From Explanation to Detection: Multimodal Insights into Disagreement in Misogynous Memes

Giulia Rizzi[1,2,*], Paolo Rosso[2] and Elisabetta Fersini[1,*]

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

Abstract

Warning: This paper contains examples of language and images that may be offensive.
This paper presents a probabilistic approach to identifying the disagreement-related elements in misogynistic memes by considering both modalities that compose a meme (i.e., visual and textual sources). Several methodologies to exploit such elements in the identification of disagreement among annotators have been investigated and evaluated on the Multimedia Automatic Misogyny Identification (MAMI) [1] dataset. The proposed unsupervised approach reaches comparable performances, and in some cases even better, with state-of-the-art approaches, but with a reduced number of parameters to be estimated. The source code of our approaches is publicly available[†].

Keywords

Disagreement, Perspectivism, Multimodal, Misogyny

1. Introduction

Hate detection has been a serious concern in recent years, penetrating internet platforms and causing harm to individuals across various communities. Users found in the online environment new modes of representation to express various types of hatred, including the more deeply rooted ideologies and beliefs with historical origins, for example towards women [2].
Detecting abusive language has become an increasingly important task. The challenges introduced by the new modes of representation, which require a multimodal analysis, are further compounded when considering the subjectivity of the task. The subjectivity of the task derives from the fact that individuals' perception of what characterizes a message of hate varies widely. Such diversification is reflected in the labeling phase in the form of disagreement among annotators. Identifying elements within the sample that can lead to disagreement is of paramount importance for several reasons. For content that can lead to disagreement, specific annotation policies might be introduced, and the number of annotators might be enlarged to capture multiple perspectives [3, 4, 5].
In this work, we propose a methodology to identify the disagreement-related elements in multimodal samples by exploring both visual and textual elements in the Multimedia Automatic Misogyny Identification (MAMI) dataset [1]. Moreover, four different strategies to exploit the presence of such elements in the identification of disagreement are investigated.

2. Related Works

Many natural language tasks, such as hate speech detection, humor detection, and sentiment analysis, involve subjectivity since they require an interpretation based on human judgment, cultural context, or personal opinion [6]. Such phenomenon is reflected in the dataset through multiple labels from different annotators or via the inclusion of a confidence level to ground truth labels. Labels derived from different interpretations are therefore able to capture multiple perspectives and understandings [6]. Information about annotators' disagreement has primarily been exploited as a means to improve data quality by excluding controversial instances [7, 8]. Alternatively, aiming at improving model performances, different strategies have been developed to exploit disagreement information in the training phase. For instance, in [9], the authors assign weights to instances to prioritize the ones with higher confidence levels. Another commonly adopted strategy [6, 10] aims at directly learning from disagreement without considering any aggregated label. While a considerable amount of research has been conducted to understand the reasons behind annotators' disagreement [11, 12, 8] and to leverage disagreement when training classification models [13, 14, 15, 16, 17, 18, 19], there has been comparatively little attention devoted to the explanation and a priori recognition of disagreement in hateful content. A taxonomy of possible reasons leading to annotators' dis-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
g.rizzi10@campus.unimib.it (G. Rizzi); prosso@dsic.upv.es (P. Rosso); elisabetta.fersini@unimib.it (E. Fersini)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-8922-1242 (P. Rosso); 0000-0002-8987-100X (E. Fersini)

†https://github.com/MIND-Lab/From-Explanation-to-Detection-Multimodal-Insights-into-Disagreement-in-Misogynous-Memes

agreement has been proposed by [12]. Such taxonomy articulates four macro categories of reasons behind disagreement: sloppy annotations, ambiguity, missing information, and subjectivity. Moreover, the authors evaluate the impact on classification performance of the different types.

