Multilingual Hate Speech Modeling by Leveraging Inter-Annotator Disagreement

Patricia-Carla Grigor*
University of Vienna
Vienna, Austria

Bojan Evkoski
evkoski_bojan@phd.ceu.edu
Central European University
Vienna, Austria

Petra Kralj Novak
novakpe@ceu.edu
Central European University
Vienna, Austria
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

As social media usage increases, so does the volume of toxic content on these platforms, motivating the Machine Learning (ML) community to focus on automating hate speech detection. While modern ML algorithms are known to provide nearly human-like results for a variety of downstream Natural Language Processing (NLP) tasks, the classification of hate speech is still an open challenge, partially due to its subjective annotation, which often leads to disagreement between annotators. This paper adopts a perspectivist approach that embraces subjectivity, leveraging conflicting annotations to enhance model performance in real-world scenarios. A state-of-the-art multilingual language model for hate speech detection is introduced, trained, and evaluated using diamond standard data with metrics that consider disagreement. Various strategies for incorporating disagreement are compared in the process. Results demonstrate that the model performs equally or better on all evaluated languages compared to respective monolingual models and drastically outperforms on multilingual data. This highlights the effectiveness of multilingual and perspectivist methods in addressing the complexities of hate speech detection. The presented multilingual hate speech detection model is available at: https://huggingface.co/IMSyPP/hate_speech_multilingual.

Keywords

hate speech detection, inter-annotator disagreement, multilingual language modeling

1 Introduction

The phenomenon of hate speech, which is typically defined as offensive or derogatory language targeting individuals or groups based on characteristics such as race, religion, ethnic origin, sexual orientation, disability, or gender [2], has become a significant problem on social networks in recent years, with communities being increasingly exposed to toxic content as the networks grow and become more interconnected [13, 3]. Consequently, the Machine Learning (ML) and computational linguistics communities have begun developing content moderation strategies using advanced algorithms and Natural Language Processing (NLP) techniques to detect hate speech [10, 11]. However, a key challenge is the subjectivity of hate speech, as annotators often disagree due to diverse backgrounds and perspectives.

To address this challenge, researchers have proposed alternative methodologies to ground-truthing, including the incorporation of diverse perspectives into the training and evaluation pipelines of ML models [1, 14]. One such approach is introduced by [7], who train monolingual hate speech classifiers in several languages directly on datasets that include disagreement. As an alternative to gold-standard data, such data is referred to as diamond standard data, based on the assumption that more than one single truth exists. In terms of evaluation, the researchers focus on the evaluation of models from the perspective of disagreement, with the ultimate goal of estimating the agreement between the annotators themselves, as well as between models and annotators by using the appropriate metrics. Their main findings indicate that disagreement between annotators represents an intrinsic limitation to the performance that can be achieved by automated systems.

This paper aims to explore the potential of training a multilingual hate speech model, as well as further explore the ideas of incorporating inter-annotator disagreement in model training. Therefore, at the basis of this paper lie the following research questions:

- *How does the performance of multilingual hate speech classifiers trained on diamond standard data compare to the performance of monolingual models?*
- *How can inter-annotator disagreement be effectively incorporated into the classifier fine-tuning process?*

In light of these research questions, the expected outcomes are twofold: (1) multilingual classifiers trained on diamond standard data are anticipated to outperform monolingual models, and (2) incorporating inter-annotator disagreement is expected to enhance sensitivity to nuanced hate speech. These findings could benefit computational linguistics research and social media providers by informing the development of more effective content moderation algorithms.

