Morality in the Mundane: Categorizing Moral Reasoning in Real-Life Social Situations

Ruijie Xi, Munindar P. Singh
North Carolina State University
Raleigh, North Carolina 27606 USA
rxi@ncsu.edu, mpsingh@ncsu.edu

Abstract

Moral reasoning reflects how people acquire and apply moral rules in particular situations. With social interactions increasingly happening online, social media provides an unprecedented opportunity to assess *in-the-wild* moral reasoning. We investigate the commonsense aspects of morality empirically using data from a Reddit subcommunity (i.e., a subreddit), *r/AmITheAsshole*, where an author describes their behavior in a situation and seeks comments about whether that behavior was appropriate. A commenter judges and provides reasons for whether an author or others' behaviors were wrong.

We focus on the novel problem of understanding the moral reasoning implicit in user comments about the *propriety of an author's behavior*. Specifically, we explore associations between the common elements of the indicated rationale and the extractable social factors. Our results suggest that a moral response depends on the author's gender and the topic of a post. Typical situations and behaviors include expressing *anger* emotion and using *sensible* words (e.g., f-ck, hell, and damn) in *work*-related situations. Moreover, we find that commonly expressed reasons also depend on commenters' interests.

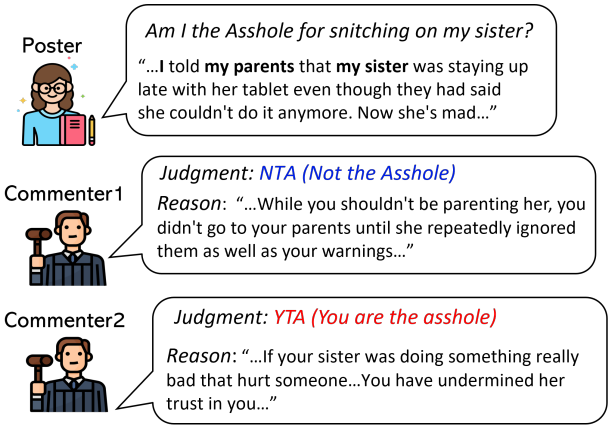


Figure 1: Sample post with comments where the final verdict (Not the Asshole) is decided by majority vote from the commenters. The post involves three parties–*I*, *my parents*, and *my sister*. Commenters provide judgments and reasons about whether the author's behavior was inappropriate.

Introduction

Moral reasoning concerns what people ought to do, which involves forming moral judgments in social or other situations (Richardson 2018). Researchers have extensively studied moral reasoning for investigating moral development in groups organized by elements of social identity, based on genders (Bussey and Maughan 1982), age (Walker 1989), and profession (Wood et al. 1988). These laboratory studies are primarily conducted using questionnaires and hypothetical social situations that make the conflicts between moral principles stark. However, real-life situations are nuanced and complex and often present a wide variety of comparatively low-stakes decisions. Social media provides an opportunity to assess the perception of normal social situations, such as understanding others' decisions on (im)morality of behaviors (Lourie, Bras, and Choi 2020).

In this work, we study in-the-wild moral reasoning by examining a popular subcommunity of Reddit (i.e., a subreddit) called /r/AmITheAsshole (AITA).[1] In AITA, a user (i.e., *author*) posts interpersonal conflicts seeking others' opinions on whether their behaviors were appropriate. Other community members (i.e., *commenters*) may comment on a post to provide moral *judgments* (i.e., a verdict or other moral assessment and a justification thereof) indicating *who is an a-hole*. Figure 1 shows a post in AITA along with comments on it.

AITA defines a few verdict codes: Of these, **NTA** indicates the author's behavior is appropriate, and **YTA** indicates the behavior is inappropriate. A verdict on a post is the verdict in the top-voted comment. Recent works focus on predicting the verdict (on a post) and the comments received by a post (Lourie, Bras, and Choi 2020; Zhou, Smith, and Lee 2021). Another line of work analyzes the community using statistical methods (Botzer, Gu, and Weninger 2022; Nguyen et al. 2022; De Candia et al. 2022; Xi and Singh 2023; Giorgi et al. 2023).

This paper focuses on the commonsense aspects of moral reasoning. To the best of our knowledge, no work has systematically analyzed the moral rationales implicit in the comments. We investigate how the authors' and com-

[1] https://www.reddit.com/r/AmItheAsshole/

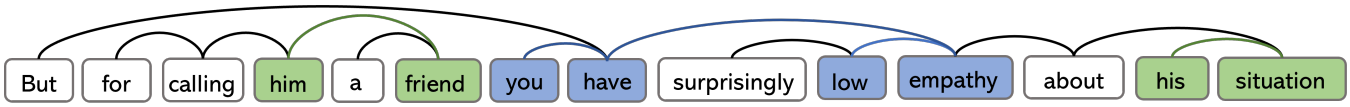


Figure 2: Dependency graph representation of an example comment. The shading shows syntactic relations.

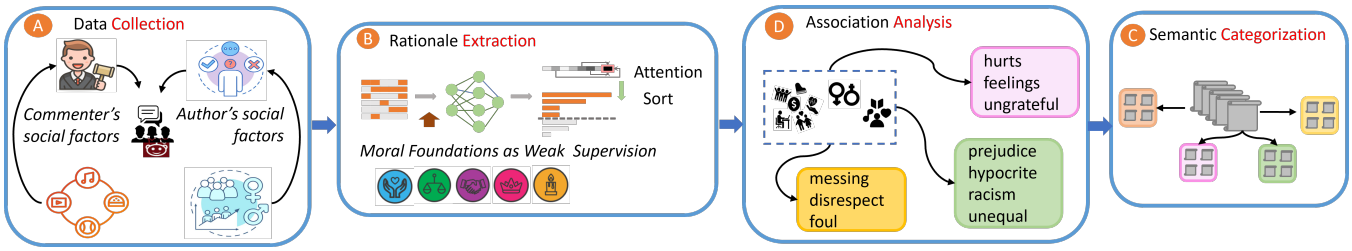


Figure 3: Flowchart depicting our research pipeline.

menters' social factors, as defined below, shape their distributions and affect moral reasoning. An author's social factors include their *self-reported genders* and the post's *topics*. We regard a post topic as a part of the author's social factors because the topic often provides social information about the author, such as whether the author has had conflicts in marriage. As in previous work, we focus on authors' self-reported genders, not the genders of others involved (Botzer, Gu, and Weninger 2022; De Candia et al. 2022). We extract authors' social factors from the posts. For commenters, we take their *interests* as the relevant social factors and we we leverage the subreddits in which they participate as proxies for their interests (De Candia et al. 2022). Extracting social factors from social media submissions has been extensively studied from various viewpoints, such as language bias (Ferrer et al. 2020) and contentious conversations (Beel et al. 2022).

The contextual content in the rationale can influence the verdict. For instance, in Figure 2, the phrase *low empathy* refers to the author's behavior and influences a **YTA** verdict. With a large corpus, these verdict-influencing factors (i.e., **rationales**) would accumulate and reveal the common elements implicit in reasoning about specific social situations. Therefore, we reformulate our task as building a computational *predict-then-extract* model for categorizing the common elements of the moral reasoning embedded in the comments. Figure 3 describes our research pipeline. Our proposed method involves predicting commenters judgments' verdicts and extracting the rationales, clustering them, and analyzing their associations with the above-mentioned factors.