Only recently, works have focused on the task of explaining disagreement [20, 21, 22, 23]. In [21], the authors propose exploratory text visualization techniques as a method for analyzing different perspectives from annotated data. In [22], the authors identify textual constituents that contribute to hateful message explanation by exploiting integrated gradients within a filtering strategy. A more recent approach [23] proposes a probabilistic semantic approach for the identification of disagreement-related constituents (e.g. textual elements) in hateful content. Overall, the findings indicate that, while LLM can yield promising results, comparable outcomes can be attained with less complex strategies and fewer computational resources. While previous research has concentrated on the analysis of textual disagreement, this study represents, to the best of our knowledge, a first insight into the explanation of multimodal disagreement. In particular, we have revised and extended to the multimodal environment the methodology proposed in [23] in order to consider not only textual elements but also visual ones.

3. Proposed Approach

3.1. Identification of Disagreement-Related Elements

The first phase of the proposed approach aims to evaluate the relationship between elements (both visual and textual) that compose a meme and annotators' disagreement. Preliminary preprocessing operations have been performed before identifying disagreement-related elements. For what concerns the textual components, preprocessing operations have been performed (i.e., tokenization, lemmatization, lower casing and stop word removal) to identify a valid set of tokens[1] that might be related to disagreement. Considering the image component, the set of 14 human readable concepts (*tags*) identified by [24] to capture specific characteristics of misogynous content has been adopted. As proposed by the authors, tags were extracted via the Clarifai API [25]. The preprocessing steps allowed us to extract a list of visual and textual elements from each meme in the dataset.

In order to measure the relationship among each element in the memes and the disagreement among annotators, the approach proposed in [23] has been extended to a multimodal scenario. In particular, [23] introduces a methodology to identify disagreement related constituents that, however, is limited to textual content. The approach includes a strategy to identify disagreement-related textual constituents and an approach for generalization towards unseen textual constituents. Both methods have been extended to a multimodal scenario in order to identify disagreement related elements both in textual and visual sources that compose a meme.

Given an element e, a corresponding Element Disagreement Score (*EDS(e)*) has been computed according to the following equation:

$$EDS(e) = P(Agree|e) - P(\neg Agree|e) \quad (1)$$

where $P(Agree|e)$ represents the conditional probability that there is agreement on a meme given that the meme contains the element e. Analogously, $P(\neg Agree|e)$ denotes the conditional probability that there is no agreement on a meme given that, that meme, contains the element e. Given that EDS represents a difference between two complementary probabilities, it is bounded within the range of -1 to +1. A higher positive score indicates stronger agreement between annotators, whereas a lower negative score suggests disagreement.

The score can be estimated on the training data and exploited to identify additional disagreement-related elements on unseen memes.

3.2. Disagreement identification

Once the Element Disagreement Scores have been estimated for each visual and textual element in the training dataset, they can be exploited to qualify the level of disagreement on unseen samples. Analogously to what was carried out in [23], different aggregation strategies have been investigated, relying on the hypothesis that the identified elements can be exploited for identifying the disagreement thanks to their different distribution in samples with and without an agreement.

For each meme in the test set, the corresponding list of elements and the corresponding Elements Disagreement Score estimated on the training data have been extracted. In particular, for each meme, the textual and visual elements have been identified and paired with the corresponding score when available. The Multimodal Disagreement Score (MDS) has been estimated according to the following strategies: **Sum**, **Mean**, **Median**, and **Minimum**. A threshold τ has been estimated according to a grid-search approach for each strategy.

A qualitative evaluation, comprehensive of a comparison with the specific misogynistic terminology and an evaluation of the keyword included in the dataset creation phase, has been performed to assess the quality of the EDS, while both the F1-score for the two considered

[1] To guarantee a more robust evaluation, tokens that appear less than 10 times in the dataset have been removed.

classes (agreement (+) and disagreement (-)) and a global F1-score have been computed to validate the MDS.

