OrigamIM: A Dataset of Ambiguous Sentence Interpretations for Social Grounding and Implicit Language Understanding

Liesbeth Allein, Marie-Francine Moens
Department of Computer Science, KU Leuven
Leuven, Belgium
{liesbeth.allein, sien.moens}@kuleuven.be

Abstract

Sentences elicit different interpretations and reactions among readers, especially when there is ambiguity in their implicit layers. We present a first-of-its kind dataset of sentences from Reddit, where each sentence is annotated with multiple interpretations of its meanings, understandings of implicit moral judgments about mentioned people, and reader impressions of its author. Scrutiny of the dataset proves the evoked variability and polarity in reactions. It further shows that readers strongly disagree on both the presence of implied judgments and the social acceptability of the behaviors they evaluate. In all, the dataset offers a valuable resource for socially grounding language and modeling the intricacies of implicit language understanding from multiple reader perspectives.

Keywords: implicit language, interpretation, ambiguity, social grounding, moral reasoning, resource

1. Introduction

A sentence frequently evokes diverse and disagreeing interpretations. Disagreement in interpretation can arise from explicit cues, such as the choice and order of words, triggering phonological, lexical, and structural ambiguities (Kennedy, 2019). This disagreement is further amplified by a diversity among readers, each guided by their unique experiences, knowledge, and viewpoints. Despite extensive exploration of ambiguity within computational linguistics (Bevilacqua et al., 2021; Haber and Poesio, 2023), little attention has been devoted to *ambiguity in the implicit layers of sentences* and the resulting *disagreement in interpretation*.

This underexposure of the implicit is surprising considering a substantial portion of human communication is inherently non-verbal. Even when using language, we convey information between the lines. Implicit communication is efficient since it obviates the need to reiterate common sense or common ground information (Stalnaker, 2002), and it is social as it can prevent a loss of face when sharing social evaluations (Dunbar, 2004). Some people also reside to the implicit layers of communication when targeting a specific audience and deceiving all others (e.g., dogwhistles (Henderson and McCready, 2017)). Achieving such human-like communication skills in computational models therefore necessitates *a transition to multi-perspective language production and understanding*, in which models are equipped with the ability to reason over implicit content from multiple angles.

To facilitate the development of such models, we curate *a first-of-its-kind dataset* of sentences, where each sentence is annotated with multiple interpretations, detailed descriptions of underlying moral judgments of people mentioned in the sentence, and measures of reader attitude describing a reader's first impression of the author upon reading the sentence[1] (Figure 1). The latter two information types socially ground the sentences from multiple perspectives. The name of the dataset, `origamIM`, refers to the analogy between the Japanese art of paper folding and the diversity of

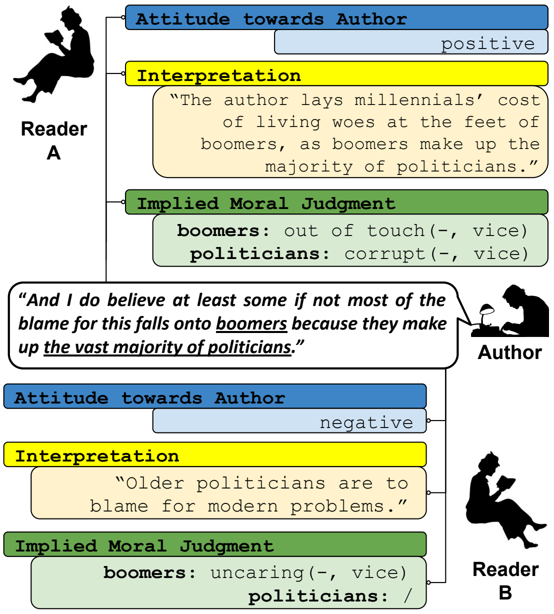
Reader A
Attitude towards Author
positive
Interpretation
"The author lays millennials' cost of living woes at the feet of boomers, as boomers make up the majority of politicians."
Implied Moral Judgment
boomers: out of touch(-, vice)
politicians: corrupt(-, vice)
"And I do believe at least some if not most of the blame for this falls onto boomers because they make up the vast majority of politicians."
Author
Attitude towards Author
negative
Interpretation
"Older politicians are to blame for modern problems."
Implied Moral Judgment
boomers: uncaring(-, vice)
politicians: /
Reader B

Figure 1: Sample taken from the `origamIM` dataset, demonstrating the diverging reader attitudes towards the author, slightly different interpretations, and disagreeing understandings of implicit moral judgments a sentence can trigger.

[1] The dataset is publicly available: https://github.com/laallein/origamIM.

