Balancing Quality and Variation: Spam Filtering Distorts Data Label Distributions

Eve Fleisig*[1], Matthias Orlikowski*[2], Philipp Cimiano[2], and Dan Klein[1]

[1]UC Berkeley, {efleisig,klein}@berkeley.edu
[2]Bielefeld University, {morlikowski,cimiano}@techfak.uni-bielefeld.de

Abstract

For datasets to accurately represent diverse opinions in a population, they must preserve variation in data labels while filtering out spam or low-quality responses. How can we balance annotator reliability and representation? We empirically evaluate how a range of heuristics for annotator filtering affect the preservation of variation on subjective tasks. We find that these methods, designed for contexts in which variation from a single ground-truth label is considered noise, often remove annotators who disagree *instead* of spam annotators, introducing suboptimal tradeoffs between accuracy and label diversity. We find that conservative settings for annotator removal (<5%) are best, after which all tested methods increase the mean absolute error from the true average label. We analyze performance on synthetic spam to observe that these methods often assume spam annotators are less random than real spammers tend to be: most spammers are distributionally indistinguishable from real annotators, and the minority that are distinguishable tend to give fixed answers, not random ones. Thus, tasks requiring the preservation of variation reverse the intuition of existing spam filtering methods: spammers tend to be *less* random than non-spammers, so metrics that assume variation is spam fare worse. These results highlight the need for spam removal methods that account for label diversity.

1 Introduction

Because spam responses are common on crowdsourcing sites, researchers need reliable ways to filter out low-quality data. Many of these methods aim to find annotators with unusual labeling behavior. However, a growing body of work has found that information from annotators with minority opinions can be a valuable source of information, since this disagreement helps to understand variability in the opinions of a population, identify cases where some annotators may be better- or worse-informed, or reveal ambiguity in the task. How can we preserve the opinions of annotators who disagree, while still removing spam annotations?

We examine the effects of applying several common methods for discounting spam annotators based on their labeling behavior. Despite the existence of spam removal methods that use attention checks or metadata (e.g., time spent on task), filtering based on labeling behavior remains common practice (Klie et al., 2024); thus, weaknesses in these methods risk affecting a wide range of common machine learning tasks. We test three of these methods—MACE (Hovy et al., 2013), CrowdTruth (Aroyo and Welty, 2014), and inter-annotator agreement metrics—on relatively subjective tasks and analyze effects on variability in the filtered data. We find that, although many methods are near-indistinguishable in terms of their *accuracy* at classifying spam annotators, some are far more likely to remove non-spam annotators who disagree. Furthermore, we find that under most tested methods, removing more annotators degrades the variety of opinions expressed, without improving accuracy at removing spam annotators; thus, these methods seem most effective only when a very low number of annotators are removed.

We also find that assumptions about the distribution of spam annotations can hinder the effectiveness of these methods. We examine performance on synthetic distributions of spam annotations to analyze whether these methods effectively remove spam annotations, or simply remove annotations farther from the mean. Performance on synthetic spam indicates that most methods perform far better for random spam (e.g., randomly clicking answers) than fixed spam (e.g., always answering "No"). Yet true spammer behavior exhibits the opposite trend: most spammers are distributionally

*Equal contribution; order determined by coin flip.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 47–62

similar to high-quality annotators; the minority that can be reliably identified tends to have fixed spamming behavior. As a result, methods that perform poorly on fixed spam tend to also perform poorly on real spam.

Our results indicate that spam detection for subjective problems flips model assumptions: spam annotators are often less random than non-spam ones. Thus, attempts to remove spam can backfire by instead removing annotators with minority opinions who are not spammers. As a result, existing methods work best when only low percentages of annotators are removed based on their labeling behavior. When over-filtering for spam, these methods risk distorting the distributions of labels.

2 Related Work

Methods for spammer removal impact the variation in resulting disaggregated datasets, as discussed in work on *spammer detection and aggregation methods* and studies underscoring the role of *subjectivity and variation in annotation*.

Defining Spammers in Annotation. Drawing a conceptual boundary between spammers and genuine annotators is complex; definitions vary on what range of intentional, inattentive, or low-effort behaviors should be filtered out; and on whether spammers are posited as too random or too fixed. Buchholz and Latorre (2011) highlight that spammers are incentivized to earn more money faster, leading them to ignore task instructions or participation requirements. Rothwell et al. (2015) argue that spammers act with intention, unlike other types of low-quality annotators, and show *repeated patterns* in an attempt to complete tasks fast. In contrast, Raykar and Yu (2012) posit that spammers *assign labels randomly*, because they do not follow labeling criteria, skip reading the instances or might use automation. Gadiraju et al. (2015) present a nuanced taxonomy of annotator types and underscore that genuine annotators' behavior might overlap with spammers, e.g., failing attention checks for innocuous reasons. The datasets used in our study excluded annotators if they failed data quality checks combining multiple sources of information, thus following a wider definition of spam (Aroyo et al., 2023; Huang et al., 2023, see Section 3). To summarize, any definition of "spammer" includes or excludes different subsets of annotators. These ambiguous boundaries suggest that different subsets of spammers may exhibit different behaviors, potentially raising challenges in distinguishing spammers from non-spammers.

