2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Noise Audits Improve Moral Foundation Classification*

Negar Mokhberian[♥] Frederic R. Hopp[♦] Bahareh Harandizadeh[♥] Fred Morstatter[♥] Kristina Lerman[♥]

[♥]*Information Sciences Institute,*
University of Southern California
{nmokhber, harandiz, fredmors, lerman}@isi.edu

[♦]*Amsterdam School of Communication Research,*
University of Amsterdam
f.r.hopp@uva.nl

***Abstract*—Morality plays an important role in culture, identity, and emotion. Recent advances in natural language processing have shown that it is possible to classify moral values expressed in text at scale. Morality classification relies on human annotators to label the moral expressions in text, which provides training data to achieve state-of-the-art performance. However, these annotations are inherently subjective and some of the instances are hard to classify, resulting in noisy annotations due to error or lack of agreement. The presence of noise in training data harms the classifier's ability to accurately recognize moral foundations from text. We propose two metrics to audit the noise of annotations. The first metric is *entropy* of instance labels, which is a proxy measure of annotator disagreement about how the instance should be labeled. The second metric is the *silhouette* coefficient of a label assigned by an annotator to an instance. This metric leverages the idea that instances with the same label should have similar latent representations, and deviations from collective judgments are indicative of errors. Our experiments on three widely used moral foundations datasets show that removing noisy annotations based on the proposed metrics improves classification performance.**[1]**

***Index Terms*—crowd-sourcing, annotation, ambiguity, subjective annotations, noisy annotations**

I. Introduction

Moral foundations theory (MFT) [1], [2] suggests that the moral values expressed in opinions, thoughts, and cultures can be explained by five universal, but contextually variable *moral foundations*. These foundations are typically described along bipolar dimensions: care vs. harm, fairness vs. cheating, authority vs. subversion, loyalty vs. betrayal, and purity vs. degradation. MFT was first introduced in social psychology and has found many applications in political science and the social sciences. For example, moral foundations motivate behaviors such as charitable donations [3], violent protests [4] and social homophily [5].

The broad adoption of MFT was driven, in part, by advances in natural language processing (NLP), which enabled researchers to quantify moral values expressed in text, including news [6]–[8], political speech [9], and social media discussions [10], at scale. Early works relied on lexicons that defined words associated with moral virtues and vices to classify moral foundations from text [1]. However, by neglecting semantic context in sentences, lexicon-based approaches fail to capture the nuances of moral expression [11]. To address this challenge, more recent approaches use large language models to capture the moral context of text [12], [13]. These approaches leverage a text corpus manually annotated for moral values to train language models to recognize examples of moral language. Several such ground truth data sets exist [7], [8], [10], [14], [15].

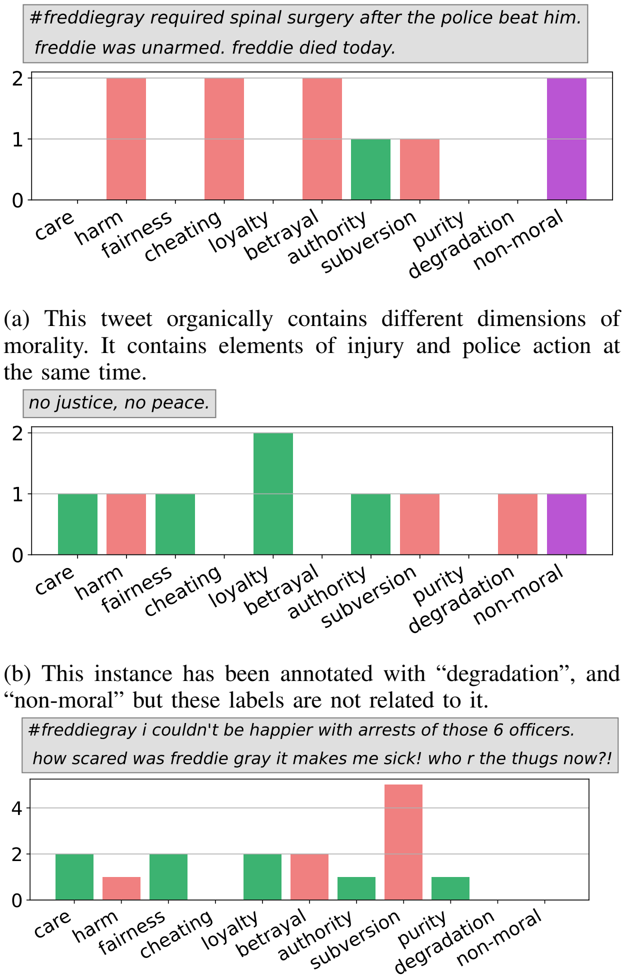
A major challenge when constructing ground truth data for training moral foundation classifiers is the subjectivity of individual moral judgments, which are prone to bias and noise [16]. Figure 1 illustrates this challenge using real examples from the Moral Foundation Twitter Corpus (MFTC) [10]. The first tweet, "#freddiegray required spinal surgery after the police beat him. freddie was unarmed. freddie died today," expresses a range of moral concerns (Fig. 1a). It was labeled as *harm*, *cheating*, *betrayal* and *non-moral* by two annotators each, and *authority* and *subversion* by one annotator each. Text can also be mislabeled due to errors or ambiguity. This is illustrated by the second example "no justice, no peace." (Fig. 1b). This instance was annotated with *degradation*, and *non-moral*, but these labels are not related to the tweet. The third example illustrates the subjectivity of moral judgments (Fig. 1c): "#freddiegray I couldn't be happier with arrests of those 6 officers. how scared was freddie gray it makes me sick! who r the thugs now?!" The moral labels assigned to this instance are subjective and depend on whether the annotators support police actions or not. As a result of these factors, individual moral labels will be noisy. To partly control for the variability of judgments, researchers use the label chosen by the majority of annotators as the correct ground truth label for each instance [10], [17]. However, the question of how annotation noise affects the performance of models learned from data has been underexplored [18], [19].

