## Analysis of Annotator Demographics in Sexism Detection

Narjes Tahaei and Sabine Bergler CLaC Lab Concordia University, Montreal n\_tahaei@encs.concordia.ca, sabine.bergler@concordia.ca

## Abstract

This study explores the effect of annotators' demographic features on labeling sexist content in social media datasets, focusing specifically on the EXIST dataset (Plaza et al., 2023), which includes direct sexist messages, reports, and descriptions of sexist experiences and stereotypes. We investigate how various demographic backgrounds correlate with annotation outcomes and examine methods to incorporate these features into BERT-based model training. Experiments show that adding demographic information improves performance in classifying sexism and assessing intention of the author.

#### 1 Introduction

According to the United Nations definition, *gender-based violence* includes violence that is directed against a woman because she is a woman or that affects women disproportionately, and, as such, is a violation of their human rights<sup>1</sup>. A report published by Amnesty International<sup>2</sup> found that 23% of women using Twitter reported experiencing online abuse or harassment at least once. Such violence and abuse on social media significantly undermines women's rights to express themselves equally, freely, and without fear.

The EXIST shared task (Plaza et al., 2023) aims to identify and categorize sexism on Twitter. Developing any dataset for topics such as sexism, sarcasm, or hate speech is challenging, since it is difficult to determine a ground truth for these topics (Gordon et al., 2022; Plaza et al., 2023). One person may find a tweet sexist, while another may find it acceptable. Traditionally, machine learning models take the majority vote over all labels, effectively ignoring differences in annotators' backgrounds, which might, however, influence their perspectives. It also disregards minority opinions when most annotators agree on one label but a few dissent. The resulting datasets have been described as partially subjective and not fully reliable for downstream applications by (Rottger et al., 2022).

In response, (Uma et al., 2021) introduced the Learning With Disagreement paradigm (LwD), which involves training systems on datasets that include all annotators' perspectives, aiming to reflect the diversity of views. Evaluation with a soft metric enables ambiguity-aware models to compare the probability distributions of labels they generate (soft labels) to the full distribution provided by annotators using cross-entropy. This step away from hard gold labels to distributions provides the possibility of better modeling both, potential ambiguity of the wording as well as disagreements in the judgments. The EXIST sexism task adopted LwD and incorporated the annotators' demographic information into the training and test sets.

We show here that for the EXIST dataset the performance of sexism detection and the assessment of the author's intentions improves when demographic information is included for classification. We explore different methods to integrate the demographic features and find that including it as additional input to a BERT model enhances performance. While we use gold labels during training, which are crowd labels aggregated into hard labels using majority voting, we include detailed information from the annotations in the training process.

The goal is to investigate whether there is a correlation between annotator judgments and demographic features in two ways. First, we exploit the demographic features from the training data for improved classification. Secondly, we give a first exploration of potential bias in the dataset. *Bias* here refers to uneven distributions and has no value judgment attached to it from the outset. While EX-

<sup>&</sup>lt;sup>1</sup>https://www.un.org/womenwatch/daw/cedaw/ recommendations/recomm.htm

<sup>&</sup>lt;sup>2</sup>https://www.amnesty.org/en/documents/amr51/ 4723/2021/en/

IST is a renowned shared task aimed at identifying sexism in social media, no paper has yet analyzed the impact of annotators' features on the labeling process in this dataset.

## 2 Related Work

#### 2.1 Sexism Detection

Recent years have seen an increase in the availability of sexism-related datasets. (Waseem and Hovy, 2016) introduced a dataset of tweets labeled with the categories *racism, sexism,* and *neutral*. Similarly, MeTwo (Rodríguez-Sánchez et al., 2020) provides a Twitter corpus in Spanish, categorizing tweets as *sexist, non-sexist*, or *uncertain*.

