Dragonfly_captain at SemEval-2023 Task 11: Unpacking Disagreement with Investigation of Annotator Demographics and Task Difficulty

Ruyuan Wan, Karla Badillo-Urquiola
University of Notre Dame
rwan@nd.edu, kbadill3@nd.edu

Abstract

This study investigates learning with disagreement in NLP tasks and evaluates its performance on four datasets. The results suggest that the model performs best on the experimental dataset and faces challenges in minority languages. Furthermore, the analysis indicates that annotator demographics play a significant role in the interpretation of such tasks. This study suggests the need for greater consideration of demographic differences in annotators and more comprehensive evaluation metrics for NLP models.

1 Introduction

Which voices should an algorithm be trained to replicate? In natural language processing, detecting and handling disagreements among annotators is a critical task. Annotators may disagree due to various reasons, including differences in personal beliefs, cultural background, or subjective interpretations. Disagreements in annotated data may lead to biased models, which can have negative implications, such as amplifying hate speech or discrimination. However, in the current supervised learning paradigm, annotation disagreements are often resolved by aggregating the opinions of majority groups, effectively excluding minority perspectives. Therefore, it is essential to develop methods that can capture and address disagreements in annotated data. This paper reports analysis of participating the SemEval-2023 Task 11, Learning with Disagreement, to promote a better understanding of disagreements among annotators and their impact on natural language processing models.

2 Background

The use of crowdsourcing for annotating data has become a popular solution for the expensive and time-consuming process of labeling large amounts of data in NLP models. However, it also brings its own set of challenges, with annotator disagreement being a primary concern (Checco et al., 2017; Kairam and Heer, 2016). Annotator disagreement can have various causes, but subjective judgment and language ambiguity are common culprits (Uma et al., 2022). Failure to consider the subjective and ambiguous nature of some instances can result in inaccurate predictions.

2.1 Annotation Disagreement

Annotation disagreement is a prevalent challenge in NLP datasets, where different annotators label the same instance with different labels. There can be various reasons why annotators disagree, such as the subjective nature of the task, the ambiguity of language, or personal biases of the annotators. Even objective tasks can have multiple acceptable answers, leading to disagreement among annotators.

Subjectivity is a significant cause of disagreement, as tasks like sentiment analysis or hate speech detection require annotators to make subjective judgments. Annotators' personal experiences, cultural backgrounds, and beliefs can influence their perception of the text, leading to conflicting labels(Wan et al., 2023). Additionally, ambiguity in language can also cause disagreement. Words and phrases can have multiple meanings or interpretations, leading to differences in labeling. Furthermore, personal biases of annotators can also play a role in annotation disagreement. Annotators may bring their own opinions and preferences to the task, leading to differences in labeling. Even with clear annotation guidelines, some annotators may interpret them differently due to personal biases.

Therefore, it is essential to consider and address annotation disagreement when developing NLP models. Understanding the causes of disagreement and developing methods to mitigate them can lead to more accurate and robust models.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1978–1982
July 13-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Task	Textual type	Lang	N. items	Disaggregated labels	Pool Annotators	Annotators' Info	Additiona Info
HS-Brexit	Hate speech detection	Tweets	En	Train/dev 952 Test 168	6	6	Ann. ID targe or control	Aggressiveness Offensiveness
ConvAbuse	Abusive language detection	Conversation with AI-systems	En	Train/dev 3210 Test 840	Variable	8	Ann. ID	Ableist Homophobic Intellectual Racist, sex harassment transphobic, target, explicit..
ArMIS	Misogyny and sexism detection	Tweets	Ar	Train/dev 798 Test 145	3	3	Ann. ID Gender Political view	
MD-Agreement	Offensive language detection	Tweets	En	Train/dev 7696 Test 3057	5	>800	Ann. ID	Domain

Table 1: Datasets overview

2.2 Prediction with Disagreement

When dealing with annotated data that contains disagreement, there are several methods that can be employed to make predictions (Davani et al., 2022). One common approach is to create an ensemble model that combines the outputs of multiple models trained on different subsets of the data. Another method is the use of multi-task models that treat predicting each annotator's judgment as a separate subtask, while sharing a common learned representation of the task. Additionally, multi-annotator models can also be used to address disagreement in annotated data. These approach can capture the nuances and different viewpoints expressed by annotators and provide a more accurate prediction. However, these approaches still have major voting in the end to generate final results.