2 Related Work

Several methods exist for incorporating disagreement into ML training pipelines [12, 5], but few focus on hate speech detection. One approach is presented in [7], where monolingual hate speech classifiers were trained for English, Italian, and Slovenian. These classifiers utilized diamond standard datasets sourced from YouTube and Twitter, employing a consistent annotation process for each language. Their main findings indicate that, according to the accuracy scores, the annotators demonstrated a high degree of agreement in approximately 80% of the cases across all three datasets. In terms of Krippendorff's ordinal alpha score, which considers both agreement by chance and the ordering of classes (from least to most severe), the agreement score is approximately

*The first author conducted the research with significant input from the second author, under the supervision and guidance of the third author. All authors contributed to writing the manuscript.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2024.sikdd.7

0.6 for all three languages. Furthermore, the evaluation results indicate that the performance of each model aligned with the inter-annotator agreement, both in terms of accuracy and the alpha score. This implies that the performance of models is inherently constrained by the level of agreement among annotators. Consequently, when trained on diamond standard data, it is unlikely that the performance of these models can significantly surpass human performance.

This work was built upon these findings through investigating the potential of multilingual models to enhance hate speech detection, with the aim of broadening their applicability across diverse linguistic contexts. Additionally, strategies for incorporating annotator disagreement were explored, with the goal of improving model performance to approach human-level accuracy and agreement.

3 Method

This section details the methodology for training and evaluating the multilingual hate speech classifier presented in this paper. It begins with a brief overview of the datasets used, followed by an explanation of the chosen pre-trained language model that serves as the foundation for fine-tuning. The section concludes with a description of the methods employed for evaluating the models.

3.1 Datasets

Three monolingual datasets, i.e. the English (Youtube), Italian (Youtube) and Slovenian (Twitter) datasets, introduced in [7] served as the basis for our multilingual model. Each item was annotated by two annotators independently, assigned to one of four available classes: [Appropriate], [Inappropriate], [Offensive], and [Violent]. In the case of conflicting labels, both annotating instances were kept.

To explore strategies for incorporating disagreement, three multilingual datasets were created. First, the *Duplicate All* (DA) dataset, which contains all instances by their respective two annotators from the three monolingual datasets. Second, the *Duplicate Disagreement* (DD) dataset, in which instances where annotators disagreed appear twice with their respective conflicting labels, while instances that they agreed upon appear only once, creating a more balanced training set that reflects both agreement and disagreement, potentially preventing the models from being biased towards instances where annotators agree. And third, the *Remove Disagreement* (RD) dataset, which consists only of instances where annotators agree. Thus, the first two datasets contain diamond standard data, while the third dataset can be considered a gold standard dataset in which disagreement has been explicitly removed.

All instances in these datasets have undergone the same preprocessing steps, such as replacing links and usernames with placeholders. This step was undertaken to mitigate any potential biases associated with certain names, as discussed in [6]. Table 1 presents an overview of the label distribution across the three multilingual training sets. The datasets used for monolingual evaluation are the unmodified evaluation sets presented in [7].

Table 1: Label distribution of the multilingual train sets

Dataset	Acceptable	Inappropriate	Offensive	Violent
DA	191,677	11,005	112,833	7,145
DD	111,324	8,346	72,706	4,992
RD	80,573	2,661	40,255	2,161

3.2 Model Selection and Fine-Tuning

Our proposed multilingual hate speech model builds on the pre-trained XLM-R transformer model [4], chosen for its proven effectiveness in cross-lingual understanding and its ability to handle a wide range of languages. This provides a robust foundation for fine-tuning and optimization, particularly since English, Italian, and Slovenian—the languages used for fine-tuning—were included in XLM-R's pre-training. To explore various strategies for incorporating annotator disagreement during training, three model variants were fine-tuned on the previously presented datasets, referred to in the tables as MDA, MDD, and MRD, respectively.

To address class imbalance and enhance model performance on minority classes, a custom training loop with a weighted cross-entropy loss function was implemented, as proposed in [9]. The class weights were calculated to be inversely proportional to the frequency of each hate speech class within the training data. The hyperparameters for the fine-tuning process included a learning rate of 6×10^{-6}, a batch size of 8, and 3 training epochs. During the training phase, the AdamW optimizer was employed to optimize the model parameters. The fine-tuning process was implemented using PyTorch.

