Perspectives on Hate: General vs. Domain-Specific Models

Giulia Rizzi[*†], **Michele Fontana**[*], **Elisabetta Fersini**[*]
[*] University of Milano-Bicocca, Milan, Italy,
[†] Universitat Politècnica de València, Valencia, Spain
{g.rizzi10, m.fontana36}@campus.unimib.it, elisabetta.fersini@unimib.it

Abstract

The rise of online hostility, combined with broad social media use, leads to the necessity of the comprehension of its human impact. However, the process of hate identification is challenging because, on the one hand, the line between healthy disagreement and poisonous speech is not well defined, and, on the other hand, multiple socio-cultural factors or prior beliefs shape people's perceptions of potentially harmful text. To address disagreements in hate speech identification, Natural Language Processing (NLP) models must capture several perspectives. This paper introduces a strategy based on the Contrastive Learning paradigm for detecting disagreements in hate speech using pre-trained language models. Two approaches are proposed: the General Model, a comprehensive framework, and the Domain-Specific Model, which focuses on more specific hate-related tasks. The source code is available at https://github.com/MIND-Lab/Perspectives-on-Hate.

Keywords: Hate Speech, Disagreement, Contrastive Learning

1. Introduction

With the widespread use of social media, the opportunities to share people's experiences and opinions have grown rapidly. As a consequence, hatred on social media is growing accordingly, with people sharing hateful content towards various targets and minorities. To ensure the continued shared of knowledge and ideas and improve individual and social well-being in the online environment, it is critical to understand the potential harm that hate content can cause on a human level. However, as people use online forums and Social Media to express themselves and engage in debate, the distinction between healthy disagreement and toxic speech becomes increasingly blurred. Moreover, individuals' susceptibility to objectionable content is substantially influenced by their cultural beliefs and origins, emphasizing the importance of considering various perceptions (Sang and Stanton, 2022; LaFrance and Roberts, 2019; Sap et al., 2021).
Addressing disagreement, especially in the context of hate speech identification has received more attention in recent years. Nevertheless, the development of Natural Language Processing (NLP) models capable of completely capturing and representing diverse perspectives is critical. Various approaches have been proposed to address disagreements in hate speech identification, and explored the area of perspectivism (Akhtar et al., 2021; Sachdeva et al., 2022; Uma et al., 2021). According to recent studies, it may be beneficial to consider the exploration of more elaborate and established techniques, such as integrated gradients or uncertainty quantification (Astorino et al., 2023; Davani et al., 2022; Rizzi et al., 2023). The identification of disagreements among hateful statements and the identification of disagreement-related aspects would lead to more reliable benchmarks. Moreover, it would allow the definition of specific annotation policies (e.g., adding more annotators, removing samples from the dataset that need annotation, etc.) to be adopted for contents that are likely to cause disagreement among readers. In this paper, we exploit the Contrastive Learning paradigm to predict Disagreement in hateful content. In particular, we exploit pre-trained large language models for hate speech detection and leverage the embedding representation derived from this model to accurately predict disagreement among annotators. We propose two different approaches with distinct characteristics:

- **General Model**: a comprehensive approach, combining multiple tasks (e.g. aggressive, offensive, and abusive language detection) under the umbrella of hate speech identification (Poletto et al., 2021). This inclusive viewpoint enables the model to effectively capture the subtle manifestations of hate across multiple linguistic dimensions and different languages, resulting in a more robust and versatile solution for identifying and treating various forms of harmful text.

- **Domain-Specific Model**: The Domain-Specific Model represent a more refined approach, focusing solely on elements that share specific characteristics. This approach focuses on instances of the same hate-related task that share homogenous aspects such as language, type of text, and hate target, recognizing the close relationship between those

NLPerspectives 2024@LREC-COLING 2024, pages 78–83

characteristics and annotator disagreement on hate speech.

