ABSTRACT

XI, RUIJIE. Decoding Morality in Social Media. (Under the direction of Dr. Munindar P. Singh).

Morality presents a complex challenge within society, encompassing a broad spectrum of ethical dilemmas and conflicting perspectives. This research focuses on three crucial aspects – blame assignment, moral reasoning, and moral incidents – to explore real-life morality through insights gained from moral scenarios unfolding within social media platforms. These scenarios involve interpersonal moral situations shared by authors and judgments of whose behaviors were inappropriate provided by other users. By adopting Natural Language Processing (NLP) technologies, we consider psychological theories and social commonsense with linguistic characteristics of the descriptive moral situations to explore how individuals perceive and respond to these situations. Through this comprehensive analysis, we aim to gain deeper insights into the intricate dynamics of real-life morality.

Taking into account psychological theories, the study uncovers compelling evidence of biases in social morality, wherein blame assignment is influenced by social identities. Moreover, the research examines the commonsense aspects of morality and finds that the moral reasoning used to judge the appropriateness of social behaviors is influenced by social factors of people. Furthermore, the research delves into the most prominent ethical touchpoints within a post that attract attention and generate discussion among commenters. Through an analysis of linguistic patterns, the study reveals that specific portions of ethical points tend to receive more attention. Intriguingly, these portions share common elements across moral situations involving different domains.

Overall, this research offers innovative computational approaches to tackle challenges related to morality. By harnessing the power of NLP techniques, we aim to contribute to theoretical psychological and social computing studies on morality by delving into real-life social situations, providing valuable insights to inform research in this domain.

Decoding Morality in Social Media

by
Ruijie Xi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina
2023

APPROVED BY:

Dr. James Lester	Dr. Christopher Healey
Dr. Kelly Lynn Mulvey	Member 4 of Committee

Dr. Munindar P. Singh
Chair of Advisory Committee

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Munindar P. Singh, for all his help and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

While morality has been extensively studied for many years through theoretical frameworks, there has been a relatively limited exploration of moral phenomena within the realm of social media. Traditional moral social theories have provided valuable insights into the principles and foundations of moral judgment, focusing on aspects such as ethical reasoning, moral emotions, and normative frameworks. However, the advent of social media platforms has introduced a new dimension to the study of morality. These platforms have become powerful arenas for moral discussions, providing a vast amount of data and an unprecedented opportunity to observe real-time moral interactions on a global scale. Despite this, the integration of social media into the study of morality has been relatively scarce until recently.

Examining morality in the context of social media presents a unique and valuable opportunity to understand how these processes unfold in a digital environment. Social media platforms have become significant arenas for public discourse, where individuals from diverse backgrounds engage in discussions, share opinions, and make moral judgments. By investigating morality on social media, researchers can gain insights into several important aspects:

Online Moral Communities: Social media platforms provide spaces for individuals to participate in moral discussions and debates. Analyzing morality within these online moral communities can offer valuable insights into how individuals negotiate moral norms

and values in a digital context. It helps us understand how social media platforms shape moral reasoning and the formation of moral judgments.

Emerging Computational Methods: Social media platforms generate vast amounts of data, offering opportunities for computational methods to analyze blame assignment and moral reasoning at scale. Natural language processing and machine learning techniques can be employed to analyze large-scale datasets and identify patterns in how blame is assigned, moral judgments are made, and moral reasoning is expressed on social media.

Implications for Public Discourse: Social media plays a significant role in shaping public opinion and influencing societal debates. Understanding morality on these platforms can inform strategies for promoting ethical and responsible discourse. It enables us to identify and address potential pitfalls, biases, and misinformation that may arise in online moral discussions, ultimately contributing to more informed and constructive public dialogue.

By bridging the gap between traditional moral theories and the dynamics of social media, researchers can deepen understanding of how individuals navigate moral dilemmas, how digital communication influences moral judgments, and how moral norms and values are shaped in the digital age. This knowledge can inform interventions, policies, and guidelines aimed at fostering ethical engagement and enhancing the quality of moral discourse on social media platforms.

1.1 Background and Research Questions

1.1.1 Blame Assignment in Social Media

Psychological studies on morality have proposed underlying linguistic and semantic factors. Malle et al. (2014) find that people assign blame to individuals they observe violating norms. Gray and Wegner (2011) show that victims can escape from being blamed, whereas heroes may cause blame. Guglielmo and Malle (2019) suggest that moral blame is more complex than moral praise. Existing studies are mainly based on surveys on stylized social situations and few participants. In contrast, online social systems enable people worldwide to share viewpoints about a spectrum of social situations, enabling deeper computational studies of blame assignment.

Moral situations posted on subreddit r/AmItheAsshole (AITA) involve multiple individuals and social identities (e.g., genders). Previous works use the posts to study crowd-sourced blame assignments and predict verdicts of who is to blame. Much attention falls on predicting verdicts

using Artificial Intelligent (AI) models to achieve high accuracy. However, these models do not shed light on what linguistic characteristics affect an audience's decisions on assigning blame and don't consider social psychology Fraser et al. (2022).

According to social science, blame is assigned to an *agent* when behaviors are causing damage to a vulnerable *patient* Schein and Gray (2018). Therefore, agents are perceived as blameworthy, where their agentiveness depends on how they are described Gray and Wegner (2011). The audience assigns blame to the individuals that they perceive as agents. However, the descriptive features that affect the audience's recognition of agentiveness are still not well understood. Therefore, this study focuses on investigating moral situations submitted to social media and addresses two research questions:

$RQ_{Feature}$: What cognitive-affective language features are crucial in blame assignment?

RQ_{Social}: What biases, if any, arise in blame assignment on social media?

1.1.2 Moral Reasoning in Social Media

Moral reasoning has been a subject of long-standing study. Bussey and Maughan (1982) find that moral decisions by males are typically based on law-and-order reasoning, while 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 on business ethics by professionals than by students. However, these studies do not provide a comprehensive understanding of moral reasoning within the context of social media.

This work considers the comments in AITA as source to investigate real-life moral reasoning towards moral situations. Community members, referred to as *commenters*, may comment on a post to provide moral *judgments* (i.e., verdicts, justifying the verdict (if any) and other moral assessment) and the *reasoning*. Recent studies have focused on predicting verdicts in both the posts and comments Lourie et al. (2020); Zhou et al. (2021); Botzer et al. (2022); Xi and Singh (2023). However, to the best of our knowledge, no empirical work has conducted a systematic analysis to understand the implicit reasoning present in the comments.

This research fills the gap in empirical work by conducting a systematic analysis of the implicit reasoning present in the comments on AITA. By examining the underlying moral reasoning of the commenters, we aim to uncover patterns and factors that shape their judgments and decision-making processes. This investigation provides valuable insights into how individuals navigate complex moral dilemmas and make ethical assessments within the online community, contributing to a deeper understanding of moral decision-making in the digital age.

1.2 Contributions and Novelty

We list contributions and novelty for each project below.

1.2.1 The Blame Game: Understanding Blame Assignment in Social Media

Our work contributes significantly by integrating Natural Language Processing (NLP) and social psychology to examine practical morality. We introduce innovative language features that combine linguistic and psychological insights, facilitating the creation of interpretable models of morality. Moreover, our research has practical implications for designing morally-aware and interpretable AI systems. Moreover, our study provides valuable contributions to theoretical research by offering language features that can enhance the precision of empirical studies in the field of morality. Detailed discussions on this project can be found in Chapter 2.

1.2.2 Moral Reasoning in the Mundane: Categorizing Moral Reasoning in Real-life Social Situations

Our research fills a gap in empirical work by analyzing implicit reasoning in comments, focusing on commonsense aspects of moral reasoning. Applying NLP tools, we investigate how social factors shape distributions and impact moral reasoning. Our novel framework has practical implications for monitoring systems, assisting commenters in reconsidering their remarks and helping moderators identify concerning comments. We find connections with social psychology, revealing societal pressures on appearance and the influence of personal interests on language use. These findings highlight the need for further research to explore language use, social factors, and moral reasoning, informing more effective online communication strategies. We discuss about this project in Chapter 3.

1.3 Proposal: Unveiling Moral Sparks: Investigating Real-life Moral Incidents

This research focuses on identifying incidents within moral narratives in AITA. These incidents are verbal events that connect entities and depict specific behaviors of the involved entities. Within a moral narrative, incidents hold varying degrees of importance and resonance for individual commenters, leading to different levels of attention. Commenters often quote incidents from the original posts to engage in moral discussions, and these selected portions are referred to as “Moral sparks” (M-sparks). M-sparks serve as prominent ethical touchpoints,

drawing attention and generating discussions among commenters. This work delves into described incidents from two distinct aspects: causal social commonsense and linguistic features. Social commonsense and morality are closely intertwined in real-life, where social norms and shared knowledge shape behavior and align with moral principles, influencing societal interactions. The incidents within moral narratives have the power to activate commenters' social commonsense, facilitating a deeper understanding of M-sparks. Linguistic features also play a crucial role in understanding moral narratives due to their nuances and complexity. Our investigation aims to explore the differentiation of M-Sparks through social commonsense and linguistic features. Additionally, we plan to examine the impact of social factors on commenters' identification of M-Sparks. We discuss about this plan in Chapter 4.

1.1 shows the timeline for the entire thesis research.

Table 1.1: Proposed plan.

	Task	Status	Estimate Time of Completion
1	The Blame Game	Complete	Mar 2023
2	Moral Reasoning in the Mundane	Complete	July 2023
3	Moral Sparks	Ideation	Jan 2024

CHAPTER

2

THE BLAME GAME: UNDERSTANDING BLAME ASSIGNMENT IN SOCIAL MEDIA

Psychological studies on morality have proposed underlying linguistic and semantic factors. However, current empirical studies often lack the nuances and complexity of real life. This paper examines how well the findings of prior studies generalize to a corpus of over 30,000 narratives of tense social situations submitted to a popular social media forum. A poster describes interpersonal moral situations or misgivings; other users judge from the post whether the poster (*protagonist*) or an opposing side (*antagonist*) is morally culpable. We extend and apply natural language processing (NLP) techniques to understand the effects of descriptions of the people involved in these posts. We conduct extensive experiments to investigate the effect sizes of features to understand how they affect the assignment of blame on social media. Our findings show that aggregating psychological theories enables understanding real-life moral situations. We also find evidence of bias blame assignment on social media, such as that males are likelier to receive blame no matter whether they are protagonists or antagonists.

2.1 Introduction

How do people judge whether someone deserves blame for their actions? This question has been extensively studied in social science. Malle et al. (2014) find that people assign blame to individuals they observe violating norms. Gray and Wegner (2011) show that victims can escape from being blamed, whereas heroes may cause blame. Guglielmo and Malle (2019) suggest that moral blame is more complex than moral praise. Existing studies are mainly based on surveys on stylized social situations and few participants. In contrast, online social systems enable people worldwide to share viewpoints about a spectrum of social situations, enabling deeper computational studies of blame assignment.

Example : Sample post and comments on it.

Title:
"AITA for snitching on my sister?"

Body:
"...I told my parents that my sister was staying up late with her tablet even though they had said she couldn't do it anymore. Now she's mad..."

Top-level Comment:
"OTHER. While you shouldn't be parenting her, you didn't go to your parents until she repeatedly ignored them as well as your warnings..."

Top-level Comment:
"AUTHOR. If your sister was doing something really bad that hurt someone ... You have undermined her trust in you..."

Flair: "OTHER"

This paper examines a popular subreddit (i.e., forum), /r/AmITheAsshole (AITA),[1] where users describe interpersonal conflicts and other users (i.e., audience) comment and judge who deserves blame. Example 2.1 shows a post and associated comments from AITA. The *title* and *body* are of the post, *top-level comment* comes from the audience (often including a verdict), and *flair* is the verdict of the top-voted comment. The most common verdicts are *author* and *other*. We use the term *blame assignment* to represent a post's verdict.

Section 2.2.1 provides additional details. Previous works take AITA as a resource for studying

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

crowd-sourced blame assignments on first-person moral situations. Much attention falls on predicting verdicts Lourie et al. (2020); Emelin et al. (2021); Jiang et al. (2021) with high accuracy. However, these models do not shed light on what linguistic characteristics affect an audience's decisions on assigning blame and don't consider social psychology Fraser et al. (2022).

Table 2.1: Summarizing recent work on moral-decision making models and AITA.

Type	Paper	Description	Dataset
Moral-decision making	Lourie et al. (2020)	Building neural models to predict moral scenarios	Scraped from AITA
	Forbes et al. (2020)	Building neural models to predict morality of social norms	Crowd-sourced dataset
	Emelin et al. (2021)	Building neural models to predict intents, actions, consequences of norms	Crowd-sourced dataset
	Jiang et al. (2021)	Building neural models to provide moral advisor	Multi-sourced datasets
Statistical Analysis	Nguyen et al. (2022)	Taxonomizing the structure of moral discussions	Scraped from AITA
	De Candia et al. (2022)	Analyzing demographic information of blame assignments	Scraped from AITA
	Zhou et al. (2021)	Analyzing linguistic features in blame assignments	Scraped from AITA
	Botzer et al. (2022)	Analyzing morality by building a moral judgment classifier	Scraped from AITA

Previous empirical work has not studied how social psychology theories generalize to real-life situations posted in AITA. According to the Theory of Dyadic Morality (TDM), blame is assigned to an *agent* when behaviors are causing damage to a vulnerable *patient* Schein and Gray (2018). Under TDM, agents are perceived as blameworthy, where their agentiveness depends on how they are described Gray and Wegner (2011). The posts in AITA are first-person narratives that involve multiple individuals' and social identities (e.g., genders). Accordingly, the audience assigns blame to the individuals that they think are described as agents. However, what descriptive features of individuals affect the audience's recognition of agentiveness remains unstudied.

This work studies the features of AITA's posts that affect the audience's blame assignment. We especially focus on social psychology research relating to language and social features.