Spammer Detection and Gold Label Aggregation. Data quality and questionable trust in non-expert raters are longstanding problems in crowdsourced annotation (Snow et al., 2008). Attempts to improve data quality may modify tasks to attract less spam before data collection (Eickhoff and de Vries, 2013) or use quality control afterwards (Difallah et al., 2012). Methods for a posteriori detection of low-quality raters and spammers often use intrinsic metrics based on the labeling behavior itself (Buchholz and Latorre, 2011). Intrinsic metrics used for spammer detection include clustering on a post-processed annotation matrix (Traganitis and Giannakis, 2021), rater similarity and agreement scores (Ak et al., 2021), or distance between sequential spamming behaviors (Ba et al., 2024), among others (Ipeirotis et al., 2010; Raykar and Yu, 2012; Gadiraju et al., 2015). Other methods analyze labeling behavior with the goal of *aggregating to the true label* while accounting for varying annotator reliability. Dawid and Skene (1979) model annotator error rates to estimate the true labels and are foundational to many subsequent aggregation methods (Whitehill et al., 2009; Welinder et al., 2010), including in NLP (Wiebe et al., 1999). Passonneau and Carpenter (2014) present a probabilistic variant of the Dawid & Skene model, and many other extensions of this basic model exist (Paun et al., 2018, 2022). In particular, Hovy et al. (2013) present MACE, a probabilistic model tailored towards estimating annotator competence by modeling spamming behaviors. In contrast, CrowdTruth, a non-probabilistic paradigm, derives quality metrics from vector space representations of annotators, annotated examples and annotations (Aroyo and Welty, 2014; Dumitrache et al., 2018b). We evaluate MACE and CrowdTruth as they underwent widespread adoption in NLP and have reference implementations available (see Sections 4.1, 4.2).

Subjectivity and Variation in Annotation. There is a growing body of work researching informative disagreement, diversity of perspectives, and label variation in human annotation (Plank, 2022; Leonardelli et al., 2023; Sandri et al., 2023; Frenda et al., 2024; Fleisig et al., 2024). These works agree that aggregating labels into a single truth is an oversimplification for many tasks (Aroyo and Welty, 2015; Uma et al., 2021; Basile et al., 2021) and might not represent perspectives fairly (Abercrom-

bie et al., 2022). Instead, studies release annotator-level labels (Prabhakaran et al., 2021) to enable alternative approaches, such as modeling individual annotators' rating behaviors (Fleisig et al., 2023; Orlikowski et al., 2023; Heinisch et al., 2023; Orlikowski et al., 2025). Our work is motivated by studies on *rating distributions in a given population* as an alternative to single ground truth prediction (Sorensen et al., 2024; Meister et al., 2025). Among these, Prabhakaran et al. (2024) study systematic disagreement using similar metrics to ours, but on the level of demographic subgroups. In this context, the issue of how capturing labeling variation intersects with annotation quality is largely unexplored. One exception is VariErr (Weber-Genzel et al., 2024), an annotation methodology to differentiate between annotation errors and plausible variation in annotation. In contrast, we study properties of methods that determine annotator reliability, not individual annotation errors.

3 Datasets

We selected two datasets for the basis of our experiments: DICES 350 (Aroyo et al., 2023) and Huang et al. (2023)'s survey of Amazon Mechanical Turk workers. We present each dataset's statistics and discuss our dataset selection process below.

DICES-350 DICES-350 (Aroyo et al., 2023), a harmful language dataset, consists of 43,050 annotations on a 3-point scale across 350 items, each of which was labeled by every participant. 123 annotators participated, of whom 19 annotators were labeled as spam (15% of annotators).

MTurk From Huang et al. (2023)'s survey, we used 16 questions on a 7-point scale, each of which was answered by every participant, for a total of 3,312 annotations. 207 annotators participated, of whom 40 were labeled as spam (19% of annotators).

Dataset selection. Our experiments require datasets that retain (a) responses from multiple annotators per question, permitting measurement of disagreement statistics, and (b) responses from known spammers. Despite increasing availability of annotator-level data,[1] most public datasets do not include spammer information. For papers that report spammer removal, we contacted authors for access to the unfiltered datasets, but spammer responses are regularly lost over time (e.g., Buchholz and Latorre, 2011; Dumitrache et al., 2017; Paun et al., 2018). Even for published data, maintaining data access is not always possible; some datasets with verified spammer information were no longer available (e.g., Soberón et al., 2013; Gadiraju et al., 2015). Similarly, many studies on spammer detection evaluate only on downstream performance or exclusively use synthetic data, so they do not provide metadata on known natural spammers (e.g., Raykar and Yu, 2012; Ak et al., 2021). See Appendix D for details on all 22 considered datasets, including spammer metadata and data availability. In summary, DICES-350 and the MTurk survey are, to the best of our knowledge, the only available datasets meeting our criteria. Nevertheless, these datasets do represent two representative use cases in which preserving rater variation is essential. DICES-350 collects annotations on AI safety to preserve variation on diverse perspectives regarding high-stakes topics; the MTurk dataset polls workers on their personal opinions about their crowdwork experiences in order to best understand the range of opinions of the community.

4 Methods for Spammer Detection

We study a number of established methods and baselines to calculate scores of annotator reliability. To perform spammer detection, we rank annotators using the respective reliability score and identify the k lowest-scoring annotators as spammers for a given value of k.

4.1 Multi-Annotator Competence Estimation (MACE)

MACE (Hovy et al., 2013) is based on a probabilistic model of annotation. We highlight a few aspects of MACE that are important to our study and refer to the original paper for full details. The model includes a parameter θ for each annotator which encodes the probability that they give the true answer (*competence*). Specifically, for each instance i and annotator j, the binary variable S_{ij} indicates whether an annotator is spamming. S_{ij} is drawn from a Bernoulli distribution with parameter $1 - \theta_j$. If the annotator is spamming, i.e., $S_{ij} = 1$, then the assigned label A_{ij} is sampled from a multinomial distribution with a parameter vector ζ_j that encodes each annotator's spamming strategy. Otherwise, if $S_{ij} = 0$, the model assumes that the

[1] For example, https://github.com/mainlp/awesome-human-label-variation

annotator simply assigns the correct label—an intentional simplification to focus on modeling spam behavior. Only the annotations A_{ij} are observed; the other parameters are inferred when updating the model from data.

