# OFFENSIVENESS AS AN OPINION: DISSECTING POPULATION-LEVEL LABEL DISTRIBUTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

**Warning:** *This paper contains language that may be offensive.*

Human annotation is an essential component for building human-in-the-loop machine learning systems (MLs). The diverse human disagreement that arises during annotation is often obscured because of majority voting label aggregation used for training MLs. When the minority opinion is removed in this process it may also extricate the sentiments held by people in minority demographics. This information is essential when MLs are used for offensive or hate speech identification as some content is offensive to only a minority. Collecting human annotations is an expensive task and it is even more challenging when collecting for minority voices. Population-level learning (PLL) utilizes unsupervised learning methods to represent populations of annotators using existing annotations. We test the viability and transparency of PLL with a large dataset of toxic content. We explore the clusters qualitatively by studying the language of the data items assigned to different clusters. In addition, we quantitatively analyze the nature of human disagreement via the data points assigned to the clusters.

## 1 INTRODUCTION - TOXICITY AS A PERSPECTIVE

A "winner-take-all" approach such as majority voting label aggregation, is often used to select each top label and can potentially hide the diversity of opinions from minority annotators (Ovesdotter Alm, 2011; Sabou et al., 2014; Waseem & Hovy, 2016; Plank et al., 2014; Kralj Novak et al., 2022; Wan et al., 2023) when models are trained. In annotation tasks like identifying offensive content, ties in opinions or a majority opinion that goes against the true nature of the content can be potentially dangerous, especially when the annotators are not representative of the population that the content targets (Sap et al., 2019).

Along with other methods in label distribution learning Dawid & Skene (1979); Geng et al. (2014); Liu et al. (2019c); Davani et al. (2021); Gordon et al. (2022), population-level label distribution learning (PLL) (Liu et al., 2019a; Weerasooriya et al., 2020; 2022) advocates for label distributions for both training and final predictions in the ML pipeline. The challenge often in such models is not having enough annotator-level data to train a model in the wild (Prabhakaran et al., 2021). In this paper, we stress test PLL (Liu et al., 2019a; Weerasooriya et al., 2020) on a publicly available dataset on toxic content collected from various social media sources. We utilize the $D_{TR}$ dataset collected by Kumar et al. (2021) (more details in Appendix A.1) for this study.

## 2 METHODOLOGY - DISSECTING DISAGREEMENT

To understand the causes of disagreements between human annotators, we perform several measurements. In this study, we audit the performance of the KMeans-based population-level clustering model. The KMeans-based model (PLL-KM) is able to semantically group data items that share the same label distribution. In our framework, we attempt to understand annotator disagreement through the perspective annotators' opinions and we evaluate our ability of understand and predict the population-level disagreement (Weerasooriya et al., 2020).

We study the level of annotator agreement within the dataset using the following methods: (1) Entropy is utilized to understand the level of human disagreement with the majority label for a

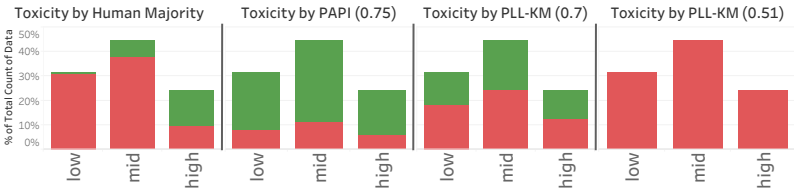


Figure 1: Cross analysis of the human toxicity classification compared the to machine model. The dataset is split into three entropy levels: "low" or majority agreeing on a single label is (0 to 0.35), "mid" is (0.35 to 0.70), and "high" or disagreement on a majority label is (0.70 to 1.05). Red is toxic and green is non-toxic data item final classification. A PAPI score $> 0.75$ is considered toxic.