NLPerspectives 2024@LREC-COLING 2024, pages 116–122
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Context	*Appropriateness*		
Sphere of Action	**Vice of Deficiency**	**Virtue of Mean**	**Vice of Excess**
Confidence, fear, uncertainty	Cowardice	Courage	Rashness
Pleasures of the body	Insensibility	Temperance	Profligacy
Giving & taking: Small money	Stinginess	Liberality	Prodigality
Giving & taking: Added value	Meanness	Magnificence	Vulgarity
Pride, honor as cause	Little-mindedness	High-mindedness	Vanity
Ambition, honor as goal	Lack of ambition	Proper ambition	Over-ambition
Anger	Spiritlessness	Gentleness	Wrathfulness
Pleasure and pain of others	Cross, contentious	Agreeableness	Flattery
Truth, honesty about oneself	Irony	Truthfulness	Boastfulness
Amusing conversation	Boorishness	Wittiness	Buffoonery

Table 1: Overview of spheres of actions and the degrees of appropriateness (Hursthouse, 1999).

interpretations and attitudes that could be obtained when presented with the same sentence.

2. A Moral Framework for Grounding

Moral judgments offer an interesting case for examining and modeling disagreement. Individuals namely look through their own lenses when judging people and interpreting judgments made by others, despite a shared understanding of moral norms and values. The judgments annotated in the dataset are grounded in Virtue Ethics (Hursthouse, 1999). The moral theory introduced by Aristotle poses that a person's moral character can be evaluated by the contextual appropriateness of their voluntary behavior within a sphere of action (see Table 1). A virtuous behavior is characterized by moderation and appropriateness within its context (e.g., considering the people involved and the severity of the situation) while contextually deficient or excessive behaviors are not celebrated in society.

The axis of appropriateness in Virtue Ethics provides a distinct advantage over other popular moral frameworks (e.g., Moral Foundation Theory (Haidt and Joseph, 2004)) as it enables individuals not only to differentiate between negative behavior based on its context, but also to annotate their understanding of the implied moral judgments given their cultural and social backgrounds.

3. Dataset Creation

3.1. Data Collection

We automatically retrieve blog posts in English from the Subreddit /r/ChangeMyView that were posted between 13 July 2020 and 3 March 2022. These posts typically present views on often controversial and polarizing topics, such as abortion and racism. We anticipate that a considerably large portion of the posts pass judgments about people given the human tendency to gossip (Dunbar, 2004; Baumeister et al., 2004; Feinberg et al., 2012). Moreover, negative judgments are expected to be conveyed implicitly due to the subreddit's moderation policies[2]. We remove duplicated and deleted blog posts and extract the title, body text, and additional metadata[3] for each post. Lastly, the body text is segmented into sentences using SpaCy.

3.2. Data Annotation

We recruit crowd workers on Amazon Mechanical Turk[4] and let them annotate the sentences in two rounds. An annotator never annotates the same sentence in both rounds. The first round distinguishes sentences that mention people and imply a character trait of at least one of them from those that either lack explicit mentions people or do not imply any character trait. A character trait presents a voluntary aspect of a person's attitude or behavior, e.g., *lazy* and *charitable*. The second round takes the first set of sentences and gathers multiple reader attitudes, interpretations, and entity-level moral judgments for each sentence.

3.2.1. First Round: People Entities

Two annotators mark all entities referring to people other than the author (i.e., 'I') in a sentence and indicate whether or not the author seems to imply a character trait of at least one highlighted entity. We show the title of the blog post from which the sentence was taken as additional context. In cases where they disagree on the presence or absence of implied traits, a third annotator is consulted and a majority vote is taken. Data quality and consistency is manually checked. A total of 6,820 sentences were annotated, of which 2,018 implied a character trait of at least one people entity. These figures confirm our expectations regarding the presence of implicit social evaluations in these posts (see §3.1).

3.2.2. Second Round: Attitudes, Interpretations, and Moral Judgments

Five annotators read the same sentence and first describe their attitude towards its author using a

[2] The moderation rules dictate that posts suggesting harm to others and hostile comments will be removed. See https://www.reddit.com/r/changemyview/wiki/modstandards/ [accessed on 4 April 2024].

[3] The metadata is not used during the annotation process.