In this paper we present and compare two approaches for auditing and removing noise from annotations used to train moral foundations classifiers. The first approach identifies difficult instances in the ground truth data. We propose *entropy*

*Preparation for this manuscript was supported, in part, by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002

[1] Our code can be found at: https://github.com/negar-mokhberian/noise-audits

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE



(a) This tweet organically contains different dimensions of morality. It contains elements of injury and police action at the same time.

(b) This instance has been annotated with “degradation”, and “non-moral” but these labels are not related to it.

(c) This instance’s moral labels are subjective, depending on whether the annotators support the police actions or are against it.

Fig. 1: Examples of high-entropy instances in MFTC dataset. The text of the tweet and the number of annotators selecting each moral foundation as the label are shown.

as a measure of annotator disagreement in §III-A and use it to identify instances with little agreement that will degrade classifier training. For each instance, we calculate *entropy* based on how many annotators have selected each of the labels (examples of such distributions are shown in Fig. 1 for three instances). By removing instances on which the annotators disagree, we hope to create better data for training classifiers.

Our second method identifies annotations that deviate from collective judgments of all annotators. We propose the *silhouette* coefficient in latent space as a measure of label quality, which leverages the idea that similarly-labeled instances should have similar latent representations. By removing annotations that deviate substantially from collective judgments, we hope to improve the quality of ground truth data.

We evaluate both approaches on three large datasets with moral annotations [7], [10], [14] (more information of the datasets in §IV-A). We show that training models on ground truth data from which noisiest annotations have been removed improves morality classification. This does not stem simply from having less data: compared to a model trained on data from which the same number of instances were removed at random, removing the noisiest instances from data significantly improves classification performance.

Our work primarily focuses on noise audit of subjective annotations to identify mislabeled or difficult instances in moral foundations classification. This could be applied in any subjective labeling setting. Identifying moral foundations in text is a representative example of a subjective task that we have selected for this paper. An important enabler for our noise audit is access to the individual judgments of annotators. Hence, we encourage the crowd-sourced annotation builders to include fine-grained details of individual annotator judgments on instances instead of the common practice [20] of reporting an aggregated judgment (e.g., majority label). Including the individual judgments gives the opportunity to refine the dataset, enhancing its utility on learning tasks.

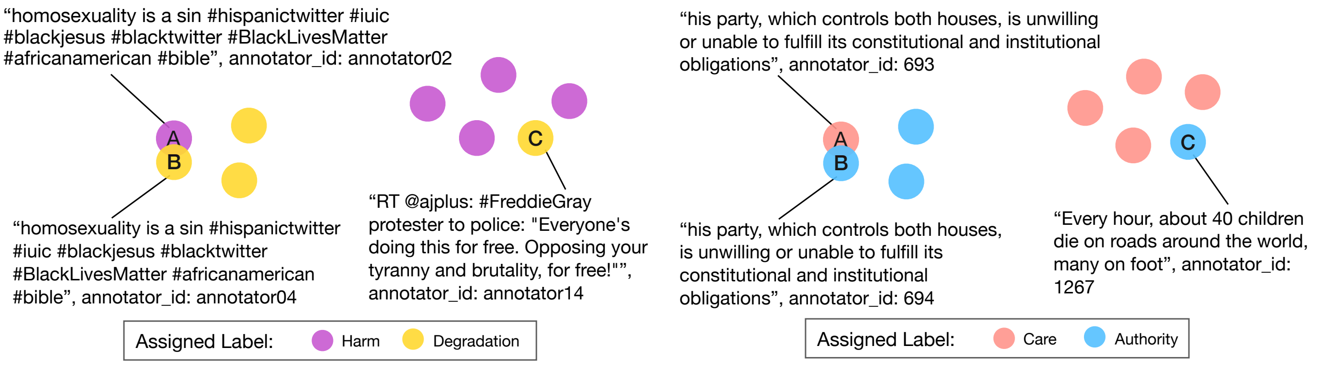
II. Related Work

Researchers have tried to incorporate individual annotator’s perspectives of the subjective tasks [18]. Using multi-annotator models, they have shown that training a separate model for each annotator and then aggregating to a majority vote performs better than aggregating labels in the data prior to training. However, their methodology is not practical in the cases where the data have been crowd-sourced and there are many annotators because 1) there are not usually enough data points per annotator to train separate modules, and 2) it is not cost-efficient to train many separate modules.

