Ground-Truth, Whose Truth? - Examining the Challenges with Annotating Toxic Text Datasets

Kofi Arhin
Lally School of Management
Rensselaer Polytechnic Institute, Troy NY
arhink@rpi.edu

Ioana Baldini
IBM Research
ioana@us.ibm.com

Dennis Wei
IBM Research
dwei@us.ibm.com

Karthikeyan Natesan Ramamurthy
IBM Research
knatesa@us.ibm.com

Moninder Singh
IBM Research
moninder@us.ibm.com

Abstract

The use of language models (LMs) to regulate content online is on the rise. Task-specific fine-tuning of these models is performed using datasets that are often labeled by annotators who provide "ground-truth" labels in an effort to distinguish between offensive and normal content. Annotators generally include linguistic experts, volunteers, and paid workers recruited on crowdsourcing platforms, among others. These projects have led to the development, improvement, and expansion of large datasets over time, and have contributed immensely to research on natural language. Despite the achievements, existing evidence suggests that Machine Learning (ML) models built on these datasets do not always result in desirable outcomes. Therefore, using a design science research (DSR) approach, this study examines selected toxic text datasets with the goal of shedding light on some of the inherent issues and contributes to discussions on navigating these challenges for existing and future projects.

1 Background

Recent advancements in the Machine Learning (ML) domain have contributed to the development and use of language models [20]. This can be attributed to the curation of large datasets that serve as training resources, among other factors [11]. The current paper focuses on toxic text datasets, which are often labeled by annotators who provide ground-truth labels for the data samples. The importance of such datasets to the growth of natural language research cannot be overemphasized. However, there are some notable challenges with these datasets. For instance, labels provided by annotators are not always reliable and consistent [13], [3], [15]. We reason that this is largely due to the fact that language is highly contextual, and its interpretation, often subjective [19]. That is, a phrase that is offensive to one person may be deemed as normal by another. These difficulties often translate into poor model performance with regards to metrics such as bias and accuracy [5]. Hence, it is important to understand these difficulties and propose solutions on reducing their impact on prediction outcomes.

Existing studies have proposed alternate approaches to resolving some of the challenges. For example, Matthew *et al.* [14] posit that training models by highlighting the portion of a particular text that people use to distinguish offensive text from normal text can improve model performance. Also, Sap *et al.* [17] show that priming annotators before annotation tasks can reduce their insensitivity to different dialects and the occurrence of bias in ground-truth labels. Similarly, Sap *et al.* [18] show how nudging annotators to provide additional information such as context inference, biased

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

implications, and targets, among others, can help to improve the quality of crowdsourced datasets. However, Ball-Burack *et al.* [1] find that solutions developed to tackle issues in one dataset may not necessarily be effective in resolving issues with out-of-sample datasets. In certain instances, annotator information may be required to improve model performance, highlighting the problem that labels may not be independent of annotators [5]. These insights highlight the need for a deeper understanding of the issues with crowdsourced toxic text datasets [4]. Therefore, in the present paper, we take first steps in shedding light on some challenges in these datasets with the hope of addressing the challenges for current and future annotation tasks.

To help achieve the goal for this study, we adopt the design science research (DSR) [10] framework as a guide. The framework provides guidelines on developing innovative solutions to existing problems, especially where people and technology are concerned [16], [7]. Using this framework as a lens for problem identification and solution development, we find additional challenges to those that have already been highlighted in the extant literature, by examining three toxic text datasets that approach ground-truth labeling differently. We use a multi-label approach to re-annotate one of the datasets, and find that 1) given different contexts, text samples can have different labels, 2) multiple labels for toxic text datasets can increase agreement with external ML annotators, but however, 3) this may not guarantee an improvement in inter-annotator agreement.

2 Data

The datasets selected for the study include the HateXplain [14], Social Bias Inference Corpus (SBIC) [18], and the Jigsaw[1] datasets. Our selection of these three datasets is founded on the basis that they address a similar problem (toxic text), yet they are diverse in how the annotations were collected. For instance, HateXplain has exactly three annotators for all text samples while SBIC and Jigsaw do not have a fixed number of annotators for samples. The SBIC dataset has additional information on annotators (i.e., data statements [2]). Further, while the HateXplain and SBIC sets use majority voting to determine the final label, the Jigsaw data uses a continuous final label (i.e, toxicity) that represents the proportion of annotators who labeled a particular sample as toxic. For example, if 4 our of 5 annotators label text sample *A* as offensive, Jigsaw's final label will be 0.8 toxicity, HateXplain and SBIC will have *offensive* or *hatespeech* as the final label. Table 1 provides a summary of the number of annotators, text samples, and the distribution of final labels. In the last column for Table 1, we include our definition for offensive text for each dataset for the purpose of this study.