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

Dataset Statistics	
# Blog posts	396
# Sentences	2,018
– Total word count	44,902
– Min/max words per sentence	2 / 107
# People entities	3,313
– # Sentences with 1/2/3/4+ entities	1,103 / 661 / 174 / 80
# Interpretations	9,851
– Total word count	155,368
– Min/max words per interpretation	1 / 113
Distribution reader attitudes	
– Very negative	813 (8.25%)
– Negative	1,971 (20%)
– Neutral	4,302 (43.67%)
– Positive	2,025 (20.56%)
– Very positive	740 (7.51%)

Table 2: Statistics of the `origamIM` dataset.

five-point Likert scale ranging from *very negative* (1) to *very positive* (5). They then write down their interpretation of the sentence. We explicitly instruct them to not copy the sentence and manually check the relatedness between sentence-interpretation pairs, removing annotations that present unrelated pairs or poorly-formulated interpretations. Going over all the people entities marked in the first annotation round, the annotators indicate for each entity whether or not the author implies a character trait. In case a trait is implied, they describe it using, preferably, an adjective, mark whether it considered a good or bad trait in society, and classify it in Virtue Ethics (see §2). A complete annotation for a single sentence interpretation looks as follows:

Title CMV: It Should Be Mandatory for Every Person to Work AT LEAST 1 Month in a Customer - Facing Hospitality Role Before Leaving School.

Sent I truly believe it would have been life changing for **[him]** to work in hospitality for a bit before leaving school, to see and experience what **[some people]** have to go through on a daily basis just to eat and have a roof.

Att Positive (4)

Int *"Real life experience is better than theory."*

Judg **[him]** ✓ *"ignorant"*, Bad, Pride/honor as cause, Vice of Deficiency.

[some people] ✓ *"hardworking"*, Good, Ambition/honor as goal, Virtue of Mean.

4. Data Analysis

Table 2 presents general statistics of the `origamIM` dataset.

4.1. Disagreement in Attitudes

Each annotater described their attitude towards the author using a five-point Likert scale. Each sentence therefore potentially evokes up to five distinct attitudes among its readers. Figure 2 illustrates the *diversity* of attitudes elicited by a sentence, revealing that the vast majority of sentences trigger at least two different attitudes among readers. As many as one in five sentences even evoke four or five distinct attitudes. This underscores the variability in reactions among readers when presented with the same sentence.

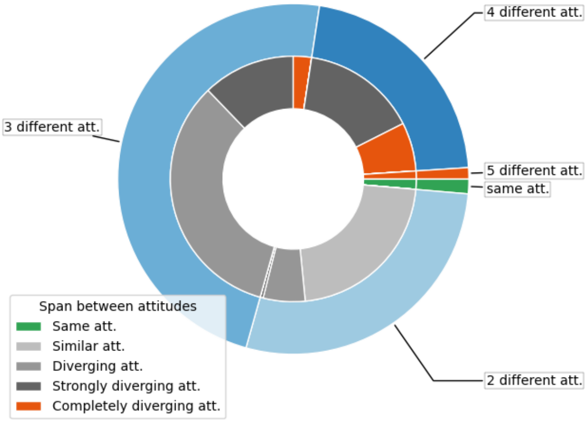

Figure 2: Donut chart representing the disagreement in reader attitude. The outer donut shows the distribution of attitude *diversity*. The inner donut shows the distribution of attitude *divergence*.

We also examine the *divergence* among those attitudes by measuring the span between the lowest and highest attitude, as indicated on the Likert scale, among the five annotators for each sentence. Figure 2 shows that more than one in four sentences evoke strongly (e.g., *very negative* - *positive*) or completely diverging attitudes (i.e., *very negative* - *very positive*). Similar attitudes are elicited for fewer than 20% of the sentences. These findings show that sentences frequently spark not only different, but also diverging attitudes.

4.2. Disagreement in the Implicit

4.2.1. Moral Judgments

We observe that the diversification in interpretation already starts when discerning the presence of implicit moral judgments as annotators exhibited high disagreement on this issue. Merely 291 sentences (14.42%) garnered unanimous agreement among all five annotators on this matter. This disagreement may arise from varying degrees of subtlety in the social evaluations, requiring more in-depth reasoning to uncover them.

The annotators also disagreed on the societal desirability of the implied character traits (i.e., whether the traits are considered 'good' or 'bad'), with Krippendorff's $\alpha = .354$ (Krippendorff, 2011) over the annotators' evaluations of each entity. This entails that often one annotator identifies a negative judgment of an entity's character while another perceives a positive one, and vice versa. Even when they agree, it does not automatically lead

Title *CMV: I don't see a problem with people valuing to defend their property over an intruders life.*
Sentence *Who knows, maybe* ***she*** *is stealing* ***his*** *last 1000 dollars that will pay his rent.*

