Integrated Gradients as Proxy of Disagreement in Hateful Content

Alessandro Astorino[1], Giulia Rizzi[1,2] and Elisabetta Fersini[1,*]

[1]University of Milano-Bicocca, Milan, Italy
[2]Universitat Politècnica de València, Valencia, Spain

Abstract

Online platforms have increasingly become hotspots to spread not only opinions but also hate speech, posing substantial obstacles to developing constructive and inclusive online communities. In this paper, we propose a novel approach that leverages the integrated gradients of pre-trained language models to automatically predict both hate speech and the potential disagreement that can arise from readers. The integrated gradient attributions are used to shed light on the model's decision-making process attributing importance scores to individual tokens and enabling the identification of crucial factors contributing to disagreement and hate speech classifications. The integrated gradients' straightforwardness allows for the recognition of fundamental causes of disagreements and hate speech content. By adopting an interpretable approach, we bridge the gap between model predictions and human comprehension. Our experimental results highlight the effectiveness of our approach, outperforming traditional BERT models and state-of-the-art methods in both prediction tasks.

Keywords

Learning with Disagreement, Integrated Gradients, Hateful Content

1. Introduction

In the modern era, human beings are constantly subject to absorbing content of various kinds generated and shared on the web. To ensure the sustainability of continuously produced information and promote individual and societal well-being in the context of online content is important to recognize where hate content can harm from a personal perspective. Different individuals, according to their cultural beliefs and backgrounds, may be more or less susceptible to potentially offensive content. It is, therefore, necessary to safeguard the perceptions of different individuals by defining Natural Language Processing (NLP) models that are able to capture and model different perceptions. How to deal with disagreement, in particular related to hate speech detection problems, is a topic that has attracted increasing interest during the last few years [1, 2, 3, 4]. Although a good number of approaches able to deal with disagreement in hate speech detection problems have been proposed [5, 6, 7, 8], only a few of them have been focused on really modelling perspectivism.

Recognizing potential disagreements within hateful content, especially in identifying controversial elements, is of paramount importance for multiple reasons. When the possibility of disagreement arises in hateful texts shared on social media platforms (e.g. Twitter), it becomes critical to have a service that recognizes if that text written in that manner causes disagreement and works as a filter for these texts based on a personal perspective

Moreover, a detoxification strategy could be implemented to notify the authors of user-generated texts, cautioning them about the potential perception of their content as hateful by certain readers, and suggesting revisions for the original message. Identifying disagreements within hateful sentences and determining the associated disagreement-related elements can significantly contribute to the creation of reliable benchmarks. Primarily, for contents prone to disagreements, specific annotation policies can be implemented (e.g., involving more annotators, excluding samples requiring annotation from the dataset, etc.). Additionally, annotators could be provided with targeted cues to focus on particular constituents that may be perceived differently by readers (e.g., underlining words, hashtags, or emojis identified as disagreement-related elements warranting careful evaluation).

In this paper, we try to connect hate speech and disagreement by determining which hateful constituents can contribute more to predicting disagreement. In particular, we combine pre-trained language models and integrated gradients providing the following main contributions:

- a *filtering strategy* of textual constituents that contributes remarkably to explain hateful messages;
- a *unified model* that, considering the prediction of the hateful contents and the selected explanations,

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.
a.astorino2@campus.unimib.it (A. Astorino); g.rizzi10@campus.unimib.it (G. Rizzi); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8987-100X (E. Fersini)

Dataset	Language	Types	Training Size	Task	Annotators	Pool Ann.	% Full Agreement
HS-Brexit [9]	En	Tweets	1,120	Hate Speech	6	6	69%
ArMis [10]	Ar	Tweets	943	Misogyny and sexism detection	3	3	86%
ConvAbuse [11]	En	User-agent dialogues	4,050	Abusive Language detection	2-7	7	65%
MD-Agreement [12]	En	Tweets	10,753	Offensiveness detection	5	>800	42%

Table 1
Datasets characteristics.

predicts if disagreement could arise when reading such contents;

The rest of the paper is organized as follows. In Section 2 an overview of the state of the art is provided. The adopted datasets are described in Section 3. In Section 4 the proposed approach is detailed. The results achieved by the proposed approaches are reported in Section 5. Finally, conclusions and future research directions are drawn in Section 6.

