A Survey on Modelling Morality for Text Analysis

Ines Reinig∗♠, Maria Becker,∗♣ Ines Rehbein∗♠, Simone Paolo Ponzetto♣

♠Data and Web Science Group
University of Mannheim, Germany
{reinig,rehbein,ponzetto}@uni-mannheim.de

♣Germanistisches Seminar
University of Heidelberg, Germany
maria.becker@gs.uni-heidelberg.de

Abstract

In this survey, we provide a systematic review of recent work on modelling morality in text, an area of research that has garnered increasing attention in recent years. Our survey is motivated by the importance of modelling decisions on the created resources, the models trained on these resources and the analyses that result from the models’ predictions. We review work at the interface of NLP, Computational Social Science and Psychology and give an overview of the different goals and research questions addressed in the papers, their underlying theoretical backgrounds and the methods that have been applied to pursue these goals. We then identify and discuss challenges and research gaps, such as the lack of a theoretical framework underlying the operationalisation of morality in text, the low IAA reported for many human-annotated resulting resources and the lack of validation of newly proposed resources and analyses.

1 Introduction

With the rise of large language models, research goals in NLP have also become more ambitious, tackling new challenges like the prediction of complex psychological constructs. A case in point is the modelling of morality in text, a task that requires a deep and comprehensive understanding of natural language. More and more studies at the interface of NLP and Computational Social Science (CSS) have addressed this task, based on different theoretical frameworks, and a variety of methods have been applied to predict moral values from text. These studies aim at modelling the moral sentiment that a person or group holds toward a certain target in order to investigate research questions from the political or social sciences. Others model morality in the context of AI applications, e.g., to study the inherent moral biases learned by language models. Many resources have been created, but are often difficult to find due to their heterogeneous, interdisciplinary research backgrounds.

Modelling morality using NLP techniques has also been at the center of recent events such as workshops or coding competitions. For example, Kiesel et al. (2023) identify human values behind arguments in a SemEval shared task (ValueEval’23) while the MP2 workshop at NeuRIPs 2023 looks at the application of theories from moral philosophy and psychology to AI practices, demonstrating the growing interest in exploring morality with computational methods in interdisciplinary settings.

This survey aims at providing a systematic overview of recent work on modelling morality in text, the resources that are available, and the methods that have been applied. We argue that the operationalisation of morality has an immense impact on (i) the annotated resources that are created, (ii) the models that are trained, based on these resources, (iii) the predictions obtained from these models, and (iv) the analyses that result from the predictions. Our survey therefore focusses on the theoretical basis and modelling part, as well as on the validation of the concept.

This is not a survey on ethics in AI, which was the main focus of Vida et al. (2023). Motivated by the confusion regarding concepts from philosophical ethics in NLP research, Vida et al. (2023) focus on concepts from philosophy and analyse literature on moral NLP with respect to their philosophical foundation and terminology. We, instead, focus on the theoretical modelling and operationalisation of morality for text analysis and the challenges that arise for applications in CSS and Cultural Analytics. These different backgrounds are also reflected in the paper selection method and the set of surveyed papers, which we briefly discuss in the
Identification

ACM (n=192)
ACL (n=77)
JCSS (n=18)
IEEE (n=71)
BRM (n=16)
Sage (n=5)
EPJ (n=3)
KBS (n=3)
PLOS (n=3)
CCR (n=1)

Screening

Papers identified through database search (n=435)
Papers screened for eligibility, by title and abstract (and if necessary by full-text) (n=425)
Papers identified through backward snowballing (n=19)
Papers excluded by incl/excl crit. (n=320)
Papers included in the review (n=135)

Inclusion

Figure 1: PRISMA-inspired flow diagram describing our paper sampling method.

Appendix, §A.2.

The survey is structured as follows. We first describe our survey methodology (§2) and outline the research objectives and background of the papers included in the survey (§3). Then we describe different operationalisations of morality in text (§4) and their impact on applications in the social sciences (§5). We discuss trends and research gaps in Section 6 before we conclude and outline some recommendations (§7). The supplementary materials together with our code can be found in our GitHub repository: https://github.com/umanlp/survey_morality.

2 Survey methodology

In the following, we briefly describe our methodology for selecting and reviewing the papers for this survey. Our paper selection method is inspired by Alturayef et al. (2023) and also follows recommendations from Moher et al. (2009). The different steps are summarized in Figure 1, more details can be found in the appendix.

Paper sampling (see §A.1.1 and §A.1.2) We first semi-automatically identify potentially relevant papers. We search for specific keywords in 15 selected journals and venues (blue boxes in Figure 1). After this paper sampling step, we obtain 435 papers, which we then deduplicate, amounting to 406 papers. We identify 19 more papers that might be relevant for the survey using backward snowballing.

Screening and reviewing (see §A.1.4 to §A.1.6) The set of papers resulting from paper sampling and snowballing is then screened following six selection criteria (cf. §A.1.4) that must all be satisfied for a paper to be included in the survey. After screening, 135 papers are kept and distributed amongst the authors for reviewing. To increase consistency across reviewers, we use a survey form with an accompanying codebook for reviewing, both refined during an initial pilot study based on eight selected papers. After the reviewing process, we identify 4 demo papers, 1 shared task paper and 14 papers that, according to our selection criteria, were not included in the survey. This leaves us with 116 papers as the basis of the analyses presented in the remainder of this work. Figure 2 shows the distribution of publication years for the papers, demonstrating the growing interest in the field (also see Table 7 in the appendix).

3 Survey overview

To get an overview of the work included in the survey, we first classify papers according to their main research objectives, as listed below (note that a paper can have more than one objective).

1. Values, Stance, Framing: investigate the moral values of a person, group, or culture;
explore the moral sentiment towards a target; identify moral rhetoric and framing.

2. **Morality in AI**: investigate morality in the context of AI systems or applications.

3. **Comparison**: compare moral values to other concepts (e.g., stance, emotions).

4. **Moral Theories**: evaluate or improve a moral theory.

5. **Other Theories**: evaluate or improve another theory (not related to morality).

The majority of the papers model morality to analyse the moral values and stances held by an individual or group, and to investigate moral framing (87 papers, see Table 1). The next frequent class focuses on morality in AI (30 papers), often with the goal of improving an LLMs understanding of moral values or to investigate moral bias encoded in LLMs. 21 papers compare moral values to other concepts, such as stances and emotions, and 12 papers aim at evaluating or improving a theoretical framework.