Other earlier works have studied the difficulty of data points through information theory [21] and disagreements in annotations [22]. A leading way to understand the difficulty of data instances is to leverage training dynamics [23], [24]. In other words, by observing how a classifier performs on an specific instance through epochs. Perhaps the best exemplar of this approach is seen in [19], which assesses each point based on the model’s *confidence* (average probability assigned to the correct label during training epochs) and *variance* (variance in probability assigned to the correct label during training epochs). While training dynamics offer a way to measure training difficulty on an instance, their main drawback is dependency on the design of the model and on training. However, we try to address this issue by proposing metrics that are only depended on the annotated dataset and training is not a requirement for their calculation. We apply our proposed metrics to refine moral foundations datasets before training. To the best of our knowledge no prior work has analysed how difficult instances of moral foundations affect classification.

III. Methodology

In this section we propose two metrics ***entropy*** and ***silhouette*** for identifying noise in instance-level and judgment-level respectively. We use *entropy* (see §III-A) to find out


"homosexuality is a sin #hispanictwitter #iuic #blackjesus #blacktwitter #BlackLivesMatter #africanamerican #bible", annotator_id: annotator02
A
B
C
"homosexuality is a sin #hispanictwitter #iuic #blackjesus #blacktwitter #BlackLivesMatter #africanamerican #bible", annotator_id: annotator04
"RT @ajplus: #FreddieGray protester to police: "Everyone's doing this for free. Opposing your tyranny and brutality, for free!"", annotator_id: annotator14
Assigned Label: ● Harm ● Degradation
"his party, which controls both houses, is unwilling or unable to fulfill its constitutional and institutional obligations", annotator_id: 693
A
B
C
"his party, which controls both houses, is unwilling or unable to fulfill its constitutional and institutional obligations", annotator_id: 694
"Every hour, about 40 children die on roads around the world, many on foot", annotator_id: 1267
Assigned Label: ● Care ● Authority

(a) MFTC dataset

(b) MFNC dataset

Fig. 2: Examples of moral judgments with low silhouette coefficients. Judgments A, B are on the same text but from different annotators. In the language model latent space this text is very close to other judgments of *Degradation* in the left figure and to *Authority* in the right figure. In judgment A the annotator has assigned it a label that does not match the labels of similar texts. Because it is far from other texts with the same label, but close to texts that have been assigned different labels, judgment A will get a low silhouette coefficient. Judgment B is close to the other judgments of the same label so it will get high silhouette coefficient. Our methodology suggests removing judgment A from the training data but to keep judgment B on the same text. In judgment C the annotator has also selected a label that is different from the label of the other similar texts, so we suggest to remove judgment C.

annotator disagreements on a given instance. An instance is a piece of text that has been shown to annotators to collect their judgments. The disagreement on an instance is high when disparate judgments have been collected from annotators; on the other hand, the disagreement is low when all annotators agree on the same label. We refine the datasets by removing the high-entropy instances (instances with high disagreement on their assigned labels) as a pre-processing step before training a classifier.

In addition, we propose using the *silhouette* coefficient (see §III-B) as a fine-grained metric to quantify how a single judgment differs from other instances of the same-label judgments. Note that unlike entropy, this metric takes into account the text of the instance. Removing judgments we deem noisy with this metric increases the inter-annotator agreement on a given instance.

Unlike prior work in machine learning [19], [23], [25], [26], our measures are calculated before the model training starts and do not depend on the training dynamics. We provide the details of one the training dynamics metrics in §III-C and in the experiments (§IV-D) will monitor their improvement as we filter data based on our metrics.

A. Entropy at the Instance Level

We use *entropy of annotations* to quantify diversity of labels gathered for a text. For a text t_i and its multi-label annotations $< l_1 : c_{i1}, ..., l_N : c_{iN} >$ in which l_j is a member of all the labels $L = \{l_1, l_2, ..., l_N\}$ and c_{ij} is the count of annotators who have assigned l_j to t_i, we calculate the *entropy* as:

$$entropy(t_i) = -\sum_{j=1}^{N} P(l_j, t_i) log P(l_j, t_i)$$

in which

$$P(l_j, t_i) = \frac{c_{ij}}{\sum_{j=1}^{N} c_{ij}}.$$

If all annotators agree on the same label, the *entropy* is zero. At the other extreme, if every annotator gave the instance a different label, entropy has its maximum value.