Table 1: Dataset Summary

Dataset	Unique Text Samples	Offensive Text	Normal Text	Total Annotators	Offensive Text Definition
HateXplain	20,148	12,334 (61%)	7,814 (39%)	253	Both *hatespeech* and *offensive* text
SBIC	45,318	25,073 (55%)	19,401 (45%)	307	Samples given the labels 1 and 0.5
Jigsaw	1,804,874	120,084 (7%)	1,684,790 (93%)	8,899	Samples with toxicity 0.5 and above

3 Findings

3.1 Problem Identification

Table 2 below provides a summary of the problems identified in the datasets. We group the issues into four main headings, namely Annotator Influence, Annotator (Im)Balance, Inconsistent Labels, Contextless Samples.

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Table 2: Some Identified Challenges in the HateXplain, SBIC, and Jigsaw Datasets

Challenge Identified	Implications
Annotator Influence: In HateXplain, one annotator contributed to the final label for 5,730 samples. 5,730 samples means a single annotator contributed to about 28% of the final labels, about 3,000 samples and 18% points more than the second-ranked annotator in the HateXplain set.	Since annotator behavior can be reflected in prediction outcomes [5], [2], this is problematic because the annotator highly influences the outcome (final label) of many samples. This can lead to one annotator's behavior being further amplified by a ML model.
In SBIC, a majority (i.e., 80%) of the annotators identified as white. We examined whether people in the minority groups were more likely to oppose the final label and found for example 17.4% of annotators who identified as black, opposed the final label at least once, while this figure was 12% for the white annotators.	Also, low diversity in annotators selected for a labeling task can result in the silencing of minority voices. Since bias is one of the key concerns for recent advances in toxic text classification, it is important to understand how low diversity can impact predictions.
Annotation (Im)Balance: In Jigsaw and SBIC, we found instances where a large pool of annotators contributed to the final label for some text samples while others received labels from relatively fewer annotators. For example, the minimum annotators per text for the Jigsaw data is 3 annotators, while the maximum annotators per text is 4,936 annotators. As such, similar samples may have different toxicity rates due largely to very different numbers of annotations.	This is a problem, especially for Jigsaw, because the final label is the proportion of annotators who label a sample as offensive. Consider text sample *B* labelled by 3 annotators and *C* labeled by 1000 annotators. For sample *B*, if even one annotator labels the text as offensive, toxicity will equal 0.33. However, for *C*, if 100 annotators label the sample as offensive, toxicity will equal 0.10, suggesting that sample *B* is more toxic than *C* which might not be the case.
Inconsistent Labels: We found instances where annotators provided different labels for similar text. For example, in HateXplain, *has stupid rich h*e* was labeled as *normal* while *...b**** *a*** back to the east* was labeled *offensive* by the same annotator.	Inconsistency has been one of the major concerns for building good models using toxic text datasets. It creates noise in the data which ultimately leads to poor model performance [6], [12].
Contextless Samples: Some text samples were difficult to place in specific contexts. This made it difficult to know which labels to assign to them. However, annotators provided labels for these sample. For example, the sample *"why Arabs lose wars"* from HateXplain is difficult to place in a specific context, making it difficult to categorize.	Contextless samples also have the tendency to lead to noisy labels because of the high level of uncertainty. This can also lead to large disagreement rates between annotators.

3.2 Objectives, Design and Development of Solution

The problems identified in the datasets point to the fact that language is difficult to label. Clear guidelines are required to inform annotators on how to handle a variety of texts. In addition to guidelines, it is important to allow annotators to skip text samples that are difficult to categorize. Therefore providing a third label such as *undecided* can help researchers identify problematic samples in the data. Finally, by going through the samples and enumerating the challenges, we found that it might be more prudent to provide context-based label columns for annotators. That is, for each sample, annotators will have multiple columns to label.