Usually, when applying MACE for label aggregation, the model would weigh all annotations to estimate the correct labels without discarding specific annotators. But the learned parameters can also be used to identify spamming annotators: the competence θ correlates more strongly than agreement measures with an annotator's fraction of correctly annotated examples (Hovy et al., 2013) and both learned annotator parameters (θ, ζ) were shown to encode characteristic spamming behaviors (Paun et al., 2018). Consequently, other studies have used MACE to exclude spammers during dataset construction based on an empirically chosen threshold for competence (Pei and Jurgens, 2023). In our experiments, we also use the competence parameter to score annotators.

4.2 CrowdTruth

The CrowdTruth framework (Aroyo and Welty, 2014) computes several interdependent quality metrics that use vector representations of annotations to measure disagreement and ambiguity, including a worker quality score. The metrics follow the aim of ambiguity-aware label aggregation, so that, for example, disagreement on ambiguous instances discounts worker quality less. The worker quality score (WQS) for an annotator i is computed as the product of two other scores $WQS(i) = WUA(i) \cdot WWA(i)$, the worker-unit agreement (WUA) and the worker-worker agreement (WWA). Conceptually, WWA measures how similar a given worker's annotations are to other workers, weighted by the workers' quality and the instances' ambiguity. WUA measures how much a worker agrees with the aggregate label over all their annotated instances, weighted by the instances' ambiguity. (See Appendix A for details on how these metrics are computed.)

The CrowdTruth metrics were explored on various tasks (Dumitrache et al., 2017, 2018a) and have been explicitly used for spammer removal (Dumitrache et al., 2021). In a related study, Soberón et al. (2013) report an accuracy of 0.88 for removing spam annotators using CrowdTruth metrics. In our experiments, we use the worker quality score (WQS) to score annotators.

4.3 Cohen's Kappa

As a representative example of using inter-annotator agreement metrics to filter annotators, we compute each annotator's pair-wise agreement as measured by Cohen's kappa (Cohen, 1968) with each other annotator. We then use the averaged agreement to score annotators.

4.4 Random Baseline

We assign scores to annotators (from 0.0 to 1.0) by drawing from a uniform distribution.

5 Results

We applied MACE, Crowdtruth, the Cohen's kappa filter, and a random baseline on both datasets, as the threshold for number of annotators removed increases. For studies in spammer detection and gold label aggregation (see Section 2) the primary metric to optimize is downstream classification performance, often based on synthetic spam annotations, whereas we focus on tasks where preserving labeling variation is key. We measured the change in standard deviation, entropy, and accuracy of spammer detection for the DICES-350 and Mturk datasets, as well as the KL-divergence and mean absolute error of the filtered labels from the labels of non-spammers.

5.1 Accuracy vs. Preserved Variation for Spam Detection

Across methods, increasing the number of removed annotators gradually decreases the accuracy of classifying annotators as spammers (Figure 1, top). For the MTurk dataset, the accuracy of spam classification never rises above the accuracy of not removing any annotators. For the DICES dataset, Cohen's kappa and MACE outperform removing zero annotators when <10% of annotators are removed, while CrowdTruth and random removal quickly fall below baseline accuracy (Figure 1, bottom). The best accuracy is achieved when only focusing on the lowest-scoring annotators (lowest 2-4%).

We also measured the change in entropy and standard deviation of the filtered dataset, finding that these methods typically reduce variance in the distribution of annotator opinions, discarding information about annotator disagreement (Figures 2 and 3). Except for the random baseline, the tested methods generally decrease the entropy of the distributions as more raters are removed. This is especially true of CrowdTruth, which quickly decreases

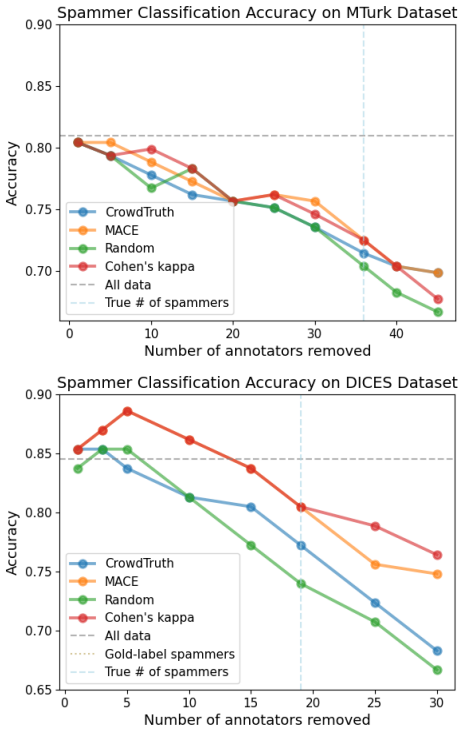

Figure 1: Across methods, increasing the number of removed annotators gradually decreases the accuracy of spam classification when over 2-4% of annotators are removed. Cohen's kappa and MACE increase the spam classification accuracy up to 4% of annotators removed on DICES; otherwise, the spam classification accuracy rarely rises above the baseline of not removing any annotators. The blue line indicates the true number of spammers in the data; the gray line indicates the baseline classification accuracy before removing any spammers.

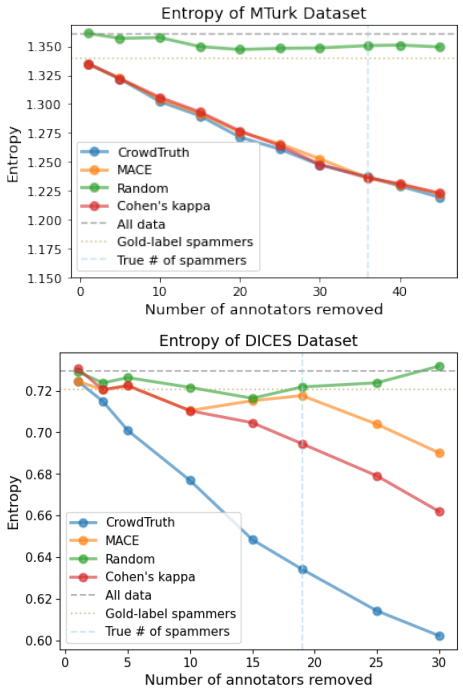

Figure 2: Entropy of each instance's label distribution, averaged over all instances. Most methods decrease the entropy of the dataset as more raters are removed. CrowdTruth especially decreases the entropy.

the entropy; MACE and Cohen's kappa also decrease the entropy to a lesser extent. CrowdTruth also consistently decreases the standard deviation of the data. MACE and Cohen's kappa decrease the standard deviation on the MTurk dataset, but not on DICES.