According to their main contributions, the papers can be broadly categorised into experimental papers, analysis papers and resource papers.3

**Experimental papers** This category includes work that focusses on the development and evaluation of methods for the identification of moral language in text. With 81 papers, it is the largest of the three categories. Regarding the machine learning methods used, we observe the following trends (see Table 2): Not surprisingly, most works use fine-tuned transformers (29 papers), followed by dictionary- or rule-based approaches (25 papers). Feature-based ML has mostly been applied in older papers, while more recent work also uses zero- and few-shot learning based on LLMs.

**Analysis papers** The second largest category includes 65 papers that present an analysis based on the application of NLP methods to predict moral values in text. Papers in this category address research questions from the CSS, for example, Zhang and Counts (2016) investigate moral values in the context of anti versus pro-abortion policies, Islam and Goldwasser (2022b) study moral messages used in COVID-19 vaccine campaigns while Ertrugrul et al. (2019) predict social protest activities from social media discussions. Most analysis papers are from the political and social sciences (46), followed by media and communication studies (11), psychology (5) and other fields (3).

**Resource papers** The last and smallest category includes 56 papers that release annotated datasets (e.g., Trager et al. (2022); Mooijman et al. (2018)), dictionaries (e.g., Zúquete et al. (2023); Araque et al. (2022)) or ontologies (e.g., De Giorgis et al. (2022); Hulpus et al. (2020)). Most of the 56 papers present a new, annotated dataset (38 papers), 11 papers create or expand a moral dictionary, while 9 others create, link or augment ontologies or knowledge graphs with moral vocabulary.4 The majority of the papers employ trained annotators (20 papers) while 14 papers use crowdsourcing. To control for annotation bias, 8 of the 14 crowdsourcing papers collect information regarding the coders’ demographics or their political or moral views.

In summary, the majority of the papers in our survey use text-based models of morality to study moral sentiment and moral framing, typically applied to research questions from the CSS. Given the strong focus on real-world applications, the question arises as to the validity of the various approaches used to model morality in text. We address this issue in the next section.

### 4 Operationalisations of morality

Next, we study how the different papers in the survey operationalise the concept of morality. We first give an overview over the task of moral value

---

3Note that the contributions are not mutually exclusive.

4These contributions are, again, not mutually exclusive.
prediction. Then we look at the theoretical frameworks that have been employed, and turn to the question on which level morality is investigated in the papers. Finally, we discuss the issue of underspecification in modelling morality and its negative impact on the reliability and validity of the analyses.

4.1 Task description
Most papers in the survey model morality by predicting the moral attitudes, beliefs, sentiment or emotions expressed in a text. Let us consider Example 4.1 below, taken from the Twitter corpus of Roy et al. (2022).

Example 4.1. “Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.”

The task in those papers is then to assign one or more labels that describe the moral values expressed in the text, where the labels depend on the theoretical framework used to model morality (see §4.2). The label could simply predict whether the text includes moral language or not (morality: yes/no), or it could include more fine-grained information describing the moral values or beliefs expressed in the text, for example, the moral foundations described in MFT (see §4.2 below) or the human values defined in Schwartz and Bilsky (1987).

Some works additionally encode the strength of the moral message by means of a continuous score (Araque et al., 2020), others additionally try to identify moral roles, such as the entity causing harm or the target of the moral message (Roy et al., 2022), as illustrated in the example below.

“Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.”

Example 4.2.

4.2 Overview of theoretical frameworks
Figure 3 provides an overview of the theoretical frameworks used in the papers. The vast majority (84 papers) uses Moral Foundations Theory (MFT) (Graham et al., 2013), which we describe below. Remarkably, 25 papers do not rely on a specific theory or framework. Out of those, 8 papers aim at modelling social norms by extracting unspoken commonsense rules, often referred to as Rules-of-Thumb (RoT) (Forbes et al., 2020).

Other theories and concepts that have been used to model morality include Schwartz’ Theory of Human Values (Schwartz and Bilsky, 1987), Hofstede’s Cultural Dimensions (Hofstede, 2001), the Theory of Contractualist Moral Decision-Making (Levine et al., 2018; Awad et al., 2022), Moral Disengagement (Bandura, 1999, 2016), the Path Model of Blame (Malle et al., 2014) and the Economics of Convention Framework (Boltanski and Thévenot, 2006).

Since MFT is the predominant theoretical framework for modelling morality in texts, below we provide a brief introduction to the main concepts of this theory.

Moral Foundations Theory (MFT) is a descriptive, pluralist theory of morality that has been highly influential within the field of moral psychology (Haidt et al., 2009; Graham et al., 2013).
While monists believe that one single dimension is sufficient to understand and explain morality, pluralists argue that the concept of morality is based on more than one such dimension, or foundation.

In MFT, moral foundations (MFs) are conceptualised as intuitions, i.e., as an “evaluative feeling (like–dislike, good–bad) about the character or actions of a person, without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion” (Haidt and Bjorklund, 2008). In short, moral foundations are intuitions or “gut feelings” that often drive moral reasoning and turn it into rationalisation. MFT assumes that the foundations have been developed during evolution as responses to several adaptive challenges. The foundations that have been proposed so far can be classified into binding foundations (ingroup LOYALTY, respect for AUTHORITY, and PURITY) and individualising foundations (CARE and FAIRNESS).

MFT does not claim to know how many of these foundations exist, instead it proposes a set of criteria for foundationhood (see Table A.4) and encourages researchers to revise and extend the set of moral foundations. Another basic assumption of MFT states the existence of an innate draft of the moral mind that is later revised by experience and cultural influences (Graham et al., 2013, p. 9), thus making it an interesting basis for cross-cultural investigations of morality.

MFT comes in different flavours We observe that the MFT-based papers differ with respect to the number of foundations considered for analysis. As shown in Figure 4, many papers either apply a set of five moral foundations (MF5: Care, Fairness, Loyalty, Authority, Purity) or use the five foundations, but with separate classes for the vice–virtue scale (MF10: Care vs. Harm, Fairness vs. Cheating etc.). This is probably due to the fact that many papers rely on existing resources that utilise the MF5 or MF10 schema (e.g. the English Moral Foundations Dictionary (MFD)).

Only three papers include new, self-defined moral foundations. One of them is Cheng and Zhang (2023) which we describe in more detail below, as the same validation method has also been applied for the creation of the English and Japanese Moral Foundations dictionaries. The authors adapt the English MFD v2.0 to Chinese and propose six new candidates for moral foundations (Liberty, Altruism, Diligence, Waste, Resilience, and Modesty). They start with a translation of the MFD2.0 to Chinese and add over 1,200 Chinese words related to morality. Then they use several rounds of expert coding to assign the dictionary entries to MFs and ask crowdworkers to write short essays about moral issues related to each foundation. This resulted in a benchmark with over 2,200 texts labelled for the different MFs that are then used to validate the dictionary, based on the average word-frequencies for each class and a word density analysis. The authors emphasise the importance of testing MFT in different cultural settings, in order to advance and refine the theory.