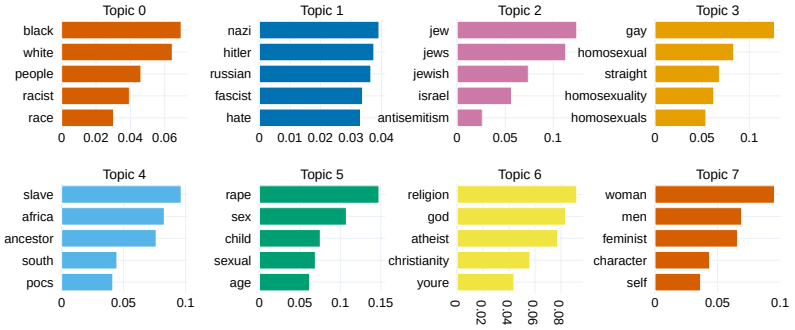


Figure 2: Word distribution for cluster #1 from extracted from PLL-KM predictions for analysis.

data item. (2) The annotator agreement against Perspective API (PAPI). An in the wild offensive language classification model. and (3) Empirical analysis into the predictions from PLL-KM model.

### 2.1 Quantitative Analysis - Entropy

We use entropy to understand the type of the disagreement between the annotators for a data point. The entropy tells us how spread out the label distribution is spread away from the majority label. Lower entropy shows agreement of the population of annotators with the majority label and a higher denotes the dissonance. We explore this in the Figure 1, where we bin the entropy into three categories and further study the disagreement. Overall in the dataset, most of the data points fall into the mid-level bin where there is some disagreement in the dataset. In Figure 1, the difference between the toxic classification of humans and PAPI outlines how unreliable the classification is for identification of the toxic content. The PAPI misclassified a significant portion of the content as non-toxic where it is identified as toxic by both humans and PLL-KM.

### 2.2 Qualitative Analysis - Language Representations

We also analyze the language of the clusters (populations of annotator pools) generated by PLL-KM using Grootendorst (2022). Figure 2 shows the most significant topics present in Cluster #1 (out of the three clusters extracted with PLL-KM). As PLL methods only cluster based on the label distributions, the language topics extracted can also assist in improving the annotator pools.

## 3 Conclusion and Future Work

In this study, we utilize entropy as a metric to understand the disagreement of the human annotators against different baselines of modeling the annotator disagreements. The Figure 1 shows how PLL-KM performs against the baselines of human majority and PAPI. And Figure 2, we explore qualitatively the nature of content that PLL-KM is able to capture for clustering annotations.

We explore clusters qualitatively and quantitatively; this framework can also be used for uncovering the significant information obscured through "winner-take-all" label aggregation methods. Our future work aims to understand the reasons why human majority, PAPI, and PLL disagree in judgment.

### URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

## REFERENCES

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *CoRR*, abs/2110.05719, 2021. URL https://arxiv.org/abs/2110.05719.

A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1):20, 1979. ISSN 00359254. doi: 10.2307/2346806. URL http://www.jstor.org/stable/2346806.

Xin Geng, Qin Wang, and Yu Xia. Facial age estimation by adaptive label distribution learning. In *Proceedings - International Conference on Pattern Recognition*, 2014.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *arXiv:2202.02950 [cs]*, February 2022. doi: 10.1145/3491102.3502004. URL http://arxiv.org/abs/2202.02950. arXiv: 2202.02950.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. Handling Disagreement in Hate Speech Modelling. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Communications in Computer and Information Science, pp. 681–695, Cham, 2022. Springer International Publishing. ISBN 978-3-031-08974-9. doi: 10.1007/978-3-031-08974-9_54.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing Toxic Content Classification for a Diversity of Perspectives. *arXiv:2106.04511 [cs]*, June 2021. URL http://arxiv.org/abs/2106.04511. arXiv: 2106.04511.