Social media platforms serve as forums for sharing both sexist content and testimonials of sexism encountered by women. Distinguishing between these two types of messages is crucial, as is understanding the definitions of sexism within these datasets. Some datasets focus on detecting misogyny or hatred towards women (Anzovino et al., 2018; Guest et al., 2021; Pamungkas et al., 2020). The EXIST dataset classifies both direct sexist tweets and reports or descriptions of sexist experiences as sexist messages. (Chiril et al., 2020) introduced a dataset of French tweets annotated to identify either reports of sexist experiences or sexist messages. In their dataset, a tweet was considered sexist if it explicitly targeted someone or described a target implicitly. For example, the tweet My boss asked me: "Who's going to cook for your husband when you're away?" is a report and might trigger different reactions from the recipient compared to a direct sexist message. The EXIST dataset covers a wide range of sexism, from explicit to other subtle or even benevolent expressions that involve implicit sexist behaviors.

#### 2.2 Learning with Disagreement

A growing body of research focuses on developing training methods that do not rely on a single label for each sample (Abercrombie et al., 2023; Mostafazadeh Davani et al., 2022; Kairam and Heer, 2016; Leonardelli et al., 2023).

According to (Uma et al., 2021), approaches to learning from disagreement in crowd annotations can be categorized into four broad methods. First, some methods automatically aggregate annotations into a single label for each instance, assuming an objective truth exists for each instance (such as majority voting). Second, other methods also assume a gold label exists but recognize it may not always be recoverable due to coder disagreement; these methods filter out or weigh items with excessive disagreement (Whitehill et al., 2009). Third, some approaches learn a classifier directly from crowd annotations by assigning probabilistic scores to each label, producing soft labels (Rodrigues and Pereira, 2018). Finally, some methods train classifiers using a combination of hard and soft labels (Fornaciari et al., 2021), or integrating gold labels with information from crowd annotations to account for item difficulty or annotator ability (Plank et al., 2014). In our study, we have adopted the latter approach, incorporating annotators' demographics into the training process for sexism identification.

Guided by the assumption that annotators' judgments build on their background, recent studies have explored this further. (Wan et al., 2023) incorporated demographic information of annotators to propose a disagreement predictor framework that gauges annotators' disagreement in subjective tasks.

(Jiang et al., 2021) found significant variations in perceptions of the harmfulness of sexually explicit language across eight countries. (Almanea and Poesio, 2022) developed an Arabic Twitter dataset on sexism and misogyny, demonstrating that annotators' religions correlate with their labels. (Sap et al., 2022) reported strong correlations between annotator identity, beliefs, and toxicity ratings.

Some studies suggest that the nature of disagreement in tasks such as sexism is not based on individual differences, but on social positions (Chulvi et al., 2023; Curry et al., 2024). Sexism is defined not at the individual level but rather based on societal norms (Curry et al., 2024). To cover all perspectives in the annotation process, (Chulvi et al., 2023) proposed considering the attitude of annotators and their behavior toward sexism.

(Curry et al., 2024) proposed that equally considering all annotations' disagreement is not sufficient. For instance, certain terms may be commonly used among African Americans but are inappropriate if used by the broader public. Thus, sexism or racism should be understood as cultural concepts formed in specific contexts, i.e. when computing disagreement, the vote of the impacted group matters more even if they are in the minority. (Gordon et al., 2022) introduced jury learning, a recommender system approach that selects a group of annotators with specified demographic characteristics from a pool of annotators to judge each text.

(Orlikowski et al., 2023) showed that even when socio-demographic information such as gender is included in their toxicity detection model as an additional group-specific linear layer, the average behavior of the female annotators does not necessarily reflect the behavior of individuals in that group.

The EXIST dataset includes a range of demographic information not typically found in other sexism datasets. However, its impact on the EXIST dataset's labeling process has not yet been explored. This study offers an investigation into such a correlation.

# **3** Data Description

### 3.1 Task Description

This study investigates the EXIST 2023 dataset (Plaza et al., 2023). EXIST stands for sEXism Identification in Social neTworks and focuses on identifying explicit and implicit sexism in social media posts.

Task 1 is a binary classification determining whether a tweet represents sexist expressions or behaviors. The sample can be sexist itself, describe a sexist situation where discrimination against women occurs, or criticize a sexist behavior.

