ARTICLE: Annotator Reliability Through In-Context Learning (Student Abstract)

Sujan Dutta[1], **Deepak Pandita**[1], **Tharindu Cyril Weerasooriya**[1], **Marcos Zampieri**[2], **Christopher M. Homan**[1], **Ashiqur R. KhudaBukhsh**[1]

[1]Rochester Institute of Technology
[2]George Mason University
{sd2516, cmhvcs}@rit.edu, {deepak, cyril, khudabukhsh}@mail.rit.edu, mzampier@gmu.edu

Abstract

Ensuring annotator quality in training and evaluation data is a key piece of machine learning in NLP. Tasks such as sentiment analysis and offensive speech detection are intrinsically subjective, creating a challenging scenario for traditional quality assessment approaches because it is hard to distinguish disagreement due to poor work from that due to differences of opinions between sincere annotators. With the goal of increasing diverse perspectives in annotation while ensuring consistency, we propose `ARTICLE`, an in-context learning (ICL) framework to estimate annotation quality through self-consistency. We evaluate this framework on two offensive speech datasets using multiple LLMs and compare its performance with traditional methods. Our findings indicate that `ARTICLE` can be used as a robust method for identifying reliable annotators, hence improving data quality.

Code — https://github.com/Suji04/ARTICLE

Introduction

Conventional approaches to distinguish *high* from *poor* quality annotators are typically based on outlier detection, where the divergence from aggregate opinions is considered a signal of poor quality annotation (Dumitrache et al. 2018; Leonardelli et al. 2021; Davani, Díaz, and Prabhakaran 2022; Ustalov, Pavlichenko, and Tseitlin 2024). However, for subjective tasks such outlier-based approaches can potentially muffle minority or unique perspectives, leading to annotation echo chambers. For example, in a war corpus with annotators from countries $\mathcal{A}$ and $\mathcal{B}$, responses to questions like *who is winning the war* can vary greatly depending on the annotator's country. If the pool is dominated by annotators from $\mathcal{A}$, perspectives from $\mathcal{B}$ may be eliminated as their responses differ from the majority.

This paper introduces an alternative method to estimate annotator quality through self-consistency. Prior work has explored annotation patterns of individual annotators (Dawid and Skene 1979) but without considering the context and annotator information. For instance, if an annotator inconsistently rates similar offensive content differently, it indicates a lack of self-consistency. Incorporating self-consistency into quality estimation bypasses the need for multiple annotators, enhancing resource efficiency, and preserving unique but self-consistent perspectives that outlier-based methods might discard. While self-consistency has been applied in other settings (Wang et al. 2023; Cooper et al. 2024), to our knowledge, this is the first application for rater quality estimation in subjective annotation tasks. Additionally, recent research has utilized LLMs as annotators (He et al. 2024), but prior work focused on replacing majority opinions rather than capturing annotator-level labels.

Methodology

We propose **`ARTICLE`** (**A**nnotator **R**eliability **T**hrough **I**n-**C**ontext **Le**arning) – a two-step framework to identify reliable annotators and model the perception of offense for different political groups.

Step 1: Identifying Inconsistent Annotators

We hypothesize that annotators who show inconsistent annotation patterns are difficult to model. We individually model each annotator using a state-of-the-art LLM, `Mistral-7B` (Jiang et al. 2023), and utilize the model's performance (ease of modeling) as a proxy for the annotator's consistency. For each annotator, we randomly split their annotations into two sets – the first set (training set) contains 10 data points, and the second (test set) contains the rest. Using the training set as ICL examples, we prompt `Mistral-7B` to predict the labels for the test set. Then, we compute the F1-score to evaluate the model's performance. A high F1 score indicates the annotator is easy to model and, hence, consistent, and a low score indicates the opposite. We define a hyperparameter (k) that acts as a threshold. If, for a given annotator, the F1-score is less than k, we mark them as inconsistent and remove them from the dataset.

Step 2: Modeling Group-level Perception

After removing the inconsistent annotators from all political groups, we recompute the aggregate labels for each group. We again use ICL to model the group-level perception of offense. For each group, we construct a training set using 70% of the data. The rest is used for testing. For each test instance, we randomly sample 15 examples from the training set and use them as in-context examples. The same `Mistral-7B` model is used in this step.

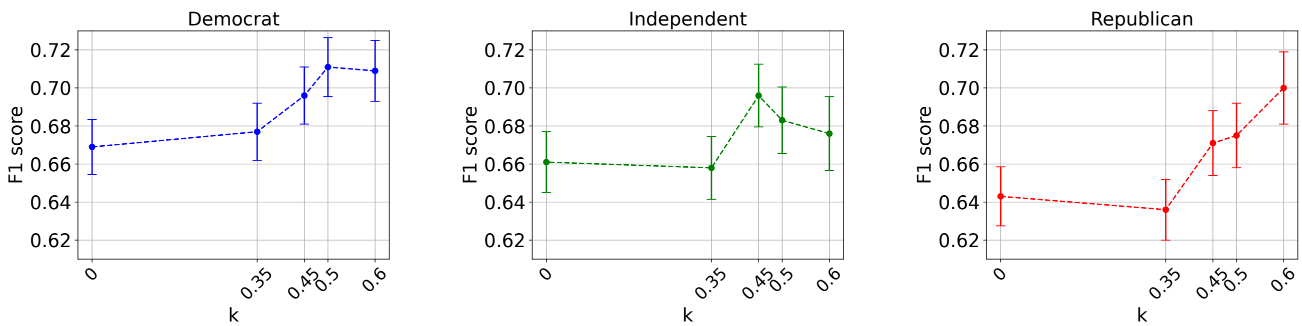

Figure 1: Group-level model performance at different k values in $\mathcal{D}_{\texttt{TR}}$. The error bars indicate 95% confidence interval.