The paper is organized as follows: Section 2 provides an overview of the state of the art. Section 3 describes the adopted datasets. Section 4 digs into the specifics of the proposed approach. The obtained results are presented in Section 5. Finally, Section 6 summarizes the findings of this study and outlines future investigations.

2. Related Works

Over the years, significant progress has been made in the development of automatic hate content detection systems, exploiting advances in Natural Language Processing (NLP), machine learning, large language models, and deep learning technologies (Mozafari et al., 2020; Alatawi et al., 2021; Saleh et al., 2023). However, hate speech detection, like many natural language tasks, is characterized by intrinsic ambiguity or subjectivity (Uma et al., 2021). These characteristics have led to datasets with multiple annotations that incorporate varied annotator perspectives and understandings or with confidence levels associated with labels. The representation of annotators' disagreement has found utility in three ways: (i) to enhance the quality of the dataset by removing instances marked by annotator disagreement (Beigman Klebanov and Beigman, 2009), (ii) to weight instances during training aiming at prioritizing those with higher confidence levels (Dumitrache et al., 2019), or (iii) to directly train a machine learning model from disagreement without considering aggregated labels (Uma et al., 2021; Fornaciari et al., 2021). While prior research focused on utilizing disagreement information, limited attention has been given to predicting and explaining annotators' disagreement. An important contribution in the field is represented by the SemEval 2023 Task 11 (Leonardelli et al., 2023) where the main goal is to model the disagreement between annotators on different types of textual messages. A first insight in explaining disagreement sources is represented by (Astorino et al., 2023). The authors leverage integrated gradients to detect both disagreement and hate speech and introduce a *filtering strategy* for textual constituents that aids in explaining hateful messages. In this paper, we investigate whether is possible to grasp disagreement from pre-trained language models fine-tuned for the hate-detection task, exploiting Contrastive Learning strategies.

3. Dataset

We employ four benchmark datasets from SemEval 2023 Task 11 focused on Learning With Disagreement (LWD) (Leonardelli et al., 2023), each exhibiting diverse characteristics such as types (social media posts and conversations), languages (English and Arabic), goals (misogyny, hate speech, offensiveness detection), and annotation methods (experts, specific demographic groups, and general crowd). In particular, we used Hate Speech on Brexit (HS-Brexit) (Akhtar et al., 2021), Arabic Misogyny and Sexism (ArMIS) (Almanea and Poesio, 2022), ConvAbuse (Cercas Curry et al., 2021) and Multi-Domain Agreement (MD-Agreement) (Leonardelli et al., 2021). A summary of the datasets is presented in Table 1.

All datasets feature hard-labels (hateful/non-hateful) and soft-labels (disagreement) for each instance. The purpose of this work is to discern agreement and disagreement rather than different levels of disagreement, therefore the number of annotators is not taken into account. The disagreement prediction is treated as a binary task. Therefore, an *agreement label* was derived from the soft-label by setting the value to (+) when there is 100% agreement between the annotators, regardless of the value of the hard label; it is set to to (-) otherwise.

4. Disagreement Estimation

The proposed approach exploits Contrastive Learning techniques that allow the comparison among multiple instances (in contrast with the pairwise comparison of the previous approach). The proposed approach includes an initial fine-tuning on hate detection task and a subsequent Disagreement predictions based on the extracted embeddings. The main phases can be summarized as follows:

1. **Fine-tuning of a pre-trained LM**: The *bert-base-multilingual-cased* has been fine-tuned to distinguish hateful content from non-hateful ones (considering the provided hard labels), proposing a loss function that is grounded on the Binary Cross Entropy and InfoNCE[1] (Khosla et al., 2020) specifically adapted for the considered problem:

$$\begin{aligned}\mathcal{L} =& \lambda L_{bce} + (1-\lambda) L_{InfoNCE} = \\ &= -\lambda \sum_{s} t(s) \log(p(s)) + \\ &+ (1-\lambda) \left(-\log \frac{e^{s \cdot k^{pos}/\tau}}{\sum\limits_{k^{neg} \in K} e^{s \cdot k^{neg}/\tau}} \right)\end{aligned} \tag{1}$$