Language features affect social media data in many ways. For instance, Beel et al. (2022) find sentiment is powerful in predicting the contentiousness of conversations on Reddit. In addition, social factors, such as gender, affect social media interactions Beel et al. (2022) and can lead to biases. For instance, De Candia et al. (2022) find that males have a higher possibility to receive blame on social media. Moreover, social scientists observe that gender and age affect morality in many ways Wark and Krebs (1996); Bracht and Zylbersztejn (2018). For instance, Reynolds et al. (2020) find that moral typecasting stereotypes females into the role of suffering patients.

Malle et al. (2014) proposes that assigning blame is a cognitive process that requires individuals to foresee the negative outcomes of agentive behaviors. Therefore, we define *cognitive-affective features* as language features that can shape the audience's blame assignment decisions. To this end, this paper aims to address two research questions:

$RQ_{Feature}$: What cognitive-affective language features are crucial in blame assignment?

RQ_{Social}: What biases, if any, arise in blame assignment on social media?

To answer $RQ_{Feature}$, we operationalize a set of novel factors using Natural Language Processing (NLP) that have explanatory power. We propose a novel **entity-centric** approach that partitions individuals involved in a situation as the protagonist (author) and antagonists (others). Then, we collect language features describing the entities based on existing social science research, such as emotions conveyed from attributive and predicative words Mohammad and Turney (2013). The language features are categorized into contextual, psycholinguistic, and linguistic features, where psycholinguistic features are entity-based and others are situation-based. Although previous research uses these features to analyze social media data Joe et al. (2020); Sap et al. (2017), it doesn't apply them to morality with entity-based approaches. We use the proposed features to build machine learning classifiers to predict whether an entity will receive blame. Using these classifiers, we examine the features' effect sizes to understand how an entity causes blame given the description.

To answer RQ_{Social}, we consider gender and age as social factors that may lead to biases in blame assignment Botzer et al. (2022); De Candia et al. (2022). We conduct qualitative and quantitative analysis using a post's textual information, which helps understand a situation in linguistic terms. We extract the demographics of the entities from the posts (they are marked with expressions such as [25F]). We apply statistical methods to measure the association strengths between blame assignment and these social factors.

2.1.1 Contributions and Findings

We characterize blame assignment with novel features inspired by social psychology. These features yield sufficient accuracy while being interpretable. Our methods go beyond theoretical models. Moreover, our findings can benefit psychological research by optimizing language used in surveys for studying morality.

We find that certain linguistic characteristics are highly correlated with blame assignment across the board: for instance, the protagonist can reduce blame by eliciting positive perspectives towards themself and can reduce blame by avoiding dominance-related words for themself, which trigger blame. Authors in the 15–45 age group are likelier to attract bias than others, as do males whether they are protagonists or antagonists, especially in situations involving *medicines and medical treatment* and *judgment of appearance*.

2.1.2 Literature Review

Tab. 2.1 summarizes recent research on improving moral decision-making through AI by understanding (im)moral social norms and behavioral rules using NLP, in two groups. The first group deals with predicting moral judgments. Lourie et al. (2020) use the title of a post in AITA as a social norm and develop a large dataset of situations based on such norms. Forbes et al. (2020) break down blame assignments of one-liner scenarios into rules of thumb and ask annotators to write moral and immoral stories based on the rules. Emelin et al. (2021) build a crowd-sourced dataset of social norms Lourie et al. (2020), with actions, intentions, and consequences of a moral situation. Others Botzer et al. (2022) build moral judgment classifiers to apply to predict blame assignment from one-line snippets from possibilities such as understandable, wrong, bad, and rude Jiang et al. (2021).

However, social scientists point out that training morality models on crowd-sourced data with no underlying moral framework is flawed Talat et al. (2021); Fraser et al. (2022). Hence, we do *not* seek the most accurate judgment predictor—and do not compete with state-of-art neural models. Instead, we expand the computational modeling of moral understanding based on how social psychology constructs are apparent in language.

The second group in Tab. 2.1 analyzes AITA using statistical methods, such as by creating a taxonomy of moral discussions Nguyen et al. (2022), analyzing the correlation between demographics and blame De Candia et al. (2022), and identifying linguistic features in judgments Zhou et al. (2021). However, no work has studied the effects of the descriptions on individuals' agentiveness reflected in social media.

2.2 Proposed Method

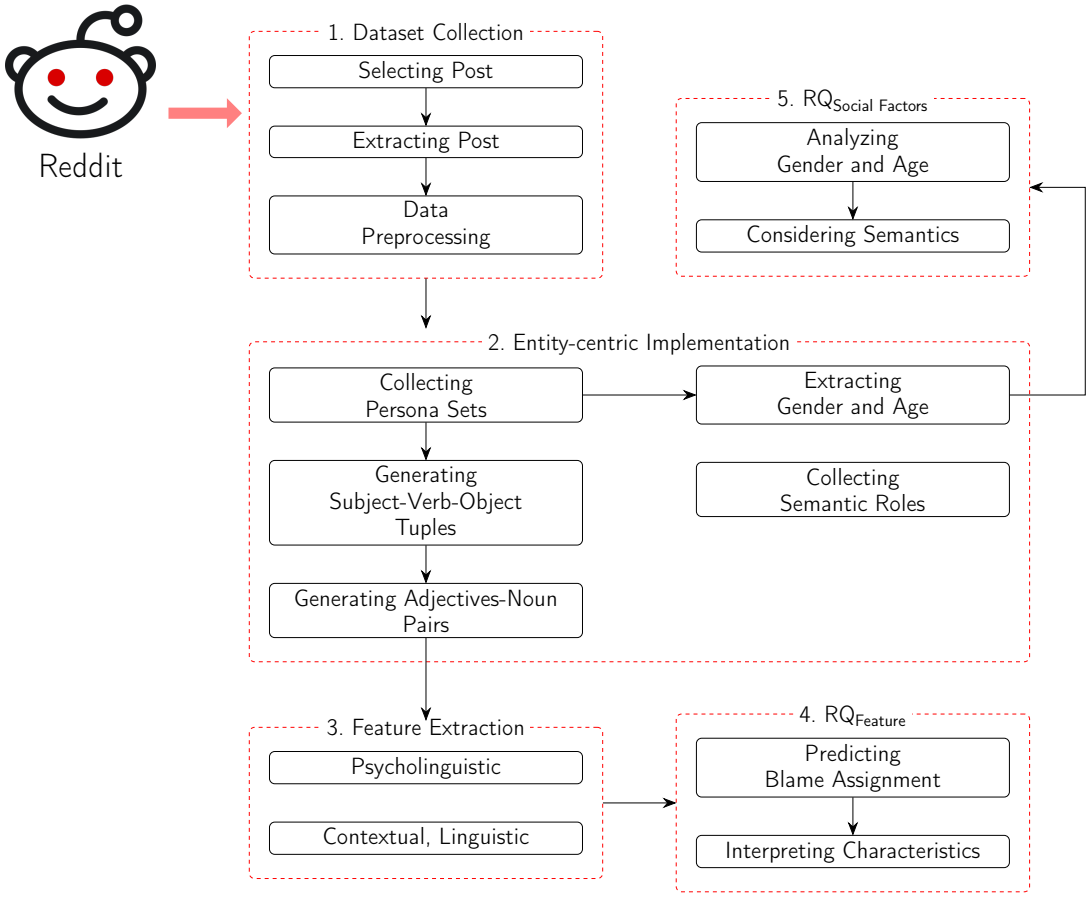
Fig. 2.1 shows the main phases of our framework:

1. *Dataset collection* involves collecting data from a subreddit and preprocessing the data into a proper format.

2. *Entity-centric implementation* involves separating entities, generating subject-verb-object (SVO) tuples, collecting semantic roles, extracting gender and age, and generating adjectives-noun pairs (ANP).

3. *Feature Extraction* includes measuring psycholinguistic, contextual, and linguistic features.

Figure 2.1: Overall pipeline of the proposed method.



2.2.1 Dataset Collection

Selecting posts from Reddit

Although AITA has been used in previous studies, they are either not public Zhou et al. (2021); De Candia et al. (2022) or insufficient for answering our research questions Lourie et al. (2020); Nguyen et al. (2022). Our work needs the selected posts to contain predicate-argument structures that previous works haven't mentioned. To improve relevance and accuracy, we constrain our dataset (FAITA) (details are in Section 2.2) to include posts that have:

- Been given flairs (the determined verdict).
- At least 50 top-level comments (judgments).
- Majority votes the same as the flair.
- At least ten extractable subject-verb-object tuples and ten extractable adjectives-noun pairs.

Extracting Post

We use the PushShift API[2] and Reddit API[3] to extract data over July 2020–July 2021. Some AITA stories may be faked to solicit outrage. The moderator deletes posts that are not truthful or not about interpersonal conflicts, which violate the subreddit rules.[4] We remove undesirable posts—those deleted, from a moderator, or too short—to ensure that the posts in our dataset decrease the conflicts between two parties and avoid discrepancies between data from Reddit and the archived data from PushShift.

Each judgment in the comments takes the form of a code: YTA, NTA, ESH, NAH, and INFO, which correspond to the classes AUTHOR, OTHER, EVERYONE, NO ONE, and MORE INFO. However, labeling the post with the majority votes may be inaccurate because morality is relative. Instead, we extract the *title*, *text*, and *flair* of each post. The *Flair* of each post is determined by the verdict of the top-voted comment 18 hours after submission (or the majority computed from its ten top-level comments if there is no flair field). We assign labels to YTA as 1 and NTA as 0 and discard other codes. This process yields 32,696 posts. We randomly split posts into 80% as the training, 10% as the development (dev), and 10% as the test sets. Tab. 2.2 shows the distributions of FAITA.

[2] `https://github.com/pushshift/api`
[3] `https://www.reddit.com/dev/api`
[4] `https://www.reddit.com/r/AmItheAsshole/about/rules/`

Table 2.2: FAITA dataset distributions.

Dataset	Train	Dev	Test
Posts	26,156	3,270	3,270
Sentences	376,846	125,766	125,332
Author Wrong (label 1)	9,874	1,182	1,238
Others Wrong (label 0)	16,282	2,088	2,031

Data Preprocessing

We combine the title and text of posts in FAITA. We preprocess the text using the NLTK toolbox.[5] We remove all emojis, punctuation (except periods for separating sentences), symbols, and special characters and replace contractions with patterns (e.g., replace *can't* with *can not*). We tokenize the sentences and lemmatize tokens using WordNet Lemmatiser Poria et al. (2012). We identify a "sentence" as *words separated by a period in the original post*.

2.2.2 Entity-Centric Implementation

We build a set of syntax-aware methods for extracting the protagonist (author) and antagonist (others) of each post using entity coreference and the syntactic dependency parse. These **entity-centric methods** require partitioning entity tokens into the protagonist and antagonist persona sets, understanding how the authors are portrayed the "cast of main characters" in the narratives, and how these characters behave. We use Semantic Role Labeling (SRL) Jurafsky and Martin (2009) to identify the protagonist and antagonist in each post.

Collecting Persona Sets

The protagonist and *antagonists* persona sets, respectively, contain first-person pronouns (e.g., I, me, and we), and third-person pronouns (e.g., she, he, and they). We add the pronouns to the persona sets as key tokens. We use the Spacy[6] dependency parser to extract more candidate terms by identifying part of speech tags (e.g., PRON, PROPN, and NOUN). We filter the nouns and proper nouns by a total of 3,125 people-related words from prior research, such as characters in history textbooks Lucy et al. (2020). Thus, we can collect all the people-related nouns and proper nouns. Then we use Huggingface[7] `neuralcoref` for coreference resolution,

[5] `https://www.nltk.org/`
[6] `https://spacy.io`
[7] `https://huggingface.co`

and append all tokens from spans that corefer to the pronouns in protagonist or antagonist persona sets, respectively.

Collecting SRL

SRL analyzes sentences with respect to predicate-argument structures such as "*who* did *what* to *whom* and *when* and *how* and *why*." We employ the AllenNLP BERT-based Semantic Role Labeller Gardner et al. (2018) to extract spans tagged `ARG0` for *agent* and `ARG1` for *patient*. As the following example shows, each sentence may have multiple tagged spans; thus, we first identify the SRL-tagged sentences.

1. **They** (`ARG0`) claimed **me** (`ARG1`) a dependent even though **I** (`ARG0`) have been financially independent for about a year.

In each post, we match the entities in persona sets with SRL labels `ARG0` or `ARG1`. This enables us to find when the author describes themselves or others as *agent* or *patient*.

Generating Subject-Verb-Object (SVO) Tuples

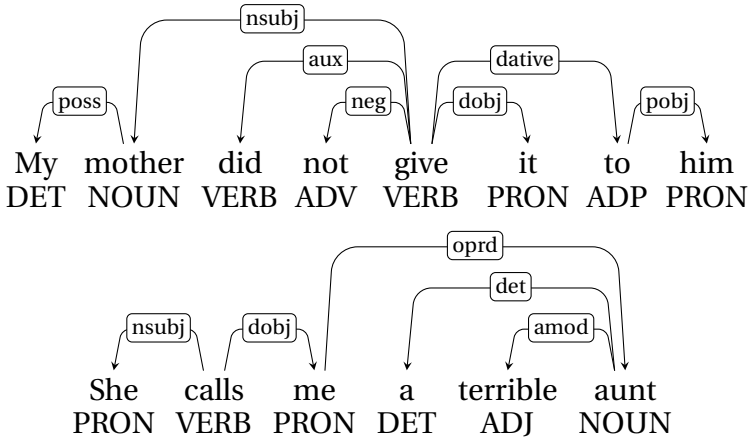
Beginning from the persona sets, we identify verbs (VERB) that have dependencies with entities in the persona sets using a syntactic dependency parse tree. We consider entities typed `nsub` (nominal subject), `nsubjpass` (passive nominal subject), `csubj` (clausal subject), `csubjpass` (passive clausal subject), `xsubj` (controlling subject), to the verbs as subjects; we consider entities typed `dobj` (direct object), and `iobj` (indirect object), to the verbs, as objects. Then, we add the SVO tuples for persona sets accordingly. In addition, we generate new SVO tuples by finding entities from spans that corefer to the subject or object. Using a dependency parser, it is possible to handle the negation of the verbs and add a "not" before the verb as shown in Fig. 2.2.