Prediction As discussed above, we distinguish the rationales that refer to authors from those that refer to others. Therefore, we consider the meaning and syntactic features (as shown in Figure 2) in references to various parties. Specifically, we build a dual-channel context-feature extractor to obtain the global and local context features of sentences in the original post. We evaluate our method in terms of its prediction performance.

Extraction We extract rationales from the comments using a rationalization process (Lei, Barzilay, and Jaakkola 2016; Bastings, Aziz, and Titov 2019; DeYoung et al. 2020). The selected rationales are small but sufficient parts of the input text that *accurately* (Jain et al. 2020) identify the most important information actually used by a neural model. Unlike previous works, we assume no human-annotated labels for rationales on social media data. Therefore, following Jiang and Wilson (2021), we "weakly" label rationales using a domain-related lexicon—here, the Moral Foundation Theory (MFT) (Haidt and Graham 2007) lexicon (Hopp et al. 2020). We then evaluate multiple methods to select plausible rationales.

Categorization and Analysis We apply k-means clustering (Lloyd 1982) on the embedding vectors of the rationales and categorize their meaning commonality using a meaning analysis system. Finally, we perform fine-grained analysis on the resulting meaning clusters.

Findings and Contributions To the best of our knowledge, this is the first study to investigate moral reasoning in AITA. Through 51,803 posts and 3,675,452 comments, we find meaning commonalities associated with the authors' and commenters' social factors. For example, female authors attract moral judgments expressing *angry* and *egoism* in *work*-related scenarios, while *politics* and *sensible* (e.g., f-ck, hell, and damn) are less likely present in such judgments. In addition, in *safety*-related situations, comments about *judgment of appearance* are more prevalent for female authors, whereas *physical/mental* (e.g., racist, homophobic, and misogynistic) are less likely to appear. Moreover, commenters interested in the *art* and *music* subreddits (e.g., r/AccidentalRenaissance) express more emotions such as *worry, concern, and confident*, than those interested in *news and politics*.

Our proposed model shows a 3% improvement in all averaged scores (F1, precision, and recall) over fine-tuned

BERT in predicting comments' verdicts. Moreover, our experiments demonstrate that additional domain knowledge improves a rationale's plausibility. The results indicate that our framework is effective in automatically understanding multiparty online discourses. Our framework is applicable in categorizing dynamic and unpredictable online discourse. For instance, the framework can be applied in automated tools, such as for moderating rule-violating comments. We have released our data, code, and supplementary material.[2]

Related Work

Moral Reasoning in Social Psychology Moral reasoning has long been studied in psychology. Bussey and Maughan (1982) find that moral decisions by males are typically based on law-and-order reasoning, whereas those by females are made from an emotional perspective. Walker (1989) observe that participants' discussions about moral situations show clear age developmental trends over a two-year period. Wood et al. (1988) report that individualism and egoism have a stronger influence on the moral reasoning about business ethics by professionals than by students. However, these studies do not provide a comprehensive understanding of moral reasoning on social media.

Morality in Social Media Social media helps ground descriptive ethics. Zhou, Smith, and Lee (2021) profile linguistic features and show that the use of the first-person passive voice in a post correlates with receiving a negative judgment. Nguyen et al. (2022) give a taxonomy of the structure of moral discussions. Lourie, Bras, and Choi (2020) predict (im)morality using social norms collected from AITA. Forbes et al. (2020) extract Rules of Thumb (RoT) from moral judgments of one-liner scenarios. Emelin et al. (2021) study social reasoning by constructing a crowdsourced dataset including moral actions, intentions, and consequences. Jiang et al. (2021) predict moral judgments on one-line natural language snippets from a wider range of possibilities. Ziems et al. (2022) build conversational agents to understand morality in dialogue systems.

Genders, Topics, and User Factors Gender differences are often relevant. De Choudhury et al. (2017) reveal significant differences between the mental health contents and topics shared by female and male users. De Candia et al. (2022) find young and male authors are likelier to receive negative judgments in AITA and society-relevant posts are likelier than romance-relevant posts to receive negative moral judgments. Giorgi et al. (2023) extract linguistic features of AITA posts and find that a positive tone reduces blame. Ferrer et al. (2020) find Reddit post topics are gender-biased; for instance, *judgment of appearances*-related posts are associated with females while *power*-related posts are associated with males. Collecting personal information by using users' submissions on online platforms is a common method to explore social media data, such as investigating conversation divisiveness through Reddit (Beel et al. 2022). Guo, Zhang, and Singh (2020) find that different topics attract different demographic bases in online arguments. Haque and Singh (2024) show that news topics and sentiment content provide useful insights into how public opinion varies.

[2]https://zenodo.org/record/7850027#.ZEGCCnaZM2w

Data

Reddit discussion structure is of a tree rooted at an initial post; comments reply to the root or to other comments.

Definitions We adopt definitions from Guimaraes and Weikum (2021) to describe instances in our dataset.

A post refers to the starting point in a discussion.

A top-level comment refers to a comment that directly replies to a post.

We focus on top-level comments because other comments in AITA may not include verdict codes based on the posts.

Collection of Posts and Comments

We require a large corpus with relevant posts and comments. Previous datasets are either nonpublic (Zhou, Smith, and Lee 2021; De Candia et al. 2022; Botzer, Gu, and Weninger 2022) or insufficient for our purposes (Lourie, Bras, and Choi 2020; Nguyen et al. 2022). Therefore, we collected our dataset using PushShift API[3] and Reddit API.[4] We scraped over 351,067 posts and the corresponding 10.3M top-level comments from AITA, spanning from its founding in June 2013 to November 2021. We collected these submissions by applying rule-based filters following the aforementioned previous works to ensure their relevance and avoid discrepancies between data from Reddit and archived data from PushShift. We excluded deleted posts and comments because they may violate AITA rules, such as by including fake content to solicit outrage. We also excluded posts and comments submitted by deleted accounts and moderators. We selected posts that have at least ten top-level comments to ensure quality. We selected top-level comments that have a predefined code indicating the judgment and fifteen or more characters representing the rationale. Reddit allows users to give positive and negative feedback to submissions in the form of *upvotes* and *downvotes*. Therefore, posts and comments in our dataset are associated with a *score*[5] representing the accumulated differences between upvotes and downvotes.

Extraction of Comments' Verdicts The judgments are predefined codes: YTA (author's behavior is inappropriate), NTA (author's behavior is appropriate), ESH (everyone's behaviors are inappropriate), NAH (everyone's behaviors are appropriate), and INFO (more information needed). Some comments use short phrases as codes (e.g., not the a-hole instead of NTA). Therefore, we applied regular expressions to match such variants. We resolved multiple matches by selecting the second match when there is a transition word such as *but*. And, we reversed the extracted codes in judgments containing negations such as *I do not think* using

[3]https://github.com/pushshift/api

[4]https://www.reddit.com/dev/api

[5]Score is the difference between upvotes and downvotes reported by Reddit: https://www.reddit.com/wiki/faq/#wiki_how_is_a_submission.27s_score_determined.3F

regular expression. We removed sentences marked with >, which indicates a quotation. To evaluate the labeling process, we checked a random sample of 500 submissions. We found 5% false positives and 6% false negatives. Following Lourie, Bras, and Choi (2020), we assigned labels to comments with YTA as 1, NTA as 0, and discard all other instances.