3.3 Model Evaluation

In terms of evaluation, the approach introduced in [7] was replicated in order to compare the performance of the multilingual classifiers to human judgment from the perspective of disagreement. This was achieved by employing identical measures to estimate the agreement between human annotators, as well as the agreement between annotators and models. Accuracy, F1 score and, most notably, Krippendorff's ordinal alpha were used to evaluate all models in this research.

Rarely used in ML applications, Krippendorff's alpha is a robust measure for assessing inter-rater reliability, accounting for agreement beyond what might occur by chance. It is applicable across various data types (nominal, ordinal, interval, and ratio scales) and is particularly effective in dealing with missing data. The value of Krippendorff's alpha ranges from -1 to 1, where 1 indicates perfect agreement and 0 suggests agreement equivalent to chance. Generally, an alpha above 0.80 is considered a strong agreement, while in hate speech datasets, the alpha values range from 0.25 to 0.65. For a detailed discussion, see Krippendorff [8].

4 Results

This section presents the evaluation results on the multilingual model and its variants. It starts with an evaluation from the perspective of inter-annotator and model-annotator agreement. Then, the class specific evaluation results, as well as a model comparison based on the models' average scores are presented. The models are also compared to monolingual baselines fine-tuned on data for their respective languages, including the BERT model for English, AlBERTo for Italian, and CroSloEngual for Slovenian, as presented in [7].

4.1 Inter-Annotator and Model-Annotator Agreement

The inter-annotator agreement was computed on the evaluation sets for each language using Krippendorff's alpha and accuracy. The same measures were also used to compute the agreement between the annotators and the models. The results are presented in Table 2.

Table 2: Inter-Annotator Agreement compared to model-annotator agreement in terms of Krippendorff's ordinal alpha (α) and Accuracy (Acc.) for the models Multilingual Duplicate All (MDA), Multilingual Duplicate Disagreement (MDD), and Multilingual Remove Disagreement (MRD) based on the language-specific evaluation sets

Dataset	Inter-Annotator Agreement		MDA		MDD		MRD	
	α	Acc.	α	Acc.	α	Acc.	α	Acc.
English	58.19	82.91	55.89	79.97	50.18	76.47	**57.90**	**81.41**
Italian	57.00	81.79	**58.29**	82.00	56.15	80.43	57.84	**82.69**
Slovenian	56.62	79.43	55.74	78.60	52.95	76.52	**55.15**	**78.84**

First, in the case of inter-annotator agreement, annotators agree around 80% of the time in terms of accuracy, with an accuracy score between 79% and 82% across all three datasets. However, accuracy does not account for class imbalance, nor the ordering of the classes. A more appropriate estimate of the agreement is computed through Krippendorff's ordinal alpha. Here, the annotators achieve an agreement score alpha in the values between 0.56 and 0.58 across the three languages.

Second, in terms of agreement between annotators and models, the same metrics were applied. The results demonstrate a consistent level of agreement between the models and annotators across all cases. Based on accuracy scores, all models align with at least one annotator approximately 80% of the time, with alpha values comparable to inter-annotator scores. In most instances, the models achieve the upper limit of inter-annotator agreement, and in some cases, even exceed it (e.g., Italian *Multilingual Duplicate All MDA*). This suggests that the models are effectively learning consistent patterns or biases that align well with one or more annotators. Such outcomes are expected in scenarios where annotator disagreement is largely due to subjective interpretation. This should not be construed as the model being inherently superior, but rather as an indication of its efficiency in modeling the predominant patterns present in the training data.

Third, a comparison between the multilingual variants shows that the *Duplicate Disagreement (DD)* strategy consistently shows worse alpha scores, meaning that emphasizing only on disagreement might be detrimental in training. No consistent difference between *Duplicate All (DA)* and *Remove Duplicates (RD)* is evident from the experiments.