B. Silhouette Coefficient at the Judgment Level

For a given judgment $x = (t_i,\ l_j,\ a_k)$ in which instance t_i has been assigned a label l_j by annotator a_k, the *silhouette* coefficient [27] is defined as a combination of its distance from the same-label (intra-cluster) judgments and from other-label (inter-cluster) judgments. Let l_j be a member of the set of all labels $L = \{l_1, l_2, ..., l_N\}$. We consider all the instances assigned to the same label to be a cluster. The intra-cluster measure $a(x)$ is defined as the average dissimilarity of $text_i$ to all other texts labeled with l_j. The inter-cluster metric is defined as:

$$b(x) = min_{y \in L, y \neq l_j} \bar{d}(t_i, T_y),$$

where $\bar{d}(t_i, T_y)$ is the average dissimilarity of t_i to all texts in the dataset labeled with y. Finally, $a(x)$ and $b(x)$ are aggregated as:

$$silhouette(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}}$$

To calculate the dissimilarities (d), we represent the text instances in a latent space using a language model and calculate the distance of vectors corresponding to texts.

The *silhouette* coefficient thus captures the consistency of an annotator's label of a specific text with other same-label texts. If the instance is too different from the content of other examples with the same label, it will have a low *silhouette* coefficient, and we consider it to be noise. Training the model on this judgment may confuse it and reduce classification performance. Hence, we suggest removing this judgment from the ground truth corpus before training. However, other high-silhouette judgements for t_i may be preserved and t_i can be part of the filtered dataset using other judgements of it.

C. Model's Training Dynamics

Recent work by [19] uses signals from epochs during training (training dynamics) as a proxy for exploratory data quality estimation. They define confidence of the model on instance i as the average of model's probability of its true label (y_i^*) across epochs:

$$\hat{\mu}_i = \frac{1}{E} \sum_{e=1}^{E} p_{\theta^{(e)}}(y_i^* | t_i)$$

where $p_{\theta^{(e)}}$ indicates to the model's probability with parameters $\theta^{(e)}$ at the end of the epoch e. Their experiments show that the examples with low confidence are likely to be mislabeled.

IV. Experiments

A. Data

In this paper we focus on three large moral foundations annotated textual datasets. In these datasets the annotations of many texts are very diverse. This diversity can appear because of the difficulty of the labeling task, subjectivity of moral judgments, the bias based on personal beliefs, ambiguity of texts, and errors made by annotators.

1) The Moral Foundations Twitter Corpus (MFTC) [10]: ***The Moral Foundations Twitter Corpus (MFTC)*** is a textual multi-label dataset containing tens of thousands of tweets related to various social movements, with each tweet annotated with categories of moral foundations. For annotating MFTC, several human annotators were trained to manually annotate 35k tweets. The tweets are drawn from 7 socially relevant topics: All Lives Matter (ALM), Black Lives Matter (BLM), the Baltimore protests, the 2016 Presidential election, hate speech and offensive language, Hurricane Sandy, and #MeToo. For each tweet several annotators selected as many moral foundations as they saw relevant, which resulted in a multi-label dataset. Note that for each tweet, the number of annotators that selected each class is known. There are in total eleven labels (ten moral foundations and one "non-moral" class indicating the text does not have moral relevance).

2) The Moral Foundations News Corpus (MFNC) [8]: In the MFNC [8], a crowd of 854 annotators was drawn from the general United States population using the crowd-sourcing platform Prolific Academic (PA[2]). Sampling was designed to match annotator characteristics to the US general population in terms of political affiliation and gender, thereby lowering the likelihood of obtaining annotations that reflect the moral intuitions of only a small, homogeneous group (see supplemental materials in [7] for detailed information on annotators) . Fifteen randomly selected news documents (from among 2,995 articles total) were assigned to each annotator. The selected corpus consisted of online newspaper articles discussing a wide range of sociopolitical topics from 11 prominent, U.S. news outlets. 557 annotators completed all assigned tasks. Each annotator underwent an online training explaining the purpose of the study, the basic tenets of MFT, and the annotation procedure. Annotators were instructed that they would be annotating news articles, and that for each article they would be (randomly) assigned *one* of the five moral foundations. Next, using a digital highlighting tool, annotators were instructed to highlight all portions of a news article that they understood to reflect their assigned moral foundation. In total, 63,958 annotations (i.e., textual highlights) were produced by the 557 annotators. Note that the coding task of the MFNC differed from the MFTC task in important ways: First, annotators were assigned to focus on the presence of *one* (randomly assigned) moral foundation per article, rather than assigning portions of the article to *any* moral foundation. Second, annotators labeled the holistic presence of a moral foundation rather than differentiating whether a foundation was upheld (e.g. care) or violated (e.g. harm). Third, annotators were free to highlight *portions* of a news article, in contrast to labeling the entire coding unit with a moral foundation.

3) The Moral Foundations Reddit Corpus (MFRC) [14]: The MFRC consists of 16,123 Reddit comments drawn from 12 different subreddits. Every instance has been labeled by at least three annotators from a set of five trained annotators.