Further, there are also methods that aim to recover the distribution of the true labels rather than predicting a single label (Sampath et al., 2022). These models attempt to estimate the probability distribution of the true label, taking into account the uncertainty and disagreement in the data. These methods can be useful in cases where there is significant disagreement among annotators, and a single ground truth label cannot be determined.

Past research has made significant progress in developing methods that enable accurate and fair prediction with disagreement in annotated data. However, subjective tasks that involve highly personal judgments, such as emotion disagreement, still pose a significant challenge(Sampath et al., 2022). While methods such as ensemble models, multi-task models, multi-annotator models, and distribution recovery models have shown promising results, further research is needed to address the challenges posed by subjectivity and ambiguity in NLP datasets.

2.3 Evaluation with Disagreement

Evaluation of NLP models is crucial to ensure their effectiveness and generalizability. Traditional evaluation metrics, such as accuracy and F1 score, assume a single "ground truth" label and do not consider the possibility of multiple acceptable answers or annotator disagreement.

However, recent work has shown that such assumptions can lead to inaccurate evaluations and biased models when dealing with annotated data that contains disagreement. Gordon et al. proposes a disagreement deconvolution method that disentangles stable opinions from noise by estimating intra-annotator consistency and compares each test set prediction to the individual stable opinions from each annotator. Applying this method to existing social computing datasets, the study found that current metrics dramatically overstate the performance of many human-facing machine learning tasks.

Additionally, a recent approach called "accurate fairness" has been proposed to align individual fairness with accuracy in machine learning models (Li et al., 2022). This approach uses a Siamese fairness in-processing approach to minimize the accuracy and fairness losses of a model under the accurate fairness constraints. This method has been shown to improve individual fairness without sacrificing accuracy, and has been applied to mitigate possible service discrimination in a real dataset. These approaches provide a promising direction for evaluating NLP models in the presence of disagreement and multiple acceptable answers.

3 Understand Disagreement from Data

The Semantic Evaluation workshop 2023 Task 11 collected a benchmark of four textual datasets with different characteristics (Leonardellli et al., 2023). The HS-brexit dataset is a new dataset of tweets on abusive language on Brexit annotated for hate speech, aggressiveness, and offensiveness by two distinct groups. The ArMIS dataset is a dataset of Arabic tweets annotated for misogyny detection by annotators with different demographics characteristics. The ConvAbuse dataset is a dataset of English dialogues annotated by at least three experts in gender studies using a hierarchical labeling scheme. The MultiDomain Agreement dataset is a dataset of English tweets from three domains annotated for offensiveness by 5 annotators via AMT, with a focus on pre-selecting tweets that are potentially leading to disagreement. Overall, the datasets provide a multiplicity of labels for each instance, and particular attention was paid to selecting instances that have the potential for disagreement among annotators, as shown in the Table 1.

HS-Brexit (Akhtar et al., 2021) consists of tweets related to hate speech on Brexit, annotated for hate speech, aggressiveness, and offensiveness by the same six annotators from two distinct groups: a target group of three Muslim immigrants residing in the UK and a control group of three other non-Muslim immigrants.

Splitting the annotation results by target group and control group in the given training set, I ran a T-test to identify whether there are significant disagreement among Muslim immigrants and other individuals. The null hypothesis is there is no difference in group means, while the alternate hypothesis is there is some difference. The T-test p-value (8.86e-27) is smaller than 0.05, so we have strong evidence to show that the target group and control group hold different stance in identifying hate speech on Brexit in the given text. Also the t-statistic is -10.91 which reflects the control group perceives more hate speech than the target group of Muslim immigrants.