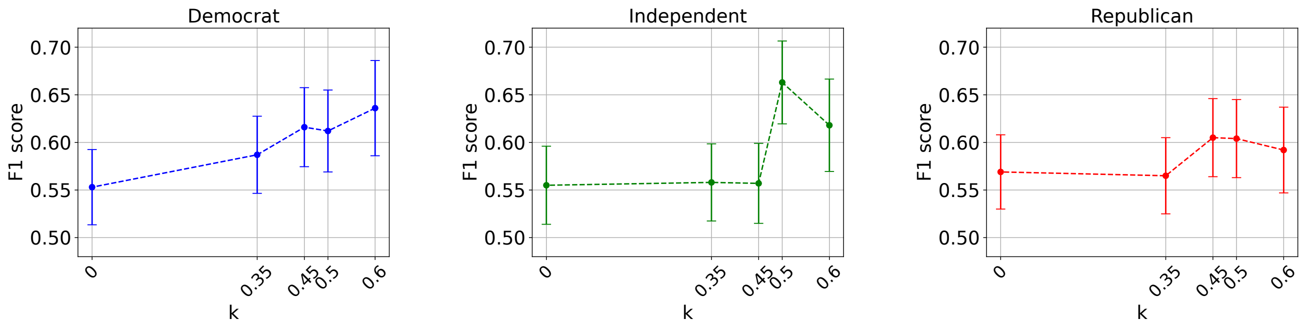

Figure 2: Group-level model performance at different k values in $\mathcal{D}_{\texttt{VOICED}}$. The error bars indicate 95% confidence interval.

Evaluation

Quantitative. We evaluate the proposed framework on two well-known English offensive speech datasets, (1) `Toxicity Ratings` (Kumar et al. 2021) ($\mathcal{D}_{\texttt{TR}}$) and (2) `VOICED` (Weerasooriya et al. 2023) ($\mathcal{D}_{\texttt{VOICED}}$). In each dataset, we model the perception of offense for each political group: Democrat, Republican, and Independent. To study the impact of the hyperparameter k, we run experiments for the following values of $k : \{0, 0.35, 0.45, 0.5, 0.6\}$. The case $k = 0$ serves as the baseline where we do not remove any annotators from the dataset. Figures 1 and 2 illustrate the performance (F1-score on the test set) at various values of k for $\mathcal{D}_{\texttt{TR}}$and $\mathcal{D}_{\texttt{VOICED}}$, respectively. In general, across all political groups, we observe an upward trend in the F1-score as the value of k increases with noticeable fluctuations for Independents. In almost all instances, the F1-score achieved with $k = 0.45$ surpassed the baseline performance, suggesting the effectiveness of the proposed method. We also note for most cases with $k > 0.5$, the performance either plateaus or declines slightly. It suggests that while increasing k generally improves model performance up to a point, there may be a threshold beyond which further increase in k does not yield additional benefits and might even be detrimental.

Qualitative. We present annotations from a few inconsistent annotators as identified by `ARTICLE`. In the first example, both the comments express that the government should not control abortion; however, the annotator labeled them differently. In the second example, the annotator labeled a seemingly harmless comment offensive; however, assigned a non-offensive label to a comment containing a racial slur.

EXAMPLE ANNOTATOR #1

Comment: `Abortion should be between the woman and God. I do agree the government should not control it. I'm not saying I agree with Abortion what so ever but the rioting and stuff is just ridiculous.` **Label:** *offensive*

Comment: `The 'state' has no business deciding what women can or cannot do to their bodies. The fact that so many in the USA feel that abortion should be banned speaks more to a moronic pandering to the right wing religious fringe crazies who want to impose THEIR beliefs on everyone.` **Label:** *non-offensive*

EXAMPLE ANNOTATOR #2

Comment: `Amazing to see some of our ex students still continuing on their passion for the arts.` **Label:** *offensive*

Comment: `yea lil ass went Mia that boy a hog. n[redacted]gg[redacted]r finna average 3 assists a game a piece.` **Label:** *non-offensive*

Ethics Statement

ARTICLE's approach to annotation quality assessment through self-consistent intends to help mitigate potential biases towards minor perspectives in NLP systems. In this work, we used two publicly available datasets referenced in the paper. No new data collection has been carried out as part of this work. The datasets used do not reveal any identifiable information about the annotators.

References

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22004–22012.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Dawid, A. P.; and Skene, A. M. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Applied Statistics*, 28(1): 20.

Dumitrache, A.; Inel, O.; Aroyo, L.; Timmermans, B.; and Welty, C. 2018. CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement. *arXiv preprint arXiv:1808.06080*.

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. K. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kumar, D.; Kelley, P. G.; Consolvo, S.; Mason, J.; Bursztein, E.; Durumeric, Z.; Thomas, K.; and Bailey, M. 2021. Designing toxic content classification for a diversity of perspectives. In *SOUPS*, 299–318.

Leonardelli, E.; Menini, S.; Aprosio, A. P.; Guerini, M.; and Tonelli, S. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *EMNLP*, 10528–10539.

Ustalov, D.; Pavlichenko, N.; and Tseitlin, B. 2024. Learning from Crowds with Crowd-Kit. *Journal of Open Source Software*, 9(96): 6227.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Weerasooriya, T.; Dutta, S.; Ranasinghe, T.; Zampieri, M.; Homan, C.; and Khudabukhsh, A. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *EMNLP*, 11648–11668.