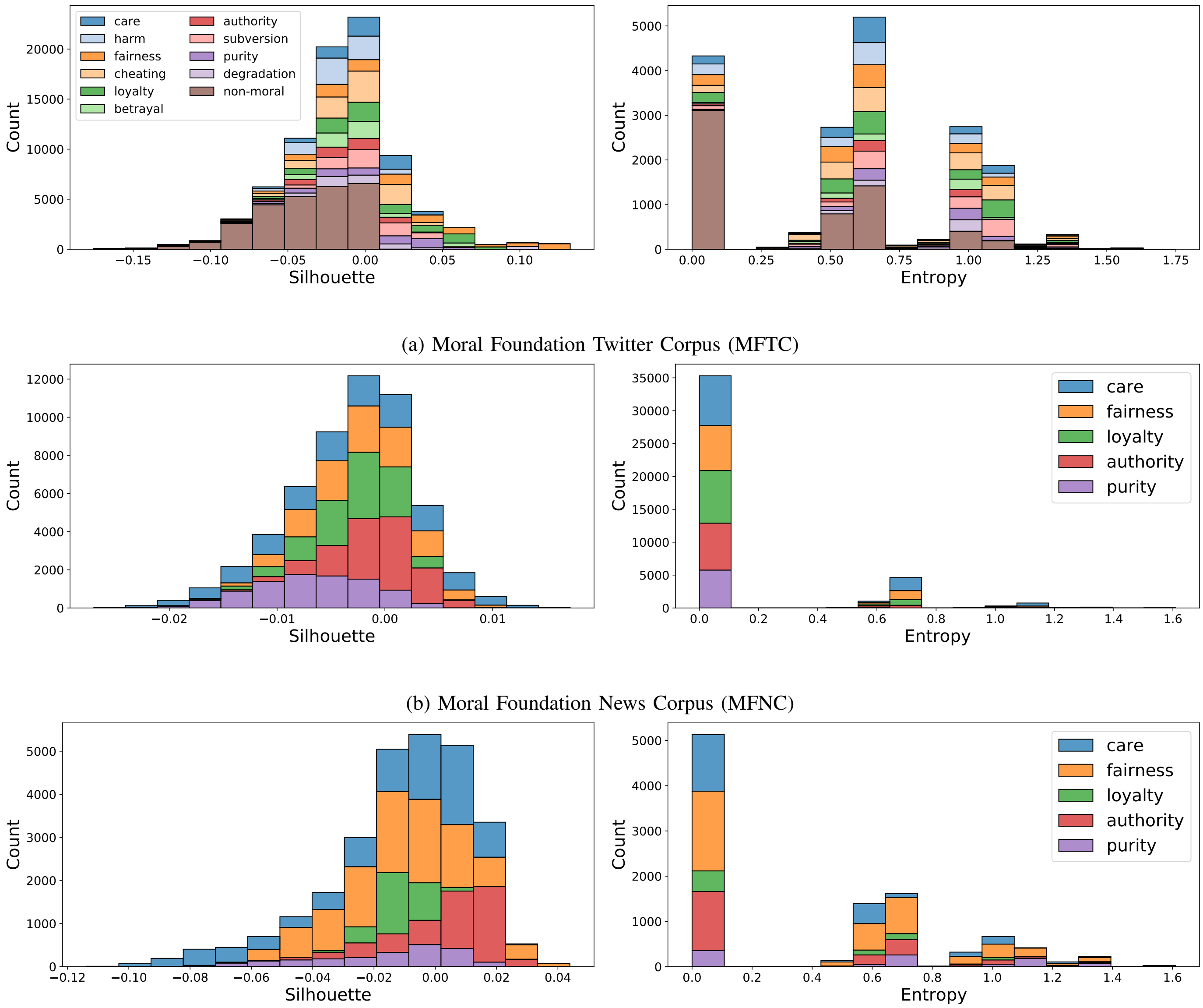
B. Metrics of Disagreement

Figure 3 shows the distributions of *entropy* and *silhouette* metrics for the three datasets. The distribution varies for each category and each dataset. For example, in MFTC, "non-moral" has the higher concentration of posts with zero entropy, meaning that in many instances all the annotators agreed on the non-moral category. On the other hand the classes purity and degradation have few samples with zero entropy of annotations. Also, in MFNC, all the categories have high concentration of posts with zero entropy which is because there are many instances with single highlights in MFNC that have no overlap with other highlights. Figure 3 also shows the distribution of silhouette coefficient of annotator judgements for each of the moral foundation categories. This metric has a consistently broad distribution for all datasets, unlike entropy.

C. Experimental Settings

1) Classifier Model: Getting a text as input, our model's task is to classify it to a moral foundation label. The classification task aims to predict majority-vote for each textual instance. Which is the label with highest number of annotators selecting it for the instance (we select randomly if there is a tie). In MFTC there are eleven labels (the labels from vices and virtues of the five moral foundations plus one label denoting non-moral category). In MFNC there are only five labels related to the main foundations regardless of the polarity. For

[2] https://www.prolific.ac/



(a) Moral Foundation Twitter Corpus (MFTC)

(b) Moral Foundation News Corpus (MFNC)

(c) Moral Foundation Reddit Corpus (MFRC)

Fig. 3: Measures of disagreement. The left column shows the distributions of the silhouette metric for the three datasets. The silhouette coefficient is calculated on annotations. The right column show the distributions of entropy metric for instances, colored by the majority-vote of moral foundation assigned to the instance.