While this approach is very much in line with the theoretical assumptions of MFT, such as the five criteria for foundationhood mentioned above (also see Table A.4), other work seems less aware of the theoretical constraints imposed by MFT, adding arbitrary new candidates to the list without providing proper validation. A case in point are candidates like Feminism–Maleness; Sustainability–Climate change; Peace–War (González-Santos et al., 2023) that clearly ignore the criteria of foundationhood.

### 4.3 Level of analysis

One perspective from which we can investigate how papers operationalise the analysis of morality in text is the level of analysis, i.e., whether morality has been analysed on the document, segment, sentence or token level, or whether the analysis is based on frames. By segment we refer to any text span that is a substring of the document, longer than a token, and not a sentence (this includes both text spans that are shorter and longer than sentences). As opposed to segment, the document level refers to an entire text; this includes tweets and other social media posts or messages, since they contain the whole text and not just a substring of it. We define frames as units that additionally encode the relations between text spans describing moral situations, actions, goals or values and their roles, such as the target or holder of a moral sentiment.

Table 3 shows that the majority of the papers analyse morality on the document level. The high
number can be traced back to the fact that most papers base their analysis on tweets and Reddit posts. Only a small number of papers (9) provide a frame-based analysis of morality that explicitly encodes the holder and target of the moral sentiment.

We argue that the practise of annotating morality at the document or sentence level instead of explicitly encoding the participants and their roles leads to underspecification in the coding scheme which not only results in low inter-annotator agreement (IAA) but also fails to capture perspective, i.e., who is the holder of the moral sentiment. We discuss these issues below.

### 4.4 The problem of low IAA

A number of the survey papers have created manually annotated resources, however, less than half of the papers make their annotation guidelines publicly available (22 out of 54 resource papers). Looking at the available guidelines, we find that about half of them have a length of less than or up to one page only. These short guidelines often present the coders with a highly underspecified task description, instructing them to assign labels to sentences or documents without an operationalisable definition of the concept of morality.

Our survey includes a number of papers dedicated to the annotation of morality. Many of them report rather low scores for inter-annotator agreement (see Table 4). Note that the scores are not comparable, given that the projects use different label sets and IAA metrics.

<table>
<thead>
<tr>
<th>IAA</th>
<th>Schema</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.61</td>
<td>MF12+Liberty</td>
<td>Pacheco et al. (2022)</td>
</tr>
<tr>
<td>0.91</td>
<td>MF+own</td>
<td>Cheng and Zhang (2023)</td>
</tr>
</tbody>
</table>

Table 4: IAA for the annotation of morality reported in the papers. Note that the scores are not comparable, as the annotations use different label sets and IAA metrics.

We now look at studies that report high IAA in order to determine the underlying reasons. The highest scores are reported in Cheng and Zhang (2023) who asked trained coders to classify moral words from the Chinese Moral Dictionary into MFs. Weinzierl and Harabagiu (2022) report high IAA for COVID-19 Vaccine Hesitancy Frames (VHFs). They identify and aggregate VHFs in tweets and annotate the aggregated and cleaned frame representations. Islam and Goldwasser (2022a) also analyse COVID-19 Vaccine Hesitancy, focussing on Facebook ads. They report a high IAA (0.74 Cohen’s $\kappa$) for the annotation of MFs, based on a small set of 110 instances. Johnson and Goldwasser (2018) collect tweets by US congress members for a number of controversial, morally charged topics (Abortion, ACA, Guns, Immigration, LGBTQ, Terrorism) and report a $\kappa$ of 0.79 for the annotation of MFs for those topics. Shahid et al. (2020) annotate moral foundations in news articles on the sentence level. They report negative IAA for the binary distinction of whether a sentence includes moral language or not. However, for sentences where the annotators agreed on the presence of moral language, IAA for the MF type was high (0.85 Krippendorff’s $\alpha$).

In sum, all studies that obtained high agreement coded moral foundations in texts that were already filtered for morally charged language. Thus, the difficult part of the annotation is not to classify moral language into MFs but to decide whether or not a text includes moral language. In other words, it seems as if the low IAA reported in many papers can be traced back to the question of what counts as moral language. This finding suggests that full-text annotation, i.e., assigning labels to each sentence in a news article or to each document in a collection of social media messages, might not be a good approach. This is further illustrated by the following example from the Moral Foundations Reddit Corpus (Trager et al., 2022):

**Example 4.3.** “Dude I think Bernie is racist. No shit I’m gonna think Le Pen supporters are racist.” (MFRC, Subreddit: neoliberal)
Coder1: EQUALITY
Coder2: CARE, PURITY, EQUALITY
Coder3: LOYALTY, AUTHORITY

Here, three trained coders assigned 5 out of 6 possible MF labels to this post, with only one label chosen by more than one coder. This inconsistency is not necessarily evidence that the coders have different moral values but could simply show that, when left without specific instructions, the coders focus on different aspects in the text and, instead of annotating the text author’s moral values, assign labels based on free word associations like racist \(\rightarrow\) EQUALITY, supporters \(\rightarrow\) LOYALTY.

4.5 The problem of perspective

Example 4.5 shows that the annotators can be tempted to code different perspectives in the same text, as illustrated below. In this case, coder 1 has chosen the AUTHORITY label, probably referring to Wilder’s moral foundations, while the other annotators use EQUALITY to encode the moral values of the text author.

**Example 4.4.** “Don’t know about Le Pen but I know Wilders was worryingly on the fascist side of things.” (MFRC, Subreddit: europe)
Coder1: AUTHORITY
Coder2: CARE, PURITY, EQUALITY
Coder3: EQUALITY

This underspecification and mix of perspectives casts considerable doubt on the construct validity of the annotations and their reliability for text analysis. As a remedy to this problem, we recommend the use of a frame-based approach, as in Roy et al. (2022). We argue that encoding moral roles like the holder and target of a moral sentiment is crucial for the analysis of morality in text. This is illustrated in Roy et al. (2022) who show that in discussions about abortion in the US both conservatives and liberals use the CARE-HARM moral foundation, however, with different targets. While conservatives focus on the protection of unborn life, liberals prioritise the well-being of the women. This example shows that the MF label on its own is not sufficient to capture the differences between conservatives and liberals. We therefore find the lack of studies that code morality on the level of frames surprising and hope that this research gap will be addressed in the near future.