3.3. Generalization towards unseen elements

The score estimation is strongly based on what is observed in the training data, resulting in the lack of scores for any elements that do not appear in the training samples. This is particularly relevant for textual components rather than visual ones. In fact, while we can assume an open-word vocabulary (where a few terms on unseen data can not appear in the training set) for the textual source, we limited the visual tags to closed-word settings (only 14 tags can be considered both in training and unseen memes). Since we need to generalize only on unseen textual constituents, for each (unseen) textual element $\hat{e}$, an approximated EDS score has been computed as follows:

- **Embeddings of the training lexicon:** the contextualized embedding representation of each textual element e has been obtained via mBert [26]. An average embedding vector representation $\vec{\mathbf{x}}_e$ is computed to jointly represent multiple embedding representations of e derived by the different contexts where it occurs. In particular, given an element e and N sentences containing it, its vector representation $\vec{\mathbf{x}}_e$ is obtained by a simple average $\vec{\mathbf{x}}_e = \sum_{i=1}^{N} \vec{\mathbf{v}}_i / N$, where $\vec{\mathbf{v}}_i$ is the constituent contextualized embedding vector related to the i^{th} occurrence of e and obtained through mBert.
- **Embeddings of unseen term:** given an unseen textual element $\hat{e}$ within a given sentence, its contextualized embedding representation has been computed via mBert [26].
- **Most similar constituent:** given an unseen textual element $\hat{e}$ with the corresponding embedding $\vec{\mathbf{v}}_{\hat{e}}$ and the average embedding of a training element e, the set D of most similar constituents to $\hat{e}$ is determined according to:

$$D = \bigcup_{e} \{e | cos(\vec{\mathbf{x}}_e, \vec{\mathbf{v}}_{\hat{e}}) \leq \psi\} \quad (2)$$

  where $cos(\vec{\mathbf{x}}_e, \vec{\mathbf{v}}_{\hat{e}})$ is the cosine similarity between the average contextualized embedding representation of element e and $\hat{e}$, and ψ is a grid search estimated threshold.
- **Unseen terms score:** the EDS score for an unseen textual element $\hat{e}$ is computed as the weighted average of the most similar constituents e of the training lexicon:

$$EDS(\hat{e}) = \frac{\sum_{e \in D} [cos(e, \hat{e}) \cdot EDS(e)]}{\sum_{e \in D} cos(e, \hat{e})} \quad (3)$$

- **Multimodal Disagreement Score with unseen constituents:** All the above-proposed strategies for MDS estimation have been extended to also include elements that do not belong to the training lexicon and for which the EDS score has been estimated. In particular, given a multimodal sample s, the aggregation functions presented in Section 3.2 will in this case consider the EDS values of both seen (by considering the $EDS(e)$) and unseen (by considering the $EDS(\hat{e})$) elements. Such generalized aggregation functions will be later referred to through the prefix $G-$.

4. Results

The proposed approach has been evaluated on the Multimedia Automatic Misogyny Identification (MAMI) Dataset [1] consisting of 10.000 memes for training and 1.000 memes for testing [2]. The dataset comprises a range of memes that exemplify various forms of misogyny, including shaming, stereotyping, objectification, and violence. Each meme has been labeled by three crowdsourced annotators for misogynistic content[3], with an estimated Fleiss-K [27] coefficient equal to 0.5767.

In particular, the proposed approach has been adopted to estimate an Element Disagreement Score (EDS) for each element and, consequently, MDS for each meme in the dataset.

Table 1 reports the top-10 highest positive and highest negative disagreement scores derived for the textual component. We can notice how terms that are rarely linked with misogynistic messages (e.g., *flu*) and terms commonly used to address women in a harmful way (e.g., *whale*) also exploiting stereotypes (e.g. *gamer* and *programmer*), achieve a high positive score, indicating a strong relation with the agreement. Additionally, some personal names of famous people (i.e., *Bernie* and *Miley*) appear within the ranking. In particular, such names

[2] Although both a training and a test dataset are provided, only the training dataset is adopted, as the proposed work is focused on the analysis and prediction of disagreement and the test dataset is constructed to include only samples with complete agreement. The training dataset, instead, is characterized by 65% of data with complete agreement. Therefore, it has been divided in order to isolate the 90% for token estimation and the remaining 10% for the evaluation.