To understand whether these methods affect how well the filtered datasets represent the true distribution of non-spam annotators' ratings, we also measured the mean absolute error (MAE) per example between the filtered annotators and the true non-spam annotators (i.e., the difference between their average labels on a given example; Figure 5) and the KL-divergence between the filtered and non-spam annotators (Figure 6). All tested methods eventually increase the mean absolute error, indicating that the mean label of the filtered data drifts away from that of the true non-spam annotators as more labels are removed. However, the

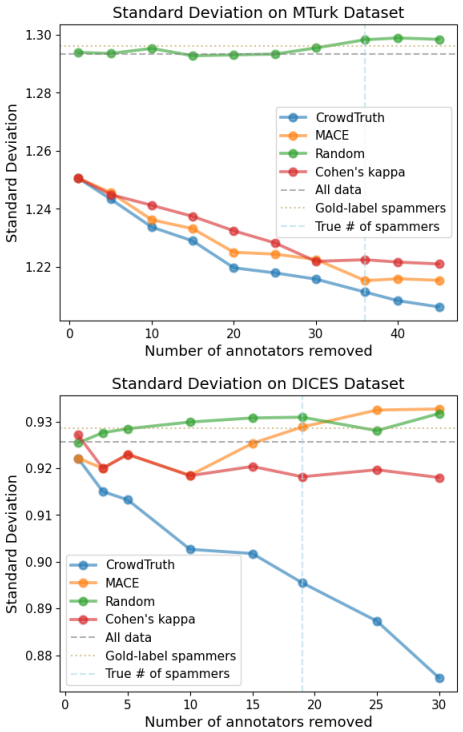

Figure 3: On the MTurk dataset, all methods except random removal decrease the standard deviation of the dataset. Among the tested methods, CrowdTruth decreases the standard deviation most.

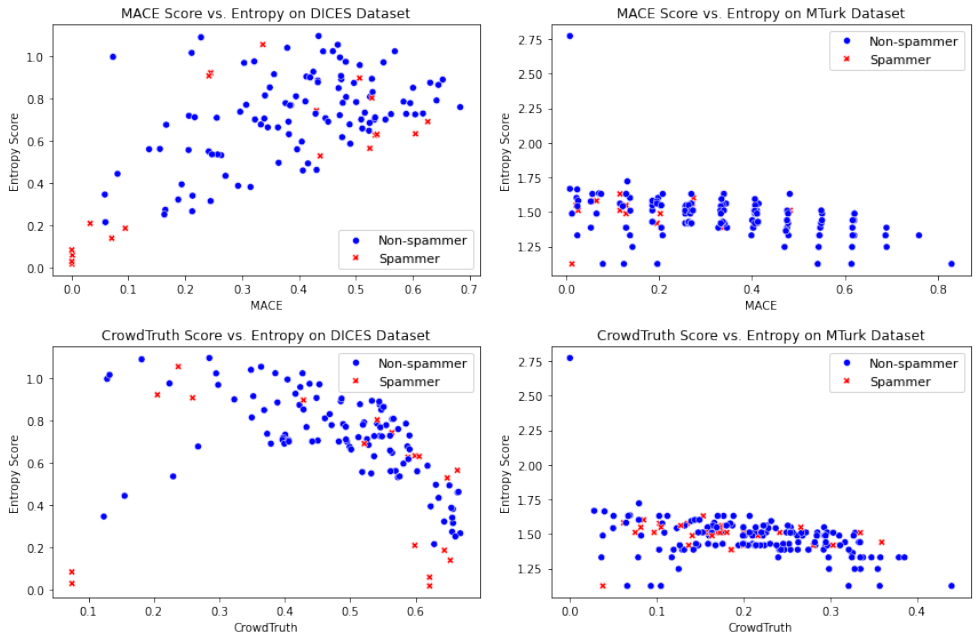

Figure 4: Entropy of each annotator's labeling distribution over all instances vs. score under filtering metrics (CrowdTruth and MACE). While many spam annotators are indistinguishable from non-spam ones under these metrics, those that are often have very low entropy: they are *less* random than non-spam annotators, not more.

extent of this varies by method and dataset: on the MTurk dataset, all non-random methods have relatively little change in MAE when <5% of annotators are removed, but increases after that; on the DICES dataset, CrowdTruth worsens the MAE much faster than other tested methods. The KL divergence remains relatively steady, but eventually increases on the MTurk dataset for all non-random methods, and fluctuates widely across methods on the DICES dataset.

Why might these methods fail to capture all spammers? Comparing the entropy of the responses given by each annotator with their scores under these metrics helps to understand where the assumed spammer behavior, as modeled by these metrics, differs from the spammer behavior seen in practice (Figure 4). Most annotators lie well within the distribution of non-spammers in terms of entropy, MACE score, and CrowdTruth score. However, a subset of annotators are distinguishable as spammers (best seen on the DICES dataset) because they have especially low entropy. MACE captures many of these annotators, but CrowdTruth only captures some of them, perhaps explaining the difference in these metrics.

Since a cluster of spam annotators that can be reliably distinguished tends to have especially fixed behavior, perhaps models perform best at capturing spam if they can identify annotators with unusually fixed annotation patterns. To investigate this, we next studied model performance using synthetic spam.