5 Applications of moral value prediction

We now focus on the application of computational methods for moral value prediction in real-world analyses. Our findings show that the thoroughness of the research methodology leaves something to be desired. Only few of the 65 analysis papers in the survey explicitly state their research questions (13 papers), and an equally small number of papers (10) formulates hypotheses and tests for statistical significance, while the vast majority of the analysis papers use data exploration or visualisations to support their findings. As dictionaries were among the most often applied method for predicting moral values in the analysis papers (also see Table 2), we next look at the available resources and their validity in real-world applications.

5.1 Validity of dictionary-based approaches

The disadvantages of dictionary-based text analysis are well known and have been discussed at length, e.g., in the context of sentiment analysis. Those drawbacks also apply to the analysis of moral values, most importantly the insensitivity of dictionaries to word meaning in context and their failure to handle compositionality, such as negation (Wiegand et al., 2010). Another crucial issue of dictionary-based methods is their failure to capture perspective, i.e., to identify the holder of the moral sentiment expressed in the text.

The most often used resources for dictionary-based moral value prediction are the English Moral Foundations Dictionary (MFD) (Graham et al., 2009) and extensions thereof. The MFD was originally developed for comparing Christian sermons from liberal and conservative churches in the US, as sermons often include moral narratives. The resulting scores were mostly in line with the predictions made by the theory, which Graham et al. (2009) take as a validation of the approach. However, the authors report that they also tried to use the MFD on a corpus of Republican and Democratic candidates’ convention speeches but were not able to extract distinctive moral values from the data, thus calling into question the general applicability of the approach to texts from other domains where moral narratives might not be as omnipresent as in the sermons.

Hopp et al. (2021) question the representativeness of the original, expert-created MFD and propose using crowdworkers to identify morally relevant passages in text, and then use the crowd-
sourced annotations to extract terms for their expanded version of the dictionary, the eMFD. They argue that this procedure is more apt to treat MFs as “the products of fast, spontaneous intuitions”, thus being superior to an expert-curated word list.

To validate their approach, Hopp et al. (2021) apply the eMFD to news articles from far-left, center-left and far-right news outlets. As the eMDF yielded more distinctive scores for the different partisan news than the MFD and MFD2.0, the authors conclude that their method is superior to the other two dictionaries. This, however, is no conclusive proof that the above dictionaries are reliable and valid approximations of measurement tools like the MFT Questionnaire (Graham et al., 2011) which has been successfully tested for internal and external validity and test-retest reliability (see §A.5), using confirmatory factor analysis.

Another question that arose from the survey concerns the assumptions and prerequisites that must be fulfilled to guarantee the reliability and validity of the results. In the following, we discuss one of the most important prerequisite for empirical analyses, namely the representativeness of the data for the population studied.

### 5.2 Representativity of the data

The question of when a corpus is representative enough to answer research questions about a certain population has been discussed at length in the field of corpus linguistics (see, e.g., Biber (1993); Egbert et al. (2022)) and these findings and best practices should be taken into account.

**Moral values across cultures** Representativity is particularly important for research questions that investigate moral values across cultures, such as Wu et al. (2023), who compare folk tales from 27 cultural backgrounds, based on a crowdsourced dictionary. In their study, they use an opportunistic collection of folk tales translated into English. The European cultures are represented at fine-grained levels in the corpus, with small countries like Denmark, Norway and Sweden regarded as separate cultures. The African continent, on the other hand, is categorised as one culture only, represented by folk tales from West Africa. This design seems rather imbalanced and Eurocentric and casts doubts on the validity of the analysis.

We therefore argue that while it is necessary to formulate hypotheses and carry out significance tests (as done by only 10 out of 65 analysis papers), it is crucial to also ensure the representativeness of the data for the respective research question. We therefore release a checklist that addresses some of the design decisions relevant for ensuring the representativity of the data.\(^9\)

### 6 Trends, research gaps and recommendations

After discussing issues regarding the modelling of morality in text and limitations of current applications to research questions in the Computational Social Sciences, we now outline some trends emerging from the survey, as well as research gaps and challenges that we would like to see addressed in future work.

**Resources for languages other than English** While several corpora and dictionaries are available for the analysis of morality in English text, only few resources exist for other languages. This is in line with the findings in Vida et al. (2023). Out of the 18 studies that work with languages other than English, only 6 release an annotated dataset for a new language. 5 papers use dictionary-based approaches, typically based on translations of one or more of the English MFDs without proper validation for the new language. A notable exception is Cheng and Zhang (2023) who introduce new moral foundations for Chinese and also test for validity (see §4.2). The remaining papers either use survey-derived measures of morality, annotate stances on moral topics, or present an annotation tool. Thus, the creation of annotated datasets and tools for languages other than English remains an important research objective.

**Modelling grounded in theory** A high percentage of the papers included in the survey (20%) are not based on any theoretical framework. We argue that grounding models of morality in theory has several advantages. First, it provides a link to previous research and can thus inform our research design and modelling decisions. Second, a sound and well-defined theoretical framework can help address the problem of underspecification in modelling. Finally, the theory can provide us with testable hypotheses and, vice versa, our results can help improve theory development. We therefore recommend researchers that aim at building new

\(^9\)The checklist is included in our GitHub repository: [https://github.com/umanlp/survey_morality](https://github.com/umanlp/survey_morality).
resources to choose an appropriate framework as the theoretical basis for their work.

**Underspecification** Another problem we found concerns the creation of annotated resources as training data. As mentioned above, annotation procedures and guidelines lack (i) transparency and reproducibility, as only half of the survey papers release their guidelines, and (ii) specificity, as the guidelines are often shorter than one page. We therefore recommend the creation and publication of more specific and detailed guidelines for the annotation of morality that provide coders with sufficient information on how to deal with the already challenging task.

In Section 4.5, we argued that it is crucial for the annotation of morality to also encode the perspective, not only to reduce inconsistencies in the annotations, but also to make the resulting resources more useful for different types of analyses. We therefore recommend annotating morality at the level of frames and roles, to explicitly capture this information.

**Representativity, reliability, validity** We highlight the importance of ensuring the representativity of the research data for the population relevant for the respective research questions, especially in comparative studies. To increase the reliability and validity of the analyses, we advise researchers to formulate their research questions, form hypotheses and test for statistical significance. In addition, more work is needed on how to test for different types of validity concerning the construct of morality.