Fig. 2.2 shows two sentences in one post and their dependency trees. We identify four referents, *my mother, him, she, me*, and two SVO tuples, i.e., *(my mother, not give, it)* and *(she call me)*. Coreference resolution finds *she* corefers with *my mother*, so we add *(my mother, call, me)* to the SVO tuples.

Generating Adjectives-Noun Pairs (ANP)

Adjective-noun pair is a semantic construct for capturing the effect of an adjectival modifier to modify the meaning of the nouns such as "cute dog" or "beautiful landscape." Similarly, we use a dependency parse tree to identify adjectives for the entities in the persona sets. We use `amod` (adjectival modifier), `acomp` (adjectival complement), and `ccomp` (clausal complement)

Figure 2.2: Dependency parsers for the example sentences.



dependencies to select the adjectives modifying the entities. As shown in Fig. 2.2, after we add *aunt* to the protagonist persona set, we find *terrible, aunt* pair because of the `amod` tag. We also add *terrible, me* because *aunt* corefers to *me*.

Extracting Gender and Age for Persona Sets

Posts on AITA are interpersonal stories; the complexity of the descriptions makes it nontrivial to extract an antagonist's social factors. Following previous works, the present work focuses on protagonist's social factors De Candia et al. (2022); Botzer et al. (2022). Gender and age identities 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) or `(?i:i|i am a)([mf]|(?:fe)?male))`, enables us to use regular expressions to extract the information. We extract age by matching on the gender modifier or the numeric age. To improve accuracy, we consider two patterns: *I [25m] and my wife [25f]*. 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 omit nonbinary genders because we do not have ground truth labels for nonbinary targets. We evaluated the regular expression via a random sample of 300 submissions. We found no false positives and 2% false negative cases. When gender and age are extracted by regular expression, it matches the manually labeled one 94% of the time.

Table 2.3: Feature categories and explanations.

Category	Feature	Explanation
Contextual	Topic	Lexicon-based topics measured by LDA Blei et al. (2003); each post has a list of topic it belongs to
	Content	TF-IDF weighted n-grams (n=1,2)
Psycholinguistic	Agent versus Patient	Ratio of author and others being an *agent* or a *patient* Schein and Gray (2018)
	Connotation Frames	Scores of connotation frames-related Sap et al. (2017) words normalized by count of subject-verb-object tuples; the scores are calculated separately as writers' perspective, value, effect, mental state
	Agency and Power	Agency and power scores Rashkin et al. (2016) normalized by count of subject-verb-object tuples
	Moral Content	Occurrences of the five virtue-vice paired Moral Foundation Theory lexicon Hopp et al. (2020) normalized by count of subject-verb-object tuples and count of adjectives-noun pairs
	Valence, Arousal, Dominance (VAD)	Occurrences of VAD lexicon Mohammad (2018) normalized by count of subject-verb-object tuples and count of adjectives-noun pairs
	Emotion	Occurrences of Emotion lexicon Mohammad and Turney (2013) normalized by count of subject-verb-object tuples or count of adjectives-noun pairs.
Linguistic	Subjectivity	Occurrences of subjectivity-related words Wilson et al. (2005) normalized by count of words
	Hedge	Occurrences of hedge words Hyland (2018) normalized by count of words
	Modal	Occurrences of modal words normalized by count of words
	Pronoun	Occurrences of first, second, and third pronouns
	Sentiment	Averaged VADER Hutto and Gilbert (2014) compound scores; nominal sentiment categories

2.2.3 Feature Extraction

We categorize the features into **contextual**, **psycholinguistic**, and **linguistic** features. We measure the psycholinguistic features for subject-verb-object tuples and adjectives-noun pairs of the protagonist and antagonist persona sets separately. We calculate scores for other features of each post. Tab. 2.3 summarizes the categories.

Contextual Features

Content is essential in analyzing social media posts Guo et al. (2020); Zhou et al. (2021). We extract the content at two levels: term frequency-inverse document frequency (TF-IDF) weighted n-grams vectors (n = 1, 2) and narrative topics. TF-IDF weights combine term frequency $tf(t,d)$ (the occurrence of a term t in a document d) and inverse document frequency $idf(t,D)$ (the rating of t in a corpus D). It reflects how important a word is to a document in a corpus.

Table 2.4: Sample topics with representative words.

Topic Label	Top Weighted Words
Relationship with family (20.8%)	life, relationship, mother, ex, child, father, life, wife, partner, son
Intimate relationship (17.3%)	girlfriend, boyfriend, relationship, dating, upset, feel, pretty, lot, love, guy
Living in shared accommodation (16.5%)	apartment, rent, live, room, living, house, lease, stay, bedroom
Money (7.3%)	pay, rent, saving, buy, job, account, car, loan, afford, cost
Pregnancy concerns in pets (5.5%)	dog, child, husband, child, pregnant, puppy, cat, law, animal, birth
Work (4.4%)	hour, work, boss, company, manager, job, employee, office, shift, week
Appearance judgment (4.2%)	hair, look, wear, white, black, comment, clothes, dress, looked, pretty
Neighborhood (3.3%)	neighbor, phone, email, post, account, people, use, street, yard, facebook

We extracted topics using Latent Dirichlet Allocation (LDA) Blei et al. (2003). LDA is a generative statistical model that assumes that each document (here, post) contains a distribution of topics; each topic is a distribution of words. We train a model on the *text* of posts in FAITA, exploring the number of topics ranging over 30–55 and finalized on 30, as it achieves the lowest perplexity. We then combine the topics that contain fewer than 200 posts. Tab. 2.4 shows a sample of eight hand-selected topics and ten example words belonging to each topic. In Tab. 2.4, the "Topic Label" column is summarized manually by the authors according to all the words they are associated with; the percentage indicates how frequently each topic occurs in the dataset.

We also show the ten most representative words for each topic. These topics show that posts in FAITA range from family to work issues. Additional topics with at least 100 posts include: driving safety (2.8%), gender communication differences (2.8%), games (2.6%), cooking (2.7%), holiday gifts (2.7%), social media (2.3%), wedding plan (2.1%), medical treatment (1.6%), and school (1.1%).

Psycholinguistic Features

These refer to the lexico-semantic analysis of the cognitive association that a word carries and its literal meaning. The scores being introduced are separated into *agent*, *connotation frames*, *power and agency*, *moral content*, and *VAD*. From SVO tuples, we calculate scores for entities as subjects and objects. From ANP, we calculate scores for entities based on their adjective modifier. We normalize the scores calculated for entities in the persona sets to capture the values of the protagonist and antagonists.

Agent The ratio of the times the protagonist and antagonists are assigned as *agents* or *patients* using SRL.

Connotation Frames A formalism for analyzing subjective roles and relationships implied by a given predicate Rashkin et al. (2016). To analyze nuanced dimensions of narratives in FAITA, we draw from a lexicon with annotations for 1,000 most frequently used English verbs across various dimensions, ranging from –1 to 1. A verb might elicit a positive sentiment for its subject but imply a negative sentiment for its object. For example, from "*Alice* betrayed *Bob*," the annotation contains the following dimensions:

- Writer's perspective. The writer (protagonist) elicits a negative perspective toward *Alice* as –0.67 (e.g., blaming) and a positive perspective toward *Bob* as 0.26 (e.g., supportive).
- Reader's perspective. (1) Values: the reader presupposes a positive value of *Bob* as 0.87 (strongly positive) and *Alice* as 0.47 (neutral to positive). (2) Effects: the reader presupposes the harms towards *Bob* as –0.93 (strongly negative) compared to *Alice* as 0.067 (neutral). (3) Mental states: the reader presupposes *Bob* is most likely to feel negative (–0.67) as a result of the event, but *Alice* it not likely to be affected (–0.03).

Power and Agency A pragmatic formalism organized using frame semantic representations Sap et al. (2017) to model how different levels of power and agency are implicitly projected on people through their actions. We use Sap et al.'s (2017) extension lexicon of Connotation Frames to measure the agency and power scores of *author* and *others*. This extension lexicon

contains more than 2,000 transitive and intransitive verbs to model how different levels of power and agency are implicitly projected on the entities through their behaviors. Entities with high agency (subjects of *attack*) are active decision-makers, whereas entities with low agency (subjects of *doubts* and *needs*) are passive. This lexicon contains binary labels of each verb, which are positive (1), equal (0), and negative (–1).

Moral Content The Moral Foundation Theory Haidt and Graham (2007) has been widely adopted in the computational social community, which is critical in understanding how the psychological influence of social content unfolds, such as quantifying moral behaviors in Twitter Joe et al. (2020) and taxonomizing the structure of moral discussions in Reddit Nguyen et al. (2022). We adopt the extended Moral Foundations Dictionary (eMFD) Hopp et al. (2020), which is a crowdsourced dictionary-based tool for extracting moral content from textual corpora. The eMFD contains 2,041 unique words, which are categorized into five broad domains based on MFT: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. Each word in the dictionary has a composite valence score ranging from –1 to 1.

VAD (Valence, Arousal, Dominance) The three affective dimensions are used to measure affective meanings from words that convey the author's attitudes toward the events and people referenced. We obtain the valence scores for 20,000 words from the NRC VAD lexicon Mohammad (2018), which contains real-valued scores ranging from 0 to 1 for each category.

Emotions Emotions conveyed in words represent sentiment from the authors toward the described entities Mohammad and Turney (2013), which may place a considerable cognitive load on the audience Dijkstra et al. (1995). The NRC Emotion lexicon Mohammad and Turney (2013) provides the emotion of around 20,000 words, indicating whether a word is associated with an emotion category. The categories are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation.

Linguistic Features

We estimate linguistic scores for *subjectivity*, *hedge*, *sentiment*, and *modal* in each post.

Subjectivity arises when people express personal feelings or beliefs, e.g., in opinions or allegations Wilson et al. (2005), which comprises the authors' perspectives towards the descriptive situations, contributing to the audience's judgments. We compute the subjectivity of a post

as the average score of words based on the Subjectivity lexicon Wilson et al. (2005) (nonneutral words of "weaksubj" = 0.5 and "strongsubj" = 1). Additionally, we count the numbers of first-person, second-person, and third-person pronouns because words such as "you" and "we" engage the audience with the discourse.

Hedge is associated with indirection in politeness theory Brennan and Ohaeri (1999), which may affect the audience's judgments.

Sentiment indicates emotions by conveying the polarity of an opinion. A negative tone may imply more immorality than a neutral tone. We calculate each post's compound sentiment scores and sentiment categories using VADER Hutto and Gilbert (2014).

Modal words affect the sentiment of the words they modify Kiritchenko and Mohammad (2016).

$RQ_{Feature}$

2.3 $RQ_{Feature}$: Blame Assignment Analysis

To answer $RQ_{Feature}$, we perform two statistical analyses: (1) prediction—can description frames predict blame assignment? (2) language characteristics analysis—can linguistic features affect blame assignment? Here, we focus on using the prediction as a tool for analyzing, not for the purpose of making an accurate prediction. Note that we do not take gender and age as features when conducting experiments as they are not available in some of the posts in FAITA.

2.3.1 Predicting Blame Assignment

Now we examine how well computational models can predict blame assignments in moral situations. For machine learning models, we explore two logistic regression models (LR) to compute the probability of a positive label for each sentence. All the models are built using scikit-learn[8] toolkit in Python. An LR classifier computes the probability of a discrete outcome given an input variable. BERT-LR is logistic regression where our features are replaced with the BERT Devlin et al. (2019) embeddings of input sentences. We evaluate the performance of different models in terms of recall, precision, and F1 scores. All computation models were run 10 times and we measure the standard deviation of the scores for each method. For LR, we

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html`

Table 2.5: Blame assignment prediction accuracy (macro-average scores).

Method	F1		Recall		Precision	
	DEV	TEST	DEV	TEST	DEV	TEST
Random	0.49	0.50	0.50	0.50	0.50	0.49
Length	0.39	0.38	0.50	0.50	0.49	0.52
LR	0.66	0.65	0.65	0.64	**0.66**	**0.65**
(×) Linguistic	0.63	0.62	0.60	0.61	0.63	0.62
(×) Contextual	0.53	0.53	0.54	0.54	0.60	0.60
(×) Psycholinguistic	0.48	0.49	0.52	0.53	0.59	0.59
BERT	**0.71**	**0.72**	**0.68**	**0.69**	**0.66**	**0.65**

set the class weights to "balanced" to account for the label imbalance. And we explore feature selection using the L1-norm and regularization using L2-norm. Other hyperparameters for LR include setting the weight ranging over $(1e-4, 1e-3, 1e-2, 1e-1)$. We propose two baseline models. Random predicts the verdict randomly. Length predicts using the lengths of the sentences in a post, which has been shown to be effective in predicting blame Zhou et al. (2021).

The quantitative results of our methods are shown in Tab. 2.5. All scores have standard deviations between 0.01 and 0.03. The best scores are in bold. We only report LR and BERT-LR results because they yield the performances of other models such as multilayer perceptron, SVM, and random forest. BERT-LR and LR outperform the baselines significantly, while BERT-LR performs best.