Interpretation	**Attitude**	**Moral Judgments**
She is taking money which does not belong to her.	**very positive**	**she:** bad, greedy woman, VE: giving and taking (money) - Vice of Excess. **his:** good, generous, morality: giving and taking (money) - Virtue of Mean
Perhaps the thief is stealing an individual's last thousand dollars that they needed for rent.	negative	**she:** bad, dishonest, VE: ambition, honour (goal) - Vice of Deficiency **his:** good, innocent, VE: pride, honour (cause) - Virtue of Mean
We never know who we are dealing with and other people have different problems that we might not be aware of.	neutral	**she:** bad, insensibility, VE: giving and taking (money) - Vice of Deficiency **his:** [`No judgment`]

Table 3: Sample from the dataset illustrating the disagreement existing between readers in terms of interpretation, attitude, and inferred moral judgments.

to similar interpretations or attitudes (see Table 3). We suspect that the latter partially stems from (dis)agreement between the beliefs held by the reader and those seemingly held by the author. One reader may find their beliefs confirmed by the author and consequently report a positive attitude while another disagrees with the author, indicating a negative attitude.

4.2.2. Interpretations

We investigate whether a difference in interpretation is linearly correlated with a difference in attitude and implicit moral judgments. We quantify the difference between two interpretations i of a sentence by two readers j and k as $di(i_j, i_k)$:

$$di(i_j, i_k) = 100 - \text{BLEU-1}(i_j, i_k) \qquad (1)$$

where BLEU-1 (Papineni et al., 2002) measures the lexical overlap at the unigram level. We specifically opt for a simple lexical metric since more complex semantic metrics (e.g., model-based metrics) do not sufficiently capture subtle semantic variations. The difference between two reader attitudes a_j and a_k is denoted as $da(a_j, a_k)$ and obtained by taking the absolute difference in Likert score:

$$da(a_j, a_k) = |a_j - a_k| \qquad (2)$$

The difference in implicit moral judgments $dm(m_j, m_k)$ is quantified as follows:

$$dm(m_j, m_k) = \frac{1}{Q}\sum_{q=1}^{Q} non\_overl(m_{j,q}, m_{k,q}) \qquad (3)$$

where Q is the number of people entities in the sentence and $non\_overl(m_{j,q}, m_{k,q})$ counts the non-overlapping moral judgment characteristics m of people entity q annotated by reader j and k. The moral characteristics include a binary indicator of the presence/absence of an implicit character trait, its description, its evaluation, its classification in a sphere of action, and its contextual appropriateness. The three difference metrics are proportional to disagreement, with high values indicating high disagreement. The lexical difference in interpretation di is positively correlated with the difference in attitude da ($r = .4375, p < .01$), and moral judgment dm ($r = .5207, p < .01$). Correlation between da and dm is also positive but weaker ($r = .3000, p < .01$). These results present promising directions for automated multi-perspective modeling of implicit language understanding.

Diversity in interpretation is especially interesting as it may lay bare various implicit layers of sentences and provide insights into the reasoning paths of readers. Take the five interpretations provided for the following sentence:
"I hear a lot about adults job jumping nowadays just to get bigger wages, and honestly?"

[1] *"Adults are changing jobs for bigger paychecks."*

[2] *"The writer describes having heard about many people changing jobs to get higher wages."*

[3] *"People switching jobs for better wages is a real awful situation nowadays."*

[4] *"People are only interested in money and not stability."*

[5] *"Capital pursuit is not worth moral sacrifice."*

Interpretation [1] and [2] reflect fairly similar understandings of the sentence that remain close to its explicit phrasing. Interpretations [3 – 5], on the other hand, dig deeper in its hidden layers, uncovering strong evaluations of the presented situation. Analyzing salient markers in the sentence guiding the different interpretations (Mastromattei et al., 2022) may here partly explain the reasoning paths taken by the annotators.

5. Related Work

The non-aggregated annotations in `origamIM` describe diverse reader understandings of implicit content. Works tackling the mining of implicit communication have looked into the retrieval of implicit sentiment (Zhou et al., 2021; Li et al., 2021), recovery of social and power implications (Sap et al., 2020), and classification of underlying abuse in statements (Wiegand et al., 2021; ElSherief et al., 2021). Despite the subjective nature of such tasks (Kanclerz et al., 2022), most of the studies relied on aggregated datasets for modeling.

The dataset also contributes to the field of automated moral reasoning, where previous work focused on judging the morality of social conduct (Hendrycks et al., 2021a; Forbes et al., 2020; Emelin et al., 2021; Jin et al., 2022; Pyatkin et al., 2023), classifying moral judgments (Botzer et al., 2022; Efstathiadis et al., 2022), presenting answers to moral dilemmas (Bang et al., 2022), and selecting morally appropriate answers (Hendrycks et al., 2021b; Ziems et al., 2022). Since debating the morality of human behavior is characterized by discord, we deliberately keep multiple ground-truth annotations of moral judgment, in contrast to the datasets supporting previous moral reasoning tasks.