[1] In order to reinforce the impact of the Contrastive Loss InfoNCE, the hyperparameter λ has been set to 0.3. The fine-tuning has been performed for 4 epochs, adopting a learning rate of 3e-5

Dataset	Language	Task	Annotators	Pool Ann.	% of items with full agr.	Agreement Distribution (Test Set)
HS-Brexit (Akhtar et al., 2021)	En	Hate Speech	6	6	69%	116/168
ArMis (Almanea and Poesio, 2022)	Ar	Misogyny and sexism detection	3	3	86%	92/145
ConvAbuse (Cercas Curry et al., 2021)	En	Abusive Language detection	2-7	7	65%	727/840
MD-Agreement (Leonardelli et al., 2021)	En	Offensiveness detection	5	>800	42%	1292/3057

Table 1: Datasets characteristics.

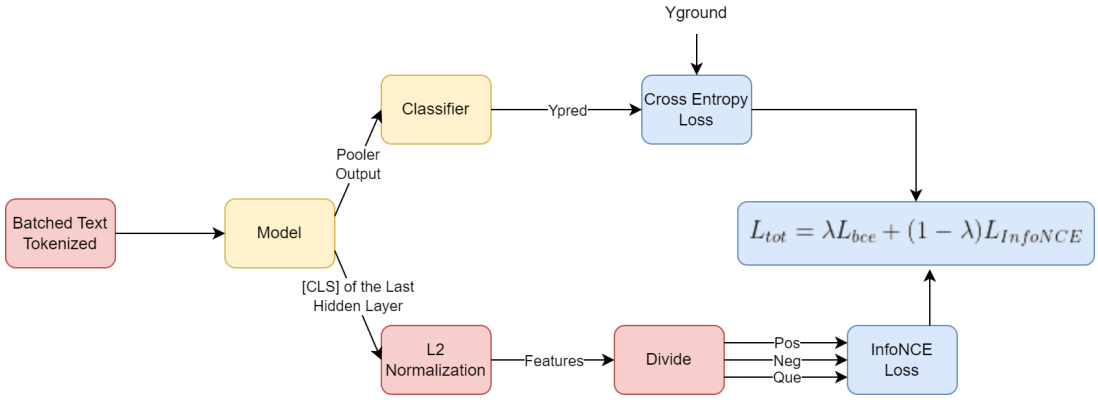

Figure 1: Schematic representation of the Fine-Tuning step.

where s indicates a given sample in the dataset, $t(s)$ denotes the target distribution, $p(s)$ represents the prediction probability distribution, k^{pos} is an instance in the dataset that has the same ground truth label of s, k^{neg} denotes an instance in the dataset that has the opposite ground truth label with respect to s, K is the set of instances in the dataset that have label opposite to s, and τ is the temperature.

The procedure is, in fact, composed of two parts. The first part allows to compute the Binary Crossentropy Loss while the second part exploits information derived from the representation of the [CLS] token in the last model layer. The Binary Crossentropy Loss has the main goal of minimizing the difference between the prediction probabilities and truth values, while the InfoNCE is aimed at maximising the agreement between positive samples and minimizing the agreement between the negative ones in the learned representation. In this way, The derived features are then normalized with L2 regularization to extract *query, positive* and *negative* features, used for computing the InfoNCE. The the fine-tuning phase is summarized in Figure 1.

2. **Similarity Matrix definition**: The fine-tuned model has been used to generate embeddings[2] for the samples in the training and test set in order to define a similarity matrix. The last contains embedding distances computed towards cosine similarity.

