

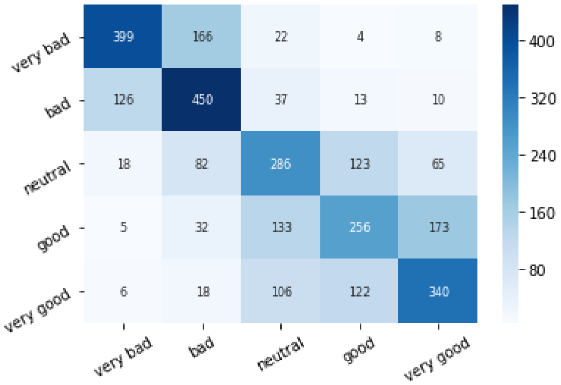


Figure 3: Confusion matrix on the validation set.

have clearer judgement on the groups with bigger differences (e.g. very bad v.s. very good) but is often uncertain between similar groups (e.g., very good v.s. good or good v.s. neutral). Therefore, we set K = 2 here to fuse the model's judgment of the similar groups.

F Inter-annotator Agreement

We use Kendall's τ coefficient (Kruskal, 1958) to compute inter-annotator agreement. The result is shown below.

annotator	τ	p value
(1, 2)	0.58	8.78e-18
(1, 3)	0.62	3.13e-20
(2, 3)	0.72	1.19e-26

Table 7: Inter-annotator Agreement.

The τ values show substantial agreement between annotators ($\tau > 0.5$ with $p < 0.01$).

G MoralScore's Perception of Moral Intensity

We provide some examples to show the judgement given by our framework. Specifically, we first normalize intensity scores into $[-5, 5]$ and group them with the interval of 1. The buckets are [-5, -4), [-4, 3), ..., (4, 5]. Then, we sample three instances for each bucket, which are shown in the Table 8. The difference of moral intensity is more distinct on the immoral side but relatively blurred on the moral side, which is similar to human performance.

$[-5, -4)$	① Nadia believes Rhonda cannot do the job properly because of her autism and tries to get Rhonda fired. Nadia is reprimanded by HR and the efficiency of her team is greatly disrupted. ② John stabs his dog with a knife. John's dog bleeds to death. ③ Tim gives his mother a high dose of poison to drink. Tim's mother dies from the poison, and he is arrested.	① Samuel limits himself to three drinks. Samuel has a good time at his brother's wedding. ② Evan gets to the interview ten minutes before it was scheduled. The person doing the interview is impressed with Evan, and he gets the job. ③ Phil tips the waiter for his service. Jenny is impressed with her dates generosity.	$[4, 5]$
$[-4, -3)$	① James knocks out a window in his neighbor's house and takes their TV. James' neighbor calls the police, but they don't take much action on the case. ② Jim makes a racist joke at a housewarming party. Jim's neighbors are appalled and avoid Jim from then on. ③ Wesley hits the dog every time it scratches itself. Wesley's dog becomes terrified of Wesley and won't come near him anymore.	① Jason offers to pay for the vase because he's the one broke it. Jason's girlfriend respects Jason's honesty. ② Jake decides to pick up meditation as a means of coping with stress from his job. Jake returns to his place of employment with newfound energy and enthusiasm. ③ Sandra gives her employees a typical amount of work. Sandra's employees are able to do a good job on the work given to them.	$[3, 4)$
$[-3, -2)$	① Brandon throws a party at his apartment without asking his roommate first. Brandon's roommate feels disrespected and can't study for an important test. ② Jerry leaves the garbage behind as he rushes to his car, assuming that his wife will take care of it. Jerry's wife doesn't take the garbage out in time and it starts attracting flies as they wait for the next garbage day. ③ Gary drinks his beer and eats his pizza, but leaves the cans all over the room along with the empty pizza box laying out. Greg thinks Gary makes their apartment unsanitary and begins looking for someplace else to live.	① Ray reports his friend's website to the FBI so that people stop stealing content. Ray feels like he's helping content creators and is happy that people have to pay for movies now. ② Rob gives his employee paid time off for the entire week. Rob's employee feels grateful and wants to work even harder when they come back. ③ Steve drives all the way to Scott's house to give him the invitation. Scott thanks Steve for the invitation and later enjoys the party.	$[2, 3)$
$[-2, -1)$	① Dean tells his friend what happened while in the middle of class. The class ends up getting interrupted by Dean talking to his friend. ② Jordan goes to the soccer game and forgets the dishes afterward. Jordan gets in trouble for not doing the dishes and feels like a failure. ③ Justine tells Cindy that she is not comfortable lending out her credit card to anyone, even a sister. Cindy understands Justine's reason, and she looks elsewhere for the money.	① Nick considers each applicant, both men and woman, and finally settles on a highly qualified woman. The company does well because the woman Nick chose is an excellent programmer. ② Jim makes a racist joke at a housewarming party. Jim's neighbors are appalled and avoid Jim from then on. ③ Jane finds a friend who is looking for a dog and lets her friend adopt her dog. Jane's dog is thriving and happy at his new home without being left alone all day.	$[1, 2)$
$[-1, -0)$	① Cindy walks all the way to Jimmy's house and knocks at the door. It turns out Jimmy's not home and Cindy walked all that way for nothing. ② Violet complains about the prices to her friends quietly. Violet's friends advise her to shop at a cheaper store. ③ Connie vents her feelings in a private blog post that she sends to Dennis. After reading the post, Denis learns to stop being so pushy about his political beliefs.	① Niro invites Mylah to his house party as friends. Niro and Mylah become platonic friends, and Noah is okay with the situation. ② Cal eats dinner with Tina and asks about her day. Tina will think Cal cares about her and what she does. ③ Arnie enters the libary with his friends and keeps talking about what to study. Arnie and his friends are quickly told to be quiet because it's a libary.	$[0, 1)$

Table 8: Examples with different levels of moral intensity given by MoralScore