4.2 Model Comparison

To evaluate the performance of the models across the four hate speech classes, the F1 score was used. Additionally, the combined (weighted) F1 score was computed for each model to assess their overall performance. To determine the best-performing model, the weighted F1 scores were averaged across all three languages. Table 3 shows the results achieved by each of the models on the English evaluation set. In the case of the English dataset, the results show that the multilingual model outperforms the baseline monolingual English model across all classes except the [Appropriate] class, a case in which it still performs competitively. The variant which achieved the highest score on the minority classes is the *MDA* model, with an F1 score of 39.16 for the [Inappropriate] class and an F1 score of 27.82 for the [Violent] class. This is most likely due to introducing the weighted cross-entropy loss function, which was effective in improving performance on underrepresented classes, a procedure which was not performed in [7].

Similar patterns emerge on the Italian dataset (Table 4). The multilingual model is competitive to the monolingual model while outperforming the Italian baseline on the minority classes. The highest scores on the most important classes [Violent] and [Offensive] were achieved by the *MDA* variant, once again showing the superiority of the *Duplicate All (DA)* strategy.

Table 3: Model evaluation results in terms of class-specific F1 scores on the English dataset. The Total score was calculated using the weighted F1 score. The first three models represent the monolingual baselines. The subsequent models represent the multilingual models

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	**89.38**	28.95	68.36	24.17	**83.44**
IT	85.25	13.81	0.41	0.00	63.39
SL	88.01	25.17	49.69	2.88	77.71
MDA	86.10	**39.16**	68.24	**27.82**	81.09
MDD	83.33	34.16	65.07	24.52	78.20
MRD	87.43	29.90	**69.02**	27.27	82.18

In the case of the Slovenian dataset, the observed phenomena slightly differ from the previous ones. The evaluation results are presented in Table 5. Here, two of the multilingual variants (*MDA* and *RD*) outperform the Slovenian monolingual model overall, despite predicting worse on the [Appropriate] class. Notably, the monolingual model outperforms all models on the [Violent] class, which has not been the case for the other languages. This could be due to language specifics that the multilingual model fail to capture, or to the specifics of the CroSloEngual BERT which is also heavily pre-trained on Croatian and Slovenian data. Once again, the *DA* disagreement strategy shows slight superiority over *RD*.

Finally, Table 6 shows the average scores of all models, achieved by averaging their combined (weighted) F1 scores across all three languages. Summarizing the multilingual superiority, these final results show how monolingual models drastically falter on unseen languages, while the multilingual models have the capacity to reach the inter-annotator agreement ceiling for all languages.

While overall results show that the *Remove Disagreement (RD)* gold standard strategy for incorporating disagreement is best, one should be cautious when making such conclusions. Class-specific results show that the Duplicate All (DA) strategy outperforms in all the classes most relevant to hate speech detection, except for [Appropriate], which is the least relevant class. Another difference is that the *MDA* model involved training longer on the same data which might have resulted in improvement on minority classes and saturation on the majority class. For a future fairer comparison, the fine-tuning process on gold standard data should be adjusted accordingly. The *MDA* variant of the model is available at: https://huggingface.co/IMSyPP/hate_speech_multilingual.

5 Discussion

In recent years, automated hate speech detection has become crucial for moderating online content and mitigating the negative impact on social dynamics within online communities. This

Table 4: Model evaluation results in terms of class-specific F1 scores on the Italian dataset

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	86.27	1.28	1.05	0.00	67.42
IT	**91.32**	**58.46**	59.02	40.34	**83.22**
SL	86.23	0.76	3.25	0.00	65.95
MDA	89.77	58.45	**60.42**	**44.97**	82.38
MDD	88.95	56.04	58.31	39.85	81.19
MRD	90.41	55.46	59.49	38.78	82.50

Table 5: Model evaluation in terms of class-specific F1 scores on the Slovenian dataset