3. **Disagreement prediction**: For each instant in the test set, disagreement is predicted starting from the distribution of samples with agreement and with disagreement in the closer neighborhood. Two different strategies have been proposed, distinguishing the definition of the neighborhood:
General Model. The General Model takes a comprehensive approach, combining multiple activities under the umbrella of hate speech identification. This framework incorporates tasks linked to aggressive, offensive, and abusive language, relying on the idea that these behaviors frequently share a common foundation in manifestations of hatred, disregarding the targetted minority. This inclusive viewpoint enables the model to effectively capture the subtle manifestations of hate across multiple linguistic dimensions, different languages, and towards several targets, resulting in a more robust and versatile solution for identifying and treating various forms of harmful speech. According to this rationale, for each instance in the test set, the corresponding neighbor is computed in order to include instances that appear in the overall training set (i.e. achieved via the union of the four training datasets).
Domain-Specific Model. The Domain-Specific Model takes a more refined approach, focusing solely on elements that share specific characteristics. This approach focuses

[2]The embedding representation has been obtained merging the last seven layers of the model.

Dataset	Approach	P+	R+	F+	P−	R−	F−	Macro F
HS-Brexit	m-BERT	0.85	0.69	0.76	0.51	0.73	0.60	0.68
	General Model	0.78	0.83	0.80	0.56	0.48	0.52	0.66
	Domain-Specific Model	0.80	0.94	0.86	0.77	0.46	0.58	**0.72**
	(Astorino et al., 2023)	0.84	0.78	0.81	0.57	0.67	0.62	0.71
ArMIS	m-BERT	0.60	0.27	0.37	0.32	0.65	0.43	0.40
	General Model	0.63	0.95	0.75	0.17	0.02	0.03	0.39
	Domain-Specific Model	0.65	0.88	0.75	0.48	0.19	0.27	0.51
	(Astorino et al., 2023)	0.67	0.75	0.71	0.47	0.38	0.42	**0.56**
ConvAbuse	m-BERT	0.87	0.99	0.93	0.33	0.03	0.05	0.49
	General Model	0.87	0.97	0.92	0.21	0.04	0.07	0.50
	Domain-Specific Model	0.71	0.13	0.22	0.88	0.99	0.93	0.58
	(Astorino et al., 2023)	0.94	0.70	0.80	0.27	0.72	0.40	**0.60**
MD-Agreement	m-BERT	0.43	0.34	0.38	0.58	0.68	0.63	0.50
	General Model	0.66	0.53	0.59	0.70	0.80	0.74	**0.67**
	Domain-Specific Model	0.66	0.53	0.59	0.70	0.80	0.75	**0.67**
	(Astorino et al., 2023)	0.54	0.52	0.53	0.66	0.68	0.67	0.60

Table 2: Comparison of the different approaches on the test set. **Bold** denotes the best approach according to the F1-Score.

on instances of the same hate-related task (i.e. aggressiveness, general hatred, or abusive language identification). Furthermore, the Domain-Specific Model focuses on data that shares homogenous aspects such as language, type of text (e.g. Tweets, discussion, etc.), and target, recognizing the close relationship between those characteristics and annotator disagreement on hate speech. A given term can be, in fact, interpreted as controversial and generate disagreement on a dataset that focuses on hate towards a specific task (e.g. misogyny identification) and neutral in different datasets with different characteristics (e.g. racism detection). As a result, when developing this strategy, the datasets have not been combined. For each instance in the test set, the corresponding neighborhood is computed in order to include only instances that appear in the respective training set in order to guarantee the comparison with samples that share similar characteristics (i.e., topic, type, language, etc.). In both cases, the hyperparameter *n* that defines the numerosity of the selected neighborhood has been estimated towards a grid search approach.