5.2 Synthetic Spam Analysis

To understand what factors affect spam detection methods' accuracy at classifying spam, and propensity to misclassify annotators who disagree as spam, we experiment with several kinds of synthetic data. *Random spam* experiments simulate spam annotators whose annotations are random; *fixed spam* experiments simulate spam annotators who always give the same answer, which is set to the mode response for the dataset.

Fixed spam. Because these methods tend to filter out annotators who are farther from the mean, most of them struggle to filter out annotators whose behavior is fixed to the mode value (e.g., answering "No" to every question). MACE performs much better than the other methods on fixed spam for DICES, but all methods are worse than the baseline for the fixed spammers on the MTurk data (Figure 7). MACE's higher accuracy on DICES can partially be explained by how well the method can capture fixed spamming behavior given how it is set up (see Section 4.1): A spammer would have low competence θ, so that the assigned label is frequently sampled from the annotator's spamming strategy ζ. As the spammer assigns always the same label,

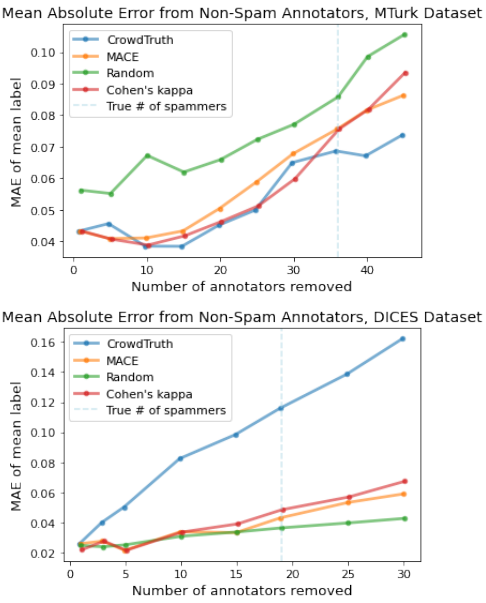

Figure 5: Mean absolute error of filtered ratings. Difference between average label on an example of non-spam annotators and filtered annotators, then averaged across examples.

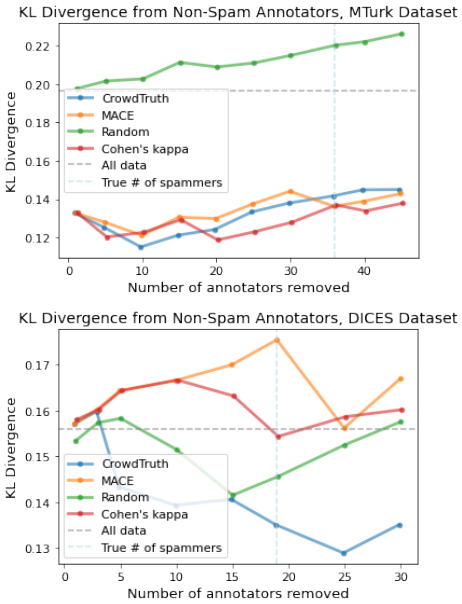

Figure 6: KL-divergence of filtered annotators per data item vs. true non-spam annotators, averaged across examples.

the parameter vector would encode high probability for that particular label and low probability for all others. In contrast, CrowdTruth factors in ambiguity but is ultimately based on agreement (see Section 4.2). As a spammer who always assigns the mode label can score relatively high agreement in subjective tasks with stronger labeling variability, fixed spam annotators are not filtered out by CrowdTruth. This result about agreement for fixed spam is in line with the accuracy scores by Cohen's kappa filtering, which are identical to CrowdTruth. Notably, this observation does not transfer to real spammer behavior (Figure 1), where CrowdTruth is often more accurate than Cohen's kappa.

MACE's poor accuracy on MTurk is surprising given its perfect accuracy on DICES. This result is likely caused by answers in the MTurk dataset mostly following a normal distribution with the same mode, so that MACE overestimates the competence of fixed spammers (in contrast to DICES; see Appendix C).

Because the spammers all give the same ratings, we expect accurate spam classification to increase the standard deviation and the entropy, as happens for MACE on DICES; by contrast, Crowdtruth and Cohen's kappa filtering on DICES (and MACE on MTurk) decrease the standard deviation and the entropy without ever increasing spam classification accuracy above the baseline (Appendix B).

Random spam. On the random data (Figure 7, right), CrowdTruth, MACE, and Cohen's kappa have similar accuracies (peaking when the number of annotators removed equals the number of spam annotators). This suggests that random spam is closest to the spam behavior for which these methods work optimally.

In this case, we expect accurate spam classification to decrease the entropy, which indeed happens for both datasets across methods (Appendix B); the standard deviation also decreases for MTurk, and is more random for DICES, likely because DICES has a smaller set of possible answer values.

Together, these results suggest that real spam annotators are less random than the imagined spammer behavior under CrowdTruth and inter-annotator agreement filtering. This makes these methods vulnerable to removing annotators who are further from the mean rather than actual spammers. MACE, which is more robust to filtering out fixed spammers, also performs better at filtering out real spammers.

