IREL at SemEval-2023 Task 11: User Conditioned Modelling for Toxicity Detection in Subjective Tasks

Ankita Maity, Pavan Kandru, Bhavyajeet Singh, Kancharla Aditya Hari and **Vasudeva Varma**
IIIT Hyderabad
{ankita.maity, siri.venkata, bhavyajeet.singh, aditya.hari}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

This paper describes our system used in the SemEval-2023 Task 11 Learning With Disagreements (Le-Wi-Di). This is a subjective task since it deals with detecting hate speech, misogyny and offensive language. Thus, disagreement among annotators is expected. We experiment with different settings like loss functions specific for subjective tasks and include anonymized annotator-specific information to help us understand the level of disagreement. We perform an in-depth analysis of the performance discrepancy of these different modelling choices. Our system achieves a cross-entropy of 0.58, 4.01 and 3.70 on the test sets of HS-Brexit, ArMIS and MD-Agreement, respectively. Our code implementation is publicly available. [1]

1 Introduction

Natural language expressions, such as sentences and phrases, can often have multiple possible interpretations depending on the context in which they are used. This ambiguity arises due to language's inherent complexity and flexibility, which can lead to different interpretations of the same expression by different individuals. Additionally, subjective tasks can lead to disagreements between annotators with different perspectives or interpretations of the same text.

The SemEval-2023 task 11 Learning With Disagreements (Le-Wi-Di) (Leonardellli et al., 2023) focuses entirely on such subjective tasks, where training with aggregated labels makes much less sense. In this task, we worked with three (textual) datasets with different characteristics in terms of languages (English and Arabic), tasks (misogyny, hate speech, offensiveness detection) and annotations' methodology (experts, specific demographic groups, AMT-crowd). We leverage this additional information in order to get more accurate estimates of each annotator's annotation.

All the datasets provide a multiplicity of labels for each instance. The focus is on developing methods able to capture agreements/disagreements rather than focusing on developing the best model. Since a "truth" cannot be assumed, "soft" evaluation is the primary form of evaluating performances, i.e. an evaluation that considers how well the model's probabilities reflect the level of agreement among annotators.

2 Related Work

This is the 2nd edition of the Le-Wi-Di task; the previous version was held in SemEval 2021 (Uma et al., 2021a). A survey paper by Uma et al. (2021b) was also released, which identified several NLP and CV tasks for which the gold-standard idealisation has been shown not to hold. It used them to analyse the extensive literature on learning from data possibly containing disagreements.

Akhtar et al. (2021) who introduced the HS-Brexit dataset, trained different classifiers for each annotator, and then took an ensemble of classifiers.

In the case of multiclass problems (for example, classifying different kinds of hate speech instead of simply distinguishing between hate speech vs non-hate speech), there have been efforts to frame it as which class is harder to classify instead of which text belongs to which class (Peterson et al., 2019).

For tasks without annotated labels to calculate soft loss, augmentation techniques like mixup, as shown by Zhang et al. (2018), could be used to distribute the probability mass over more than one class.

3 Data

The three datasets we worked with all deal with Twitter data - HS-Brexit (Akhtar et al., 2021),

[1] https://github.com/Ankita-Maity/LeWiDi

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2133–2136
July 13-14, 2023 ©2023 Association for Computational Linguistics