2. Related Work

The fast rise of social media and online communication platforms has changed the way people communicate, exchange information, and express their ideas, while simultaneously increasing the spread of hate speech. Hateful content includes a wide range of various forms of offensive, abusive, and discriminatory language targeted at individuals or groups based on their race, religion, ethnicity, gender, or other protected characteristics. The propagation of hate speech online has major implications, perpetuating discrimination, stoking antagonism, and instigating violence, necessitating the urgent need for effective anti-hate speech solutions. Over the years, significant progress has been made in developing automatic hate content detection systems that leverage advancements in Natural Language Processing (NLP), machine learning, and deep learning techniques. In this section, we highlight some of the state-of-the-art approaches and methodologies employed in hate speech detection. The dominant approach for hate speech detection is represented by supervised learning [13, 14]. In particular, the approaches based on Language Models (LM) [15, 16, 17] have shown promising results in capturing contextual information and semantic relationships, leading to improved classification performance.

One of the key challenges in hate speech detection is the ability to make sense of the context in which the offensive language is used. Researchers have explored context-aware models [18, 19] that consider the surrounding text or conversation to make more accurate predictions. This can exploit speaker attributes, or discourse patterns to better grasp the intended meaning and differentiate between hate speech and non-hateful expressions. In recent years, hate speech detection has extended to encompass multimodal data analysis to keep up with the increasing usage of images and videos in online communication. Combining textual information with visual cues from images and videos has shown promise for improving the accuracy and granularity of hate speech identification systems [20, 14]. An increasing number of datasets are collecting multimodal examples of hate content ranging from memes [20, 21] to advertisements [22] and videos [23].
The latest datasets are addressing the problem of hate speech under the Learning with Disagreements paradigm reporting information both on the *hard label* (usually obtained through majority voting) and on the *soft label* (with all the annotators' labels or a confidence level attached to the labels). The inclusion of different perspectives allows us to address the subjectivity of the task by representing the multiple perceptions of the annotators with different points of view and understanding [24]. The information that represents annotators' disagreement is not only used to improve the quality of the dataset [25] but also in the training process by weighting the samples according to their disagreement values [26] or by directly training from disagreement, without considering any aggregates label [27, 28].

3. Dataset

The four benchmark datasets provided by SemEval 2023 task 11 related to Learning With Disagreements [29] have been considered in order to address the problem of predicting disagreement in hateful content. The datasets have different characteristics for what concerns language, type, and goal as summarized in Table 1. All the datasets have been adapted by the challenge organizer to share a common structure for what concern the textual input and the hard and soft labels (additional dataset-specific attribute are present). Since in this work, the disagreement prediction is addressed as a binary task, an *agreement label* has been derived from the *soft label*. This is because taking the levels of disagreement into account requires knowledge of the number of annotators, which is not taken into account at this time since the objective is to distinguish agreement and disagreement and not the various levels of disagreement. In particular, the agreement label is set equal to $(+)$ when there is a 100% agreement between the annotators, regardless of the value of the hard label, while equal to $(-)$ in all the other cases.

4. Proposed Approach

The proposed approach aims at addressing the tasks of predicting both disagreement and hate speech while maintaining the method fully interpretable through the adoption of integrated gradients. Integrated gradients are used to shed light on the model's decision-making process attributing importance scores to individual tokens and enabling the identification of crucial factors contributing to the model's decision.

In particular, the proposed approach is composed of four main steps:

1. **Fine-tuning of a pre-trained LM:** the multilingual BERT (m-BERT) has been fine-tuned to distinguish hateful content from non-hateful ones. The textual input (i.e. the tweet or the conversation depending on the dataset) has been given as input to the m-BERT model with a final sigmoid layer. Additionally, to overcome the datasets' class imbalance, in the training phase, the loss function has been penalized accordingly to the class distribution. The optimal decision threshold has been determined according to the Youden's J statistics [30]. The statistics, which is a linear combination of sensitivity and specificity, is maximized by evaluating several cut-offs.
2. **Estimation of the attribution score:** the attribution score for each textual constituent has been estimated using the integrated gradients presented in [31] on the fine-tuned model. This attribution score assumes values from -1 to 1, 1 means that that token has a high contribution to the prediction of the model and -1 the opposite. A visual representation of the integrated gradient on two available samples is reported in Figure 1. On one hand, each attribution score allows us to identify those tokens that contribute more to the final prediction, and on the other hand, those compositions of tokens characterized by divergent values make the content controversial potentially leading to disagreement. The variability and the magnitude of attribution values within a text are subsequently exploited to detect a potential disagreement.
3. **Filtering constituents:** the integrated gradient's attribution scores have been used to filter out those tokens that do not bring a significant contribution to explain the target label. In particular, let t_{im} be the i-th token within a text m and s_{im} the corresponding attribution score. The token t_{im} is considered significant and maintained for the subsequent disagreement model if $s_{im} \geq \tau$, otherwise the token is removed from the original input text. In our case study, τ is a specific threshold estimated according to a grid search approach.
4. **Extraction of latent representations:** the tokens considered significant according to the previous step are used to extract the corresponding latent representation of the filtered sentence from the fine-tuned m-BERT model.
5. **Creation of the disagreement input space:** the latent representation obtained at the previous step is used according to the following strategies:
   - *Filtered Embeddings:* the embedding of the filtered sentence is obtained by fine-tuned model on hate and used to train the subsequent disagreement model.
   - *Predicted Label:* the Boolean labels predicted by the model fine-tuned to distinguish hateful from non-hateful messages are included in the input space for training the disagreement model.
   - *Distribution values:* the distribution probability obtained through the sigmoid layer of the fine-tuned models has been alternatively considered.
6. **Training of the disagreement model:** the derived input space (latent representation of the selected token, concatenated with the predicted label or probability distribution) is given as input to a trivial Neural Network with the following structure to predict disagreement labels:
   - *Input layer*: layer that reflects the shape of the input, with Relu as activation function and dropout of 0.7;
   - *Hidden layer*: layer that halves the size of the input with Relu and dropout of 0.7;
   - *Output layer*: one output neuron with a sigmoid function to predict the final agreement/disagreement.

The entire proposed approach is synthesized in Figure 2.

5. Experimental Results

In this section, the results obtained by the proposed approach are reported. We measured Precision (P), Recall (R) and F-Measure (F), distinguishing between hateful $(+)$ and not hateful $(-)$ labels and reporting also the

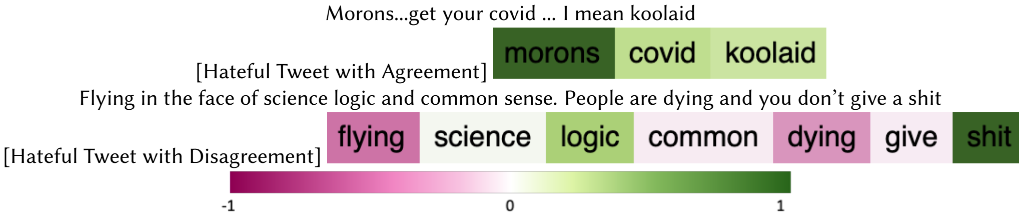

Figure 1: Visual representation of the integrated gradients on sentences from the MD-Agreement dataset. Positive values are represented with the green colour, negative values are associated with the pink colour, while the white colour is used for attribution values equal to zero.

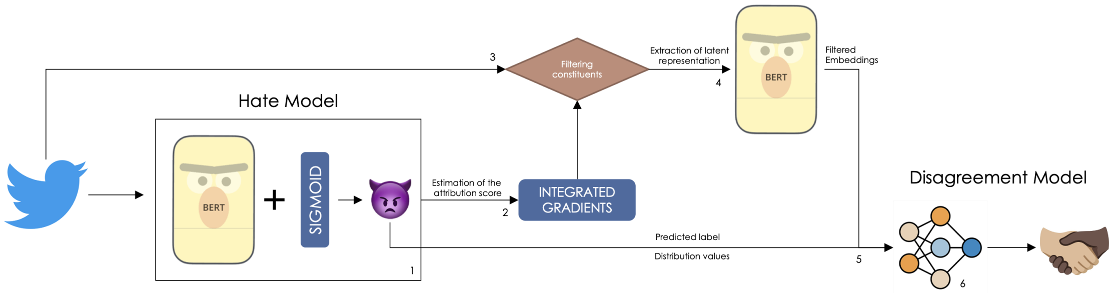

Figure 2: Proposed Approach

Macro F-Measure. We show in Table 2 the performance achieved by the fine-tuned model on the hate speech detection task. The achieved results denote good prediction capability, especially for the negative class (non-hateful). This behaviour is mainly due to the unbalanced nature of the datasets and in some cases to the limited number of instances available.