**Morality in AI systems** While not in the center of our interest, we find 20 papers that probe large language models (LLMs), either to search for biases or to investigate what LLMs have learned about morality. Some papers explore whether language models such as BERT or ChatGPT capture moral norms or include a “moral compass” (cf. e.g. Hendrycks et al. (2021); Schramowski et al. (2022)). Most of the approaches try to assess a language model’s knowledge of moral concepts by analyzing the semantic space of the underlying word or sentence representations or use prompts to provide the models with morality related questions or scenarios. While these papers mainly try to uncover the moral and ethical biases of AI systems, we expect to see more research in the future which goes one step further, by not only making the (mostly western centric) moral biases in LLMs transparent but by developing methods for removing the undesirable properties of these models.

7 Conclusion

In the survey, we present a systematic review of work on modelling morality in text. We highlight problems that need to be addressed, one of them being the lack of resources for languages other than English. Another issue concerns the low IAA for the annotation of morality in text. We argue that one reason for this, besides the inherent subjectivity of the task, results from underspecified task instructions, witnessed by a) very short or unavailable guidelines and b) the attempt to code morality on the sentence or document level, making it hard for the annotators to know which aspects of the text they are supposed to code. To address this issue, we recommend annotating morality at the level of frames, to make it clear what is being annotated and from whose perspective.

Another problem we found is the lack of validation of both the resources and the analyses. While many papers use NLP methods to investigate research questions in CSS, only few studies formulate hypotheses or use significance testing. Equally important is the representativeness of the data, especially in comparative studies, and the extent to which it is justified to replace carefully validated methods such as questionnaires with automated dictionary-based text analysis procedures. This should be investigated in future studies. With the paper, we release a checklist that addresses crucial design decisions for text-based analysis that we hope will help researchers to identify some of these issues.

**Limitations**

While we took great care in selecting the papers for our survey, we are aware that there might be relevant papers that we did not consider as they were published in venues not included in our list. We also did not search for broad terms such as “value”, as they resulted in hundreds of thousands of hits for some of the journals and venues that we were not able to screen. Furthermore, the page limit did not allow us to include all references that might be relevant for this topic.

In addition, the discussion about the validity of automated dictionary-based text analysis techniques or other NLP methods for predicting the
moral values of an individual, social group or culture has only scratched the surface. However, a more in-depth discussion requires far more space and is therefore beyond the scope of this survey.

**Ethics Statement**

We strictly adhere to the ACL Ethics Policy. We make all of our approaches, resources, and methods transparent and publicly available and we ensure that the findings and conclusions of our survey are reported accurately and objectively. We believe that our work is helpful for researchers who work on the computational analysis of morality in text and discuss research gaps and possible future research directions, e.g. by addressing the problem of validity of automated text analyses for the prediction of moral values or the lack of resources for non-English languages. We do not anticipate any ethical concerns arising from the research presented in this paper.

**Acknowledgements**

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (PO1900/7-1 and RE3536/3-1). We thank Rainer Gemulla for his constructive and helpful comments on an earlier version of this paper.

**References**


Rezvaneh Rezapour, Saumil H. Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 35–45, Minneapolis, USA. Association for Computational Linguistics.


A Appendix

A.1 Methodology details

In this section of the Appendix, we provide more details about each step of our methodology for selecting and reviewing papers in this survey. The different steps are summarized in Figure 1 (main paper).

A.1.1 Databases

We include ACL, ACM and IEEE as these are large and well-known databases\(^{10}\) containing a vast variety of literature related to NLP and Computer Science from peer-reviewed journals, conferences and workshops. Additionally, we identified relevant journals and publication venues from the Moral Foundations homepage. We manually scanned the publication page for relevant papers by reading titles, abstracts and keywords and then added the corresponding journals to our list. In total, we consider 15 venues and journals.

A.1.2 Search strings

We employ different search strings to detect relevant papers. For some NLP-related venues such as ACL, we opt for a more general search term, e.g. "moral", in order to increase recall. Whenever possible, we search in the fields title, abstract and keywords.

Table 5 provides an overview of search strings and results for each database. For transparency and reproducibility, we provide additional details on the search\(^{11}\) and the exhaustive search results\(^{12}\) the GitHub repository.

---

\(^{10}\)By database, we mean any source used to search for papers. For example, this can be a search on a website in a particular journal or venue.

\(^{11}\)https://github.com/umanlp/survey_morality/search_queries.txt

\(^{12}\)https://github.com/umanlp/survey_morality/search_results/
<table>
<thead>
<tr>
<th>Database</th>
<th>Fields</th>
<th>Keywords</th>
<th>File/URL</th>
<th>No. papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Digital Library</td>
<td>title, abstract, keywords</td>
<td>S&lt;sub&gt;search&lt;/sub&gt;</td>
<td>Query link</td>
<td>192</td>
</tr>
<tr>
<td>ACL Anthology</td>
<td>title, abstract</td>
<td><em>moral</em></td>
<td>anthology+abstracts.bib</td>
<td>77</td>
</tr>
<tr>
<td>Journal of Computational Social Science</td>
<td>any</td>
<td>moral</td>
<td>Query link</td>
<td>18</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>title, abstract, author</td>
<td>S&lt;sub&gt;search&lt;/sub&gt;</td>
<td>Query link</td>
<td>71</td>
</tr>
<tr>
<td>Behavior Research Methods</td>
<td>any</td>
<td>S&lt;sub&gt;search&lt;/sub&gt;</td>
<td>Query link</td>
<td>16</td>
</tr>
<tr>
<td>Sage Journals</td>
<td>abstract, keywords</td>
<td>S&lt;sub&gt;search&lt;/sub&gt; (abstract) and S&lt;sub&gt;extra&lt;/sub&gt; (keywords)</td>
<td>Query link</td>
<td>5</td>
</tr>
<tr>
<td>EPJ Data Science</td>
<td>any</td>
<td>moral foundation</td>
<td>Query link</td>
<td>3</td>
</tr>
<tr>
<td>Knowledge-Based Systems</td>
<td>any</td>
<td>S&lt;sub&gt;search&lt;/sub&gt;</td>
<td>Query link</td>
<td>3</td>
</tr>
<tr>
<td>PLOS One</td>
<td>title, abstract</td>
<td>moral</td>
<td>Query link (CL), Query link (NLP)</td>
<td>3</td>
</tr>
<tr>
<td>Computational Communication Research</td>
<td>any</td>
<td>moral</td>
<td>Query link</td>
<td>1</td>
</tr>
<tr>
<td>AAAI Conference on Artificial Intelligence</td>
<td>title</td>
<td>moral</td>
<td>Query link</td>
<td>12</td>
</tr>
<tr>
<td>AIES</td>
<td>title</td>
<td>moral</td>
<td>Query link</td>
<td>22</td>
</tr>
<tr>
<td>ICWSM</td>
<td>title</td>
<td>moral</td>
<td>Query link</td>
<td>9</td>
</tr>
<tr>
<td>NeurIPS</td>
<td>title</td>
<td>moral</td>
<td>Query link</td>
<td>2</td>
</tr>
<tr>
<td>ICLR</td>
<td>title</td>
<td>moral</td>
<td>Query link</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total (w/ duplicates)** 435