MFRC we convert the labels “proportionality” and “equality” to fairness and only keep the labels from the main foundations to keep consistency with MFNC.

We use pretrained language model RoBERTa [28] and finetune it on our datasets. This is done by adding a multi-class classification layer on top of the language model and with a cross-entropy loss updating all the parameters end-to-end for five epochs. We minimize cross-entropy with the Adam optimizer [29] with learning rate 2×10^{-5}. Our experiments use a batch size of 50. We run each experiment with five random seeds and split data into 70% train and 30% test sets. Each experiment is performed on a single GTX 1080Ti GPU. In our implementation we use the Huggingface Transformers library [30].

2) Silhouette Coefficient Calculation: To calculate the silhouette coefficient of the judgments we need to use a language model to represent the text of instances in a latent space. We use sentenceBERT framework [31], which has been shown to be helpful for capturing semantic textual similarity. We use the pretrained model ”all-mpnet-base-v2”[3]. We use the default parameters of sklearn implementation of the silhouette

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

coefficient[4] which uses euclidean distance as a metric to capture semantic distance of texts.

D. Results

We study the effect of removing noisy annotations on the performance of the trained language model, discarding from each dataset either 1) highest-entropy instances (and all the labels assigned to them), or 2) lowest-silhouette coefficient labels. After filtering the data with a specific ratio, we split the dataset to train and test set. We compare performance to models trained on data from which the same ratio of instances have been removed at random. Figure 4 shows performance on the three datasets in terms of the F1 Macro (left subplot), and accuracy (middle) as a function of the fraction of data removed. These measures are aggregated over five runs with different random seeds and calculated on held-out test data.

Discarding data at random shows a non-increasing trend in the performance of the model. On the other hand, discarding portions of judgments with lowest silhouette coefficients improves the performance of moral classification on all three datasets. These results suggest that our proposed method identifies better-quality ground truth data that helps model performance.

Although removing highest-entropy instances improves performance of the classifier on MFTC, it does not help improve classification performance on the MFNC dataset and offers only a small improvement on MFRC. This result can be explained by the way each dataset has been collected. In MFTC, each tweet was shown to several annotators and their judgments about all the labels were collected. However, in MFNC, annotators were shown a news document and asked to identify segments of text relevant to a specific moral foundation. For example, annotator $a1$ was asked to highlight any part of document d related to the *care* foundation, but annotator $a2$, $a3$ were asked to highlight portions of d relevant to *authority* and *loyalty*. Their highlighted texts might overlap at sentence s, and s at the end will have annotations $< care : 1, fairness : 0, authority : 1, loyalty : 1, purity : 0 >$. At first glance, it seems there is no agreement on s and its entropy is very high. However, if the annotators were given the opportunity to choose any label, we could have seen more agreement in the distribution of labels. In crowd-sourcing moral annotations, it helps to simplify the task to finding only one moral foundation in a document to reduce confusion [7]. However, this results in a proliferation of instances labeled by only one annotator. The low entropy score will deceptively indicate low disagreement, even though the single annotation could have been made in error.

It is important to understand how the dataset was constructed when choosing between using instance-level entropy and judgment-level silhouette coefficient when preprocessing the data with filtering.

The rightmost panels in Figure 4 show model confidence as a function of the fraction of data discarded by the entropy or silhouette methods. The figures show that the distribution of the model's confidence on instances shifts toward higher values when we discard labels with lowest silhouette values. As a reminder, higher confidence means that the model is more likely to correctly classify the instances. The same improvement occurs when we discard high-entropy instances in MFTC. Similar to previous results, removing high-entropy instances in MFNC or MFRC does not improve confidence. Moreover, the distribution of confidence shifts toward lower values or stays the same when we remove instances at random from all datasets. Prior work [19] has shown that if a trained model has low confidence on an instance, it means that the instance was likely mislabelled. The observed positive shift in the distribution of confidence values suggests that the subset of the data we kept has fewer mislabelled instances.

V. Conclusions

In this work we show that auditing annotated data for noise can improve morality classification. We propose two metrics—*entropy* and *silhouette coefficient*—for refining annotated datasets in a pre-processing step, i.e. before training a classifier. The metrics leverage annotator judgments in order to identify instances that are difficult to label or have been mislabeled. We show that removing these instances reduces the noise in the ground truth data, improving classification performance of models trained on the remaining data. We also show that refining annotations improves the training dynamics of the model. As a result, the average of confidence increases when the model is trained on the instances that are recognized as less noisy with our proposed metrics.

