Characterising and Mitigating Aggregation-Bias in Crowdsourced Toxicity Annotations

Agathe Balayn
IBM Netherlands
Center for Advanced Studies
and TU Delft WIS Group
a.m.a.balayn@student.tudelft.nl

Panagiotis Mavridis
TU Delft
WIS Group
p.mavridis@tudelft.nl

Alessandro Bozzon
TU Delft
WIS Group
a.bozzon@tudelft.nl

Benjamin Timmermans
IBM Netherlands
Center for Advanced Studies
b.timmermans@nl.ibm.com

Zoltán Szlávik
IBM Netherlands, Center for Advanced Studies
zoltan.szlavik@nl.ibm.com

Abstract

Training machine learning (ML) models for natural language processing usually requires lots of data that are often acquired through crowdsourcing. The way this data is collected and aggregated can have an effect on the outputs of the trained model such as ignoring the labels which differ from the majority. In this paper we investigate how label aggregation can bias the ML results towards certain data samples and propose a methodology to highlight and mitigate this bias. Although our work is applicable to any kind of label aggregation for data subject to multiple interpretations, we focus on the effects of the bias introduced by majority voting on toxicity prediction over sentences. Our preliminary results point out that we can mitigate the majority-bias and get increased prediction accuracy for the minority opinions if we take into account the different labels from annotators when training adapted models, rather than rely on the aggregated labels.

1 Introduction

When using crowdsourcing to gather training data for Machine Learning (ML) algorithms, several workers work with the same input samples and the annotations are aggregated into a unique one like the majority vote (MV) to ensure its correctness (elimination of annotation mistakes and spammers mainly). Although this data collection method is designed to get high-quality data, we expect that certain tasks involving subjectivity such as image aesthetic prediction, hate speech detection, detection of violent video segments, sentence sentiment analysis, cannot be tackled this way: samples should not be described with unique labels only since they are interpretable differently by different persons.

The use of hate/toxic speech has increased with the growth of the Internet (Tsesis 2001). Predicting whether a sentence is toxic is highly subjective because of its multitude of possible interpretations. The sentence "I agree with that and the fact that the article needs cleaning. Some of these paragraphs [..] seem like they were written by 5 year olds." is judged negative or positive by different readers, but this perceptions' diversity is ignored when selecting one unique label as done in recent research (Schmidt and Wiegand 2017).

(Dixon et al. 2017) studied the existence of identity term biases resulting from the imbalance of a toxicity dataset content, we show with the example of MV-aggregation that crowdsourcing processing methods on the same dataset also create an algorithmic bias here towards the majority opinion. When annotations differ but are all valid for certain annotators, aggregation loses information and leads to decrease of accuracy and unfairness in ML results, thus we hypothesize that the bias can be mitigated by using disaggregated data. In this study, we first exhibit the presence of the majority-bias and its consequences, then we propose a methodology to expose and counter its algorithmic effects.

2 Majority-biased dataset and consequences

We show on the toxicity dataset (Wulczyn, Thain, and Dixon 2017) that in usual crowdsourcing aggregations of annotations, certain worker contributions are ignored for the majority and that it affects the fairness of ML algorithms' results. The dataset consists of 159686 Wikipedia page comments for which 10 annotations per sample are available. A large number of annotators (4301) that we have their personal information rate the phrases with 5 labels of toxicity ranging from -2 (very toxic) to 2 (very healthy) with 0 being neutral.

Subjectivities in the dataset. For each worker, we compute the average disagreement rate (ADR) with the ground truth (percentage of annotations different from the MV here), and plot the distribution over the dataset after removing the annotations of the lowest quality workers (spammers) (fig. 1). The quality score for each worker (WQS) is computed with the CrowdTruth framework (Aroyo and Welty 2013) using binary labels ([-2;-1]:toxic, [0;2]:non-toxic), along a unit quality score (UQS) to represent the clarity of each sentence. Without removing low-quality workers, the proportion of high agreement is high because most spammers constantly use one positive label and the dataset is unbalanced with more samples with non-toxic MV. The more possible spammers are removed, the more the disagreement increases until the distributions stabilize. Only 0.09% of the workers always agree with the MV for 50 spammers removed: MV-aggregation is not representative of most individuals but only of a sentence-level common opinion.

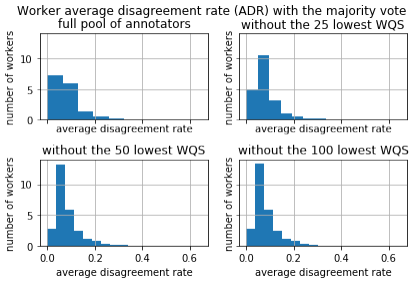

Figure 1: Normalized distribution comparison of the ADR with the MV with and without low quality worker filtering.