6. Conclusion

This work introduces a novel, non-aggregated dataset of sentences from social media annotated with diverse sentence interpretations, reader attitudes, and implicit moral judgments. It presents a valuable resource for investigating and modeling ambiguity in the implicit layers of sentences and grounding language in society. Possible NLP tasks include perspective modeling, sentiment analysis, and opinion mining. Lastly, future work may look into techniques for dealing with disagreement in the ground truth in the modeling and evaluation phase (Lovchinsky et al., 2019; Uma et al., 2021; Davani et al., 2022; Leonardelli et al., 2023).

7. Ethics Statement

We follow the recommendations in Pater et al. (2021) for reporting annotator selection, compensation and communication. Regarding selection, workers were allowed to work on our annotation task immediately after passing an initial annotation instruction test, which was automatically corrected. They were paid a fixed amount per accepted HIT through the Amazon MTurk platform within three working days after completion and could earn between the U.S. legal minimum wage of $7.5 and $15/hour depending on their annotation flow and experience with the task. In case we rejected a HIT, we provided instructive motivations and gave additional feedback upon request. The majority of rejections originated from incorrect following of explicit instructions. We personally replied to all messages from the workers, most of them within one working day. We did not discriminate between the annotators in terms of gender, race, religion, or any other demographic feature.

8. Acknowledgments

This work was realized with the collaboration of the European Commission Joint Research Centre under the Collaborative Doctoral Partnership Agreement No 35332. It is also funded in part by the Research Foundation - Flanders (FWO) under grant G0L0822N through the CHIST-ERA iTRUST project and in part by the European Research Council (ERC) under the Horizon 2020 Advanced Grant 788506. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

9. Bibliographical References

Yejin Bang, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Dan Su, Elham J Barezi, Andrea Madotto, Hayden Kee, and Pascale Fung. 2022. Aisocrates: Towards answering ethical quandary questions. *arXiv preprint arXiv:2205.05989*.

Roy F Baumeister, Liqing Zhang, and Kathleen D Vohs. 2004. Gossip as cultural learning. *Review of general psychology*, 8(2):111–121.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. Analysis of moral judgment on Reddit. *IEEE Transactions on Computational Social Systems*, pages 1–11.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Robin IM Dunbar. 2004. Gossip in evolutionary perspective. *Review of general psychology*, 8(2):100–110.

Ion Stagkos Efstathiadis, Guilherme Paulino-Passos, and Francesca Toni. 2022. Explainable patterns for distinction and prediction of moral judgement on Reddit. *arXiv preprint arXiv:2201.11155*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Matthew Feinberg, Robb Willer, Jennifer Stellar, and Dacher Keltner. 2012. The virtues of gossip: reputational information sharing as prosocial behavior. *Journal of personality and social psychology*, 102(5):1015.

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

Janosch Haber and Massimo Poesio. 2023. Polysemy-evidence from linguistics, behavioural science and contextualised language models. *Computational Linguistics*, pages 1–67.

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

R Henderson and Elin McCready. 2017. How dogwhistles work. In *New Frontiers in Artificial Intelligence. JSAI-isAI 2017. Lecture Notes in Computer Science*, pages 231–240.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *International Conference on Learning Representations*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021b. What would jiminy cricket do? towards agents that behave morally.

Rosalind Hursthouse. 1999. *On Virtue Ethics*. OUP Oxford.

Zhijing Jin, Sydney Levine, Fernando Gonzalez, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pages 28458–28473.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Christopher Kennedy. 2019. Ambiguity and vagueness: An overview. *Semantics-Lexical Structures and Adjectives*, page 236.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *Computing*, 1.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Igor Lovchinsky, Alon Daks, Israel Malkin, Pouya Samangouei, Ardavan Saeedi, Yang Liu, Swami Sankaranarayanan, Tomer Gafner, Ben Sternlieb, Patrick Maher, et al. 2019. Discrepancy ratio: Evaluating model performance when even experts disagree on the truth. In *International Conference on Learning Representations*.

Michele Mastromattei, Valerio Basile, and Fabio Massimo Zanzotto. 2022. Change my mind: How syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 117–125, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. Standardizing reporting of participant compensation in HCI: A systematic literature review and recommendations for the field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

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

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. Implicitly abusive comparisons – a new dataset and linguistic analysis. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit sentiment analysis with event-centered text representation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.