Comment Corpus

Our corpus selection criteria require that selected comments: (1) have scores higher than 100, (2) have a token length between 20 and 200, (3) have commenters who were previously awarded by a *flair* (i.e., to select comments submitted by reputable users), and (4) replied to posts that contain authors' self-reported genders. A flair is awarded by AITA and represents how many times a user's judgments have become the most upvoted comments, thus, reflecting the commenter's reputation. As a result, our corpus includes 51,803 posts and 120,760 out of 3,675,452 total comments that belong to the selected posts. The label distribution of NTA to YTA is 60–40. We randomly selected 45,505 instances labeled as 0 and all instances (i.e., 45,505) labeled as 1. We split our corpus as 80/10/10 for training, development, and testing. Table 1 summarizes our dataset.

	Total	NTA	YTA	Mean # Tokens
Training	72,808	36,405	36,405	184
Development	9,101	4,550	4,550	162
Testing	9,101	4,550	4,550	178

Table 1: Dataset summary.

Method

This section introduces the processes of extracting *social factors*, *verdicts*, and *rationales*. There are two advantages to using rationalization for summarizing common patterns of moral reasoning: (1) rationalization can be trained with neural networks in an unsupervised manner (DeYoung et al. 2020),[6] and (2) it provides appropriate rationales for social media data (Jiang and Wilson 2021).

Extraction of Topics, Genders, and Interests

We adopt Nguyen et al.'s (2022) topic model (with topics named by experts) to identify topics for posts in our corpus. Then, we use regular expressions to extract authors' self-reported genders. We leverage commenters' participation on Reddit to proxy their interests.

Topic Modeling for Posts' Topics Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is widely applied for clustering text. Nguyen et al. (2022) find 47 named topics in AITA posts via LDA models. These topics are associated with clusters of words sorted by the probability of belonging to that topic. We found that our corpus and Nguyen et al.'s (2022) corpus have 34,098 posts in common; the remaining 17,705 posts were submitted after April 2020 (the ending time of their dataset). We follow their method to assign each post the topic that has the highest prior probability.

Authors' Genders Extracting demographics from social submissions using regular expressions is common in analyzing Reddit data, e.g., (Beel et al. 2022). Gender and age are not typically available on Reddit, allowing for anonymous posting. Fortunately, the social media template for posting gender and age, e.g., [25f] (25-year-old female) enables us to use regular expressions to extract the information. Note that authors typically report the demographic information of multiple parties in the situation described, such as *I [25m] and my wife [25f].* Therefore, we extract authors' self-reported genders by filtering first-person pronouns (i.e., I). Besides, we consider gendered alternatives where available; for example, male can be estimated by `\b(boy|father|son|...)\b`) and female by (`\b(girl|mother|daughter|...)\b`). We do not match nonbinary genders because we do not have ground truth labels for nonbinary targets. As a result, we find the female to male split of 90–100 in our dataset. We took a random sample of 300 submissions to evaluate our regular expressions. We found that gender extracted using our regular expression matches the manually labeled one 94% of the time.

Commenters' Interests Following De Candia et al. (2022), we proxy commenters' interests via the subreddits they participated in by making at least one submission (i.e., post or comment) within a six-month period (three months before and after the comment timestamp) based on the timestamp of the comment found in our corpus. We focus on the commenters because they have made quality judgments. We chose a six-month window based on users' prolificity Beel et al. (2022): a user is prolific if they submit more than 25 posts or comments in their subreddit of interest. We manually checked 100 users and found that a six-month timeframe prompts greater user activity compared to four months, and is closely comparable to twelve months in terms of prolificity. We discard deleted user accounts, which restricts our analysis to 46,519 commenters with 104,915 comments. Unlike De Candia et al. (2022), we map the collected subreddits following Reddit's predefined subreddit categories.[7] Commenters may have interests in various categories. Therefore, we set their interest as the subreddit to which they submit posts or comments most frequently.

Predicting Verdicts and Extracting Rationales

We now introduce the rationalization process, followed by how our predict-then-extract model operates.

Introduction to Rationalization Given a pretrained model $\mathcal{M}$, each instance is of the form of (x, y), where $x = [x^i]$ are the input tokens and $y \in \{0, 1\}$ is the binary

[6] Our social media data is inherently without human annotated rationales as in previous works (Jain and Wallace 2019; Jain et al. 2020; Atanasova et al. 2020).

[7] https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits/

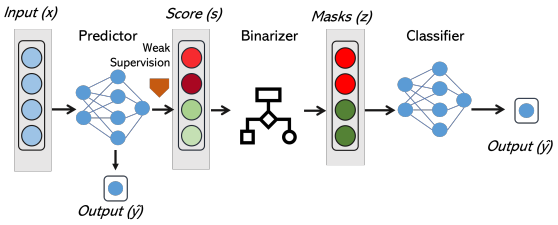


Figure 4: Soft rationalization is a three–phased process. The predictor outputs $\hat{y}$ and importance scores s. The binarizer assigns masks to tokens z. The classifier predicts y again with unmasked tokens to evaluate a rationale's accuracy.

label. Rationalization outputs a predicted $\hat{y}$ with a binary mask $z = [z^i] \in \{0, 1\}$ of input length, indicating which tokens are used to make the decision (i.e., $z^i = 1$ if the ith token is used). The tokens masked with ones are called rationales (R). They are considered accurate explanations of the model's decisions and can be used alone to make correct predictions (Jain et al. 2020). Binarization methods are hard and soft according to (DeYoung et al. 2020). Hard selection uses the Bernoulli distribution to sample binary masks (i.e., $z \sim$ Binarizer(x)). In contrast, soft selection (Jain et al. 2020) outputs multivariable distributions over tokens derived from features, e.g., self-attention values. We adopt **soft selection** because hard selection faces performance limitations (Jain et al. 2020) and soft selection is more appropriate when there is no ground truth of rationales (Jiang and Wilson 2021).

Prediction then Extraction Figure 4 illustrates the architecture of a soft rationalization model.[8] The predictor in Figure 4 is a standard text classification module that predicts a verdict. We omit the last classifier module because we need the rationales instead of accurately predicting y. The importance scores z are computed via feature-scoring methods using the parameters (e.g., gradients) learned during training. Therefore, the extracted rationales can capture the most salient contextual information used by a neural model when predicting a verdict.

We aim to empirically collect plausible rationales for categorizing the commonality of moral reasoning reflected in social media, instead of building an accurate predictor (Botzer, Gu, and Weninger 2022) or improving rationalization extraction (Atanasova et al. 2020; Chrysostomou and Aletras 2022).

Figure 5 shows our predictor. We first weakly label tokens that appear in the moral lexicon. We then obtain embeddings of input instances by adopting the pretrained `bert-base-uncased` model using Huggingface.[9] Next, we prepare global and local representations of a sentence by a stacked Bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997) and a Syntactic Graph Convolutional Network (SGCN) (Bastings et al. 2017; Li et al. 2021).