It is worth noting that our features, despite the lower performance than BERT-LR, are clearly informative of morality prediction because they directly capture the information contributing to the audience's decision on blame. Transformer models such as BERT encode linguistic characteristics in a more sophisticated manner and may include additional information. But it is less clear exactly what transformers capture and whether they capture irrelevant statistics. To examine the contribution of each feature category, we conducted ablation tests based on the LR model. Regarding F1 scores, psycholinguistic features have the highest contribution, followed by contextual and linguistic. This result reaffirms the importance of analyzing the lexical semantics of attributive and predicative words in first-person moral narratives.

2.3.2 Interpreting Characteristics

We measure the effect size and statistical significance of each feature. The effect of each feature is conditioned on the domain of each post using logistic regression. For interpretation purposes,

we use the Odds Ratio (OR) (the exponent of the effect size). Odds represent the ratio of the probability of an author being blamed to the probability of not being blamed; OR is the ratio of odds when the effect size increases by one unit. The OR is calculated using the equation $OR = exp(\beta_i), i \in N$, where β_i is the coefficient of attribute i obtained by the trained LR model and N denotes the attribute set. Moreover, we estimate statistical significance by Spearman's Rank Correlation Coefficient to avoid assuming normality or other distributions for FAITA.

Contextual Features

We begin by looking at OR between topics and blame assignments. Table Tab. 2.6 shows the OR values and correlation coefficients for the authors being blamed corresponding to Tab. 2.4. These results show that posts related to relationships, pregnancy concerns in pets, and games are positively correlated with blame assignment. Other topics mentioned in Section 2.2.3 that may decrease the probability of an author being blamed are school, holiday gifts, and cooking; the rest are positively correlated.

Table 2.6: Odds ratio (OR) and Spearman's correlation coefficient of topics calculated from the test set. An effect is positive (blue) if OR > 1 and negative (red) if OR < 1.

Topic	Moral Blame	p-value
Relationship with family	1.11	0.02
Intimate relationship	1.07	0.07
Living in shared accommodation	0.79	0.02
Money	0.82	0.20
Pregnancy concerns in pets	1.46	0.03
Work	0.98	0.03
Appearance judgment	1.16	0.07
Neighborhood	0.71	0.14

Psycholinguistic Features

Tab. 2.7 shows the features that influence at least a 1% probability of the author being blamed. Tab. 2.7 reveals that psycholinguistic features are informative. The results for agent and patient are consistent with social psychology that "being a victim can help escape blame" Gray and Wegner (2011). These features do not store lexical information but affect the audience's judgments in the cognitive aspect. In addition, our analysis indicates the protagonist can reduce blame by eliciting positive perspectives (e.g., supportive) towards themselves. Additionally,

Table 2.7: Odds ratio (OR) and Spearman's correlation coefficient of psycholinguistic features calculated from the test set. WP represents *writer's perspective*. An effect is positive (blue) if OR > 1 and negative (red) if OR < 1.

Feature	Protagonist		Antagonist	
	Moral Blame OR	p-value	Moral Blame OR	p-value
Agent	1.93	0.05	0.93	0.002
Patient	0.53	0.03	1.01	0.001
WP	1.01	0.006	0.81	0.031
Value	0.99	0.13	1.02	0.13
Power	1.04	0.006	0.97	0.003
Agency	2.00	0.003	0.96	0.002
Care	0.99	0.03	1.03	0.03
Harm	0.97	0.08	1.00	0.07
Betrayal	0.95	0.06	1.11	0.16
Loyalty	0.97	0.08	1.03	0.17
Valence	0.99	0.14	1.22	0.13
Arousal	1.04	0.11	1.21	0.15
Dominance	1.23	0.14	1.09	0.15
Joy	0.98	0.09	0.98	0.13
Sadness	0.31	0.05	1.28	0.005
Anger	1.05	0.01	0.11	0.03
Fear	2.33	0.01	1.06	0.03
Trust	1.10	0.08	0.20	0.09
Disgust	1.06	0.07	2.16	0.02
Anticipation	1.34	0.05	1.74	0.04

authors can reduce blame when they describe themselves as suffering more from harm than the antagonist. The Agency and Power features are consistent with the above findings because the high values imply the agent's high-level authority and powerful capability, which can trigger blame.

Care and harm are opposite concepts in Moral Foundation Theory, whereas increasing the use of words from the lexicon reduces the probability of the author being blamed. Moreover, we find that VAD features do not have significant p-values. However, they increase the probability of the author being blamed by 23% when increasing the use of the dominance lexicon when describing the protagonist. Different emotion categories have different effects on blame assignment. Specifically, using sadness-related words when describing the protagonist lowers the probability of the authors being blamed to one-third. Additionally, increasing the use of

disgust-related words when describing the antagonist more than doubles the probability of the author being blamed. We highlight a possible explanation: the description frames of the protagonist and antagonist need to be captured as a whole, not as individual components.

Linguistic Features

As shown in Tab. 2.8, subjectivity is positively correlated to blame assignment in contrast to hedging, which indicates that subjective descriptions increases the possibility of the author being blamed with greater certainty. The frequent use of third-person pronouns triggers blame

Table 2.8: Odds ratio (OR) and Spearman's correlation coefficient of linguistic features calculated from the test set. An effect is positive (blue) if OR > 1 and negative (red) if OR < 1.

Feature	Moral Blame	p-value
Subjectivity	1.09	0.006
Hedge	0.66	0.04
First pronoun	1.45	0.10
Second pronoun	1.01	0.0009
Third pronoun	1.96	0.003
Sentiment score	1.78	0.001
Sentiment: positive	1.18	0.005
Sentiment: neutral	0.99	0.08
Sentiment: negative	3.18	0.01

because the audience may think that the author is trying to escape from blame by avoiding describing themselves. Although second-person pronouns have a small p-value, they increase the probability only by 1% of the author being blamed. However, the negative sentiment category strongly affects blame assignment with an OR of 3.18, which may explain that extreme sentiment triggers blame assignment.

2.4 RQ_{Social}: Social Factors Analysis

In this section, we examine gender and age features in FAITA to investigate whether audiences exhibit differences in their assessments of moral situations.

2.4.1 Analyzing Gender and Age Association

This section investigates whether the authors' self-reported gender and age lead to an imbalance in blame assignment. Using the method of Section 2.2.3, 13,935 posts describe the genders of all entities involved, and 6,079 posts state the authors' age is between 15 and 65. To determine the association between blame and social factors, we perform the χ^2 significance test and compute Cramer's ϕ as the effect size. Here, 0.07–0.21, 0.21–0.35, and >0.35 respectively indicate small, moderate, and strong association Cohen (1988).

We aggregate occurrences of entities being blamed when they are protagonists and antagonists. The overall χ^2 test result between genders and blame assignment is ($\chi^2(13,935) = 515.02, p < 0.001$) with $\phi = 0.17$. Whereas the effect size indicates a small association between gender and blame assignment in FAITA, the evidence indicates there is an association between the two ($p < 0.001$). Besides, we observe that males are 53% (the log-odds-ratio of occurrences when authors of different genders receive blame) likelier to receive blame. The results allow us to discern the direction of the biases due to gender: male authors are likelier to be considered agentive no matter their position. The observation coheres with previous psychological research that some sets of biases stereotype females into the role of suffering *patient* on social media Reynolds et al. (2020).

To further investigate the correlation between blame assignment and age, we divide authors' ages (antagonists' ages are scarce) into four groups ranging from 15 to 55 as the range accounts for almost 80% of active Reddit users. Tab. 2.9 illustrates blame assignment is associated with protagonists' ages when they are in the 15–45 age group ($p < 0.05$), especially when authors are in the 36–45 age group ($\phi = 0.18$).

Table 2.9: The columns are age ranges. N represents the number of corresponding posts. $p < 0.05$ indicates the age group and blame assignment are associated.(** : $p < 0.05$, ***: $p < 0.001$.)

	Age Ranges			
Metrics	15-25	26-35	36-45	46-55
N	3,554	1,951	410	136
χ^2	76.56 (***)	50.89 (***)	13.46 (**)	2.96 ()
ϕ	0.15	0.16	0.18	0.15

2.4.2 Considering Semantics with Social Factors

We now examine how blame assignment differs between female and male protagonists in similar situations. We employ pretrained sentence-BERT models Reimers and Gurevych (2019) to cluster the 13,935 posts based on semantic similarity. We learn embeddings of the posts' *titles* as they serve as summaries of posts. To remove the effect of gender-related tokens, we replace gender-identified words with "someone" using the resources mentioned in Section 2.2.2.

We adopt Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) McInnes and Healy (2017) because no external references identify topic numbers in FAITA. Then we perform dimension reduction with Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) McInnes et al. (2018) to alleviate the problem of sparse embeddings. We fine-tune parameters for HDBSCAN and UMAP models Angelov (2020) by increasing the model's Density-Based Clustering Validation (DBCV) score Moulavi et al. (2014). To enhance the semantic similarity in each cluster, we exclude the clusters containing fewer than 50 posts. This process yields 7,248 posts that clustered into 47 groups, where the counts of posts in a cluster range from 51 to 712. We measure χ^2 and ϕ to select the clusters where gender is strongly associated with blame assignment ($p < 0.001$ and $\phi > 0.35$) and find six clusters include 1,162 posts.

To categorize the semantics of the clusters, we use the UCREL Semantic Analysis System (USAS) USAS, a framework for automatic semantic analysis and tagging of text, which is based on McArthur's Longman Lexicon of Contemporary English Summers and Gadsby (1995). USAS has a multitier structure with 21 major discourse fields subdivided in fine-grained categories such as *People*, *Relationships*, and *Power*. Using USAS, we label each cluster with the most frequent tag (or tags) among the highest TF-IDF-scored nouns from the posts. We notice that the topics of the most gender-associated situations in FAITA corroborate previous work on categorizing language biases in Reddit. For example, the most frequent gender-polarized situations on FAITA (ordered by frequency) are *kin*, *relationship: intimate/sexual*, *groups and affiliation*, *anatomy and physiology*, *work and employment*, *sports*, *games*, *money*, *medicines and medical treatment*, and *judgment of appearance*. These tags account for the most discussed topics, as Tab. 2.4 shows. It is important to note that our analysis of social factors in blame assignment is more suggestive than conclusive. Our analysis suggests that social biases exist in social media posts, which influences blame assignment, at least in some topics.

2.5 Discussion

This paper contributes to the study of morality by assessing social psychology insights on descriptive real-life situations. We incorporate a novel set of language features to predict blameworthiness. Statistical methods help visualize the effects of the features. The effective prediction performance confirms the linkage between blame and psychological theories. For example, entities described using fewer *care*-related words Haidt and Graham (2007) are likelier to receive blame. Furthermore, our findings suggest that gender and age are associated with blame assignment. For example, males are likelier to receive blame than females; biases in blame assignments are likelier to be presented when protagonists are aged 15–45.

Our results can be explained by the fact that people perceiving themselves as deserving blame are subject to feelings of guilt Scott (1971). These feelings conveyed from social media posts may affect the audience's decision on who is blameworthy. In addition, the psychological literature observes that social media might have typecasting towards male and female individuals Reynolds et al. (2020). However, people of different genders might be subject to different social pressures and thus be different in choosing the language to describe conflict Asher et al. (2017). Our results agree with these observations and support the use of social psychology instruments in computational methods to understand morality Fraser et al. (2022).

2.5.1 Implications

Our work contributes a new framework that combines NLP and social psychology and suggests theoretical and empirical guidelines for the study of practical morality. First, our research provides novel language features that combine linguistic and psychological insights into morality. Second, these features provide a basis for interpretable models of morality. Practically, this work could motivate the design of future AI systems that act morally and interpretably so.

Our study contributes by providing language features that can undergird theoretical research. For example, our analysis shows how individuals' agentiveness is affected by the associated descriptions. Our work can help design empirical studies through a more careful use of the language used in vignettes.

2.5.2 Limitations and Future Work

As in any study dealing with social media data, there are some limitations. Although our study gains from a high ecological validity, it presents critical causal inference challenges. Hidden confounders include the demographics of the audience (including age, cultural background, and education) and other events they may have observed, e.g., through viral videos, that affect

their judgment. Therefore, the coefficients we find cannot be interpreted as causal predictors.

This work suggests interesting extensions. One direction is to focus on deeply moral situations such as MeToo posts Garg et al. (2023). Another is to incorporate language features from the comments accompanying each post to extract cognitive-affective features directly from the audience. The accompanying comments may help explain *what*, *why*, and *how* language features affect the audience's cognitive processes. In this light, the relationship of moral language and other dialogue phenomena such as derailment would be relevant Yuan and Singh (2023). In addition, to provide a causal explanation of how social factors appear in blame assignments and how they function, future work can leverage explicit and implicit social factors in the narratives.

CHAPTER

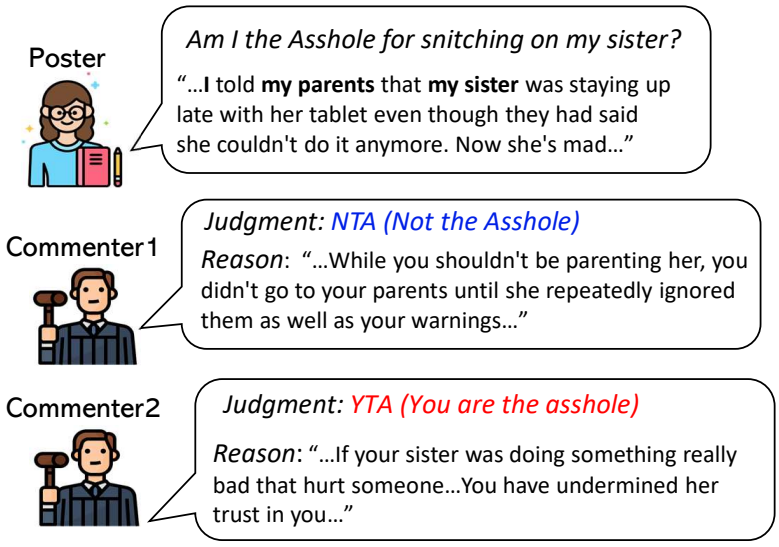
3

MORALITY IN THE MUNDANE: CATEGORIZING MORAL REASONING IN REAL-LIFE SOCIAL SITUATIONS

Moral reasoning reflects how people acquire and apply moral rules in particular situations. With social interactions increasingly happening online, social media data 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) where an author may describe their behavior in a situation to seek comments about whether that behavior was appropriate. A situation may decide other users' comments to provide *judgments* and *reasoning*.