[3] Additionally, a boolean disagreement label has been derived to represent complete agreement among annotators. In particular, this last label is set to 1 if all the annotators have indicated the same label, to 0 otherwise.

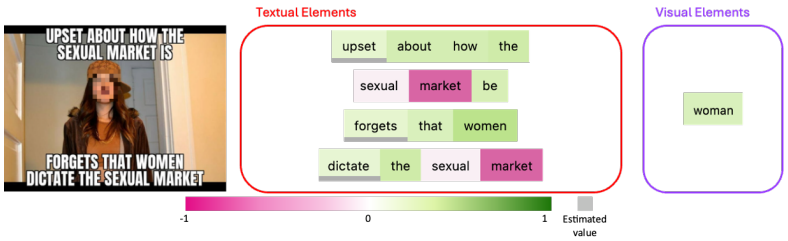

Figure 1: Visual representation of disagreement scores distinguishing among textual and visual elements. Positive and negative scores are represented with green and pink respectively. The gray bar denotes elements for which the EDS has been estimated, while the white color represents elements with an EDS equal to zero.

Term	EDS	Term	EDS
flu	1.00	market	−0.64
folk	1.00	fetish	−0.60
bug	1.00	nut	−0.57
Bernie	1.00	hotel	−0.50
whale	1.00	apologize	−0.45
feeling	0.90	Miley	−0.45
gamer	0.87	lonely	−0.43
rest	0.87	award	−0.43
programmer	0.87	coke	−0.43
san	0.83	blowjob	−0.43

Table 1
Terms with the highest positive and lowest negative scores

Tag	EDS	Tag	EDS
crockery	0.49	dishwasher	0.00
nudity	0.46	broom	0.14
cat	0.46	dog	0.20
car	0.43	child	0.23
kitchenutensil	0.41	woman	0.26

Table 2
Tags with the highest positive and lowest negative scores

might appear in memes as the target of a hateful message, referring to their personal life, physical appearance, or specific events that involved them. As a consequence, depending on the reasons that lead to such criticism (gender, physical appearance, and personal choices for Miley Cyrus vs. political stance and career, without the same gendered connotations, for Bernie Sanders) there might be disagreement about misogyny.

Table 2 reports the top-5 highest positive and highest negative disagreement scores derived for the visual component. It is easy to notice how all the scores are positive and achieve small values, denoting a tendency of such tags to be weakly related to the agreement label.

Figure 1 reports an example of a meme with disagreement along with the visual representation of the EDS of its textual and visual elements. Moreover, as highlighted with a grey bar, some of the reported scores have been estimated. Such scores correspond, in fact, to constituents that are not present in the training dataset and for which it was not possible to calculate the ESD score. The visual representation of the scores related to such elements corresponds to the score obtained through the estimation strategy. Overall, it is easy to notice the presence of elements strongly related to disagreement (i.e., *sexual* and *market*), highlighted in pink.

The concept of the "sexual marketplace" is often the subject of debate, particularly in relation to its intersection with misogynistic ideologies [28, 29]. Some supporters, often aligned with "manosphere" or "red pill" ideologies, argue that the sexual marketplace disproportionately empowers women, giving them more control over sexual selection and relationships, which can disadvantage men. On the other hand, critics assert that this perspective reduces human relationships to transactional exchanges and objectifies both genders, ultimately reinforcing misogynistic attitudes. This last viewpoint asserts that framing relationships in market terms devalues emotional connection and perpetuates harmful stereotypes about women's worth being tied solely to their sexual desirability. Achieved results suggest the ability of the approach to detect such variety in interpretations and reflect them within the EDS scores.