Model	Appropriate	Inappropriate	Offensive	Violent	Total
EN	79.93	3.98	2.34	0.00	53.84
IT	79.84	3.80	1.24	0.00	53.43
SL	**85.70**	43.69	65.26	**29.12**	78.39
MDA	84.30	**45.22**	**69.69**	24.79	**78.88**
MDD	82.33	43.39	68.59	23.84	77.19
MRD	84.98	38.47	68.40	15.50	78.80

Table 6: Average performance of models based on class-weighted F1 scores across three languages

Model	Avg. Weighted F1 Score (all languages)
EN	68.23
IT	66.68
SL	74.02
MDA	80.78
MDD	78.86
MRD	**81.16**

research proposes a novel multilingual hate speech model to address these challenges on a broader scale. The following discusses the main findings.

First, the inter-annotator agreement and the agreement between annotators and models suggest that inter-annotator agreement sets an intrinsic limit on model performance. Models are limited by the quality and consistency of the annotated data, which directly affects their ability to accurately predict unseen data. However, incorporating areas of disagreement into model development can lead to more robust models capable of handling ambiguous cases by employing one of the several available strategies for incorporating disagreement.

Second, the multilingual model consistently surpassed the monolingual baselines, achieving the inter-annotator agreement ceiling across all languages. This success can be attributed partly to the ability to leverage patterns learned from multiple languages, partly to vast amounts of data incorporated into state-of-the-art pre-trained multilingual models, and partially to the class weighting scheme employed in the fine-tuning. These findings support the first research question, demonstrating that a multilingual hate speech classifier trained on diamond standard data outperforms its monolingual counterparts.

Finally, this research contributes substantially to hate speech classification in a multilingual context by introducing a novel multilingual hate speech detection model and making it available on the Hugging Face platform. Our model underscores the importance of incorporating inter-annotator disagreement into model development, challenging the reliance on gold standard data in subjective tasks, such as hate speech detection.

6 Conclusions

This paper advances automatic hate speech detection by introducing a novel multilingual model fine-tuned on the state-of-the-art XLM-R transformer. By leveraging multilinguality, the model significantly outperforms monolingual baselines, demonstrating its effectiveness across diverse linguistic contexts. This highlights the potential of multilingual approaches in improving hate speech detection, especially in scenarios where content spans multiple languages.

Additionally, this research incorporates inter-annotator disagreement into the fine-tuning process using diamond standard data, offering a valuable alternative to traditional gold-standard models. By embracing rather than ignoring annotator disagreement, the model better reflects the nuances of subjective annotations, enhancing its real-world applicability. However, while this approach shows promise, annotator disagreement continues to present challenges, indicating that further work is needed to fully address its impact on model performance.

Future research could extend this work by evaluating the models on additional languages, exploring alternative baseline models, refining strategies for incorporating annotator disagreement and handling minority classes. As online hate speech extends its impact, developing robust, multilingual content moderation systems is crucial to maintaining safe and inclusive digital environments.

7 Acknowledgments

The authors acknowledge partial financial support from the Slovenian Research Agency (research core funding no. P2-103).

References

[1] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: a closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 45–54.

[2] Alexander Brown. 2017. What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, 36, 561–613.

[3] Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40, 108–118.

[4] Alexis Conneau et al. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR, abs/1911.02116*.

[5] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

[6] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115, 16, E3635–E3644.

[7] Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 681–695.

[8] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[9] Andraž Pelicon, Syrielle Montariol, and Petra Kralj Novak. 2023. Don't start your data labeling from scratch: opsala-optimized data sampling before labeling. In *International Symposium on Intelligent Data Analysis*. Springer, 353–365.

[10] Juan Manuel Pérez et al. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11, 30575–30590.

[11] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55, 477–523.

[12] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: a survey. *Journal of Artificial Intelligence Research*, 72, 1385–1470.

[13] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, 19–26.

[14] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, e598.