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 reasoning 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 3.1: 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.



3.1 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 developments 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 experiments 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 provide an opportunity to assess the perception of normal social situations, such as understanding others' decisions on (im)morality of behaviors Lourie et al. (2020).

In this work, we study in-the-wild moral reasoning by examining a popular subcommunity of Reddit (i.e., subreddit) called /r/AmITheAsshole (AITA) AmITheAsshole. In AITA, a user (i.e., *author*) post interpersonal conflicts seeking others' opinions on whether their behaviors were appropriate. AITA defines a few verdict codes, such as NTA indicate authors' behaviors are appropriate, whereas YTA indicate authors' behaviors are inappropriate. Other community members (i.e., *commenters*) may comment on a post to provide moral *judgments* (i.e., verdicts, justifying the verdict (if any) and other moral assessment) and the *reasoning*. Fig. 3.1 shows a

post along with comments on it. Each comment includes a predefined community code along

Figure 3.2: Dependency graph representation of an example comment. The shading shows syntactic relations.

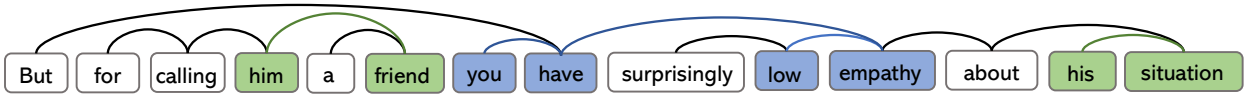
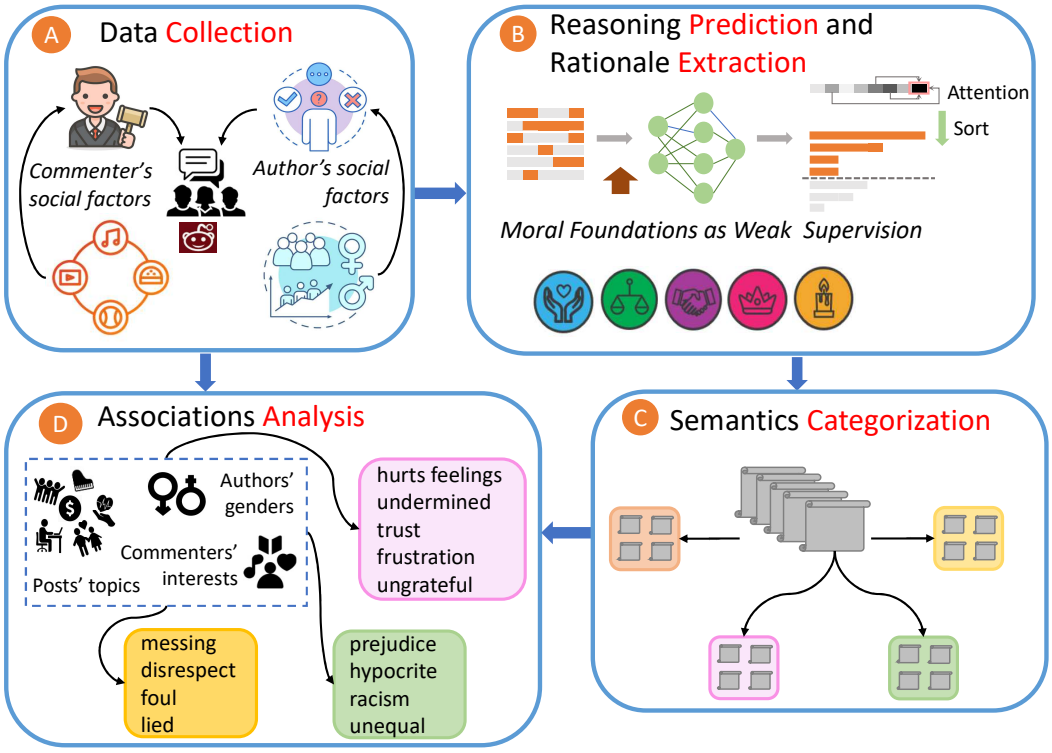


Figure 3.3: Flowchart depicting our research pipeline.



with an explanation for it. A verdict of a post is decided by the top-voted comment's verdict. Recent works focus on predicting verdicts of the posts and comments Lourie et al. (2020); Zhou et al. (2021); Botzer et al. (2022); Xi and Singh (2023). Another line of work analyzes the community using statistical methods Botzer et al. (2022); Nguyen et al. (2022); De Candia et al. (2022).

However, to the best of our knowledge, no empirical work has conducted a systematic analysis to understand the reasoning implicit in the comments. This paper focuses on the commonsense aspects of moral reasoning. We apply Natural Language Processing (NLP) tools to investigate how the authors' and commenters' social factors shape their distributions and affect moral reasoning. We extract authors' social factors from the posts. These include *authors' self-reported genders* and *posts' topics*. Here, we regard a post topic as a part of the author's social factors because the topic 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 et al. (2022); De Candia et al. (2022). For commenters' social factors as proxies of their *interests*, we leverage the subreddits in which they participate 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 reasoning can determine the verdict. For instance, in Fig. 3.2, the phrase *low empathy* refers to the author's behavior and determines a YTA verdict. With a large corpus, these verdict-determining 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. Fig. 2.1 describes our research pipeline. Our proposed method involves predicting reasoning and extracting the rationales, categorizing them by meaning similarities, and analyzing their associations with the abovementioned 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 Fig. 3.2) 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 apply the rationalization process Lei et al. (2016); Bastings et al. (2019); DeYoung et al. (2020) to extract rationales from the reasoning. The selected rationales are small but sufficient parts of the input texts 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, we follow Jiang and Wilson (2021) to "weakly" label rationales using a domain-related lexicon, the Moral Foundation Theory (MFT) Haidt and Graham (2007). 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 explore 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 in the judgments. 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 finetuned BERT in predicting verdicts of the reasoning. Moreover, our experiments demonstrate that with additional domain knowledge improve 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.

3.2 Related Work

Moral Reasoning in Social Psychology Moral reasoning has long been studied. Bussey and Maughan (1982) find that moral decisions by males are typically based on law-and-order reasoning, while 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 on 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 et al. (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 et al. (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 crowd-sourced 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 to receive negative moral judgments than romance-relevant posts in AITA. 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).

3.3 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 judgments and reasoning based on the posts.

3.3.1 Collection of Posts and Comments

We require a large-scared corpus with relevant posts and comments. Previous datasets are either nonpublic Zhou et al. (2021); De Candia et al. (2022); Botzer et al. (2022) or insufficient for our purposes Lourie et al. (2020); Nguyen et al. (2022). Therefore, we collected our dataset using

PushShift API PushShiftAPI and Reddit API Reddit API. 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 including fake content to solicit outrage. We also exclude 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 reasoning. 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*, which is an aggregate of number reported by Reddit representing the accumulated differences between upvotes and downvotes Reddit score.

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 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 et al. (2020), we assigned labels to comments with YTA as 1, NTA as 0, and discard all other instances.

3.3.2 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 reputed 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. Tab. 3.1 summarizes our dataset.

Table 3.1: Dataset summary.

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

3.4 Method

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

3.4.1 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 et al. (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, and the rest 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, such as in exploring contentious conversations

[1] 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).

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 [25f] and my wife [25m]*. 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/male split of 90–10 in our dataset We took a random sample of 300 submissions to evaluate the regular expression. 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, as defined by Beel et al. (2022), considering a user prolific if they submit more than 25 times in their interested subreddit. We manually checked 100 users and found that a six-month period makes users more prolific than four, eight, and twelve months. We discard deleted user accounts, which restricts our analysis to 46,519 commenters with 104,915 comments. Unlike De Candia et al.'s (2022) work, we map the collected subreddits following Reddit's predefined subreddit categories, collected from . Commenters may have interests in various categories. Therefore, we set their interest as their most frequently submitted subreddit.

3.4.2 Predicting Verdicts then Extracting Rationales

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

Introduction to Rationalization Process Given a pretrained model $\mathscr{M}$, 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 label. The rationalization process 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), and considered accurate explanations

Figure 3.4: 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 unmasked tokens; it predicts y again to evaluate a rationale's accuracy.

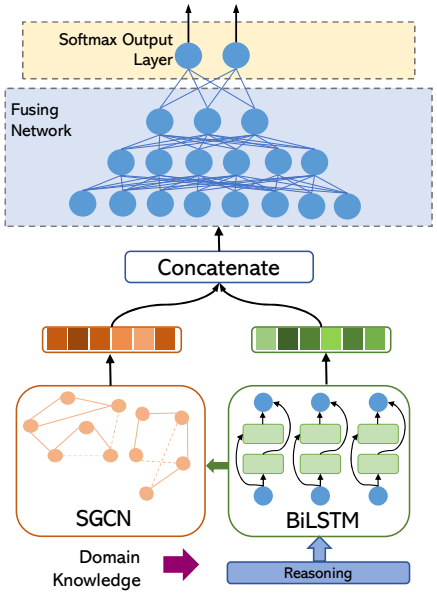
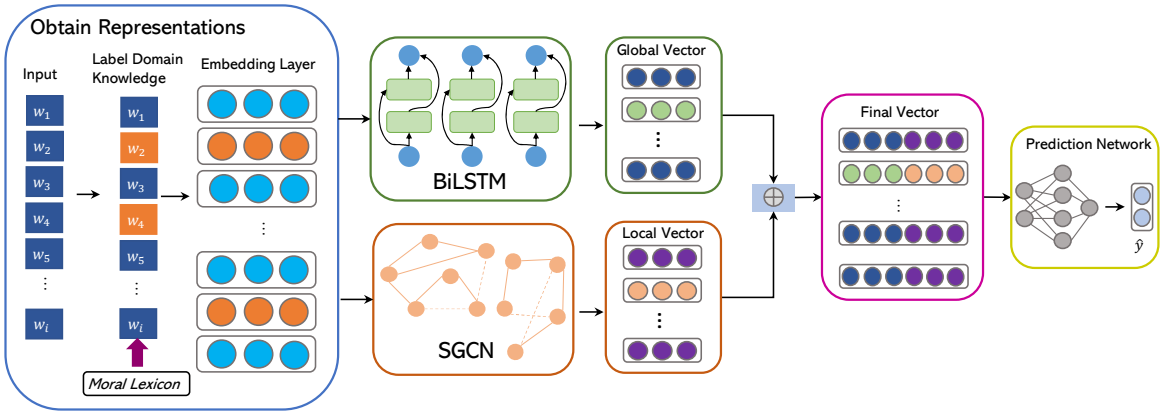


Figure 3.5: Architecture of the predictor in Fig. 3.4. Here, w_i represents a token in an input instance and $\widehat{y}$ is a predicted verdict. Tokens in ■ are labeled by an additional moral lexicon.



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., $z \sim \text{Binarizer}(x)$). 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 Fig. 3.4 illustrates the architecture of a soft rationalization model.[2] The predictor in Fig. 3.4 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.

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

Fig. 3.5 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 Huggingface. 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). 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. To obtain extended contexts, we use a Bidirectional LSTM (BiLSTM) Hochreiter and Schmidhuber (1997). We compute the hidden states by passing the BERT-encoded embeddings to a stacked BiLSTM:

$$\overleftarrow{h_{g,i}}; \overrightarrow{h_{g,i}} = \text{BiLSTM}(S), i = 1, 2, \tag{3.1}$$

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.

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

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 Spcay. 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 $(j+1)$ layer is:

$$h_v^{j+1} = \sum_{u \in \mathcal{N}(v)} W^j h_u^j + b^j, \tag{3.2}$$

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 $(j+1)$ and the j 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:

$$h_v^{j+1} = \sum_{u \in N(v)} g_{u,v}^j (W_{d_{u,v}}^j h_u^j + b_{u,v}^j), \tag{3.3}$$

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.

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 determine 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:

$$\arg\min L(y, \widehat{y}) + \lambda L_d(z, z_d), \tag{3.4}$$

where λ controls the weight of domain knowledge loss.

3.5 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.

3.5.1 Experimental Settings

Baseline Methods for Prediction We evaluate these machine learning models: the state-of-art transformer model BERT Devlin et al. (2019), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). For LR, we use the lengths of the instances as a baseline model. For traditional machine learning methods, we use GloVe Pennington et al. (2014), a text encoder that maps a word into a low-dimensional embedding vector, for textual classification to generate vector representations. For BERT baseline, we apply the "Obtain Representations" and "Prediction Network" modules as shown in Fig. 3.5. 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 et al. (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 $\mathscr{M}$ 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 drop 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 producing accurate classifiers. Therefore, we do not conduct human experiments 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 2e−5, $\epsilon = 1e{-}8$, 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 Eq. 3.3 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 from Huggingface.

3.5.2 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.

Table 3.2: 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).

Methods	F1 (%)	Precision (%)	Recall (%)
LR-Length	53.4	53.9	53.8
LR-GloVe	57.0	57.7	56.2
Random Forest	61.6	60.8	62.4
SVM	63.8	63.2	65.3
BERT	83.1	83.7	82.6
BERT-Domain	82.6	82.8	82.5
Global	83.0	83.0	82.8
Global-Domain	82.6	83.7	81.5
Local	83.5	82.9	84.2
Local-Domain	83.7	83.6	83.6
Global-Local	86.2	85.6	**86.9**
Global-Local-Domain	**86.4**	**86.8**	86.1

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 Tab. 3.2. We observe that 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. 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

Table 3.3: 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, results with "No Domain" are without the module. The best scores (the revF1 are the lowest; the NS and NC are the highest) in each column are shown in bold. The under-waved numbers are the highest NS and NC scores and lowest revF1 score among the three metrics.