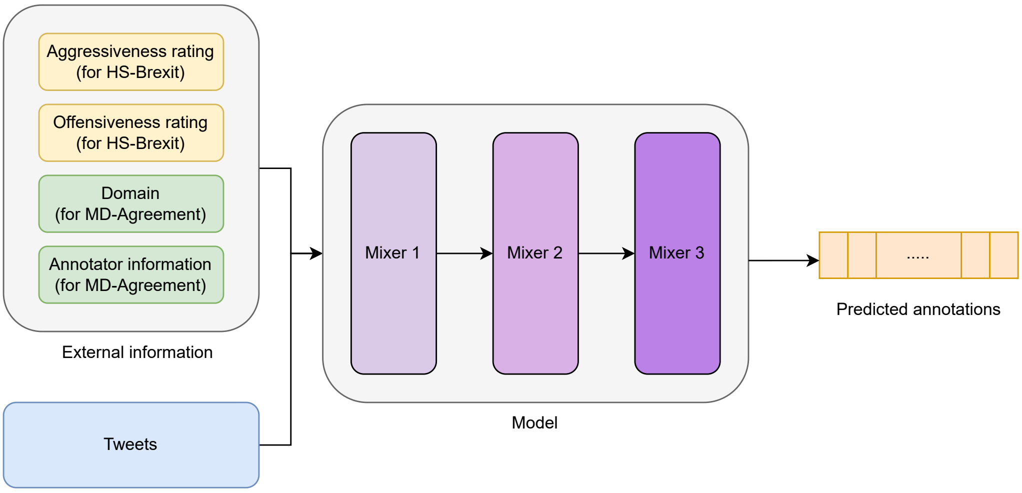
Aggressiveness rating (for HS-Brexit)
Offensiveness rating (for HS-Brexit)
Domain (for MD-Agreement)
Annotator information (for MD-Agreement)
External information
Tweets
Mixer 1
Mixer 2
Mixer 3
Model
.....
Predicted annotations

Figure 1: A high level overview of the model

ArMIS (Almanea and Poesio, 2022) and MultiDomain Agreement dataset (Leonardelli et al., 2021). While HS-Brexit and MultiDomain Agreement deal with English tweets, ArMIS deals with Arabic tweets. Details of these datasets are given in Table 1.

The "HS-Brexit" dataset is a new dataset of tweets on abusive language on Brexit and annotated for hate speech (HS), aggressiveness and offensiveness by six annotators belonging to two distinct groups: a target group of three Muslim immigrants in the UK, and a control group of three other individuals.

The "ArMIS" dataset is a dataset of Arabic tweets annotated for misogyny and sexism detection by annotators with different demographic characteristics ("Moderate Female", "Liberal Female", and "Conservative Male").

The "MultiDomain Agreement" dataset comprises around 10k English tweets from three domains (BLM, Election, Covid-19). Each tweet is annotated for offensiveness by five annotators via AMT. Particular focus was put on annotating preselected tweets that could lead to disagreement.

4 Methodology

Since the datasets we used have both textual information and external information, we needed to combine this information to derive meaningful insights. This combination can be done in various ways - from simply concatenating the information to using the attention-based summation of the information.

The main models we used were BERTweet (Nguyen et al., 2020) for HS-Brexit and MD-Agreement, and AraBERT (Antoun et al., 2020) for ArMIS. LM-based text embeddings were common across all datasets, but other embeddings we used varied based on the datasets. For HS-Brexit, we used embeddings for aggressiveness and offensiveness information. For MD-Agreement, we first concatenated the tweet's domain information to the tweet's text. Also, since there were over 800 annotators for MD-Agreement, we needed additional embeddings to capture this. For this, we used one-hot vectors and let the model learn information about the annotators based on their annotations. In our experiments, we concatenate the embeddings and then use attention-based mixing as a combining module. We use auxiliary losses to improve model performance. Figure 1 shows a high-level overview of this structure.

For our experiments, we use Adam optimizer with a learning rate of 1e-6 and cyclicLR scheduler with triangular2 mode. We train the model for 30 epochs with a batch size of 16.

Initially, we use hard labels (for the submission) but later also experiment with soft labels and apply softmax over the logits produced by the classifier. Uma et al. (2020) found that soft-loss training systematically outperforms gold training when the objective is to achieve a model whose output mimics most closely the distribution of labels produced by the annotators. We compare the effects of using soft loss training with respect to hard loss training on the given datasets, as explained in the results section.