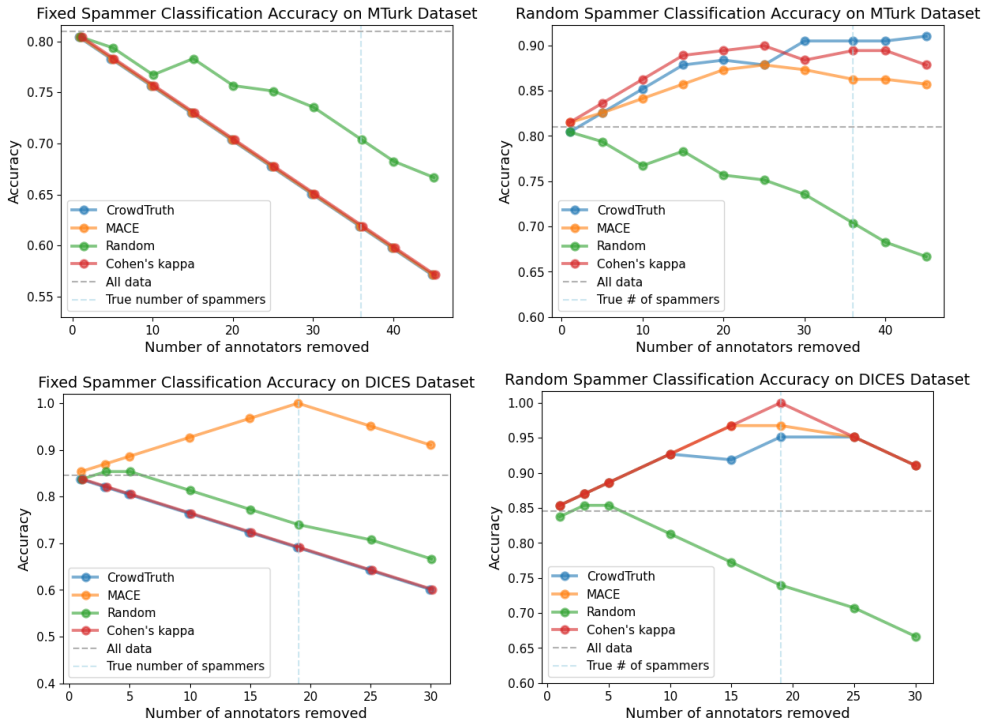

Figure 7: Accuracy with fixed-spam and random-spam synthetic annotators. For DICES, MACE performs best on fixed data; the other methods universally struggle. For random spam, all methods outperform the baseline, with Cohen's kappa performing optimally on DICES.

6 Discussion and Conclusion

Spam detection for subjective problems flips model assumptions: spam annotators are often *less* random than non-spam ones. Longstanding paradigms of annotation, focused on improving downstream model accuracy under the assumption of a single ground truth, often assume that disagreement indicates low-quality annotations. However, in problems where disagreement is expected, and preserving this variation *is* the goal, this intuition is flipped. We find that many spam annotators are indistinguishable from non-spam annotators, and those that are identifiable are in fact those with very low entropy. Examining the performance of tested methods on completely random vs. completely fixed spam reveals that many methods struggle to identify fixed spam. In particular, as a fixed mode response results in relatively high agreement in datasets with substantial variation. These models also struggle on real-world spam in our tested datasets, suggesting that, where preserving variation is paramount, models assuming that spam annotators are more random are not as well suited.

Existing methods work best only when removing few annotators, and distort distributions afterwards. Tested methods (particularly MACE) are effective at identifying spam annotators for low n (<2-4% of tested annotators). When more annotators are removed, we see issues across a range of metrics: increased mean absolute error; lower accuracy at spam detection, lower standard deviation, and lower entropy. These issues mean that over-filtering data can lead to labels that do not fully represent the variation in the original distribution.

Detecting spammers vs. detecting low-quality raters. Since different types of low-quality annotators behave differently, annotators that need to be excluded can exhibit varied behaviors beyond simple patterns such as always selecting the same answer. Consequently, while annotator reliability scoring can often single out spammers showing these stereotypical behaviors, many genuine annotators will be scored similarly to low-quality raters. This result highlights that in addition to the labeling behavior, additional signals should be included in spammer removal. These can be *metadata*, such as when and how much time is spent on annotation (Rothwell et al., 2015) or previous acceptance rates of annotators (Difallah et al., 2012). Similarly, *verifiable test questions* could be used, that is, unambiguous cases where comparison to known answers is possible (gold standard or attention checks,

Difallah et al., 2012; Rothwell et al., 2015).

Future work. Existing methods struggle to distinguish spam from non-spam annotators in contexts where variation in opinion is expected and desirable. This gap highlights the need for spam filtering methods that are robust to variation in labeling behavior.

In addition, the scarcity of available metadata on removed spam data makes it difficult to characterize spammer behavior across a range of contexts. Difallah et al. (2012) highlight a "need for new benchmarks on which to evaluate and compare existing and novel spam detection techniques for crowdsourcing platforms" that still persists. Datasets often do not report spam filtering techniques or preserve the spam responses; however, this data is extremely helpful for more fine-grained characterization of spam behavior, especially in complex contexts where variation is expected. Thus, making this data available would be a valuable resource for future research.

Limitations

Due to data scarcity, we only used a narrow range of datasets. While the used datasets represent two important use cases where capturing variation matters (AI safety annotations, survey questions), more datasets are needed, especially with different levels of subjectivity, languages and use cases. As such our results represent only a fraction of relevant scenarios.

Categorizing raters as "spammers" is based on varying definitions and procedures. So "gold spammers" are not ground truth the same way that other data might be. Importantly, self-reported spammer information, where spammers disclose themselves, is largely not even gathered (see for an exception, Paun et al., 2018) and not publicly available. Consequently, the "gold spammer" labels used in our study are based on external categorizations. While these are reported to be based on manual checks and multiple data types (labeling behavior, metadata, and attention checks), there remains a risk of wrong categorizations.

We scoped to spam filtering methods that only look at the labeling behavior, given our research question on how this (widely adopted) type of filtering changes the captured variation in labeling. However, there are approaches based on metadata that we could expect to be more effective, perhaps in combination with the evaluated methods using intrinsic metrics based on labeling behavior.