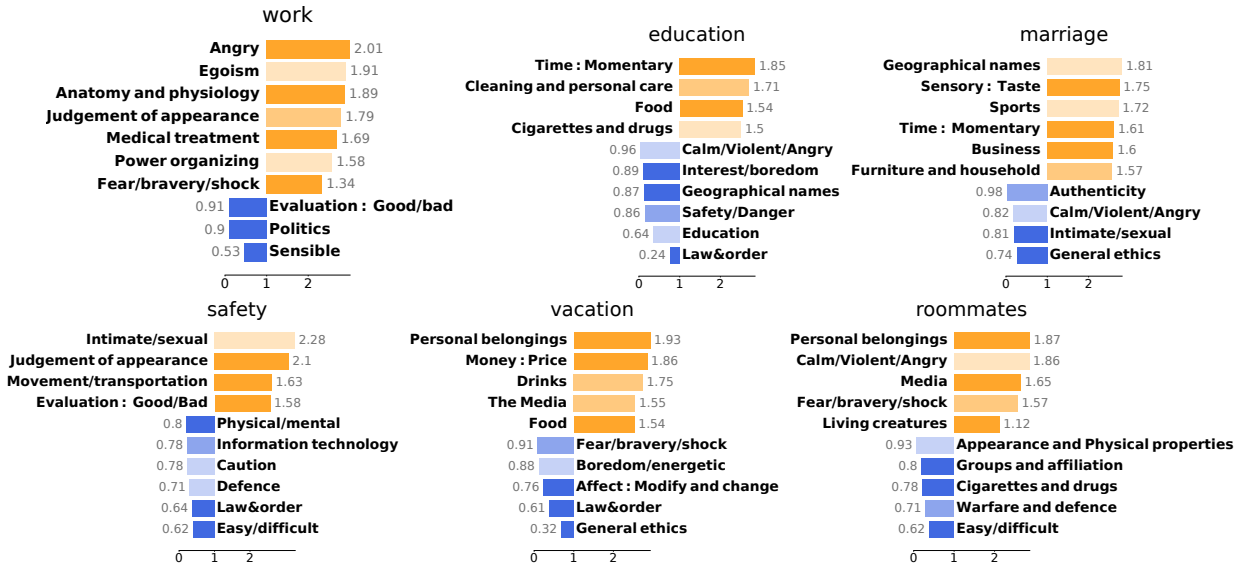
		Global			Local			Global-Local		
	Methods	revF1	NS	NC	revF1	NS	NC	revF1	NS	NC
Domain	RAND	85.9	0.26	0.27	79.3	0.25	0.27	84.0	0.20	0.34
	α	59.2	0.31	0.42	56.0	0.33	0.53	52.5	0.45	0.64
	$\alpha\nabla\alpha$	58.1	**0.47**	0.61	45.8	0.50	0.77	42.9	0.47	**0.81**
	IG	66.8	0.31	0.54	65.2	0.35	0.54	65.9	0.37	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

compared to a finetuned BERT. We are unable to compare our results with Botzer et al. (2022) because of different research purposes and lack of their dataset and experimental details.

Ablation Studies for Prediction We perform ablation studies to understand how global and local representations affect prediction performance. Tab. 3.2 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 Tab. 3.3 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 Tab. 3.2 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. Moreover, 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

(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.

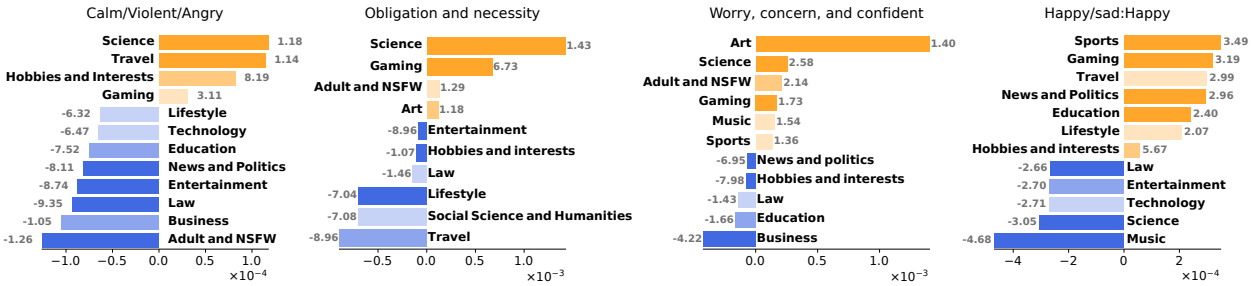


Figure 3.6: Results of OR values and social factors. We use orange rectangles ■ to indicate odds ratio greater than one and effects greater than zero (on the right), and blue rectangles ■ indicate the opposite (on the left). The shade shows the p-values: ■ and ■ (darkest): ≤ 0.0001, ■ and ■ (middle): ≤ 0.001, ■ and ■ : ≤ 0.05.

is **10 out of 15**, for Local **9 out of 15**, and for Global-Local **13 out of 15**. Combining the results of Tab. 3.2 and Tab. 3.3, we use Global-Local-Domain to predict and the FLX feature-scoring method to extract appropriate rationales.

Table 3.4: Examples of meaning clusters embedded in moral reasoning 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.

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, sexual, 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

3.6 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.

3.6.1 Clustering and Tagging

We apply pretrained GloVe embeddings Pennington et al. (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 Multilingual-USAS, 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.

3.6.2 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. Fig. 3.6 and Tab. 3.4 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. Fig. 3.6a shows the topics with OR, indicating polarized comments for authors' genders. Our analysis reveals consistent gender effects in certain categories, such as *Calm/Violent/Angry* in *education* and *marriage*, and *Judgment of appearance* in *work* and *safety*. Moreover, as indicated in Tab. 3.4, 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. In addition, adverbs used in *Evaluation: Good/Bad* are more prevalent than adjectives used in *Judgment of Appearance*. 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 fact that commenters have 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 societal pressure on women to conform to certain 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, which is a common method for analyzing social variables Stolzenberg (1980). The following model captures the linear effects:

$$b = \beta_0 + \beta_i x_i + \epsilon_i, i \leq n, \tag{3.5}$$

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.

Fig. 3.6b 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).

3.7 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 reasoning 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 moral reasoning 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 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.

CHAPTER

4

UNVEILING MORAL SPARKS: EXPLORING MORAL NARRATIVES IN REDDIT COMMUNITY

Moral narratives shared on social media platforms provide readers with a wide range of information using diverse descriptions. This research examines real-life moral narratives posted on subreddit r/AmItheAsshole, where users present interpersonal conflicts, and other community members (commenters) comment to vote for whose behaviors were inappropriate. This paper focuses on commenters' quoted sentences, referred to as "Moral-sparks" (M-sparks), which serve as pivotal points for moral decision-making.

We delve into two aspects: (1) posts' linguistic features that differentiate M-sparks and (2) comments' reasoning for highlighting M-sparks. From posts and comments submitted between 2013 to 2021, our results reveal that social commonsense in moral narratives can vary based on moral narrative domains. We also demonstrate the impact of moral and emotion-related linguistic features on the likelihood of quoting. In addition, the language used to describe entities influences the probability of excerpts being highlighted, with passive tense descriptions being more likely to be quoted.

4.1 Introduction

We propose to examine real-life moral narratives from the subreddit, r/AmItheAsshole (AITA), where users (poster) present interpersonal conflicts, and other community members (commenters) comment to vote for whose behaviors were inappropriate. Previous studies use AITA as a valuable resource for analyzing the entities and their described behaviors. Previous works use AITA as the resource to study Lourie et al. (2020) employ one-line texts extracted from the titles of AITA posts to investigate what behaviors are considered inappropriate. Giorgi et al. (2023) consider event chains as one of the key features for studying the characters of authors in moral narratives. Following these research, we identify *incidents* within the moral narratives as verbal events that connect entities, depicting specific behaviors of involved entities. Within a moral narrative, incidents hold varying degrees of importance and resonance for individual commenters, evoking different captures of their attention. Therefore, commenters may quote incidents from the original posts to engage in moral discussions. We name the quoted sentences as *Moral sparks* (M-sparks), suggesting that the selected portions of the original post serve as prominent ethical touchpoints, drawing the attention and generating discussion among commenters.

In this research, we narrow down our attention on these highlighted excerpts. To advance our understanding, we propose the following research questions:

$\mathbf{RQ_{linguistic}}$: What differentiate M-sparks from other sentences?

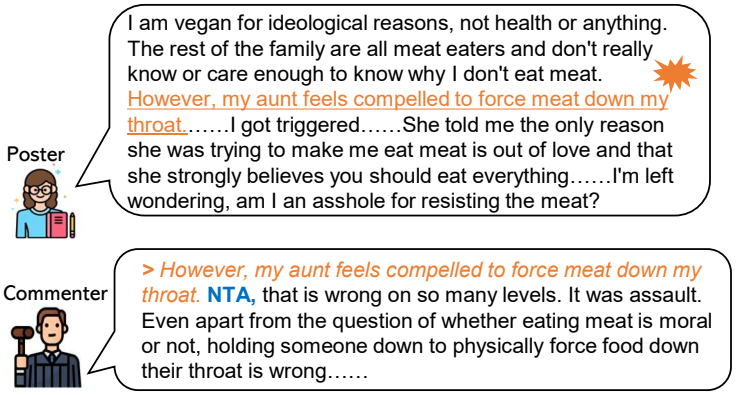
$\mathbf{RQ_{reasoning}}$: What is the underlying reasoning for highlighting M-sparks?

To address the research questions, our investigation delves into the described incidents from two distinct aspects: (1) causal social commonsense and (2) linguistic features. Social commonsense and morality in real-life are intertwined, where social norms and shared knowledge play a guiding role in shaping behavior and aligning with moral principles, influencing societal interactions. The incidents within moral narratives have the power to activate commenters' social commonsense, facilitating a deeper understanding of M-sparks. For example, Fig. 4.1a depicts a post and its comment that identifies an M-spark. The example describes an incident of "forces". By integrating commonsense knowledge, as shown in Fig. 4.1b, one may infer that the poster's aunt may be perceived as *controlling*.

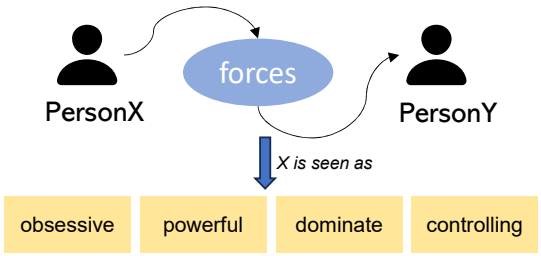
On the other hand, linguistic features play a crucial role in understanding moral narratives due to their nuances and complexity De Candia et al. (2022); Xi and Singh (2023); Giorgi et al. (2023). For instance, considering a incident as "I had to force myself to wake up early every morning to study for my exams." Here, the use of the word "force" describes the personal effort and discipline involved in motivating oneself, without any external coercion or unethi-

Figure 4.1: Example of a post, a comment, and a M-spark highlighted by the commenter. The attributes of PersonX can be inferred from social commonsense knowledge.

(a) An example of a post and it comment. The M-spark quoted by a commenter (orange italic with ">") is highlighted. And the verdict "NTA" indicates the commenter voted for the poster' behaviors were appropriate.



(b) Inferred attributes of PersonX from commonsense knowledge Hwang et al. (2021) when the M-spark indicate a persons' behavior as "forces."



cal behavior. The semantics understanding is essential to discern that the incident refers to *self-control* rather than exerting control over others. Moreover, different language features used to describe similar incidents elicits contrasting M-sparks. For instance, considering "He forcefully criticized my performance during the meeting," and "He constructively criticized my performance during the meeting." The linguistic nuances in these examples also play a pivotal role in shaping the interpretation of individuals interpretation of incidents. Therefore, varied descriptions in moral narratives shape commenters' interpretation of incidents, leading to diverse moral decision-making outcomes, even when similar social commonsense is activated.

We start from differentiating M-sparks from others among over 30k posts and 10M comments in AITA, spanning from June 2013 to November 2021. Next, we convert posts' sentences into textual representations of triples as *(subject, predicate, object)* to filter out sentences without incidents. We then filter the incidents that do not matched with social commonsense knowledge. Then, we employ semantic similarity measures to filter out incidents generated from sentences that lack semantic relatedness to others within the corpus. Finally, we identify linguistic features of the sentences contain the selected incidents to investigate underlying components for M-sparks.

4.2 Methodology

We now introduce our framework as shown in Fig. 4.2, which involves distinguishing M-sparks, selecting instances, and extracting linguistic features. The steps of selecting instances are to ensure we compare the instances exhibit similar social commonsense but different linguistic features. The numbers of instances after each step is shown in Tab. 4.1b. To improve accuracy, we consider quoted instances as M-sparks when they are quoted at least twice by more than one comment. Finally, we label 46,618 M-sparks as one and 182,122 other instances as zero. All the embedding-related implementation employs `based-base-uncased` via HuggingFace.