**Total (w/o duplicates)** 406

Table 5: Number of papers (last column) found in each database (first column). The column **Fields** indicates in which fields we searched. The column **Keywords** lists the keywords we searched for. For all databases except ACL Anthology, we conducted a web search. The column **File/URL** shows either the file in which we searched (for ACL Anthology) or a link to the exact query that was used to retrieve documents. The cut-off for all searches is 31.12.2023. Note that query links might produce different results when visited at a later date.

**ACL Anthology** We search the anthology.bib file for the string "moral," in lowercased titles and abstracts: any string containing the substring "moral" will result in a match, e.g. "Exploring Morality in Argumentation". If the string appears in any of the two fields, we select it to be included in the screening phase.

**DBLP** We also search the proceedings of NeurIPS, ICLR and three AAAI-related venues: (i) AAAI Conference on Artificial Intelligence, (ii) AIES AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society and (iii) ICWSM International AAAI Conference on Web and Social Media. To increase the number of search results, we search for the string "moral" without any refinements, as we did for the ACL Anthology.

We conduct the search in the DBLP computer science bibliography, since the aforementioned conferences do not offer a search function that includes all proceedings, nor a way to download all years' proceedings of a conference at once. At the time of writing, the DBLP web interface does not allow us to exclude the author field from the search, so we also obtain numerous matches where the substring "moral" is part of the authors' names. We thus exclude them from the search results, keeping only publications with a match in the title. For example, the query string for the AAAI Conference on Artificial Intelligence shown in Table 5 originally returned 26 results, but we only count the 12 results that contain "moral" in the title. In short, we search for the case-insensitive substring "moral" in titles only.

**Search strings for other sources** For any other source, whenever possible, we opt for a detailed search string:\rm 14

\[ S_{search}: "moral foundation" \text{ OR } "moral foundations" \text{ OR } "moral value" \text{ OR } "moral values" \text{ OR } "moral sentiment" \text{ OR } "morality frame" \text{ OR } "morality frames" \text{ OR } "morality rhetoric" \]

We limit the search to the paper’s metadata fields **title**, **abstract** and **keywords** (or **author keywords** in the case of IEEE).

Additionally, if the search using only \( S_{search} \) re-
turns too many (unrelated) results, we further con-
straint the search using the following string in the
keywords metadata field:
\( S_{extra} : \) "content analysis" OR "text
analysis" OR "discourse analysis" OR "semantic analysis" OR "machine learning"
OR "deep learning" OR "NLP"

For some databases, however, (i) a complex
string search is too restrictive, yielding little to
no results, or (ii) the web interface does not allow
for an advanced search. In these cases, we relax the
search by only searching for the word "moral" or
the string "moral foundation" in any metadata
field.

A.1.3 Filtering and supplementing

After filtering for duplicates, the number of papers
resulting from the search is reduced from 435 to
406 (see Table 5). After screening these papers,
123 relevant papers are left. We then supplement
this set of papers with backward snowballing. 19
papers are considered as potential candidates, from
which 12 remain after screening, amounting to a
total of 135 kept for reviewing.

A.1.4 Screening process

We closely follow the inclusion and exclusion cri-
teria proposed in Alturayef et al. (2023). To be
included in the survey, the paper must satisfy all of
the inclusion criteria below:

1. The methodology must rely on text or speech
data. For example, we do not include pa-
pers that analyze morality using only meth-
ods from psychology (e.g. psychometric ques-
tionnaires) or data science (e.g. purely de-
mographic attributes of users). This criterion
enables to only include papers that are related
to text analysis using computational methods,
and more generally to NLP.

2. The paper proposes a new resource (such as
a dataset), or an experiment, or an analysis.
This means that we exclude papers that are
(solely) surveys or reviews.

3. The paper is either a short or long paper in
conference findings, a journal or workshop
findings. Publications that are only available
as posters, abstracts or other short-form or
visual formats are not included.

4. The paper must be written in English.

5. The paper must be peer-reviewed.\(^{15}\)

6. The paper must be accessible.\(^{16}\)

A.1.5 Pilot study

After completing the screening, we conduct a pilot
review study with the goal of refining the survey
form that we use for reviewing the papers. We
select eight papers (see Table A.1.5) from different
venues and years in order to obtain a more diverse
and informative set of papers for the pilot.

<table>
<thead>
<tr>
<th>ID</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Matsuo et al. (2019)</td>
</tr>
<tr>
<td>2</td>
<td>Haemmerl et al. (2023)</td>
</tr>
<tr>
<td>3</td>
<td>Zhao et al. (2022)</td>
</tr>
<tr>
<td>4</td>
<td>Ziems et al. (2022)</td>
</tr>
<tr>
<td>5</td>
<td>Mahajan and Shaikh (2020)</td>
</tr>
<tr>
<td>6</td>
<td>Orizu and He (2016)</td>
</tr>
<tr>
<td>7</td>
<td>Xu et al. (2023)</td>
</tr>
<tr>
<td>8</td>
<td>Liu et al. (2022)</td>
</tr>
</tbody>
</table>

Table 6: List of papers included in the pilot review study
of the survey.

These eight papers are then reviewed indepen-
dently by the three authors of this paper using a test
version of the survey form. We discuss open ques-
tions, refine and enrich the survey form, and create
a codebook with precise definitions and examples of
the survey categories.

A.1.6 Review process

Based on the test version of the review form, we
create a browser-based survey form to collect re-
views for each paper. The final selection of 135
papers is distributed amongst the authors for the
final reviewing. The codebook ensures that all vari-
ables are well defined and are used consistently by
the reviewers.\(^ {17}\)

Survey form Figure 5 illustrates part of the sur-
vey form that was used to collect reviews for all
papers. Reviewers proceed by starting a server,
which launches the survey form application. After
entering the bibkey of a paper, metadata such as
title, authors and year are entered automatically in
the corresponding fields. The survey form consists of
radio buttons, multiple choice buttons and free

\(^{15}\)Note that papers obtained via backward snowballing are exempt from this constraint; we implement it during the paper sampling phase in order to ensure publication quality of semi-
automatically collected papers.