Figure 2 reports two memes that share the same text and a different image. Despite such commonalities, the memes have been labeled differently: while the first meme has been labeled as misogynous by 2 annotators out of 3, the second one has been unanimously labeled as non-misogynous. Since such memes share a common textual representation, the derived textual elements and textual-EDS are also equal, resulting in an indistinguishable representation that is ineffective for disagreement identification. Moreover, although the memes differ in the visual content, resulting in different tags and, therefore, different textual-EDS, as previously mentioned, such a component alone is not sufficient for disagreement prediction.

The findings demonstrate the necessity of joint considera-

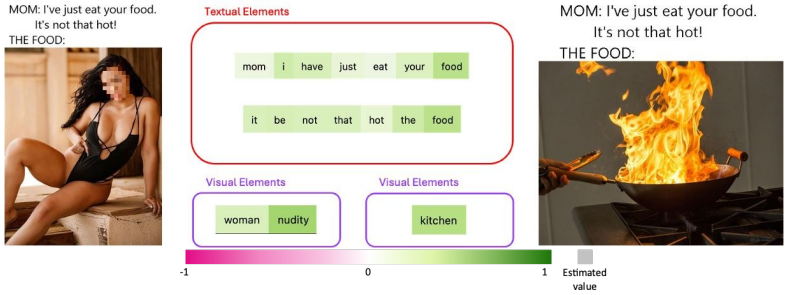

Figure 2: Visual representation of disagreement scores distinguishing among textual and visual elements for two samples in the dataset. Positive and negative scores are represented with green and pink respectively. The white color represents elements with EDS equal to zero.

tion of both visual and textual modalities for the purpose of predicting disagreements.

All the proposed aggregation strategies have been implemented, both considering the modalities individually and jointly. Table 3, and Table 4 summarise achieved results on disagreement identification considering only the score related to elements derived from the textual component (i.e., terms) and only the scores of elements derived from the visual component (i.e., tags) respectively. Table 5 instead summarises results achieved by the aggregation of the scores derived from all the elements (i.e., terms and tags). Results achieved on the textual component only highlight G-Mean as the most performing approach. Overall, the estimation strategy results in an improvement of performances up to 6%, confirming the ability of the proposed strategy to capture disagreement relationships for unseen terms. Furthermore, BERT [30][4] has been reported as a state-of-the-art baseline for unimodal textual classification. Achieved results show how BERT performs better on the majority class, struggling in predicting the disagreement class. The proposed approach, instead leads to performance more balanced among the two classes.

Table 4 reports the performances of the different approaches for disagreement identification considering the visual component only. However, while the Sum approach (i.e., the most performing approach among the tag-based) demonstrates satisfactory performance in identifying positive instances (achieving an F1+ of 0.69), it exhibits considerable difficulty in accurately identifying negative instances.

Finally, Table 5 reports the performances of the different approaches for disagreement identification jointly considering both modalities. Furthermore, for a better comparison of the performance achieved by the proposed approach, a state-of-the-art baseline for multimodal classification has been implemented: CLIP [31][5].

Approach	ψ	τ	F1+	F1-	F1 Score
Sum	-	3.1	0.61	0.39	0.50
Mean	-	0.2	0.78	0.20	0.49
Median	-	0.2	0.07	<u>0.79</u>	0.43
Minimum	-	-0.1	0.29	0.75	0.52
G-Sum	0.8	3.1	0.65	0.37	0.51
G-Mean	0.8	0.2	0.73	0.34	**0.53**
G-Median	0.8	0.2	0.77	0.21	0.49
G-Minimum	0.8	-0.1	0.75	0.30	0.52
BERT [30]	-	-	0.80	0.00	0.40

Table 3
Comparison of the different approaches for disagreement detection considering the textual component only. The agreement label (+) indicates complete annotator agreement, regardless of the misogyny value, while the agreement label (-) denotes samples without complete agreement. **Bold** denotes the best approach in terms of F1-score, and <u>underline</u> represents the best approach according to the disagreement label. ψ and τ represent the best hyperparameters estimated via a greed search approach.