Task 2 considers the author's intention motivating the tweet, i.e. it explicitly distinguishes between a sexist tweet and one describing or reporting a sexist experience to criticize sexism. Task 2 is a three-way classification into the categories

Direct The tweet itself is sexist, e.g.

A woman needs love, to fill the fridge, if a man can give this to her in return for her services (housework, cooking, etc), I don't see what else she needs.

*Reported* The tweet reports a sexist situation experienced by a woman or women, e.g. *Today, one of my year 1 class pupils could not believe he'd lost a race against a girl.* 

*Judgemental* The tweet describes sexist situations or behaviors to condemn them, e.g.

As usual, the woman was the one quitting her job for the family's welfare...

#### 3.2 Dataset statistics

The dataset is collected from a wide range of Spanish and English tweets including annotations provided by a diverse group of annotators from various countries.

Each tweet has been annotated by six annotators. The EXIST 2024 dataset includes demographic information for each annotator, such as *id*, *age*, *gender*, *level of study*, *country*, and *ethnicity*. The training dataset includes 6920 tweets which have been annotated by 725 distinct annotators from 45 different countries. The demographic features (scalar representations in our models are in parentheses<sup>3</sup>) provided are

Gender male (1) or female (0)

Age ranges are 18-22 (1), 23-45 (2), and >46 (3)

- Levels of study is less than a high school diploma, high school degree or equivalent, bachelor's degree, master's degree, and doctorate
- *Ethnicity* is *Black or African American, Hispanic or Latino, White or Caucasian, Multiracial, Asian, Asian Indian, or Middle Eastern*

## 4 Methodology

The dataset includes annotations from six annotators for each sample, along with their demographic features. We compare here performance when including and when omitting these demographic features as input for a simple pre-trained model (finetuned RoBERTa (Loureiro et al., 2022) or Multilingual Cased BERT (Devlin et al., 2019)) followed by two feedforward networks for classification.

When no demographic features are introduced to the model:

- **Encoding**<sub>base</sub> Input to BERT is the sample text, input to the classifier is BERT's CLS token.
- **Encoding**<sub>preprocessed</sub> Four tokens ([url], [user], [spanish], [english]) were added as special tokens to the BERT tokenizer. In each sample, URLs and mentions were replaced with [url] and [user] tokens, respectively. The dataset identifies for each tweet whether it is in English or Spanish, thus either [english] or [spanish] were added at the beginning of each instance. Based on findings that word embeddings capture the meaning of emojis better

<sup>&</sup>lt;sup>3</sup>No scalar representations are shown here for the level of study and ethnicity and we do not report on scalar runs for these as they did not improve performance.

than symbols (Tahaei et al., 2022; Mostafavi and Porter, 2021), emoji word embeddings were added to the BERT vocabulary as special tokens, enabling BERT to map a specific identifier to each emoji. The input to BERT is the sample text preprocessed this way and the input to the classifier is BERT's CLS token.

The following describes two models that incorporate annotators' demographic features. In both configurations, each preprocessed sample is replicated six times, corresponding to each annotator. Consequently, the original training set of 6,920 tweets expands to 41,520 instances.

- **Encoding**<sub>features</sub> Input to BERT is the preprocessed sample text. Input to the classifier includes demographic features in a stack of one-hot vectors of each feature (*age, gender, study level,* and *ethnicity*) and the CLS vector. We send additive attention to the classifier. Additive attention (Bahdanau et al., 2015) obtains a weighted average of the representations, where the weights are determined based on the importance of each representation.
- **Encoding**<sub>annotations</sub> Input to BERT concatenates demographic features to the end of each preprocessed sample. Each sample is replicated six times (Uma et al., 2022; Sheng et al., 2008), each with details on gender, age, level of study, and ethnicity for each annotator (see Section 5.1). Both the hard label, which is the aggregated label for that tweet (see Section 5.2), and the annotator's label for each instance were added to each instance. Input to the classifier is the CLS token.