\(^{16}\)Besides open source publications, we include papers that
we could access through our university libraries (all except 10
papers).

\(^{17}\)The codebook is available on GitHub.
text fields. A codebook accompanying the survey form ensures consistent reviews.

**Consistency checks**  After completing the reviewing process, we run semi-automated consistency checks on the output of the survey forms. In order to find potential mistakes in the reviewing process, we check different variables for consistency, for example:

- The content length of a paper must be lower or equal the total length.
- Certain obligatory fields cannot be left empty.
- Certain variables require a specific truth value for other variables: for instance, in a paper that includes supervised classification of morality, LLMs without fine-tuning cannot be the only method related to experiments that is true.

The script used to run these consistency checks can be found in the supplementary materials.

### A.2 Methodological differences to Vida et al. (2023) in the paper selection method

Vida et al. (2023) review 92 papers, while our survey reviews 135 papers. There is an overlap of 48 papers that have been included in both surveys, however, our survey covers 88 additional papers not considered in Vida et al. (2023). This shows the different focal points of our work, which is reflected in the design of the selection criteria used in each survey. While we search in a range of 15 selected journals and venues, Vida et al. (2023) consider ACL, ACM and the 100 most relevant search results on the search engine Google Scholar. Our search method is not only independent of the ranking of a search engine, but also reproducible and reliable. Additionally, search strings are defined differently in both surveys, reflecting the different backgrounds of both works. For example, Vida et al. (2023) also search for broader concepts related to ethics, such as "utilitarianism" or "deontology"; our work, on the other hand, has a narrower focus on the operationalization of morality in NLP and CSS works.

### A.3 Distribution of paper types per publication year

Tables 7 and 8 respectively show the distribution of paper types and research goals over the years.

### Table 7: Distribution of paper types over time.

<table>
<thead>
<tr>
<th>Year</th>
<th>Analysis</th>
<th>Exp</th>
<th>AI</th>
<th>Demo</th>
<th>Res</th>
<th>NotRel</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2015</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2016</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2017</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2018</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2019</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2020</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>2021</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2022</td>
<td>10</td>
<td>21</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>2023</td>
<td>10</td>
<td>15</td>
<td>9</td>
<td>0</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

### Table 8: Distribution of research goal over time (V/S/F: Values/Stance/Framing; Theory: moral + other).

<table>
<thead>
<tr>
<th>Year</th>
<th>V/S/F</th>
<th>Comparison</th>
<th>Theory</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2015</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2016</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2017</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2018</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2019</td>
<td>17</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2020</td>
<td>12</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2021</td>
<td>28</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2022</td>
<td>33</td>
<td>7</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>2023</td>
<td>41</td>
<td>4</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

### A.4 Criteria for “Foundationhood”

Table 9 lists the MFT criteria for foundationhood, as described in Graham et al. (2013).

<table>
<thead>
<tr>
<th>ID</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A common concern in third-party normative judgments</td>
</tr>
<tr>
<td>2</td>
<td>Automatic affective evaluations</td>
</tr>
<tr>
<td>3</td>
<td>Culturally widespread</td>
</tr>
<tr>
<td>4</td>
<td>Evidence of innate preparedness</td>
</tr>
<tr>
<td>5</td>
<td>Evolutionary model demonstrates adaptive advantage</td>
</tr>
</tbody>
</table>

### Table 9: The five criteria for foundationhood, as detailed in Graham et al. (2013).

### A.5 Validation of the MFT Questionnaire

Measurement tools like the MFT Questionnaire (Graham et al., 2011) are thoroughly tested for different types of validity and reliability, such as:

1. **Internal validity** assesses to which extend we can be sure that the measured effect has been caused by the variable of interest and not by some other factors, such as external variables or alternative explanations.
2. **External validity** assesses how well the findings generalize to other settings (e.g., other time periods, another population, etc.)
3. **Test-retest reliability** assesses the reliability of results when applying the same experimental treatment twice to a group of subjects over a period of time. Results are reliable when both treatments give the same or similar results.

Dictionaries do not undergo any such validation procedure and can thus only be considered as a very rough approximation of the above measurement tools.

### A.6 List of resources

We provide detailed lists of resources for morality in text, including annotated corpora (see Table 11), dictionaries (Table 10) and ontologies/knowledge graphs (Table 12).

**Table 10: Available dictionaries. Please note that the sizes are not comparable, as some dictionaries include word forms while others include lemmas or regexes. Some dictionaries also include a generic MORAL category (not included in the counts above).**