The estimated configurations are summarized in Table 3.

dataset	n
ArMIS	22
HS-Brexit	50
ConvAbuse	19
MD_Agreement	105
Overall Datasets	59

Table 3: Estimated Hyperparameter

Once the neighbor has been selected, the final disagreement label is predicted evaluating the number of samples with agreement and the number of samples with disagreement in the selected neighborhood. In particular, if the difference between the number of samples with agreement and the number of samples with disagreement in the selected neighbor is smaller than τ[3], then the predicted label is set to disagreement. On the other hand, if the difference between samples with agreement and samples with disagreement in the selected neighbor is bigger than τ the prediction is computed toward majority voting (i.e., Agreement if the majority of samples in the selected neighbor are labeled as agreement, Disagreement otherwise).

5. Results

In this section, the results obtained by the proposed approaches are reported. We measured Precision (P), Recall (R) and F-Measure (F), distinguishing between Agreement (+) and Disagreement (-) labels and reporting also the Macro F-Measure.

Table 2 summarized the achieved results. We also report results achieved by (Astorino et al., 2023) for a state-of-the-art comparison. This last approach exploits integrated gradients from pretrained language models in the recognition of disagreements' causes and hate speech contents. One of the main contribution is given by the introduction of a filtering strategy that contributes to explain hateful messages via textual constituents. It can be easily noted that, in the majority of the considered datasets, the proposed approach "Domain-Specific Model" outperforms the considered baseline m-BERT and achieves competitive results with (Astorino et al., 2023). It is also interesting to highlight that the *Domain-Specific Model* outperforms the *General* one in all the proposed datasets. The *Domain-Specific Model* is designed to concentrate on a single dataset, allowing it to define its representation based on its unique characteristics, such as

[3] n has been estimated via Grid Search. It has been set to 7 for the General approach and to 2 for the Domain-Specific approach.

the type of text, target of hate, language, and more. This leads to a better understanding of the terms in relation to the hate task at hand, and therefore to higher performance with respect with the *General* approach. More important, although the proposed approach is comparable or in some cases even better than (Astorino et al., 2023), it has the great advantage of being computationally less complex than (Astorino et al., 2023) thanks to presence of a simpler objective function compared to the two fine-tuning losses in the considered baseline model.

6. Conclusions and Future works

The proposed paper introduces a novel approach for detecting disagreement in hateful content. The method exploits contrastive learning techniques applyed to pre-trained language models to predict both hate speech and potential disagreement arising from different readers. The propoesed approach outperforms m-BERT and achieve competitive results on four benchmark datasets from the Learning With Disagreement (LeWiDi) task at Semeval (Leonardelli et al., 2021). Overall, the proposed approach demonstrates the potential to encaplulate Contrastive Learning tecnique in Natural Language tasks. Future work could focus on exploring the applicability of the proposed approach to other datasets in different domain and expanding the scope to include multimodal data analysis.

Acknowledgments

The work of Elisabetta Fersini has been partially funded by MUR under the grant REGAINS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca and by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR))

7. Bibliographical 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.

Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

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.

Alessandro Astorino, Giulia Rizzi, and Elisabetta Fersini. 2023. Integrated gradients as proxy of disagreement in hateful content. In *CEUR WORKSHOP PROCEEDINGS*, volume 3596. CEUR-WS. org.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Anca Dumitrache, FD Mediagroep, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of NAACL-HLT*, pages 2164–2170.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Marianne D LaFrance and Sarah J Roberts. 2019. The role of bias in hate speech detection. *Journal of Language Aggression and Conflict*, 7(1):1–20.

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 Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi).

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

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.

Giulia Rizzi, Alessandro Astorino, Paolo Rosso, and Elisabetta Fersini. 2024. Unraveling disagreement constituents inhateful speech. In *Advances in Information Retrieval*, pages 21–29, Cham. Springer Nature Switzerland.

Giulia Rizzi, Alessandro Astorino, Daniel Scalena, Paolo Rosso, and Elisabetta Fersini. 2023. Mind at semeval-2023 task 11: From uncertain predictions to subjective disagreement. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 556–564.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, pages 425–444. Springer.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Alexandra 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.