The inclusion of both modalities leads to a slight improvement in performances that, however, remain quite poor, highlighting the difficulty of the task. The inclusion of the unseen constituents estimation leads to an improvement of performance (except for the sum-based method) up to 8% for the mean-based approach. However, the best performances are achieved by the minimum and G-minimum approaches, for which the estimation methodology is not effective. Such behavior may be attributed to the imbalance in the dataset. The larger the number of samples with agreement, the greater the num-

[4]BERT has been implemented and finetuned using the hugging-face framework with default hyperparameters. We adopted "bert-base-cased" available at https://huggingface.co/google-bert/bert-base-cased.

[5]CLIP has been implemented and finetuned using the huggingface framework with default hyperparameters. In particular, we used the version available at https://huggingface.co/openai/clip-vit-large-patch14 to which we concatenated a linear layer for binary classification.

Approach	ψ	τ	F1+	F1-	F1 Score
Sum	-	0.3	0.69	0.34	**0.52**
Mean	-	0.3	0.41	0.48	0.45
Median	-	0.3	0.41	0.49	0.40
Minimum	-	0.3	0.35	0.49	0.40

Table 4
Comparison of the different approaches for disagreement detection considering the visual component only. The agreement label (+) indicates complete annotator agreement, regardless of the misogyny value, while the agreement label (-) denotes samples without complete agreement. **Bold** denotes the best approach in terms of F1-score, and underline represents the best approach according to the disagreement label. ψ and τ represent the best hyperparameters estimated via a greed search approach.

Approach	ψ	τ	F1+	F1-	F1 Score	Param.
Sum	-	3.4	0.63	0.36	0.50	|E|
Mean	-	0.2	0.79	0.13	0.46	|E|
Median	-	0.2	0.80	0.05	0.42	|E|
Minimum	-	0	0.69	0.42	**0.55**	|E|
G-Sum	0.8	3.6	0.64	0.35	0.49	179M
G-Mean	0.9	0.2	0.70	0.39	0.54	179M
G-Median	0.9	0.2	0.77	0.21	0.49	179M
G-Minimum	0.1	0	0.69	0.42	**0.55**	179M
CLIP [31]	-	0.5	0.63	0.42	0.52	428M

Table 5
Comparison of the different approaches for disagreement detection considering both textual and visual components. The agreement label (+) indicates complete annotator agreement, regardless of the misogyny value, while the agreement label (-) denotes samples without complete agreement. **Bold** denotes the best approach in terms of F1-score, and underline represents the best approach according to the disagreement label. ψ and τ represent the best hyperparameters estimated via a greed search approach, and E is the set of elements.

ber of agreement-related terms that impact the estimation phase. Consequently, the estimation of scores for unseen elements is likely to be positive due to the aforementioned imbalance. Overall, the findings suggest that achieving a balanced performance remains challenging.

5. Conclusion and Future Works

This paper proposes a probabilistic approach to identify disagreement-related elements in multimodal content. The proposed approach allows for the identification of elements that could be used as a proxy to identify samples that might be perceived differently by the annotators, and therefore, that could lead to disagreement. Achieved results highlight the difficulty of the task, denoting the need for a more advanced approach. Future work will include different strategies for image analysis in order to provide a better description of the image itself in all the elements that compose it. Furthermore, a study of the compositionality might be carried out to better represent the relationship among such elements inside the meme. The sense of a meme is often derived from the meanings of its individual parts (i.e. the image and text) and the way they are combined. By analyzing how different elements interact and contribute to the overall message, it is possible to gain a deeper understanding of how the meaning is represented within the different modalities. This will help in identifying complex patterns and improve the accuracy of classification models.

Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

[1] 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.

[2] L. Fontanella, B. Chulvi, E. Ignazzi, A. Sarra, A. Tontodimamma, How do we study misogyny in the digital age? a systematic literature review using a computational linguistic approach, Humanities and Social Sciences Communications 11 (2024) 1–15.