Tong Liu, Pratik Sanjay Bongale, Akash Venkatachalam, and Christopher M. Homan. Learning to predict population-level label distributions. In *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, WWW '19, pp. 1111–1120. ACM, 2019a. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3317082. URL http://doi.acm.org/10.1145/3308560.3317082.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. Learning to predict population-level label distributions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):68–76, Oct. 2019b. URL https://ojs.aaai.org/index.php/HCOMP/article/view/5286.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding, 2019c. URL https://arxiv.org/abs/1901.11504.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

Cecilia Ovesdotter Alm. Subjective natural language problems: Motivations, applications, characterizations, and implications. 2011. URL https://aclanthology.org/P11-2019.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 507–511, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-2083. URL https://aclanthology.org/P14-2083.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark 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*, pp. 133–138, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.14. URL https://aclanthology.org/2021.law-1.14. _eprint: 2110.05699.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. 2014.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. 2023. doi: 10.48550/ARXIV.2301.05036. URL https://arxiv.org/abs/2301.05036.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. San Diego, California, 2016. doi: 10.18653/v1/N16-2013. URL https://aclanthology.org/N16-2013.

Tharindu Cyril Weerasooriya, Tong Liu, and Christopher M. Homan. Neighborhood-based Pooling for Population-level Label Distribution Learning. In *Proceedings of the 24th European Conference on Artificial Intelligence 2020*, April 2020. URL http://arxiv.org/abs/2003.07406. arXiv: 2003.07406.

Tharindu Cyril Weerasooriya, Alexander G Ororbia, and Christopher M Homan. Improving Label Quality by Joint Probabilistic Modeling of Items and Annotators. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pp. 5. European Language Resources Association (ELRA), 2022. URL http://lrec-conf.org/proceedings/lrec2022/workshops/NLPerspectives/pdf/2022.nlperspectives-1.12.pdf.

# A Appendix

## A.1 Experimental Dataset - Toxic Ratings ($D_{TR}$) Dataset

Kumar et al. (2021) collected the dataset containing 107,620 items that are annotated by 17,280 participants. There are at least five annotators per data item. In the dataset, the content sources are Twitter (67%), Reddit (15%), and 4chan (18%) comments. The dataset contains the scores from the Perspective API[1] (PAPI) and granular annotator demographic information. The authors used the toxicity score of $> 0.75$ to indicate a comment as toxic. We utilize the same categorization in this study. The authors use five levels of toxicity in the study, (1) extremely toxic, (2) very toxic, (3) moderately toxic, (4) slightly toxic, and (5) not at all toxic.

For the analysis in our Methodology 2, we condense the five label choices as (a) toxic - (1) extremely toxic and (2) very toxic, (b) moderate toxic - (3) moderately toxic and (4) slightly toxic, and (5) not at all toxic.

## A.2 JS Divergence

We analyze the JS-divergence (Menéndez et al., 1997) of the dataset against the empirical result and predicted result for understanding how label distributions changed during the prediction. Liu et al. (2019b); Weerasooriya et al. (2020) utilized KL-divergence in their analysis. However, since KL is not a symmetrical measure, we utilize JS. The JS analysis is utilized as a way to understand how the PLL-KM models are able to predict the human labels.

[1] https://www.perspectiveapi.com

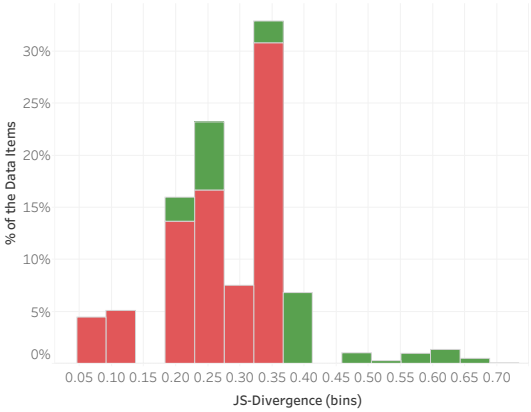


Figure 3: Histogram of the distribution of the closest 1000 items to the PLLKM predicted cluster centroid. The colors denote the human classification for the data item, where red denotes toxic and green denotes non-toxic.