For the final experiment, we improve our mixers and add more complexity to the model. We experiment with multi-head attention-based mix-

	Task	Language	Size	Disaggregated labels	Pool annotators	Additional information
HS-Brexit	Hate speech detection	English	Train/Dev: 952 Test: 168	6	6	Aggressiveness, Offensiveness
ArMIS	Misogyny and sexism detection	Arabic	Train/Dev: 798 Test: 145	3	3	
MD-Agreement	Offensive language detection	English	Train/Dev: 7696 Test: 3057	5	>800	Domain

Table 1: Overview of the datasets used

ing for this. The final embedding is obtained after three layers of multi-head attention-based mixing followed by feed-forward layers. Since, for ArMIS, no additional information was present, this dataset was excluded from this experiment.

5 Results

5.1 Overall Performance

We summarize the results from our experiments in Table 2. Cross entropy was used as the primary evaluation metric, but we also show micro F1 scores alongside cross-entropy. Using hard loss gives the results that were submitted for the competition, and we compare that with our other experiments.

The addition of soft loss most helped MD-Agreement results, but the results were mixed for HS-Brexit and ArMIS. In fact, the best performance of ArMIS for cross-entropy came from hard loss. This may be because ArMIS used no additional information besides text and had the least number of annotators (just three) to distribute the probability mass.

Our architectural improvements, which included designing better mixers, gave better cross entropy and micro F1 results for both HS-Brexit and MD-Agreement datasets. The best results for cross entropy for test sets of these two datasets resulted from this.

5.2 Error Analysis

Some tweets have a larger amount of disagreement than others. Two cases are of particular interest to us. We wanted to check how many obvious cases (annotators agreed with 75 per cent certainty over the class the tweet belonged to) our system was missing. We also wanted to check how many less obvious cases (there was less than 35 per cent agreement between annotators) our system could predict correctly.

For the obvious cases our system was missing,

- HS-Brexit had 26 predictions wrong in total, out of which 10 were obvious cases.
- ArMIS had 61 predictions wrong in total, out of which 37 were obvious cases.
- MD-Agreement had 1275 predictions wrong in total, out of which 877 were obvious cases.

For the less obvious cases our system was able to predict correctly,

- HS-Brexit had 33 less obvious instances, out of which our system correctly predicted 17 instances.
- ArMIS had 53 less obvious instances, out of which our system correctly predicted 29 instances.
- MD-Agreement had 856 less obvious instances, out of which our system correctly predicted 458 instances.

Thus, on average, our model was able to predict 53.24 per cent of controversial/less obvious cases correctly, which seems promising. However, more work is needed since 55.97 per cent of incorrect predictions were obvious cases.

6 Conclusion

When there is no clear answer that all annotators can agree with, it is crucial to consider the factor of their disagreement over just getting a single label for a data point. Our results highlight the benefits of using soft loss over hard loss for such controversial cases. We also find that using better ways to

Testing strategy	HS-Brexit				ArMIS				MD-Agreement			
	val		test		val		test		val		test	
	CE	F1	CE	F1	CE	F1	CE	F1	CE	F1	CE	F1
Majority baseline	2.71	0.89	5.62	0.89	8.23	0.60	8.91	0.57	7.74	0.65	7.38	0.67
Hard loss	0.47	0.86	0.75	0.84	4.55	0.57	**4.01**	0.58	7.50	0.51	9.92	0.42
Soft loss	0.65	0.88	1.07	0.86	3.82	0.58	4.70	0.56	6.42	0.57	8.73	0.50
Better mixers with multi-head attention	0.58	0.88	**0.58**	0.84	-	-	-	-	3.40	0.59	**3.70**	0.58

Table 2: Results for cross entropy and micro F1 across the three datasets

combine multiple channels of information, which can potentially help us model the annotators and predict their choices, can lead to the best results. However, the deep learning models of today are primarily encouraged to focus on hard evaluation scores like F1 and disregard the noise in the data, which leads to excellent results in constrained lab environments but fail in real-world scenarios. Finding more ways to incorporate the subjectivity of real-world data and people's opinions could help make these models more robust and generalizable.

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.

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, Marseille, France. European Language Resources Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8:173–177.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.