Now, we report in Table 3 the performance on the disagreement prediction, distinguishing however between agreement $(+)$ and disagreement $(-)$. The results of the proposed method are shown according to the input space previously described. In particular, we report:

- **m-BERT**: a baseline m-BERT model fine-tuned according to the disagreement label;
- **NN + Filt**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores and trained on the disagreement label. This configuration corresponds to the one described in step 5(a);
- **NN + Pred**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores with an additional Boolean feature denoting the label predicted by the fine-tuned model on the hate. This configuration corresponds to the one described in step 5(b);
- **NN + Dist**: a neural network that takes as input the embedding representation of the sentence composed of the tokens selected according to the attribution scores with two additional features denoting the probability distribution associated with the labels predicted by the fine-tuned model on the hate. This configuration corresponds to the one described in step 5(c);

In order to understand whether the proposed approaches obtain significant results compared with m-BERT, a McNemar Test has been performed. In particular, the McNemar Test has been adopted to perform a pairwise comparison between the m-BERT predictions and each of the proposed strategies according to a confidence level equal to 0.95. If a given model outperforms m-BERT and its error distribution is different compared to m-BERT, then the corresponding F1-Score is marked with a wildcard symbol $(*)$ in Table 3.

It can be easily noted that, in the majority of the considered datasets, all of the proposed approaches significantly outperform the considered baseline m-BERT. It is also interesting to highlight that, considering the datasets are even more unbalanced and with a very limited number of samples, the proposed approach NN-Dist tends to achieve more balanced performance between the two labels than the other methods. The McNemar test confirms that the NN-Dist strategy is not only the best-performing one but also that the predictions are different with respect

Dataset	P+	R+	F+	P−	R−	F−	Macro F
HS-brexit	0.37	0.78	0.58	0.97	0.84	0.90	0.70
ArMIS	0.55	0.76	0.64	0.75	0.54	0.63	0.63
ConvAbuse	0.77	0.50	0.60	0.90	0.97	0.93	0.77
MD-Agreement	0.73	0.56	0.63	0.80	0.90	0.85	0.74

Table 2
Model performance on the hate speech detection task on the test set.

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
	NN + Filt	0.69	0.86	0.83	0.62	0.50	0.55	0.69
	NN + Pred	0.79	0.86	0.83	0.62	0.50	0.55	0.69
	NN + Dist	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
	NN + Filt	0.64	0.93	0.76	0.50	0.11	0.18	0.47*
	NN + Pred	0.66	0.84	0.73	0.46	0.25	0.32	0.53*
	NN + Dist	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
	NN + Filt	0.94	0.59	0.73	0.23	0.76	0.25	0.54*
	NN + Pred	0.92	0.67	0.78	0.24	0.65	0.35	0.56*
	NN +Dist	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
	NN + Filt	0.47	0.71	0.57	0.67	0.43	0.53	0.55*
	NN + Pred	0.53	0.52	0.52	0.66	0.66	0.66	0.59*
	NN + Dist	0.54	0.52	0.53	0.66	0.68	0.67	**0.60***

Table 3
Comparison of the different approaches on the test set for disagreement detection. **Bold** denotes the best approach according to the F1-Score, while underline represents the best approach according to the disagreement label. (*) denotes that model outperforms M-BERT and obtains results that are statistically different.

to the ones given by m-BERT. This implies that the performance of the proposed approach could be considered statistically significant.

An additional remark concerns the relationship that exists between the disagreement prediction model and the model able to predict hateful content. The performances of the proposed models are strictly related to the recognition capabilities of the model fine-tuned to distinguish hateful content from non-hateful ones. Improving the recognition capabilities of the hateful model is expected to increase the recognition potential of the proposed disagreement models.

For what concerns the errors of the most promising approach, i.e., NN-Dist, we can highlight that on the HS-Brexit dataset, most of the misclassifications are due to the absence of relevant information. In particular, in 70% of the misclassified samples, there are references to users and links that have been omitted, making the understanding of the context even more complex. Regarding the ArMis dataset, most of the errors are related to the implicit language used to express hateful content against women (no explicit insults or sexist expressions are used, but more subtle misogynous samples are reported). In ConvAbuse, the misclassification of the proposed approach is mainly due to the reduced number of tokens of the text. In fact, 40% of the original text contains less than 3 tokens, making difficult the prediction of disagreement. Finally, in MD-Agreement the error rate is quite higher (42.79%) compared to the other datasets. In this scenario, the misclassified samples are almost balanced between the two classes, (i.e., 0.45% for the agreement and 55% for the disagreement class). The main reason behind the high classification error can be found in the different arguments covered by the dataset. This suggests that disagreement is not only related to different beliefs or backgrounds but also to specific discussed topics.