[8]Compared to Lei, Barzilay, and Jaakkola (2016) we simplify the names of *encoder* as *predictor* and *generator* as *binarizer*. And we name *extractor* (Jain et al. 2020) as *binarizer*.

[9]https://huggingface.co/

Then, we feed the concatenated final hidden representation vectors into a fully connected prediction network. The prediction network uses softmax to output the probabilities of a particular verdict. We adopt cross-entropy in the network to measure loss.

Global context features are multidimensional embeddings encoded using BERT (Devlin et al. 2019), which maps a token into a vector based on its context. We adopt the pretrained `bert-base-uncased` model from Huggingface to obtain embeddings. We use a Bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber 1997) to obtain extended contexts. We compute the hidden states by passing the BERT-encoded embeddings to a stacked BiLSTM:

$$\overleftarrow{h_{g,i}}; \overrightarrow{h_{g,i}} = \text{BiLSTM}(S), i = 1, 2, \quad (1)$$

where S represents the encoding output of the last layer of BERT and i denotes the direction. We compute the global context representations $h_{g,1}$ and $h_{g,2}$ by averaging the hidden outputs in both directions.

Local context features are obtained using a Syntactic Graph Convolutional Network (SGCN) (Bastings et al. 2017; Li et al. 2021), representing the local syntactic context of each token. We capture words and phrases modifying the parties in input instances by using dependency graphs, which are obtained by applying the Stanford dependency parser (Chen and Manning 2014) using Spacy.[10] The dependency graphs are composed of vertices (tokens) and directed edges (dependency relations), which capture the complex syntactic relationships between tokens.

SGCN operates on directed dependency graphs based on Graph Convolutional Network (GCN) (Kipf and Welling 2016). GCN is a multilayer message propagation-based graph neural network. Given a vertex v in G and its neighbors $\mathcal{N}(v)$, the vertex representation of v on the $(j+1)$st layer is:

$$h_v^{j+1} = \sum_{u \in \mathcal{N}(v)} W^j h_u^j + b^j, \quad (2)$$

where $W^j \in \mathbb{R}^{d^{j+1} \times d^j}$ and $b^j \in \mathbb{R}^{d^{j+1}}$ are trainable parameters, and d^{j+1} and d^j denote latent feature dimensions of the $(j+1)$st and jth layers, respectively. SGCN improves GCN by considering the directionality of edges, separating parameters for dependency labels, and applying edge-wise gating (Bastings et al. 2017; Li et al. 2021). Edge-wise gating can select impactful neighbors by controlling the gates for message propagation through edges. Therefore, the SGCN module takes word embeddings and syntactic relations to compute local representations. The local representation for a vertex (token) v is:

$$h_v^{j+1} = \sum_{u \in N(v)} g_{u,v}^j (W_{d_{u,v}}^j h_u^j + b_{u,v}^j), \quad (3)$$

where j represents a layer, g is the gate on the jth layer to select impactful neighbors $u \in N$ of v, W is the weight, and b represents bias. For each sentence, we use a pooling layer to convert tokens' local representations into a single hidden vector.

[10]https://spacy.io/

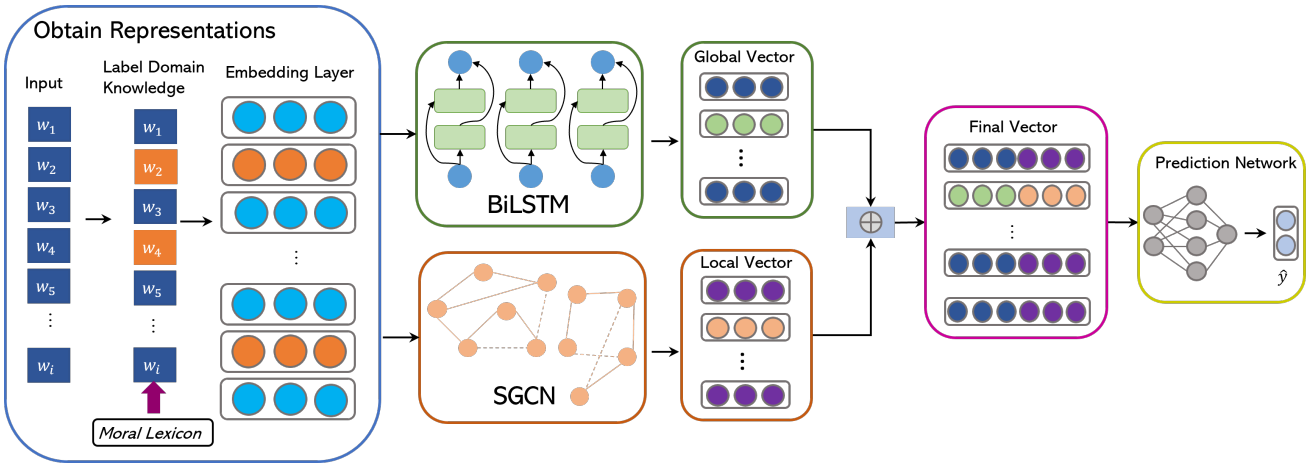


Figure 5: Architecture of the predictor in Figure 4. Here, w_i represents a token in an input instance and $\widehat{y}$ is a predicted verdict. Tokens in ■ are labeled by an additional moral lexicon.

Domain knowledge is used to weakly label rationales, following (Jiang and Wilson 2021). Such unsupervised rationalization favors informative tokens to optimize losses. However, our dataset's highly informative and frequent tokens such as gendered words (e.g., wife, boyfriend, and mother) may not influence the verdict. Therefore, we use the popular (Nguyen et al. 2022; Ziems et al. 2022) psychological theory, Moral Foundation Theory (MFT) (Haidt and Graham 2007), to effectively select moral rationales. MFT refines morality into five broad domains: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. We adopt the extended version of the MFT lexicon (Hopp et al. 2020), containing 2,041 unique words, to label comments in our corpus. We reprocess the input instances and generate weak labels for rationales $z_d = [z_d^i] \in \{0, 1\}$, where $z_d^i = 1$ if x^i is in the lexicon. We include a loss term $L_d(z, z_d) = -\sum_i |a^i| z_d^i$ for the soft selection process (Jiang and Wilson 2021), where a^i denotes the attention weight for token z^i. The term L_d lowers the loss when the tokens selected by feature–scoring methods are morality-related; otherwise, it has no effect. For prediction loss, we apply cross-entropy to optimize the network by calculating $L(y, \widehat{y})$ using the last hidden layer's output. Combining the loss items, the objective of our model is:

$$\arg\min L(y, \widehat{y}) + \lambda L_d(z, z_d), \tag{4}$$

where λ controls the weight of domain knowledge loss.

Experiments and Results

We now evaluate of our predict-then-extract model. For extraction performance, we first adopt multiple features scoring methods from previous works, followed by verifying the plausibility of the extracted rationales.

Experimental Settings

Baseline Methods for Prediction We evaluate the following machine learning models: Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and the transformer model BERT (Devlin et al. 2019). For LR, we use the lengths of the instances as a baseline model. For LR, RF, and SVM, we use GloVe (Pennington, Socher, and Manning 2014), a text encoder that maps a word into a low-dimensional embedding vector, for textual classification to generate vector representations. For the BERT baseline, we apply the "Obtain Representations" and "Prediction Network" modules as shown in Figure 5. We feed CLS representations generated from the embedding layer into the prediction network instead of passing them through the BiLSTM and SGCN networks.