ConAbuse (Curry et al., 2021) is a dataset of English dialogues between expert users and conversational agents. It is important to note that this data is collected during experimental setups, which may be significantly different from real-life conversations, such as those that take place on social media platforms.

ArMIS (Almanea and Poesio, 2022) is a collection of Arabic tweets annotated for misogyny detection. The annotations were provided by annotators with different demographic characteristics, including "Moderate Female", "Liberal Female", and "Conservative Male". Like the HS-Brexit dataset, the annotators represent compound features of their gender and political views. Additionally, this dataset presents its unique challenge of the Arabic language.

To investigate how different demographics of annotators may identify misogyny/sexism differently in the ArMIS dataset, three t-tests were conducted to compare the set of annotations between each pair of annotators. The results revealed that the opinions of the moderate female and the conservative male were not significantly different from each other (p-value=0.318). However, both groups were significantly different from the liberal females' identification of misogyny(p-value=1.14e-06 and 0.0001)). Specifically, the liberal female annotators identified much less misogyny in the dataset compared to the other two groups. However, only three annotators couldn't show whether the distinction of these three people will also apply to other moderate female, liberal female, conservative male groups that they were representing in the task.

MD-Agreement (Leonardelli et al., 2021) contains English tweets from three different domains: Black Lives Matter (BLM), Election in 2020, and Covid-19. The dataset has been annotated for offensiveness via Amazon Mechanical Turk (AMT). Compared with previous datasets, MD-agreement has a more diverse annotator pool.

4 Experiments

The experiments involves using RoBERTa-base (Liu et al., 2019) to predict soft agreement label in regression setup with evaluation using cross-entropy loss. Given that subjective disagreement tasks lack a definitive "truth," the primary metric for evaluating the model's performance will be a soft evaluation using cross-entropy. This metric assesses the degree to which the model's predicted probabilities align with the agreement level among annotators.

Data	Cross Entropy	Rank
HS-Brexit	0.415	17
ConvAbuse	0.347	19
ArMIS	0.688	17
MD-Agreement	0.528	15
Average	0.494	15

Table 2: Cross Entropy results across datasets

5 Results

Although my results were in the middle of the leaderboard[1], I did not implement any further modifications to the basic model. Instead, I focused on understanding how the disagreement prediction works across datasets. This allowed me to gain insights into the nature of the disagreement in each dataset and identify potential areas for improvement in future work.

Based on the results, it shows that the basic RoBERTa model performs best on the ConvAbuse task with a cross-entropy score of 0.347, and worst on the ArMIS task with a score of 0.688. This suggests that the ConvAbuse task is relatively easier task while ArMIS is more challenging. Also becasue I used RoBERTa base model that is pretrained on English corpus. The ranking also follows this pattern. The best performed ConvAbuse task is the lowest rank which shows ConvAbuse is the easist task that advanced language techniques can easily learn to identify agreement/disagreement pattern. Further, because ConvAbuse data is collected from experimental setup which also make it easier task than identifying agreement/ disagreement in complex real social context.

6 Conclusion

In conclusion, this study highlights the importance of considering disagreement in NLP tasks and the impact of annotator demographics on the interpretation of such tasks. My results suggest that machine learning models can perform well on experimental datasets but may struggle with minority languages. As a future direction, it is important to investigate more effective approaches for addressing disagreement in subjective tasks in complicated societal context, as well as to explore ways to improve the diversity and representativeness of the annotator pool.

[1] Due to lack of time, we submitted only predictions for one dataset valid for the competition. The results reported in this paper are conducted after the competition.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Dina Almanea and Massimo Poesio. 2022. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational ai. *arXiv preprint arXiv:2109.09483*.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. 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.

Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Xuran Li, Peng Wu, and Jing Su. 2022. Accurate fairness: Improving individual fairness without trading accuracy. *arXiv preprint arXiv:2205.08704*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Aneesha Sampath, Victoria Lin, and Louis-Philippe Morency. 2022. Seedbert: Recovering annotator rating distributions from an aggregated label. *arXiv preprint arXiv:2211.13196*.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *arXiv preprint arXiv:2301.05036*.