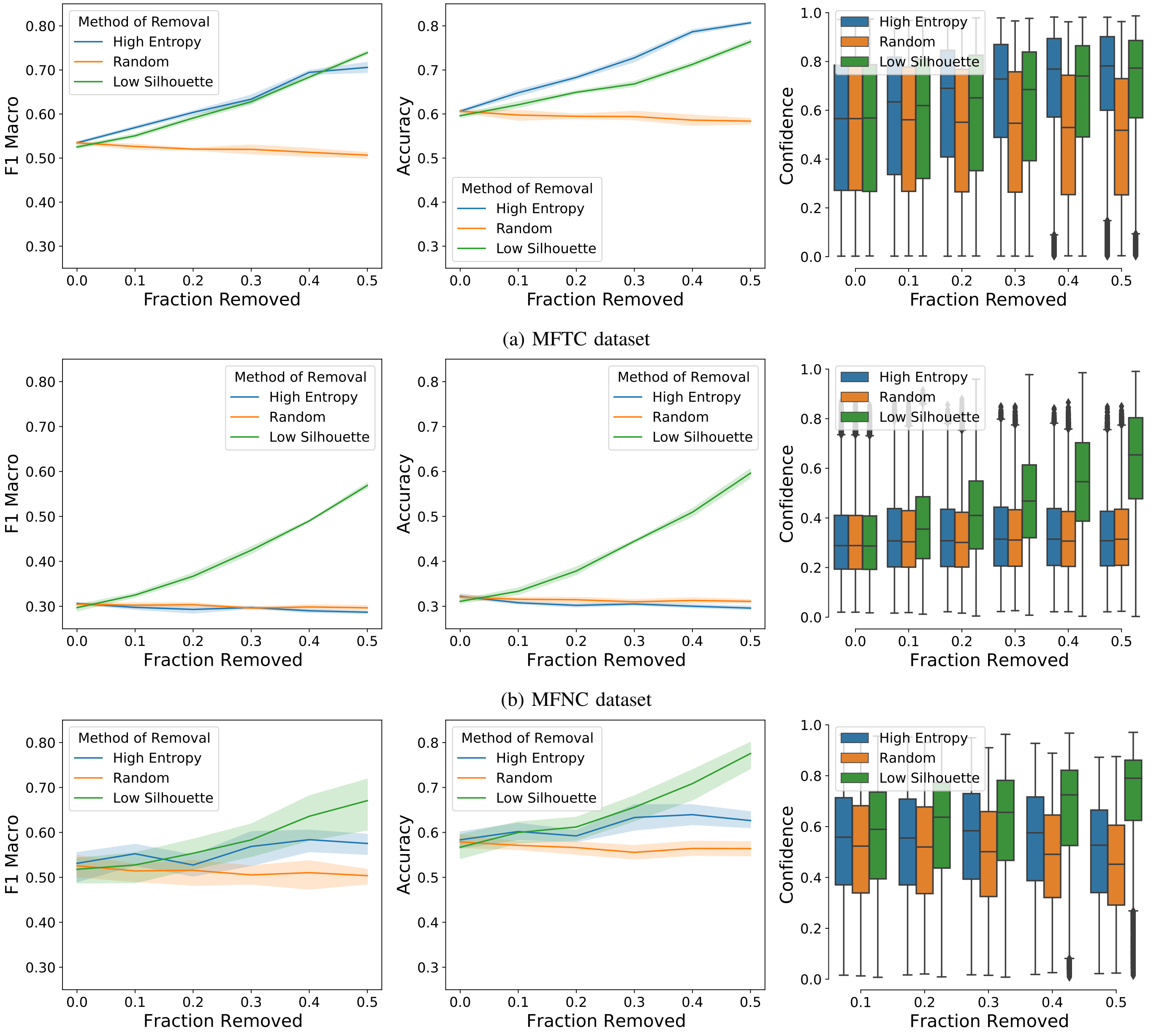
We validated our approach on three datasets where multiple annotators were asked to label the moral values expressed in text. Classifying morality is inherently a subjective and difficult task, and the resulting ground truth data from human annotation will naturally contain a large amount of disagreement and noise, which can degrade the performance of single-task models trained on the data. We showed that a classifier trained on refined data, from which the potentially noisy samples have been removed, can learn better models that more accurately recognize new instances of moral foundations. Our approach is not specific to moral annotations and can be applied to other datasets constructed from subjective judgments of multiple annotators.

VI. Limitations and Ethical Considerations

Other works [17], [32]–[34] have shown annotator demographic features and annotators' life experiences can impact their judgments. In this work we focus on single-task classifier models that need an aggregated label (e.g. to majority vote or averaging) for training. However, we encourage future work to design models beyond single-task classifiers in order to overcome the need for aggregating the labels and move to subjective models that can give predictions based on different beliefs, demographics, and backgrounds.

Also, considering the use-case of a gathered dataset, the removal criteria described in section III can be enhanced to

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_samples.html



(a) MFTC dataset

(b) MFNC dataset

(c) MFRC dataset

Fig. 4: Morality classification after de-noising annotations. Comparison of F1 (left sub-figures), accuracy (middle sub-figures), and distributions of model confidence on instances (right sub-figures) when removing high-entropy instances, low-silhouette judgments, or removing randomly. Removing low-silhouette judgments helps with model's performance on all MFTC, MFNC, and MFRC datasets. However, removing high-entropy instances is only effective on MFTC and doesn't help with learning on MFNC or MFRC due to the high ratio of instances with zero entropy.

make sure we are not removing the samples from specific demographics or groups of people. A possible remedy is to discard weighted portions of data from each demographic group in a way to keep a more balanced subset of data or to prioritize keeping the data gathered by minorities.