4.2.1 Distinguishing M-sparks

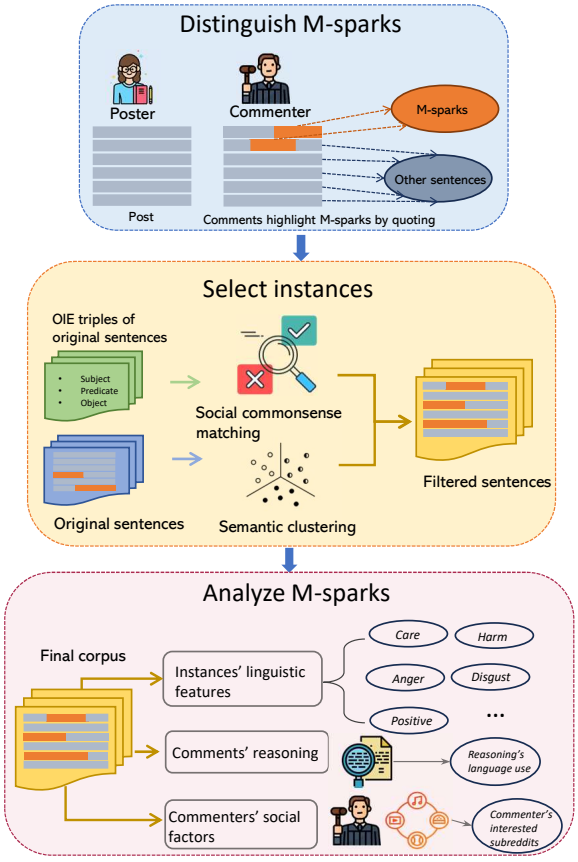
Reddit discussion structure is of a tree rooted at an initial post; comments reply to the root or to other comments.

Definitions We adopt definitions from Guimaraes and Weikum (2021) to describe 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.

Figure 4.2: Framework.



An instance refers to a sentence that is parsed and separated by Stanford Part-Of-Speech Tagger [1] in a post.

We focus on top-level comments because other comments in AITA may not include judgments and reasoning based on the posts.

Tab. 4.1 displays our data curation processes. We use the same dataset described in Section 3.3.1, as shown in the first two rows in Tab. 4.1a. Reddit allows users to ">" symbol to quote lines in posts, so we extract M-sparks by adopting regular expression to match ">" (encoded as `>`) symbols in the comments. Therefore, we have a total of 175,988 comments that quoted at least one sentence in the 24,672 posts. For generating instances, we use Stanford POS Tagger to separate posts. As a result, there are 486,583 instances in the selected 24,672.

[1] `https://nlp.stanford.edu/software/tagger.html`

Table 4.1: Data collection stages.

(a) Collecting posts and comments.

Stage	#posts	#comments
Rule-based selection	351,067	10,296,086
Comment quality filter	51,803	3,675,452

(b) Instance selecting processes.

Stage	#instances
Distinguishing M-sparks	483,583
Deconstructing instances	432,067
Matching commonsense	420,975
Semantic clustering	228,740

4.2.2 Selecting Instances

Answering our research questions requires deconstructing instances into described verbal-driven incidents. To do so, we adopt an Open Information Extraction (OpenIE) system to convert instances into textual representations of triples as *(subject, predicate, object)*, where each triple element is composed of phrases or words. We identify the triples as **incidents**, which indicate verbal events representing entities' behaviors. Then, we match social commonsense knowledge graph with the generated triples to filter incidents do not activate commonsense. As a result, the instances do not contain the selected incidents are removed.

Deconstructing Instances by OpenIE Systems In this study, we adopt COMPACTIE Fatahi Bayat et al. (2022) to generate OpenIE triples. COMPACTIE offers the advantage of producing more compact textual extractions compared to traditional OpenIE systems, which prioritize maximizing the information coverage in extractions over compactness of their constituents. For example, given a sentence *I am a white, born and raised Australian man.*, traditional OpenIE generates a triple *subject: I, predicate: am, object: a white, born and raised Australian man)*, while COMPACTIE generates two triples as: *subject: I, predicate: am, object: a white* and *subject: I, predicate: am, object: born and raised Australian man*. Instances with an empty *predicate* are removed.

Matching Commonsense with Incidents We then employ the causal social commonsense knowledge, ATOMIC Hwang et al. (2021), to match the left instances with incidents. ATOMIC serves as a graph for if-then reasoning, connecting incidents through nine different relation

edges. For our research, we specifically focus on the "xAttr" edge, which describes the perceived attributes of the subject entities involved in an incident. For instance, using the "xAttr" edge, the ATOMIC event "PersonX forces people" provides insights into the perceived attributes of PersonX, such as being perceived as *controlling, aggressive, obsessive, and powerful*. By leveraging this knowledge, we can extract and analyze the perceived values associated with different incidents.

Inspired by previous work Bauer et al. (2023), our selection consists of two phases aimed at balancing speed and precision. In the first phase, we employ a coarse-grained filter to efficiently collect an initial pool of ATOMIC knowledge candidates (E_i) for each data point ($d_i \in D$). We add an ATOMIC candidate incident to E_i if there is word overlap between d_i and the candidate. However, we observed variations in ATOMIC data, such as "PersonX abandons the ___ altogether" and "PersonX abandons ___ altogether." To address this, we use TF-IDF to identify the top n (we set $n = 100$) least similar ATOMIC candidates (considering the ATOMIC event only) to d_i in E_i. These candidates are added to a smaller pool, denoted as P_i.

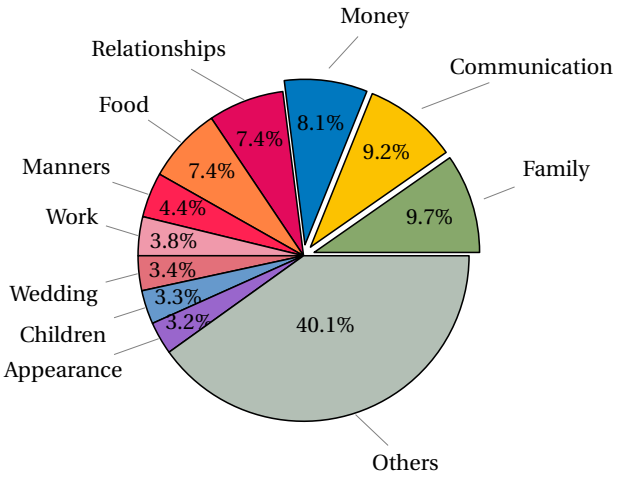
In the second phase, we prioritize precision and assess the semantic match between each ATOMIC candidate in P_i and d_i using BERTScore, an evaluation metric for text similarity Zhang et al. (2019). We take into account both the ATOMIC incident and attributes to ensure that the causal social knowledge of each incident is captured. The candidates in P_i are scored, and we rank them based on their scores. Finally, we select the top k candidates ($k = 3$) from P_i as the chosen ATOMIC candidates for each data point.

Clustering Instances So far we have selected instances that contain incidents matched with social commonsense knowledge. It is essential to select semantically similar instances that exhibit distinct linguistic features for comparison purpose. We further refine the selection of instances by eliminating those that do not demonstrate semantic similarity to any other instances in the corpus. To ensure semantic coherence between instances, we employ semantic clustering to eliminate instances that extracted from instances that lack semantic relevance to others within the corpus.

We first conduct topic modeling using Latent Dirichlet Allocation (LDA) Blei et al. (2003) to cluster instances with moral narratives' domains. We leverage trained bag-of-words representations for AITA topics from previous work Nguyen et al. (2022), which have undergone human evaluation and encompass 47 named topics associated with over 100,000 posts. We follow their method to assign each post the domain that has the highest prior probability. Fig. 4.3 shows domain distributions in our corpus.

To address the lack of external references for identifying instances' cluster labels with each domain, we employ the Hierarchical Density-Based Spatial Clustering of Applications with

Figure 4.3: Topic names and their percentages in our corpus. The three most frequent topics are Family, Communication, and Money.



Noise (HDBSCAN) algorithm McInnes and Healy (2017). This clustering method allows us to discover instances' clusters in an unsupervised manner. To alleviate the issue of sparse embeddings, we perform dimension reduction using the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) technique McInnes et al. (2018). To optimize the performance of the algorithms, we fine-tune the parameters using Bayesian optimization [2]. This process helps us identify the best parameter settings that lead to improved Density-Based Clustering Validation (DBCV) scores Moulavi et al. (2014). Subsequently, we eliminate instances that are clustered as -1 within each domain. The presence of a -1 value indicates that these instances are not relevant to any other instances within the same domain. This step ensures that the clustering analysis focuses exclusively on instances that align with the identified moral situation domains.

4.2.3 Extracting Linguistic Features

We have completed several stages to eliminate irrelevant instances to the greatest extent. Our next step involves extracting linguistic features for the instances in our corpus. We categorize these features into two groups: post (POST) features and entity-centric (EC) features. POST features are calculated based on each post and normalized on the number of words for each instance. EC features are calculated based on the descriptions of entities (i.e., subjects and objects) in each instance and are normalized on the number of the describing words.

[2] `http://hyperopt.github.io/hyperopt/`

EC: Connotation Frames This is a verb-centric formalism for analyzing subjective roles and relationships implied by a given predicate Rashkin et al. (2016). To analyze nuanced dimensions of narratives in AITA, we draw from a lexicon Rashkin et al. (2016) with annotations for 1,000 most frequently used English verbs across various dimensions, ranging from –1 to 1. A verb might elicit a positive sentiment for its subject but imply a negative sentiment for its object. For example, from "*Alice* betrayed *Bob*," the annotation contains the following dimensions:

- Writer's perspective. The writer elicits a negative perspective toward *Alice* as –0.67 (e.g., blaming) and a positive perspective toward *Bob* as 0.26 (e.g., supportive).
- Reader's perspective. (1) Values: the reader presupposes a positive value of *Bob* as 0.87 (strongly positive) and *Alice* as 0.47 (neutral to positive). (2) Effects: the reader presupposes the harms towards *Bob* as –0.93 (strongly negative) compared to *Alice* as 0.067 (neutral). (3) Mental states: the reader presupposes *Bob* is most likely to feel negative (–0.67) as a result of the event, but *Alice* it not likely to be affected (–0.03).

EC: Power and Agency This is a pragmatic formalism organized using frame semantic representations Sap et al. (2017) to model how different levels of power and agency are implicitly projected on people through their actions. We use Sap et al.'s (2017) extension lexicon of Connotation Frames to measure the agency and power scores of *author* and *others*. This extension lexicon contains more than 2,000 transitive and intransitive verbs to model how different levels of power and agency are implicitly projected on the entities through their behaviors. Entities with high agency (subjects of *attack*) are active decision-makers, whereas entities with low agency (subjects of *doubts* and *needs*) are passive. This lexicon contains binary labels of each verb, which are positive (1), equal (0), and negative (–1).

EC: Passive Voice Passive voice causes the subjects to seem more responsible for the incidents Niemi and Young (2016), which has been widely studied in social computing research, such as moral blame assignment Zhou et al. (2021). We measure the number of times a subject or object is described in passive voice for each instance.

POST: Moral Content The Moral Foundation Theory (MFT) Haidt and Graham (2007) has been widely adopted in the computational social community, which is critical in understanding how the psychological influence of social content unfolds, such as quantifying moral behaviors in Twitter Joe et al. (2020) and taxonomizing the structure of moral discussions in Reddit Nguyen et al. (2022). We adopt the extended Moral Foundations Dictionary (eMFD) Hopp et al. (2020), which is a crowdsourced dictionary-based tool for extracting moral content from

textual corpora. The eMFD contains 2,041 unique words, which are categorized into five broad domains based on MFT: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. Each word in the dictionary has a composite valence score ranging from –1 to 1.

POST: VAD (Valence, Arousal, Dominance) The three affective dimensions are used to measure affective meanings from words that convey the author's attitudes toward the events and people referenced. We obtain the valence scores for 20,000 words from the NRC VAD lexicon Mohammad (2018), which contains real-valued scores ranging from 0 to 1 for each category.

POST: Linguistic Inquiry and Word (LIWC) LIWC Pennebaker et al. (2015), a dictionary developed in psycholinguistic field, has been widely used in psychological social computing research Xu et al. (2021); Beel et al. (2022). Prior works suggest that information gained from examining these lexical patterns can be useful in measuring moral situation in real-life Xi and Singh (2023); Giorgi et al. (2023). We use 76 LIWC (i.e., POS categories such as *adj* are excluded) features to extract a count vector of occurrences of words from the LIWC lexicon and then normalized by total number of words in each instance.

POST: Subjectivity Subjectivity arises when people express personal feelings or beliefs, e.g., in opinions or allegations Wilson et al. (2005), which comprises the authors' perspectives towards the descriptive situations, contributing to the audience's judgments. We compute the subjectivity of a post as the average score of words based on the Subjectivity lexicon Wilson et al. (2005) (nonneutral words of "weaksubj" = 0.5 and "strongsubj" = 1). Additionally, we count the numbers of first-person, second-person, and third-person pronouns because words such as "you" and "we" engage the audience with the discourse.

POST: Hedge Hedge is associated with indirection in politeness theory Brennan and Ohaeri (1999), which may affect the audience's judgments.

4.2.4 Regression Model

We use a statistical model for measuring linguistic features. For each feature, we use the following logistic regression model:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_X X + \alpha_1 D_1 + \ldots + \alpha_{|D|} D_{|D|}, \tag{4.1}$$

where X is a variable that takes the value of a characteristic that we are interested in. $D_d(d = 1,\ldots,|D|)$ is a binary variable that takes 1 if the sentence belongs to the d-th domain. Y is a binary response variable that takes 1 if the sentence is a M-spark. β_X is the regression coefficient of the characteristic X, which is the main value of our interest for examining the association between the characteristic and the response; $\exp(\beta_X)$ is the odds ratio (OR) that is interpreted as the change of odds (i.e., the ratio of the probability that a sentence is a M-spark to the probability that a sentence is not a M-spark) when the value of the characteristic increases by one unit. If β_X is positive, we can infer that the characteristic and the response have positive association, and vice versa. To enhance reliability Jafari and Ansari-Pour (2019), we apply a Benjamini-Hochberg False Discovery Rate (FDR) correction Benjamini and Hochberg (1995). Since we have no a priori hypotheses, we simply examine whether or not β_X has a p-values lower than 0.001.

4.3 Preliminary Results

We report top three of the frequently occurring incidents within each domain in Tab. 4.2.