Acknowledgments

Matthias Orlikowski and Philipp Cimiano were funded by Volkswagen Foundation as part of the "Bots Building Bridges (3B)" project in the "Artificial Intelligence and the Society of the Future" programme. Eve Fleisig is partly supported by an NSF GRFP grant.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Ali Ak, Mona Abid, Matthieu Perreira da Silva, and Patrick Le Callet. 2021. On Spammer Detection in Crowdsourcing Pairwise Comparison Tasks: Case Study on Two Multimedia QoE Assessment Scenarios. In *ICME 2021 - First International Workshop on Quality of Experience in Interactive Multimedia*, Virtual, China.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Human Computation*, 1(1). Number: 1.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24. Number: 1.

Yang Ba, Michelle V. Mancenido, Erin K. Chiou, and Rong Pan. 2024. Data Quality in Crowdsourcing and Spamming Behavior Detection. *arXiv preprint*. ArXiv:2404.17582 [cs].

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Sabine Buchholz and Javier Latorre. 2011. Crowdsourcing Preference Tests, and How to Detect Cheating. In *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, Florence, Italy, August 27-31, 2011*, pages 3053–3056. ISCA.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2016. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2039–2046, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28. Publisher: Royal Statistical Society, Oxford University Press.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, volume 842 of *CEUR Workshop Proceedings*, pages 26–30. CEUR-WS.org.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. False Positive and Cross-relation Signals in Distant Supervision Data. In *6th Workshop on Automated Knowledge Base Construction (AKBC)*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018a. Crowdsourcing Semantic Label Propagation in Relation Classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium. Association for Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018b. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *Preprint*, arXiv:1808.06080.

Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. 2021. Empirical methodology for crowdsourcing ground truth. *Semantic Web*, 12(3):403–421. Publisher: SAGE Publications.

Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2):121–137.

Şeyda Ertekin, Cynthia Rudin, and Haym Hirsh. 2014. Approximating the crowd. *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1631–1640, New York, NY, USA. Association for Computing Machinery.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. Association for Computing Machinery.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator

attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

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.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2025. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200, Vienna, Austria. Association for Computational Linguistics.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are Large Language Models Reliable Argument Quality Annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Spencer Rothwell, Ahmad Elshenawy, Steele Carter, Daniela Braga, Faraz Romani, Michael Kennewick, and Bob Kennewick. 2015. Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proceedings of Interspeech 2015*, pages 2784–2788. ISCA.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy.

2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring crowd truth: disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, pages 45–58, Aachen, DEU. CEUR-WS.org.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Panagiotis A. Traganitis and Georgios B. Giannakis. 2021. Identifying Spammers to Boost Crowdsourced Classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2840–2844. ISSN: 2379-190X.

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.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2025. HelpSteer2: open-source dataset for training top-performing reward models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NeurIPS '24*, pages 1474–1501, Red Hook, NY, USA. Curran Associates Inc.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2022. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 2035–2043, Red Hook, NY, USA. Curran Associates Inc.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA. Association for Computational Linguistics.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

A Computing the CrowdTruth Worker Quality Score

As highlighted in Section 4.2, WWA measures how similar a given worker's annotations are to other workers, weighted by the workers' quality and the instances' (or units') ambiguity. WUA measures how much a worker agrees with the aggregate label over all their annotated instances, weighted by the instances' ambiguity. WWA and WUA are roughly computed as follows (ignoring the normalization terms for clarity, full details in Dumitrache et al., 2018b):

$$WWA(i) = \sum_{j,u} sim(i,j,u) \cdot WQS(j) \cdot UQS(u)$$

$$WUA(i) = \sum_{u \in units(i)} sim(i,u) \cdot UQS(u)$$

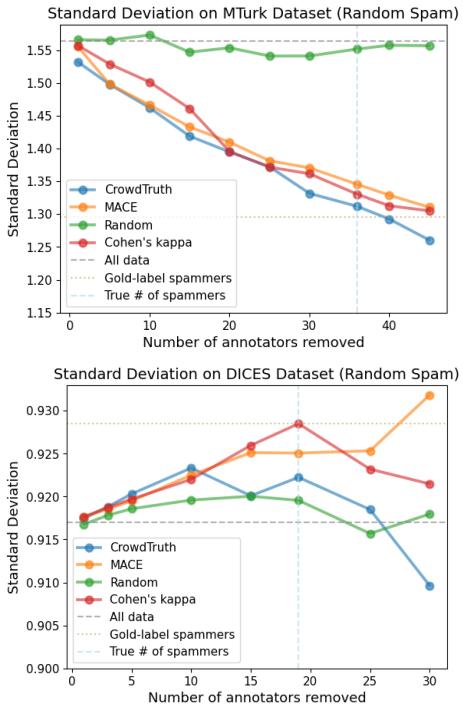

Figure 8: Standard deviation on random spammers.

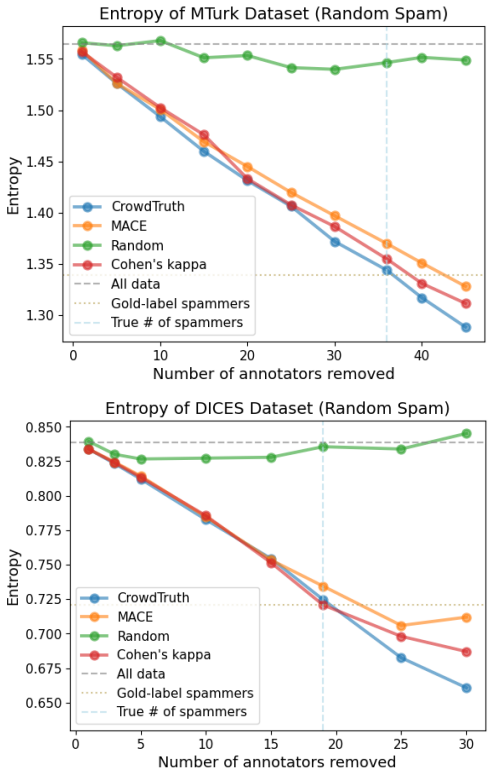

Figure 9: Entropy on random spammers. Entropy generally decreases as more spammers are removed, as expected for accurate spam classification.