[3] 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.

[4] C. van Son, T. Caselli, A. Fokkens, I. Maks, R. Morante, L. Aroyo, P. Vossen, GRaSP: A multilayered annotation scheme for perspectives, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Por-

torož, Slovenia, 2016, pp. 1177–1184. URL: https://aclanthology.org/L16-1187.

[5] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: a survey, Language Resources and Evaluation (2024) 1–28.

[6] 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.

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

[8] Y. Sang, J. Stanton, 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, Springer, 2022, pp. 425–444.

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

[10] 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.

[11] L. Han, E. Maddalena, A. Checco, C. Sarasua, U. Gadiraju, K. Roitero, G. Demartini, Crowd worker strategies in relevance judgment tasks, in: Proceedings of the 13th international conference on web search and data mining, 2020, pp. 241–249.

[12] M. Sandri, E. Leonardelli, S. Tonelli, E. Ježek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2428–2441.

[13] S. Shahriar, T. Solorio, Safewebuh at semeval-2023 task 11: Learning annotator disagreement in derogatory text: Comparison of direct training vs aggregation, arXiv preprint arXiv:2305.01050 (2023).

[14] E. Gajewska, eevvgg at SemEval-2023 task 11: Offensive language classification with rater-based information, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of 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. doi:10.18653/v1/2023.semeval-1.24.

[15] 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 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 978–985.

[16] A. de Paula, G. Rizzi, E. Fersini, D. Spina, et al., Ai-upv at exist 2023–sexism characterization using large language models under the learning with disagreements regime, in: CEUR WORKSHOP PROCEEDINGS, volume 3497, CEUR-WS, 2023, pp. 985–999.

[17] J. Erbani, E. Egyed-Zsigmond, D. Nurbakova, P.-E. Portier, When multiple perspectives and an optimization process lead to better performance, an automatic sexism identification on social media with pretrained transformers in a soft label context, Working Notes of CLEF (2023).

[18] M. E. Vallecillo-Rodríguez, F. del Arco, L. A. Ureña-López, M. T. Martín-Valdivia, A. Montejo-Ráez, Integrating annotator information in transformer fine-tuning for sexism detection, Working Notes of CLEF (2023).

[19] G. Rizzi, M. Fontana, E. Fersini, Perspectives on hate: General vs. domain-specific models, in: Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024, 2024, pp. 78–83.

[20] M. Michele, V. Basile, F. M. Zanzotto, et al., Change my mind: How syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints, in: 1st Workshop on Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference, LREC 2022 Workshop, European Language Resources Association (ELRA), 2022, pp. 117–125.

[21] L. Havens, B. Bach, M. Terras, B. Alex, Beyond explanation: A case for exploratory text visualizations of non-aggregated, annotated datasets, in: G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 73–82. URL: https://aclanthology.org/2022.nlperspectives-1.10.

[22] A. Astorino, G. Rizzi, E. Fersini, Integrated gradients as proxy of disagreement in hateful content, in: CEUR WORKSHOP PROCEEDINGS, volume 3596, CEUR-WS. org, 2023.

[23] G. Rizzi, A. Astorino, P. Rosso, E. Fersini, Unrav-

eling disagreement constituents in hateful speech, in: European Conference on Information Retrieval, Springer, 2024, pp. 21–29.
[24] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474.
[25] Clarifai, Clarifai guide, ???? URL: https://docs.clarifai.com/.
[26] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[27] J. L. Fleiss, Measuring nominal scale agreement among many raters., Psychological bulletin 76 (1971) 378.
[28] D. Ging, A. Neary, Gender, sexuality, and bullying special issue editorial, 2019.
[29] E. Ignazzi, A. Sarra, L. Fontanella, et al., Exploring misogyny through time: From historical origins to modern complexities, Philosophies of Communication (2023) 195–214.
[30] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.