Similar to Guest et al. [8], we develop new guidelines to guide annotators for this task. We propose three context-based label columns for the annotation task: *strict label*, *relaxed label*, and *inferred group label*. For the *strict label*, we ask annotators to consider the task as a bag-of-words approach where the appearance of certain words in a text makes the entire text either offensive or normal. For the *relaxed label*, we ask annotators to consider contexts where an offensive text could be labeled as normal. For the *inferred group label* we ask annotators to consider whether an offensive text can be considered normal if it was uttered by a member of the target group in the text. We found that, providing different contexts can lead to different labels for a particular sample.

3.3 Demonstration and Evaluation

To demonstrate the outcome of the proposed solution, we randomly selected and annotated 100 samples from the HateXplain dataset using the developed guidelines. The samples were annotated separately by the five authors and final labels were determined by majority vote. We found that for each of the three columns, the final labels did not align well with the original HateXplain labels. Table 3 shows that there is on average a 22% disagreement between the new labels and the original HateXplain labels. The agreement rate between the new labels were 87% between the strict and relaxed labels, 85% between the strict and inferred group labels, and 97% between the inferred group and relaxed labels. To complement these results, we used two external ML tools as annotators, namely Perspective AI[2] and Detoxify [9]. The agreement rates with Perspective AI and Detoxify for all three new labels are comparatively higher than that of HateXplain. This suggests that providing multiple labels can increase agreement with existing ML predictions, which in turn suggests the future possibility of training more targeted ML models.

Table 3: Label Agreement from Team Annotation Task

Labels	HateXplain	Perspective AI	Detoxify Original	Detoxify Unbiased
Strict	0.77	0.75	0.74	0.75
Relaxed	0.80	0.68	0.65	0.66
Inferred Group	0.78	0.66	0.63	0.66
HateXplain	-	0.61	0.60	0.62

In addition to these results, the annotation task provided the team with three important insights. First of all, contrary to existing studies, we believe that annotator disagreement could be an indication of annotator diversity, which is a desirable attribute. Hence, it is important to understand why annotators disagree rather than trying achieve high agreement which may lead to biased outcomes. Secondly, one must reckon with the fact that disagreement in a general sense increases as the number of annotators increases. Taking the rate of unanimous agreement as a simple measure, for the 100 samples, the agreement rate at two annotators was 81% but gradually decreased to 56% when three additional annotators were added. Finally, regardless of the inter-annotator disagreement rate, intra-annotator consistency is an important metric because it can be an indicator of annotation guideline clarity.

4 Conclusion, Limitations and Future Work

In this study, we highlight the need to review existing toxic text datasets for NLP tasks. We enumerate the challenges in some selected datasets and add to conversations pertaining to addressing them. We find that while language is difficult to annotate, using multiple annotation labels can help to reduce some of the identified challenges.

One of the main limitations of this study is the fact that although we reviewed three toxic text datasets, we only annotated samples from one of them. Also, 100 samples might not provide enough observations to cover the context of the entire dataset. Furthermore, three out of the many toxic text datasets may not be representative enough. However, we believe that the insights shared provide useful implications for theory and practice.

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

References

[1] A. Ball-Burack, M. S. A. Lee, J. Cobbe, and J. Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128, 2021.

[2] E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[3] Q. Chen, D. S. Weld, and A. X. Zhang. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *arXiv preprint arXiv:2108.01799*, 2021.

[4] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang. "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, pages 1–32, 2021.

[5] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*, 2019.

[6] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[7] S. Gregor and A. R. Hevner. Positioning and presenting design science research for maximum impact. *MIS quarterly*, pages 337–355, 2013.

[8] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, 2021.

[9] L. Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[10] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, pages 75–105, 2004.

[11] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[12] E. Ishita, S. Fukuda, Y. Tomiura, and D. W. Oard. Using text classification to improve annotation quality by improving annotator consistency. *Proceedings of the Association for Information Science and Technology*, 57(1):e301, 2020.

[13] I. Martin-Morato and A. Mesaros. What is the ground truth? reliability of multi-annotator data for audio tagging. *arXiv preprint arXiv:2104.04214*, 2021.

[14] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*, 2020.

[15] C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[16] K. Peffers, M. Rothenberger, T. Tuunanen, and R. Vaezi. Design science research evaluation. In *International Conference on Design Science Research in Information Systems*, pages 398–410. Springer, 2012.

[17] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

[18] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

[19] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

[20] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.