Here, $sim(i, j, u)$ is the cosine similarity between the annotation vectors of workers i and j on an instance u. Similarly, $sim(i, u)$ is the cosine similarity between the annotation vector by worker i and the instance vector for instance u (i.e., summed annotation vectors of all other annotators). It is computed over all instances annotated by annotator i, denoted $units(i)$. Additionally, $UQS(u)$ measures how much workers agree on an instance u (how ambiguous it is) and is also connected to the workers' quality scores. Due to their inter-dependent nature, the CrowdTruth metrics are re-calculated iteratively until convergence.

B Details of Synthetic Spam Results

Standard deviation and entropy for the random and fixed spammers are shown in Figure 8, Figure 9, Figure **??**, and Figure **??**.

C Why does MACE fail to recognize fixed spammers on the MTurk dataset?

On fixed spammers, who always respond with the mode (the most frequent label in each dataset), MACE gets perfect accuracy on DICES, while on MTurk it performs as poorly as all other methods, failing to reach baseline performance (see Section 5.2). This result is likely due to the peculiarities of the survey data in the MTurk dataset, where answers follow a normal distribution and the mode is the same for most questions. Here, fixed spammers' seem competent because they always respond with the ground truth as estimated by MACE. Because of this perfect answering behavior of spammers, their average difference to the estimated ground truth is zero, as shown in Figure 10, so that naturally non-spammers are further away from the estimated ground truth, looking less competent to MACE. In contrast, on DICES, which has more varied examples of labeling behavior, non-spammers are on average closer to the estimated ground truth than spammers (see Figure 10).

D Dataset Selection Table

A total of 22 datasets were considered to be included in our study, mostly informed by related work. Table 1 lists all of these datasets, including the corresponding references. The table details for each dataset if gold spammer data was collected in principle and if that data was still available. As described in Section 3, we were only able to include two out of these 22 datasets in our experiments.

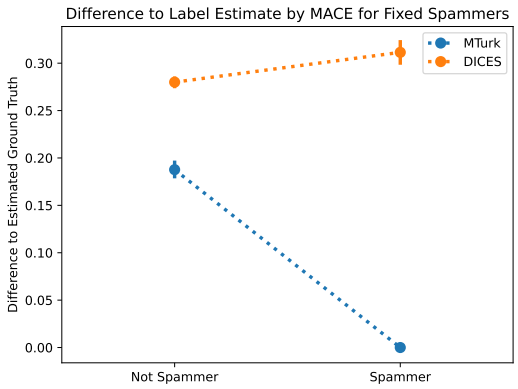

Figure 10: Distance to the ground truth estimated by MACE on fixed spammers vs non-spammers (lower is better). Shows the averaged absolute difference between annotations and the estimated ground truth label. Before averaging, distances are normalized using min-max normalization for each dataset, scaling distances into the range of zero to one.

Dataset	Reference	Gold spammers?	Included?	If excluded, why?
DICES	Aroyo et al. 2023	Yes	Yes	
MTurk Survey	Huang et al. 2023	Yes	Yes	
MHS corpus	Sachdeva et al. 2022	Yes	No	Raters excluded (details in their paper), but data not available.
AdultContent3 (”Get Another Label“ datasets)	Ipeirotis et al. 2010	No	No	No gold spammers. Experiments in paper use synthetic data and simply report impact on the collected dataset
HITspam	Discussed in Ertekin et al. 2014	No	No	Despite the name, does not contain spammers. Instead, the task is to judge whether a task on MTurk itself should be considered spam (e.g., because it asks workers to follow a specific social media account).
EDOS-DOM	Jiang et al. 2024	Yes	No	Only one annotator removed after labels were collected. That annotator had annotated only 8 examples (first author vial email).
Argument Quality	Mirzakhmedova et al. 2024	No	No	Excludes a number of disagreeing annotations per example. Does not exclude on the level of the annotator.
MultiPref	Miranda et al. 2025	No	No	No gold spammers.
HelpSteer2	Wang et al. 2025	No	No	No gold spammers.
CrowdTruth Corpus for Open Domain Relation Extraction	Dumitrache et al. 2017	No	No	Emailed first author, full data not available anymore.
AMR / Sentence Similarity Data	Wein and Schneider 2022	Yes	No	Only one annotator removed out of three in total.
Phrase Detectives	Chamberlain et al. 2016	Yes, self-reported	No	Spammer data not available anymore according to authors.
Crowd-Sourced Preference Tests	Buchholz and Latorre 2011	Yes, inferred	No	Data not available anymore according to first author.
VariErr NLI	Weber-Genzel et al. 2024	No	No	Data has annotator IDs and individual decisions plus error judgments (error = no self-validations), but no excluded raters. Not a crowd-sourced study (four annotators).
Malicious Worker Survey Dataset	Gadiraju et al. 2015	Yes	No	Dataset is not available anymore.
Dog (Imagenet Subset)	Deng et al. 2009	No	No	No spammer information. Used by Traganitis and Giannakis (2021), but only evaluated by accuracy of resulting classifier.
ImageNetV2	Recht et al. 2019	No	No	No spammer information.

Bluebird	Welinder et al. 2010	No	No	No spammer information. Used by Traganitis and Giannakis (2021), but only evaluated by accuracy of resulting classifier.
Web	Zhou et al. 2012	Unlikely	No	Could not find reference to data.
WSD	Snow et al. 2008	Unlikely	No	Data not available anymore
RTE	Snow et al. 2008	Unlikely	No	Data not available anymore
TEMP	Snow et al. 2008	Unlikely	No	Data not available anymore
POPQUORN	Pei and Jurgens 2023	Yes	No	Only a single annotator was removed.

Table 1: Dataset Selection. Shows which datasets where considered and why 20 out of 22 datasets were not included in our study.