References

[1] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations." *Journal of personality and social psychology*, vol. 96, no. 5, p. 1029, 2009.

[2] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in experimental social psychology*. Elsevier, 2013, vol. 47, pp. 55–130.

[3] J. Hoover, K. Johnson, R. Boghrati, J. Graham, and M. Dehghani, "Moral framing and charitable donation: Integrating exploratory social media analyses and confirmatory experimentation," *Collabra: Psychology*, vol. 4, no. 1, 2018.

[4] M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani, "Moralization in social networks and the emergence of violence during protests,"

Nature human behaviour, vol. 2, no. 6, pp. 389–396, 2018.

[5] M. Dehghani, "Purity homophily in social networks - invited talk," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego, California: Association for Computational Linguistics, Jun. 2016, p. 16. [Online]. Available: https://aclanthology.org/W16-0405

[6] N. Mokhberian, A. Abeliuk, P. Cummings, and K. Lerman, "Moral framing and ideological bias of news," in *International Conference on Social Informatics*. Springer, 2020, pp. 206–219.

[7] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," *Behavior research methods*, vol. 53, no. 1, pp. 232–246, 2021.

[8] R. Weber, J. M. Mangus, R. Huskey, F. R. Hopp, O. Amir, R. Swanson, A. Gordon, P. Khooshabeh, L. Hahn, and R. Tamborini, "Extracting latent moral information from text narratives: Relevance, challenges, and solutions," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 119–139, 2018.

[9] S.-Y. N. Wang and Y. Inbar, "Moral-language use by u.s. political elites," *Psychological Science*, vol. 32, no. 1, pp. 14–26, 2021, pMID: 33306432. [Online]. Available: https://doi.org/10.1177/0956797620960397

[10] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen *et al.*, "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[11] F. R. Hopp and R. Weber, "Reflections on extracting moral foundations from media content," *Communication Monographs*, vol. 88, no. 3, pp. 371–379, 2021.

[12] B. Kennedy, M. Atari, A. M. Davani, J. Hoover, A. Omrani, J. Graham, and M. Dehghani, "Moral concerns are differentially observable in language," *Cognition*, vol. 212, p. 104696, 2021.

[13] J. Y. Xie, G. Hirst, and Y. Xu, "Contextualized moral inference," *arXiv preprint arXiv:2008.10762*, 2020.

[14] J. Trager, A. S. Ziabari, A. M. Davani, P. Golazazian, F. Karimi-Malekabadi, A. Omrani, Z. Li, B. Kennedy, N. K. Reimer, M. Reyes *et al.*, "The moral foundations reddit corpus," *arXiv preprint arXiv:2208.05545*, 2022.

[15] K. Johnson, D. Jin, and D. Goldwasser, "Modeling of political discourse framing on twitter," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[16] D. Kahneman, O. Sibony, and C. R. Sunstein, *Noise: A flaw in human judgment*. Little, Brown, 2021.

[17] V. Prabhakaran, A. Mostafazadeh Davani, and M. Diaz, "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*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 133–138. [Online]. Available: https://aclanthology.org/2021.law-1.14

[18] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 2022.

[19] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 9275–9293. [Online]. Available: https://aclanthology.org/2020.emnlp-main.746

[20] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, 2014, pp. 859–866.

[21] K. Ethayarajh, Y. Choi, and S. Swayamdipta, "Information-theoretic measures of dataset difficulty," *arXiv preprint arXiv:2110.08420*, 2021.

[22] E. Pavlick and T. Kwiatkowski, "Inherent disagreements in human textual inferences," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 677–694, 2019. [Online]. Available: https://aclanthology.org/Q19-1043

[23] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger, "Identifying mislabeled data using the area under the margin ranking," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 044–17 056, 2020.

[24] V. Pulastya, G. Nuti, Y. K. Atri, and T. Chakraborty, "Assessing the quality of the datasets by identifying mislabeled samples," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 18–22.

[25] M. Toneva, A. Sordoni, R. T. des Combes, A. Trischler, Y. Bengio, and G. J. Gordon, "An empirical study of example forgetting during deep neural network learning," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=BJlxm30cKm

[26] P. Rodriguez, J. Barrow, A. M. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber, "Evaluation examples are not equally informative: How should that change NLP leaderboards?" in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4486–4503. [Online]. Available: https://aclanthology.org/2021.acl-long.346

[27] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[31] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[32] E. Denton, M. Díaz, I. Kivlichan, V. Prabhakaran, and R. Rosen, "Whose ground truth? accounting for individual and collective identities underlying dataset annotation," *arXiv preprint arXiv:2112.04554*, 2021.

[33] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection," *arXiv preprint arXiv:2111.07997*, 2021.

[34] Z. Waseem, "Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 138–142. [Online]. Available: https://aclanthology.org/W16-5618