Feature Scoring Methods for Extraction We use random selection as a baseline and multiple feature–scoring methods to compute importance scores s:

- Random (RAND): Randomly allocate importance scores.
- Attention (α): Normalized attention weights (Jain et al. 2020).
- Scaled Attention ($\alpha\nabla\alpha$): Attention weights multiplied by the corresponding gradients (Serrano and Smith 2019).
- Integrated Gradients (IG): The integral of the gradients from the baseline (zero embedding vector) to the original input (Sundararajan, Taly, and Yan 2017).
- Flexible (FLX): A flexible instance-level rationale selection method (Chrysostomou and Aletras 2022), under which each instance selects different scoring methods and lengths of rationales.

We compare only the above methods because they yield better performance than others, e.g., (Atanasova et al. 2020).

Evaluation Metrics for Rationales' Plausibility For prediction, we use macro F1-scores. For extraction, we adopt metrics from previous works (Jain et al. 2020; Chrysostomou and Aletras 2022):

- Reverse-Macro F1 (revF1): The performance of $\mathcal{M}$ in predicting y when using full input and rationale-reduced input. The predicted label with full input is used as the

gold standard. Masking rationales should reduce the prediction performance; lower is better.

- Normalized Sufficiency (NS): Reversed and normalized differences between predicting full input text and rationales: $max(0, 1 - (p(\hat{y}|x) - p(\hat{y}|R)))$; higher is better.
- Normalized Comprehensiveness (NC): Normalized differences between predicting full input text and rationale-reduced text: $p(\hat{y}|x) - p(\hat{y}|(x \notin R))$; higher is better.

Note that we are interested in generating plausible rationales, not in producing accurate classifiers. Therefore, we do not conduct a human study to evaluate the accuracy of the generated rationales but to evaluate their plausibility (Chrysostomou and Aletras 2022), which in practice does not correlate with accuracy (Atanasova et al. 2020).

Hyperparameters For generating global representations, we use Adam optimization with an initial learning rate of $2\text{e}{-5}$, $\epsilon = 1\text{e}{-8}$, a batch size of 16, 500 training steps, and a maximum sequence length of 256. For generating local representations, the initial input to the first graph convolutional layer is the 768-dimensional global model representation. These vectors are processed by the subsequent graph convolutional layer and output 128-dimensional vectors. The pooling layer for a vertex in Equation 3 is a dense linear layer with $\tanh$ activation, whose input vectors are stacked vectors of all vertices and output is a single 128-dimensional vector. We concatenate the global and local representations and obtain 896-dimensional vectors to feed into a prediction network. The prediction network is a three-layer, fully connected, dense neural network, which comprises 512, 256, and 128 units, respectively, with ReLu activation. To avoid overfitting, we regularize the prediction network using the dropout technique; at each fully connected layer, we apply a dropout level of $d = 0.5$. Finally, the prediction network output is fed into the last neural network of two units; with softmax to obtain probability distributions of the verdicts. We train five epochs for all the transformer-based models. All the experiments are implemented using Huggingface.

Results

The prediction performance of a model indicates its ability to distinguish commenters' evaluations of the various parties' behaviors. The extraction performance of a model indicates the plausibility of the rationales it generates.

Performance of Predicting Verdicts We use five-fold stratified cross-validation for the aforementioned classifiers. For the transformer-based models, we ran each model with five epochs. The reported performance score averages are shown in Table 2. The scores with domain knowledge are calculated with $\lambda = 0.1$, which yields the best performance. We observe that neural models outperform traditional machine learning models. The Global+Local-Domain method shows an average of 3% improvement among all the scores compared to a fine-tuned BERT. We are unable to compare our results with Botzer, Gu, and Weninger (2022) because of different research purposes and lack of their dataset and experimental details.

Methods	Precision (%)	Recall (%)	F1 (%)
LR-Length	53.9	53.8	53.8
LR-GloVe	57.7	56.2	57.0
Random Forest	60.8	62.4	61.6
SVM	63.2	65.3	64.2
BERT	83.7	82.6	83.1
BERT-Domain	82.8	82.5	82.6
Global	83.0	82.8	83.0
Global-Domain	83.7	81.5	82.6
Local	82.9	84.2	83.5
Local-Domain	83.6	83.6	83.6
Global+Local	85.6	**86.9**	86.2
Global-Local-Domain	**86.8**	86.1	**86.4**

Table 2: Macro F1 (the F1-scores calculated based on precision and recall scores), Precision, and Recall on the test set. The best scores are shown in bold (highest). Global+Local improves BERT by an average of 3.1% on the three scores. Although BERT with domain knowledge does not outperform its counterpart without domain knowledge, the Global+Local-Domain method demonstrates an average of 3% improvement on all three scores compared to BERT.

Ablation Studies for Prediction We perform ablation studies to understand how global and local representations affect prediction performance. Table 2 shows that separately using global or local representations does not improve prediction performance over BERT, while combining both representations achieves the best performance.

Performance of Extracting Rationales Table 3 illustrates the performance of various feature–scoring methods with and without weakly supervision through domain knowledge. We experiment on two scored token-selection methods (Jain et al. 2020): (1) selecting the K highest scoring (TopK) tokens for each instance and (2) selecting highest overall K-gram scoring tokens in the span of input tokens. We adopt TopK for further analysis because it yields the best performance. Although Table 2 shows that considering domain knowledge may not improve prediction performance for all neural models (e.g., BERT), we observe that using the instance-level rationale extraction method (FLX) with domain knowledge improves a rationale's plausibility. Combining the results of Table 2 and Table 3, we use Global+Local-Domain to predict and the FLX feature–scoring method to extract appropriate rationales.

Analysis

We leverage the 17,808 identified rationales from the 18,202 instances of our corpus's development and test sets as a lexicon. We apply this lexicon on the total 3,675,452 comments in our corpus. We introduce how we identify and cluster the extracted rationales' meanings. Then, we perform a fine-grained analysis to investigate how the clusters are associated with the authors' and commenters' social factors.