6. Conclusions and Future works

The proposed paper introduces a novel approach for detecting disagreement in hateful content. The method leverages integrated gradients from pre-trained language models to predict both hate speech and potential disagreement arising from different readers. The approach is evaluated on four benchmark datasets related to Learning

With Disagreements, and the results show that the proposed method outperforms the baseline m-BERT model in disagreement prediction tasks. One of the proposed strategies, namely NN + Dist, performs particularly well and achieves statistically significant improvements compared to a baseline model based on m-BERT. Overall, the proposed approach demonstrates the potential to predict disagreement in hateful content compared to bert. Future work could focus on exploring the applicability of the proposed approach to other languages and expanding the scope to include multimodal data analysis, considering the increasing use of images and videos in online communication.

Acknowledgments

The work of Elisabetta Fersini has been partially funded 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), and by MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

[1] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[2] P. Kralj Novak, T. Scantamburlo, A. Pelicon, M. Cinelli, I. Mozetič, F. Zollo, Handling disagreement in hate speech modelling, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2022, pp. 681–695.

[3] P. Fortuna, M. Domínguez, L. Wanner, Z. Talat, Directions for nlp practices applied to online hate speech detection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11794–11805.

[4] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, M. Poesio, SemEval-2023 task 11: Learning with disagreements (LeWiDi), in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2304–2318.

[5] D. Grötzinger, S. Heuschkel, M. Drews, CICL_DMS at SemEval-2023 task 11: Learning with disagreements (le-wi-di), in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1030–1036. URL: https://aclanthology.org/2023.semeval-1.141.

[6] M. Sullivan, M. Yasin, C. L. Jacobs, University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 978–985.

[7] S. Shahriar, T. Solorio, SafeWebUH at SemEval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 94–100. URL: https://aclanthology.org/2023.semeval-1.12.

[8] E. Gajewska, eevvgg at SemEval-2023 task 11: Offensive language classification with rater-based information, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 171–176. URL: https://aclanthology.org/2023.semeval-1.24.

[9] S. Akhtar, V. Basile, V. Patti, Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021. arXiv:2106.15896.

[10] D. Almanea, M. Poesio, ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2282–2291. URL: https://aclanthology.org/2022.lrec-1.244.

[11] A. Cercas Curry, G. Abercrombie, V. Rieser, 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, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7388–7403. URL: https://aclanthology.org/2021.emnlp-main.587. doi:10.18653/v1/2021.emnlp-main.587.

[12] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10528–10539. URL: https://aclanthology.org/2021.emnlp-main.822. doi:10.

18653/v1/2021.emnlp-main.822.

[13] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[14] A. Chhabra, D. K. Vishwakarma, A literature survey on multimodal and multilingual automatic hate speech identification, Multimedia Systems (2023) 1–28.

[15] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation in social media based on bert model, PloS one 15 (2020) e0237861.

[16] H. S. Alatawi, A. M. Alhothali, K. M. Moria, Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert, IEEE Access 9 (2021) 106363–106374.

[17] H. Saleh, A. Alhothali, K. Moria, Detection of hate speech using bert and hate speech word embedding with deep model, Applied Artificial Intelligence 37 (2023) 2166719.

[18] M. Fernandez, H. Alani, Contextual semantics for radicalisation detection on twitter (2018).

[19] M. Bilal, A. Khan, S. Jan, S. Musa, Context-aware deep learning model for detection of roman urdu hate speech on social media platform, IEEE Access 10 (2022) 121133–121151. doi:10.1109/ACCESS.2022.3216375.

[20] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.

[21] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[22] F. Gasparini, I. Erba, E. Fersini, S. Corchs, et al., Multimodal classification of sexist advertisements., in: ICETE (1), 2018, pp. 565–572.

[23] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee, Hatemm: A multi-modal dataset for hate video classification, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1014–1023.

[24] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, The Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:https://doi.org/10.1613/jair.1.12752.

[25] B. Beigman Klebanov, E. Beigman, From annotator agreement to noise models, Computational Linguistics 35 (2009) 495–503.

[26] A. Dumitrache, F. Mediagroep, L. Aroyo, C. Welty, A crowdsourced frame disambiguation corpus with ambiguity, in: Proceedings of NAACL-HLT, 2019, pp. 2164–2170.

[27] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[28] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, et al., 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, 2021.

[29] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), 2023. arXiv:2304.14803.

[30] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35.

[31] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.