Table 1: Accuracy performances of the model on the ambiguity balanced dataset.

	agg. testing	disagg. testing
agg. training	0.76	0.70
disagg. training	0.77	0.71
disagg. training with user	0.77	0.70

Algorithmic effect of the bias. We consider the task of predicting binary labels. Training traditional algorithms to predict the MV, annotations of only maximum 0.09% of annotators would be entirely correct: the majority-bias is not consistent with the worker's individual opinions. We evaluate traditional models (sec. 3) trained and tested on aggregated and disaggregated labels (table 1). In both cases accuracy is higher when measured on aggregated data, what shows that classical input data's treatment makes usual models' predictions biased towards one type of opinion, here the majority opinion, instead of representing each subjectivity.

3 Method to measure and mitigate the bias

We claim that a fairer algorithm should return different outputs for a same sample depending on its reader. Here, we propose measures of the majority-bias' algorithmic effect and a method to counter its unfairness.

Bias measure. Global metrics are usually used to optimize the algorithms' parameters and evaluate them. However, they do not inform on the bias' effects since most samples' labels have a high-agreement: the slight improvement when training on disaggregated data hints only lightly at label disaggregation (table 1, fig. 2). To identify the effects, we propose to measure sentence-level and worker-level accuracies on the annotations spread in the following bins: we divide the sentences along their ambiguity score (AS) (percentage of agreement in annotations) or UQS, the workers with their ADR, WQS or demographics categories; and also plot histograms of the per-user and per-sentence errors to identify potential unfairness among all workers or sentences.

Bias mitigation: ML. To account for the full range of valid opinions, we propose to modify the inputs to the ML models. After removing low-quality workers, instead of the aggregated labels we feed them with the annotations augmented with the available worker demographics (age, gender, education, with a continuous or one-hot encoded representation) that psychology literature (Cowan and Hodge 1996) gives as the most influencing factors of offensiveness perception (along with ethnicity not available here). Each (sentence, demographics, annotation) tuple is considered as one data sample. We employ the Logistic Regression (LR) classifier, and encode sentences with term frequency-inverse document frequency (tf-idf). The optimal hyperparameters for each set-up are chosen by performing a grid search.

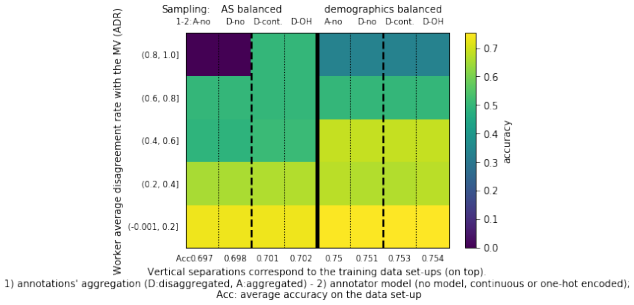

Figure 2: Average and ADR-binned accuracies for two resamplings of the dataset.

Bias mitigation: dataset balancing. We define 4 data set-ups to help the algorithms learn the individual annotations. Sentence AS and MV-toxicity are computed, and we resample the dataset following the original distribution or balancing the distribution on these 2 criteria, to obtain a dataset whose majority-bias is decreased by equally representing samples with high and low agreement between workers. We also resample the annotations along the MV-toxicity and demographics categories (removing the least frequent ones) into one dataset following the distributions and a balanced one, to foster performance fairness in-between populations.

Results. Binned metrics like the user-level ADR-binned accuracy (fig. 2 with bins along the y-axis) enable to show that models are more suited to workers who agree with the MV (bottom of the y-axis), and highlight the benefit of using disaggregated data with adapted ML models. On the AS-balanced dataset (left part of the x-axis), the user representation increases accuracy for workers with a high disagreement with the majority over using aggregated data or no user-model. The resampling choice also helps understanding and mitigating bias' effects: balancing on demographics neither clearly shows the performance gap between minority and high-ADR workers nor improves accuracy with the user representation, contrary to the AS dataset in which MV-consensus' presence is reduced.

4 Conclusion and Discussion

Disaggregating the annotations decreases the majority-bias' effects with adapted ML models' inputs and dataset resamplings. Binning the evaluation metrics enables to understand and verify the existence of these effects. We only reported results using the LR classifier but we now investigate adaptations of Deep Learning algorithm's architectures which are better suited to the large dataset (10 times more annotations than labels) and to the size of the ML inputs.

References

Aroyo, L., and Welty, C. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM* 2013.

Cowan, G., and Hodge, C. 1996. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology* 26(4):355–374.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2017. Measuring and mitigating unintended bias in text classification.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Tsesis, A. 2001. Hate in cyberspace: Regulating hate speech on the internet. *San Diego L. Rev.* 38:817.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399. International World Wide Web Conferences Steering Committee.