Table 4.2: Frequently discussed social commonsense related incidents within each domain.

Domain	Social commonsense incidents
Family	PersonX will always love PersonY PersonX has passed away PersonX does not want to cook
Communication	PersonX will always love PersonY PersonX is afraid it would hurt PersonX shoulds always
Money	PersonX would pay for it PersonX needs to pay rent PersonX shoulds always

Fig. 4.4 displays the attributes of the described entity that are most frequently indicated by selected incidents within each moral narrative domain.

(a) Family (b) Communication (c) Money

Figure 4.4: Top 500 attributes inferred from social commonsense within each domain.

4.3.1 $RQ_{linguistic:}$ What features differentiate M-sparks from other sentences?

We use the regression model to analyze the 234,021 selected instances after selecting instance as introduced in Section 4.2.2, where 86,618 instances are M-sparks. Fig. 4.5 and Fig. 4.6 report the OR values of narrative and entity linguistic features that has p-values lower than 0.001.

4.3.2 $RQ_{reasoning:}$ What is the underlying reasoning for highlighting M-sparks?

We propose the following factors to understand the reasoning.

Comments' textural information: We plan investigate the relationship between social commonsense and the factors that capture commenters' attention regarding why incidents arise.

Commenters' social factors: To explore the relationship between commenters' social interests and their involvement in social commonsense incidents, we plan aggregate their social factors based on their participation in various subreddits.

Figure 4.5: The odds ratio values of narrative linguistic features. We use ■ to indicate odds ratio greater than one and effects greater than zero (on the right), and blue rectangles ■ indicate the opposite (on the left). The shade shows the FDR-adjusted p-values: ■ and ■ (darkest): ≤ 0.005, ■ and ■ (middle): ≤ 0.002, ■ and ■: ≤ 0.001. Features are labeled in the figure using the format: "lexicon: category (three example words)."

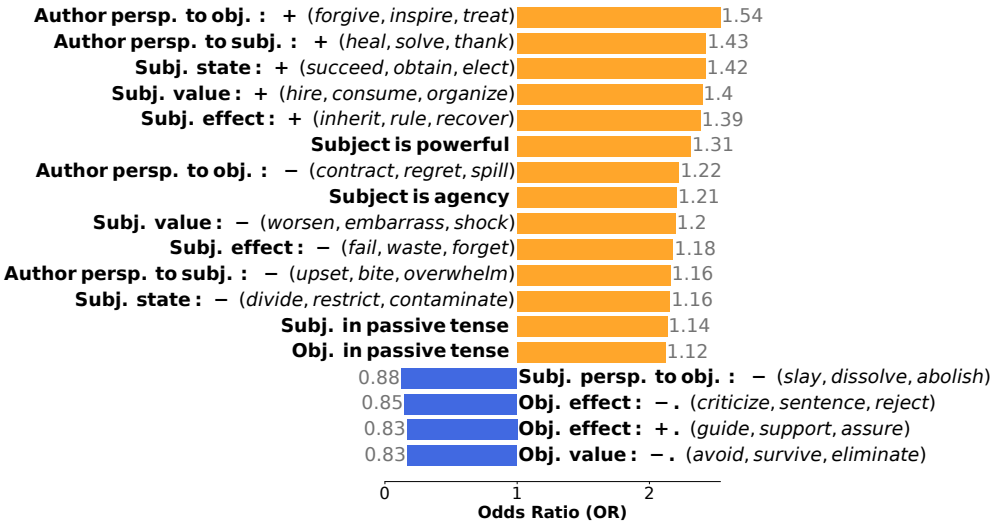


Figure 4.6: The odds ratio values of entity linguistic features. All of the features have FDR-adjusted p-values lower than 0.01 indicated by the colors ■ and ■. Features are labeled in the figure using the format: "Category: sentiment (three example words)" (+: positive, −: negative, *subj.*: subject, *obj.*: object, *persp.*: perspective).

REFERENCES

AmITheAsshole. https://www.reddit.com/r/AmItheAsshole/, note = Online; accessed 30 November 2022 „ 2023.

Dimo Angelov. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

Maya Asher, Anu Asnaani, and Idan M Aderka. Gender differences in social anxiety disorder: A review. *Clinical psychology review*, 56:1–12, 2017.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.263.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. *arXiv preprint arXiv:1704.04675*, 2017. doi: arXiv:1704.04675.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1284.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3727–3742, 2023.

Jacob Beel, Tong Xiang, Sandeep Soni, and Diyi Yang. 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, pages 32–42, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

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

Nicholas Botzer, Shawn Gu, and Tim Weninger. Analysis of moral judgment on Reddit. *IEEE Transactions on Computational Social Systems (TCSS)*, 2022.

Juergen Bracht and Adam Zylbersztejn. Moral judgments, gender, and antisocial preferences: An experimental study. *Theory and Decision*, 85(3):389–406, 2018.

Susan E. Brennan and Justina O. Ohaeri. Why do electronic conversations seem less polite? The costs and benefits of hedging. In *Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration*, pages 227–235, New York, 1999. Association for Computing Machinery. doi: 10.1145/295665.295942.

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

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1082.

George Chrysostomou and Nikolaos Aletras. Flexible instance-specific rationalization of NLP models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10545–10553, Online, February 2022. AAAI Press.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, 1988.

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

Sara De Candia, Gianmarco De Francisci Morales, Corrado Monti, and Francesco Bonchi. Social norms on Reddit: A demographic analysis. In *14th ACM Web Science Conference 2022*, WebSci, pages 139–147, NY, 2022. Association for Computing Machinery. doi: 10.1145/3501247.3531549.

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 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*, pages 353–369. Association for Computing Machinery, 2017. doi: 10.1145/2998181.2998220.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, Minneapolis, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models, July 2020.

Katinka Dijkstra, Rolf A Zwaan, Arthur C Graesser, and Joseph P Magliano. Character and reader emotions in literary texts. *Poetics*, 23(1-2):139–157, 1995.

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

of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 698–718, Online and Punta Cana, November 2021. doi: 10.18653/v1/2021.emnlp-main.54.

Farima Fatahi Bayat, Nikita Bhutani, and H. Jagadish. CompactIE: Compact facts in Open Information Extraction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pages 900–910, Seattle, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.65.

Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. Discovering and categorising language biases in Reddit. *arXiv preprint arXiv:2008.02754*, 2020.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48.

Kathleen C Fraser, Svetlana Kiritchenko, and Esma Balkir. Does moral code have a moral code? Probing Delphi's moral philosophy. *arXiv preprint arXiv:2205.12771*, 2022.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic Natural Language Processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2501.

Vaibhav Garg, Jiaqing Yuan, Rujie Xi, and Munindar P. Singh. Extracting incidents, effects, and requested advice from MeToo posts. *arXiv preprint arXiv:2303.10573*, 2023.

Salvatore Giorgi, Ke Zhao, Alexander H Feng, and Lara J Martin. 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, pages 233–244, 2023.

Kurt Gray and Daniel M. Wegner. To escape blame, don't be a hero-be a victim. *Journal of Experimental Social Psychology*, 47(2):516–519, 2011. doi: 10.1016/j.jesp.2010.12.012.

Steve Guglielmo and Bertram F Malle. Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS one*, 14(3):e0213544, 2019.

Anna Guimaraes and Gerhard Weikum. 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)*, pages 163–172, 2021.

Zhen Guo, Zhe Zhang, and Munindar P. Singh. In opinion holders' shoes: Modeling cumulative influence for view change in online argumentation. In *Proceedings of the 29th Web Conference (WWW)*, pages 2388–2399, Taipei, April 2020. ACM. doi: 10.1145/3366423.3380302.

Jonathan Haidt and Jesse Graham. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. In *Social Justice Research*, volume 20, pages 98–116. Springer, March 2007. doi: 10.1007/s11211-007-0034-z.

Sepp Hochreiter and JÃ¼rgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

Frederic R. Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 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, 2020. doi: 10.3758/s13428-020-01433-0.

Huggingface. `https://huggingface.co/`, note = Online; accessed 30 November 2022 „ 2023.

Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225, Ann Arbor, 2014.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6384–6392, 2021.

Ken Hyland. *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Publishing, London, UK, 2018.

Mohieddin Jafari and Naser Ansari-Pour. Why, when and how to adjust your p values? *Cell Journal (Yakhteh)*, 20(4):604, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4459–4473, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.409.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021.

Shan Jiang and Christo Wilson. 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)*, pages 617–631, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.51.

Hoover Joe, Portillo Wightman Gwenyth, Yeh Leigh, Havaldar Shreya, Mostafazadeh Davani Aida, Lin Ying, Kennedy Brendan, Atari Mohammad, Kamel Zahra, Mendlen Madelyn, Moreno Gabriela, Park Christina, E. Chang Tingyee, Chin Jenna, Leong Christian, Leung Jun Yen, Mirinjian Arineh, and Dehghani Morteza. Moral foundations Twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020. doi: 10.1177/1948550619876629.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Upper Saddle River, 2009.

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

Svetlana Kiritchenko and Saif Mohammad. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–52, San Diego, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0410.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011.

Zeyu Li, Yilong Qin, Zihan Liu, and Wei Wang. Powering comparative classification with sentiment analysis via domain adaptive knowledge transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6818–6830, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.546.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv preprint arXiv: 2008.09094*, 2020.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *American Educational Research Association*, 6(3):2332858420940312, 2020. doi: 10.1177/2332858420940312.

Bertram Malle, Steve Guglielmo, and Andrew Monroe. A theory of blame. *Psychological Inquiry*, 25:147–186, 04 2014. doi: 10.1080/1047840X.2014.877340.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, February 2018.

Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42, 2017. doi: 10.1109/ICDMW.2017.12.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 174–184, Melbourne, July 2018. doi: 10.18653/v1/P18-1017.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J.G.B.Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, pages 839–847, 2014. doi: 10.1137/1.9781611973440.96.

Multilingual-USAS. `https://github.com/UCREL/Multilingual-USAS`, 2022. Online; accessed 30 November 2022.

Tuan Dung Nguyen, Georgiana Lyall, Alasdair Tran, Minjeong Shin, Nicholas George Carroll, Colin Klein, and Lexing Xie. Mapping topics in 100,000 real-life moral dilemmas. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 16, pages 699–710, 2022.

Laura Niemi and Liane Young. When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and social psychology bulletin*, 42(9): 1227–1242, 2016.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin, 2015.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio, and Paul Rayson. Development of the multilingual semantic annotation system. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1268–1274, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1137.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Peipei Yang, Amir Hussain, and Tariq Durrani. Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis. In *IEEE 11th International Conference on Signal Processing*, volume 2, pages 1251–1255, 2012. doi: 10.1109/ICoSP.2012.6491803.

PushShiftAPI. `https://github.com/pushshift/api`, note = Online; accessed 30 November 2022 „ 2023.

Hannah Rashkin, Sameer Singh, and Yejin Choi. Connotation frames: A data-driven investigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 311–321, Berlin, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1030.

Reddit API. `https://www.reddit.com/dev/api`, note = Online; accessed 30 November 2022 „ 2023.

Reddit score. `https://www.reddit.com/wiki/faq/#wiki_how_is_a_submission.27s_score_determined.3F`, 2023. Online; accessed 30 November 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992, Hong Kong, 11 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.

Tania Reynolds, Chuck Howard, Hallgeir Sjåstad, Luke Zhu, Tyler G Okimoto, Roy F Baumeister, Karl Aquino, and JongHan Kim. Man up and take it: Gender bias in moral typecasting. *Organizational Behavior and Human Decision Processes*, 161:120–141, 2020. doi: 10.1016/j.obhdp.2020.05.002.

Henry S. Richardson. Moral Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition, 2018.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2329–2334, Copenhagen, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1247.

Chelsea Schein and Kurt Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018. doi: 10.1177/1088868317698288.

John Finley Scott. *Internalization of norms: A sociological theory of moral commitment*. Prentice-Hall, 1971.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282.

Spcay. `https://spacy.io/`, note = Online; accessed 30 November 2022 „ 2023.

Ross M. Stolzenberg. The measurement and decomposition of causal effects in nonlinear and nonadditive models. *Sociological Methodology*, 11:459–488, 1980.

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

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 3319–3328. JMLR.org, 2017.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. A word on machine ethics: A response to Jiang et al. (2021). *arXiv preprint arXiv:2111.04158*, 2021.

USAS. `http://ucrel.lancs.ac.uk/usas/`, 2002. Online; accessed 30 November 2021.

Lawrence J. Walker. A longitudinal study of moral reasoning. *Child Development*, 60(1):157–166, 1989.

Gillian R Wark and Dennis L Krebs. Gender and dilemma differences in real-life moral judgment. *Developmental Psychology*, 32(2):220, 1996.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354, Vancouver, October 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220619.

John A Wood, Justin G Longenecker, Joseph A McKinney, and Carlos W Moore. Ethical attitudes of students and business professionals: A study of moral reasoning. *Journal of Business ethics*, 7(4):249–257, 1988.

Ruijie Xi and Munindar P. Singh. The blame game: Understanding blame assignment in social media. *IEEE Transactions on Computational Social Systems (TCSS) (TCSS)*, 10:1–10, 2023. doi: 10.1109/TCSS.2023.3261242.

Anbang Xu, Haibin Liu, Liang Gou, Rama Akkiraju, Jalal Mahmud, Vibha Sinha, Yuheng Hu, and Mu Qiao. Predicting perceived brand personality with social media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 10, pages 436–445, 2021. doi: 10.1609/icwsm.v10i1.14733.

Jiaqing Yuan and Munindar P. Singh. Conversation modeling to predict derailment. *arXiv preprint arXiv:2303.111840*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019.

Karen Zhou, Ana Smith, and Lillian Lee. Assessing cognitive linguistic influences in the assignment of blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 61–69, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.socialnlp-1.5.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3755–3773, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.261.

APPENDICES