<table>
<thead>
<tr>
<th>Authors</th>
<th>Lang.</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graham et al. (2009)</td>
<td>en</td>
<td>323</td>
</tr>
<tr>
<td>Frimer et al. (2019)</td>
<td>en</td>
<td>2,103</td>
</tr>
<tr>
<td>Rezapour et al. (2019)</td>
<td>en</td>
<td>4,636</td>
</tr>
<tr>
<td>Araque et al. (2020)</td>
<td>en</td>
<td>487</td>
</tr>
<tr>
<td>Hopp et al. (2021)</td>
<td>en</td>
<td>689</td>
</tr>
<tr>
<td>Mather et al. (2022)</td>
<td>en</td>
<td>8,468</td>
</tr>
<tr>
<td>Araque et al. (2021)</td>
<td>en</td>
<td>3,074</td>
</tr>
<tr>
<td>Carvalho et al. (2020)</td>
<td>pt</td>
<td>790</td>
</tr>
<tr>
<td>Cheng and Zhang (2023)</td>
<td>zh</td>
<td>6,138</td>
</tr>
<tr>
<td>Matsuo et al. (2019)</td>
<td>ja</td>
<td>718</td>
</tr>
<tr>
<td>Authors</td>
<td>Lang.</td>
<td>Size</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>Alhassan et al. (2022)</td>
<td>No info</td>
<td>175000 documents</td>
</tr>
<tr>
<td>Alshomary et al. (2022)</td>
<td>en</td>
<td>230k texts, 60 arguments</td>
</tr>
<tr>
<td>Beiró et al. (2023)</td>
<td>en</td>
<td>457065 documents</td>
</tr>
<tr>
<td>De Giorgis et al. (2022)</td>
<td>en</td>
<td>unknown</td>
</tr>
<tr>
<td>Emelin et al. (2021)</td>
<td>en</td>
<td>12k moral stories</td>
</tr>
<tr>
<td>Feyen et al. (2023)</td>
<td>en, fr</td>
<td>430 documents (newspaper articles)</td>
</tr>
<tr>
<td>Forbes et al. (2020)</td>
<td>en</td>
<td>292k RoTs</td>
</tr>
<tr>
<td>Garten et al. (2018)</td>
<td>en</td>
<td>3000 Tweets</td>
</tr>
<tr>
<td>Guan et al. (2022)</td>
<td>en, zh</td>
<td>4209 Chinese documents, 1779 English</td>
</tr>
<tr>
<td>Hendrycks et al. (2021)</td>
<td>en</td>
<td>130,000 examples (elicited moral scenarios and comments from Reddit)</td>
</tr>
<tr>
<td>Hoover et al. (2020)</td>
<td>No info</td>
<td>35108 tweets</td>
</tr>
<tr>
<td>Huang et al. (2022)</td>
<td>en</td>
<td>500 Tweets</td>
</tr>
<tr>
<td>Islam and Goldwasser  (2022a)</td>
<td>en</td>
<td>557 documents</td>
</tr>
<tr>
<td>Jin et al. (2022)</td>
<td>No info</td>
<td>148 vignettes</td>
</tr>
<tr>
<td>Johnson and Goldwasser (2018)</td>
<td>en</td>
<td>2050 documents</td>
</tr>
<tr>
<td>Karami et al. (0)</td>
<td>fa</td>
<td>6000 Tweets</td>
</tr>
<tr>
<td>Kiesel et al. (2022)</td>
<td>en</td>
<td>5270 arguments</td>
</tr>
<tr>
<td>Kobbe et al. (2020)</td>
<td>en</td>
<td>320 documents</td>
</tr>
<tr>
<td>Lin et al. (2018)</td>
<td>en</td>
<td>4191 tweets</td>
</tr>
<tr>
<td>Mather et al. (2022)</td>
<td>en</td>
<td>8473 dictionary entries</td>
</tr>
<tr>
<td>Mossijman et al. (2018)</td>
<td>en</td>
<td>4800 Tweets</td>
</tr>
<tr>
<td>Orizu and He (2016)</td>
<td>en</td>
<td>7600 text segments</td>
</tr>
<tr>
<td>Qian et al. (2021)</td>
<td>No info</td>
<td>1514 stories</td>
</tr>
<tr>
<td>Pacheco et al. (2022)</td>
<td>en</td>
<td>750 annotated tweets and 85,000 unlabeled tweets</td>
</tr>
<tr>
<td>Pavan et al. (2023)</td>
<td>pt-br</td>
<td>4080 documents</td>
</tr>
<tr>
<td>Pyatkin et al. (2023)</td>
<td>en</td>
<td>clarification questions for 6425 situations</td>
</tr>
<tr>
<td>Rao et al. (2023)</td>
<td>No info</td>
<td>20537</td>
</tr>
<tr>
<td>Rizzoli (2023)</td>
<td>it</td>
<td>2381 Tweets</td>
</tr>
<tr>
<td>Roy and Goldwasser (2021)</td>
<td>No info</td>
<td>161k documents (tweets)</td>
</tr>
<tr>
<td>Roy et al. (2022)</td>
<td>en</td>
<td>1599 tweets</td>
</tr>
<tr>
<td>Ruskov et al. (2023)</td>
<td>en</td>
<td>not specified</td>
</tr>
<tr>
<td>Sánchez-Rada et al. (2023)</td>
<td>en, it, es</td>
<td>113817 prompt-answer pairs</td>
</tr>
<tr>
<td>Santos and Paraboni (2019)</td>
<td>pt-br</td>
<td>5242 texts</td>
</tr>
<tr>
<td>Shahid et al. (2020)</td>
<td>No info</td>
<td>400 documents, 100 documents</td>
</tr>
<tr>
<td>Solans et al. (2021)</td>
<td>No info</td>
<td>2565 instances (mostly sentences)</td>
</tr>
<tr>
<td>Trager et al. (2022)</td>
<td>en</td>
<td>16123 documents</td>
</tr>
<tr>
<td>Weinzierl and Harabagiu (2022)</td>
<td>No info</td>
<td>14180 documents</td>
</tr>
<tr>
<td>Zhang et al. (2023)</td>
<td>en</td>
<td>474 news articles</td>
</tr>
<tr>
<td>Zheng et al. (2022)</td>
<td>en</td>
<td>24425 entities</td>
</tr>
<tr>
<td>Ziems et al. (2022)</td>
<td>en</td>
<td>38k chatbot replies to human-authored prompts, 113817 prompt-answer pairs</td>
</tr>
</tbody>
</table>

Table 11: Overview of available datasets. The column Annot. setup indicates whether there was a manual annotation (crowd annotators, trained annotator or mix of both). The column Annot. schema indicates whether annotation guidelines were made available. IAA indicates whether inter-annotator agreement was reported. The last column Available indicates whether the created dataset is publically available. The reported information is not always available in the paper; in this case, we note "No info".
<table>
<thead>
<tr>
<th>Authors</th>
<th>Lang.</th>
<th>Size</th>
<th>Annot. setup</th>
<th>Annot. schema</th>
<th>IAA</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ruskov et al. (2023)</td>
<td>en</td>
<td>not specified</td>
<td>mixed</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Garten et al. (2018)</td>
<td>en</td>
<td>3000 Tweets</td>
<td>trained</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Lin et al. (2018)</td>
<td>en</td>
<td>4191 tweets</td>
<td>trained</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Pyatkin et al. (2023)</td>
<td>en</td>
<td>clarification questions for 6425 situations</td>
<td>crowd</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>De Giorgis et al. (2022)</td>
<td>en</td>
<td>unknown</td>
<td>No annotation</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Yes</td>
</tr>
<tr>
<td>Zhang et al. (2023)</td>
<td>en</td>
<td>474 news articles</td>
<td>trained</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Feyen et al. (2023)</td>
<td>en, fr</td>
<td>430 documents (newspaper articles)</td>
<td>trained</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sánchez-Rada et al.</td>
<td>en, it, es</td>
<td>unknown</td>
<td>No annotation</td>
<td>Not relevant</td>
<td>Not relevant</td>
<td>Yes</td>
</tr>
<tr>
<td>Weinzierl and Harabagiu (2022)</td>
<td>No info</td>
<td>14180 documents</td>
<td>trained</td>
<td>No</td>
<td>Yes</td>
<td>Partly</td>
</tr>
</tbody>
</table>

Table 12: Available ontologies. Columns are analogous to Table 11.