		Global			Local			Global+Local		
	Methods	revF1	NS	NC	revF1	NS	NC	revF1	NS	NC
Domain	RAND	85.9	0.26	0.27	79.3	0.25	0.27	84.0	0.20	0.34
	α	59.2	0.31	0.42	56.0	0.33	0.53	52.5	0.45	0.64
	$\alpha\nabla\alpha$	58.1	**0.47**	0.61	45.8	0.50	0.77	42.9	0.47	**0.81**
	IG	66.8	0.31	0.54	65.2	0.35	0.54	65.9	0.37	0.50
	FLX	**42.3**	0.52	**0.72**	**41.3**	**0.59**	0.77	**38.9**	**0.50**	0.80
No Domain	RAND	79.0	0.25	0.30	86.6	0.24	0.29	88.0	0.21	0.33
	α	62.1	0.29	0.39	62.5	0.37	0.61	63.6	0.38	0.61
	$\alpha\nabla\alpha$	57.2	0.37	0.65	56.9	0.45	0.64	54.1	0.45	0.70
	IG	68.2	0.32	0.53	63.9	0.30	0.53	62.2	0.28	0.45
	FLX	44.6	0.44	0.69	42.6	0.46	**0.78**	41.3	0.49	0.77

Table 3: The Normalized Sufficiency (NS) and Normalized Comprehensiveness (NC) scores range over $[0, 1]$. Results with "Domain" are with the domain knowledge module when predicting verdicts, and results with "No Domain" are without the module. The best scores (the revF1 score is lowest; the NS and NC scores are highest) in each column are shown in bold. The underlined numbers are the highest NS and NC scores and the lowest revF1 scores among the three metrics. The averaged performance scores ($\lambda = 0.1$) on the testing and development sets are similar. Among the three scores for the five feature–scoring methods (total fifteen for each prediction model), the number of times the Domain beats the No Domain for Global is 10 out of 15, for Local 9 out of 15, and for Global+Local 13 out of 15.

Clusters	Topics	Examples
Judgment of appearance	Work	skinny, curly, chubby, lean, eat, meat, slim, bodied, blonde, sickly
	Safety	underwear, panties, bikini, clingy, thong, boudoir, swimsuit, bras, headband, earrings
Evaluation: Good/Bad	Work	derogatory, *extremely terrible*, *horrible*, derisive, *awful*, *awesome*, pejorative
	Safety	*awful*, *horrible*, *extremely terrible*, incredible, nightmare, *awesome*, amazing
Calm/Violent/Angry	Marriage	kick, spitting, stomped, slapping, wasted, missed, fight, punching, cheating on, lied to
	Education	picky, lived, mortified, resentful, nauseous, baffled, inconvenienced, conflicted
Law&order	Education	punish, punishment, jails, prison, inability, failure, lack, fault, mistakes, error
	Safety	abuse, harassment, bullying, sexual, neglect, rape, abusers, cruelty, humiliation
Fear/bravery/shock	Work	insecurities, humiliating, exhausting, miserable, sad, doubly, stressful, messy
	Roommates	cruel, vile, despicable, cowardly, inhumane, atrocious, abhorrent, brutal, aggression

Table 4: Examples of meaning clusters embedded in moral rationales for topic-specific posts. Italics show the words that are common between the topics of the same cluster. The results indicate that words used in comments are different based on posts' topics. In addition, adverbs used in *Evaluation: Good/Bad* are more prevalent than adjectives used in *Judgment of Appearance*.

Clustering and Tagging

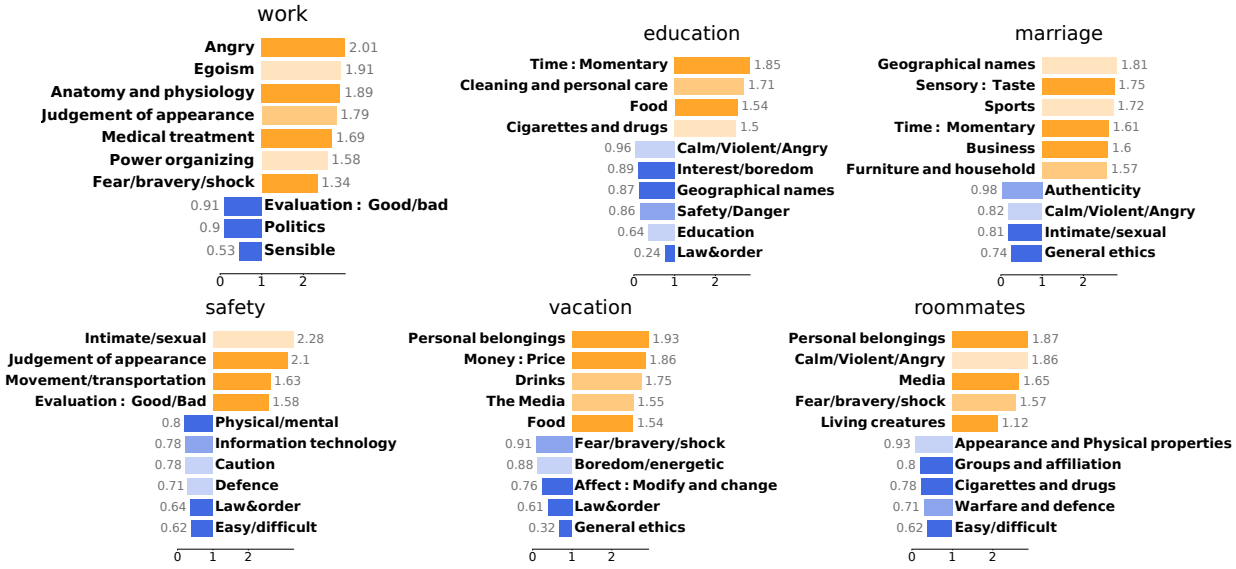
We apply pretrained GloVe embeddings (Pennington, Socher, and Manning 2014) to cluster the meaning similarities of rationales (i.e., averaged embeddings for phrases). After manually checking the clustered results, we select GloVe embeddings as they provide more detailed and informative clusters than other embedding methods, such as word2vec (Mikolov et al. 2013). We exclude the rationales that have negative dependencies in the original sentences to avoid ambiguity. To aggregate the most similar embeddings into clusters, we employ the well-known k-means clustering algorithm. We tag the resulting clusters with USAS,[11] a framework for automatic meaning analysis and tagging of text, which is based on McArthur's Longman Lexicon of Contemporary English (Summers and Gadsby 1995). We use this lexicon (Piao et al. 2015) to name the tags. Because the generated rationales contain phrases and words, we filter the phrases composed of words belonging to the same USAS categories, such as `extremely awful`. We discard clusters of pronouns and prepositions. As a result, we find 86 unique meaning clusters in our dataset.

[11] http://ucrel.lancs.ac.uk/usas/

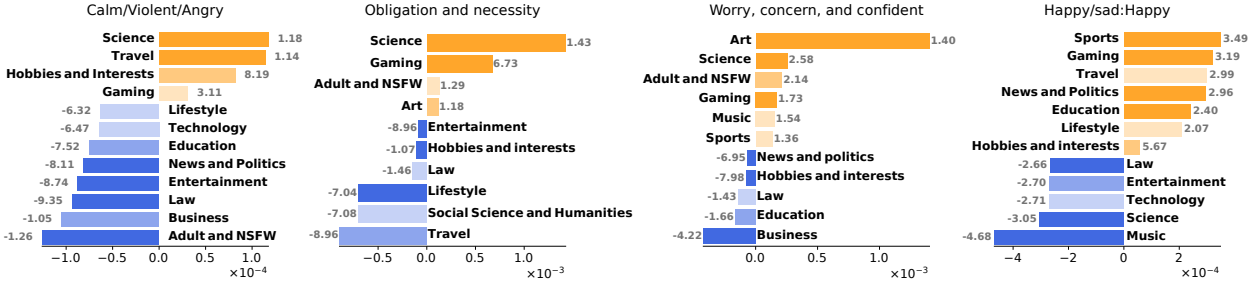
Associations between Comments and Factors

We measure the Odds Ratio (OR) to assess the associations between posts' topics, authors' genders, and meaning clusters. For commenters' interests, we apply linear regression to compute their effects on the judgments. Figure 6 and Table 4 report the results.

Common but Distinct Reasoning in Topic-Specific Situations Our corpus includes six common post topics: work, education, safety, vacation, roommates, and marriage. Figure 6a shows the topics with OR, indicating polarized comments for authors' genders. Our analysis reveals consistent gender effects in some categories, such as *Calm/Violent/Angry* in *education* and *marriage*, and *Judgment of appearance*



(a) The odds ratio values of authors' gender and meaning clusters in different topics. An odds ratio greater than one indicates the category is more likely to appears in the comments when the posters are females compared to males. And an odds ratio smaller than one indicates the opposite.



(b) Regression results for the effects of the proxied commenters' interests. An effect that is greater than zero indicates positive effect. And an effect smaller than zero indicates the opposite.

Figure 6: We use orange rectangles to indicate odds ratio greater than one and effects greater than zero (on the right), and blue rectangles indicate the opposite (on the left). With the deepening of the shades, the respective representations are 0.05, 0.001, and 0.0001.

in *work* and *safety*. Moreover, as indicated in Table 4, we also observe distinct preferences for word usage to convey identical meanings even among categories with similar gender effects, suggesting nuanced and context-specific usage of language in moral reasoning. However, the *Evaluation: Good/bad* category shows controversial effects, showing biases towards females in *safety* but towards males in *work*, indicating polarizing opinions. Interestingly, the most contentious topics, such as relationships (De Candia et al. 2022; Ferrer et al. 2020), do not show typical gender biases in our analysis. This could be due to the commenters having specific evaluation standards in different moral scenarios.

We find distinctive gender effects in *work* and *safety* (i.e., the maximum difference between OR scores is over 1.5). In such topics, words in *Sensible* and *Law&order* are less likely used in comments towards female authors, and words related to *Judgement of appearance* are more likely to be. The observation can be explained by the reflection of the persistent pressure on women to conform to societal beauty standards (Stuart and Donaghue 2012). Moreover, commenters use different adjectives, verbs, and nouns to emphasize their concerns based on a given situation, while employing similar adverbs to express their emotions. For instance, the adjectives, verbs, and nouns used in *Judgment of appearance* for *work* and *safety* are dissimilar, whereas the adverbs employed in *Evaluation: Good/Bad* are common.

Commenters' Interests Matter We now investigate how the commenters' interests (as proxied by the subreddits they participate in) affect their moral reasoning. There are eighteen categories of subreddits that the commenters participated in (ordered in frequency): lifestyle, science, locations, technology, hobbies and interests, law, adult and NSFW, business, social science and humanities, music, sports, entertainment, news and politics, gaming, architecture, art, travel, and education. These categories exhibit high popularity and diversity. For example, *news and politics* includes subreddits, such as r/PoliticalHumor and r/antiwork, each with over

a million users and *lifestyle* includes r/baking (over 1.6M members) and r/relationships (over 3.4M members).

The proxied commenters' interests are confounded with each other. Therefore, we investigate them simultaneously to measure the causal effects of their interests. We use an Ordinary Least Square (OLS), model[12] which is a common method for analyzing social variables (Stolzenberg 1980). The following model captures the linear effects:

$$b = \beta_0 + \beta_i x_i + \epsilon_i, i \leq n, \quad (5)$$

where x_i denotes the frequency of the ith cluster appearing in judgments b, β represents the constant effect of x_i, n is the total number of clusters, and $\epsilon_i \sim \mathcal{N}(\mu, \sigma^2)$ is normally distributed noise centered at 0.

Figure 6b shows the effects of the interest categories on emotion-relevant clusters. We observe that some categories such as *Sports* and *Lifestyle* are likelier to positively affect optimistic clusters *Happy/sad:Happy* than neutral clusters such as *Music*. In addition, *Gaming* and *Science* positively affect using *Obligation and Necessity* words, such as *would*, *should*, and *must*. Conversely, *Social Science and Humanities* and *Entertainment* have a negative effect. The results may be explained by the distinctive personality traits of the social groups the commenters belong to. For example, commenters interested in *Art* (e.g., r/AccidentalRenaissance) are the most likely to use *worry, concern, and confident* words and commenters interested in *Music* (e.g., r/NameThatSong) are the least likely to use *Happy/sad:Happy* words. A possible explanation may be that personalities of people interested in art are more emotionally sensitive than others (Csikszentmihalyi and Getzels 1973).

Discussion and Conclusion

Our research introduces a new framework for analyzing language on social media platforms. We focus on judgments of social situations and examine how social factors, such as a poster's gender and a commenter's interests, influence the distributions of common elements in the language used in comments. We employ NLP tools and a predict-then-extract model to collect these common elements.

Our study demonstrates that the language used in moral rationales on AITA is influenced by users' social factors. For instance, consistent gender effects are observed in the *Calm/Violent/Angry* category in *education* and *marriage*, with posts authored by males more likely to receive such comments. Interestingly, our analysis reveals nuanced word usage within identical clusters, with verbs such as "kick" and "spitting" being frequently used in the *Calm/Violent/Angry* category in *marriage*, whereas adjectives such as "picky" and "mortified" were more common in *education*. Conversely, the *Evaluation: Good/bad* category in *work* and *safety* elicited conflicting opinions.

Our observations corroborate social psychology findings (Csikszentmihalyi and Getzels 1973; Stuart and Donaghue 2012). For example, comments about *Judgment of Appearance* in *work* and *safety* exhibit prevalence for female authors, which indicate the societal pressure on women to conform to beauty standards (Stuart and Donaghue 2012). Moreover, commenters interested in *Music* and *Art* are likelier to express emotions, which may be caused by their personalities (Csikszentmihalyi and Getzels 1973). Overall, these findings highlight the context-specific and nuanced nature of language usage in discussions morality on social media platforms, and contribute to a better understanding of the influence of social factors on language use.

Broader Perspectives Our research presents a novel framework for analyzing language usage in online media, which has practical implications for the design of monitoring systems to identify biased submissions in specific communities. This framework can assist commenters in reconsidering their comments and moderators in flagging concerning comments. In addition, the proposed methods can explain why a submission is considered biased and can inform the design of better features to educate new community members about problematic aspects of their submissions.

Our findings align with social psychology research and shed light on societal pressures, such as the gendered pressure on appearance in work and safety contexts. Additionally, our study reveals the impact of personal interests on language use. These broader perspectives suggest potential implications for the development of more effective communication strategies online and underscore the need for further research exploring the relationship between language use and social factors in moral reasoning.

Limitations and Future Work Our empirical method inherently shares limitations with observational studies, e.g., susceptibility to bias and confounding. There is a limit to how much we can tease apart social factors of the posters and commenters. We acknowledge some of the boundaries are unclear. For example, we treat genders as a social factors, but do the genders also affect posters' writing styles? In addition to our single dataset analysis on AITA, there may be potential for further exploration on other data sets such as the *r/relationship_advice* subreddit. Additionally, creating new datasets with crowd-sourced moral judgments could be beneficial in expanding the scope of analysis.

Although our prediction model takes into account the syntactic relations of input sentences, it is possible for some parties mentioned in a post to be background characters rather than active participants. Moreover, the rationales we extract are not validated with ground-truth labels, mainly due to the complexity of the instances in our dataset. In future work, we plan to leverage our framework to construct annotation guidelines to obtain human-evaluated clusters for analysis.

Ethics Statements Reddit is a prominent social media platform. We scrape data from a subreddit using Reddit's publicly available official API and PushShift API, a widely used platform that ingests Reddit's official API data and collates the data into public data dumps. None of the commenters' information was saved during our analysis. The human evaluation mentioned, such as the evaluation of comments' labels, was performed by the authors of this paper and colleagues. One potential negative outcome of this research is that it may reinforce stereotypes and biases that

[12]https://www.statsmodels.org

already exist. Additionally, the research may not generalize to all populations, and may not account for other factors such as age, culture, and education that could be influencing moral reasoning on social media.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. We thank the NSF (grant IIS-2116751) for partial support for this research.

References

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–3274. Online: Association for Computational Linguistics.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2963–2977. Florence, Italy: Association for Computational Linguistics.

Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; and Sima'an, K. 2017. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*.

Beel, J.; Xiang, T.; Soni, S.; and Yang, D. 2022. Linguistic Characterization of Divisive Topics Online: Case Studies on Contentiousness in Abortion, Climate Change, and Gun Control. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 16, 32–42.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning research*, 3: 993–1022.

Botzer, N.; Gu, S.; and Weninger, T. 2022. Analysis of Moral Judgment on Reddit. *IEEE Transactions on Computational Social Systems*, 10.

Bussey, K.; and Maughan, B. 1982. Gender differences in moral reasoning. *Journal of Personality and Social Psychology*, 42(4): 701.

Chen, D.; and Manning, C. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740–750. Doha, Qatar: Association for Computational Linguistics.

Chrysostomou, G.; and Aletras, N. 2022. Flexible Instance-Specific Rationalization of NLP Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10545–10553. Online: AAAI Press.

Csikszentmihalyi, M.; and Getzels, J. W. 1973. The personality of young artists: An empirical and theoretical exploration. *British Journal of Psychology*, 64(1): 91–104.

De Candia, S.; De Francisci Morales, G.; Monti, C.; and Bonchi, F. 2022. Social Norms on Reddit: A Demographic Analysis. In *Proceedings of the ACM Web Science Conference*, WebSci, 139–147. NY: Association for Computing Machinery.

De Choudhury, M.; Sharma, S. S.; Logar, T.; Eekhout, W.; and Nielsen, R. C. 2017. Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. In *In the ACM Conference on Computer Supported Cooperative Work and Social Computing*, 353–369. Association for Computing Machinery.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. Minneapolis: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models.

Emelin, D.; Le Bras, R.; Hwang, J. D.; Forbes, M.; and Choi, Y. 2021. Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 698–718. Online and Punta Cana.

Ferrer, X.; van Nuenen, T.; Such, J. M.; and Criado, N. 2020. Discovering and Categorising Language Biases in Reddit. *arXiv preprint arXiv:2008.02754*.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Giorgi, S.; Zhao, K.; Feng, A. H.; and Martin, L. J. 2023. Author as character and narrator: Deconstructing personal narratives from the r/amitheasshole Reddit community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 233–244.

Guimaraes, A.; and Weikum, G. 2021. X-Posts Explained: Analyzing and Predicting Controversial Contributions in Thematically Diverse Reddit Forums. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 163–172.

Guo, Z.; Zhang, Z.; and Singh, M. P. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of the 29th Web Conference (WWW)*, 2388–2399. Taipei: ACM.

Haidt, J.; and Graham, J. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. In *Social Justice Research*, volume 20, 98–116. Springer.

Haque, A.; and Singh, M. P. 2024. NewsSlant: Analyzing Political News and Its Influence Through a Moral Lens. *IEEE Transactions on Computational Social Systems*, 1–9.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hopp, F. R.; Fisher, J. T.; Cornell, D.; Huskey, R.; and Weber, R. 2020. The Extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text. *Behavior Research Methods*, 53: 232–246.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 4459–4473. Online: Association for Computational Linguistics.

Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Forbes, M.; Borchardt, J.; Liang, J.; Etzioni, O.; Sap, M.; and Choi, Y. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Jiang, S.; and Wilson, C. 2021. Structurizing Misinformation Stories via Rationalizing Fact-Checks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 617–631. Online: Association for Computational Linguistics.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 107–117. Austin, Texas: Association for Computational Linguistics.

Li, Z.; Qin, Y.; Liu, Z.; and Wang, W. 2021. Powering Comparative Classification with Sentiment Analysis via Domain Adaptive Knowledge Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6818–6830. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137.

Lourie, N.; Bras, R. L.; and Choi, Y. 2020. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes. *arXiv preprint arXiv: 2008.09094*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Nguyen, T. D.; Lyall, G.; Tran, A.; Shin, M.; Carroll, N. G.; Klein, C.; and Xie, L. 2022. Mapping Topics in 100,000 Real-Life Moral Dilemmas. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 16, 699–710.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Piao, S.; Bianchi, F.; Dayrell, C.; D'Egidio, A.; and Rayson, P. 2015. Development of the Multilingual Semantic Annotation System. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 1268–1274. Denver, Colorado: Association for Computational Linguistics.

Richardson, H. S. 2018. Moral Reasoning. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Stolzenberg, R. M. 1980. The Measurement and Decomposition of Causal Effects in NOnlinear and Nonadditive Models. *Sociological Methodology*, 11: 459–488.

Stuart, A.; and Donaghue, N. 2012. Choosing to conform: The discursive complexities of choice in relation to feminine beauty practices. *Feminism & Psychology*, 22(1): 98–121.

Summers, D.; and Gadsby, A. 1995. *Longman Dictionary of Contemporary English: The Complete Guide to Written and Spoken English*. Longman Group Limited.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference on Machine Learning*, 3319–3328. JMLR.org.

Walker, L. J. 1989. A Longitudinal Study of Moral Reasoning. *Child Development*, 60(1): 157–166.

Wood, J. A.; Longenecker, J. G.; McKinney, J. A.; and Moore, C. W. 1988. Ethical Attitudes of students and Business Professionals: A Study of Moral Reasoning. *Journal of Business ethics*, 7(4): 249–257.

Xi, R.; and Singh, M. P. 2023. The Blame Game: Understanding Blame Assignment in Social Media. *IEEE Transactions on Computational Social Systems*, 10: 1–10.

Zhou, K.; Smith, A.; and Lee, L. 2021. Assessing Cognitive Linguistic Influences in the Assignment of Blame. In *Proceedings of the International Workshop on Natural Language Processing for Social Media*, 61–69. Online: Association for Computational Linguistics.

Ziems, C.; Yu, J.; Wang, Y.-C.; Halevy, A.; and Yang, D. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 3755–3773. Dublin, Ireland: Association for Computational Linguistics.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? No, our framework has no potential negative societal impacts as far as we know.
   (g) Did you discuss any potential misuse of your work? No, our framework has no potential misuse as far as we know.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, the data has no persoanl identifications.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.
   (d) Did you discuss how data is stored, shared, and deidentified? NA.