An initial baseline system predicted both sexism and the three intention categories *direct, reported, judgmental* within a single model. However, we observe better results when we reduce the number of non-sexist tweets for Task 2 by first running our Task 1 system in order to predict the sexist tweets and subsequently categorizing only those with our different approaches.

## **5** Experiments

#### 5.1 Feature Representation

To experiment with Encoding<sub>annotations</sub>, we concatenated the demographic features to the sample in different formats.  As usual, the woman was the one quitting her job for the family's welfare., F, 18-22, Hispano or Latino, Bachelor's degree

Example 1 shows the input vector representation, where the meta-labels on the sample correspond to the respective word vectors. The first four demographic features are appended represented by the strings used in the competition data (the last two demographic features are omitted here).

(2) As usual, the woman was the one quitting her *job for the family's welfare.*, 0, 1, 3, 1

Example 2 shows the first four demographic features represented by scalar values (see Section 3.2).

(3) As usual, the woman was the one quitting her job for the family's welfare. <g>female</g>
<a>GenX</a> <e>Hispano or Latino</e>
<s>Bachelor's degree</s>

Example 3 shows the demographic features enclosed in dedicated special tokens that have been added to BERT's special tokens vocabulary.

(4) As usual, the woman was the one quitting her job for the family's welfare. female, GenX, Hispano or Latino, Bachelor's degree

Example 4 shows a variation on Example 1, where 'F' was replaced by the word 'female' and '18-22' was replaced by 'GenX', namely where words replaced symbols. Example 4 representations outperform Example 1 representations and the word encodings of Example 4 are used for the following examples.

(5) As usual, the woman was the one quitting her job for the family's welfare. [SEP] female [SEP] GenX [SEP] Hispano or Latino [SEP] Bachelor's degree

Example 5 is a version of 4 where the demographic features are separated by the pre-trained [SEP] token in the input to BERT.

 (6) As usual, the woman was the one quitting her job for the family's welfare., female, GenX, Hispano or Latino, Bachelor's degree, sexist, direct

Example 6 is similar to 5, but the individual labels of the annotator for the two tasks are also concatenated to the end of the sample.

Models based on representations shown in Example 4, Example 5, and Example 6 perform better than the first three models, validating the use of word representations instead of symbols, pretrained instead of custom separating special tokens, and appending the annotator decision to all training samples as well as the test samples. Our results, detailed in Section 6, are based on these representations.

#### 5.2 Evaluation

Each instance has seven labels in the competition data: one given by each annotator and the hard (majority) label.

For Task 1, a binary classification task, the hard label is determined by a majority vote of the annotators, following the shared task evaluation setting. The class annotated by more than three annotators is selected. If no class receives a majority, meaning no class exceeds the threshold, those instances are removed from the training process. We evaluated the binary task based on the F1 measure.

Task 2 is a three-way classification task for sexist tweets, i.e. annotators label tweets as *direct*, *reported*, *judgmental*, or *non-sexist*<sup>4</sup>. The hard label is assigned if two or more annotators agree. In case of a tie, instances are removed from the training process. We evaluate Task 2 using the average F1 measure (Macro F1).

### 5.3 Training

For this multilingual dataset, we used the multilingual version of BERT (mBERT) (Devlin et al., 2019), a model pre-trained on 104 languages with Wikipedia pages using a masked language modeling objective. We also employed the 'cardiffnlp/twitter-roberta-basesentiment-latest' model (Loureiro et al., 2022), a RoBERTa-base model trained on approximately 58 million tweets and fine-tuned for sentiment analysis. The experiments yielded better performance with the fine-tuned RoBERTa-based model for Task 1 and improved results with mBERT for the bi-lingual Task 2. For comparison, we also experimented with XLM-RoBERTa, but since the performance did not significantly improve, we opted to continue our experiments with the lighter model. We use the HuggingFace implementation of BERT, trained with a batch size of 1 for 8 epochs.

### 6 Results

Table 1 shows results for the best performance in the 2023 competition and seven of our experiments with different demographic feature encodings.

The demographic features consistently improve performance in all their encodings reported here. This indicates that incorporating annotators' demographics generally improves the performance of detecting both sexism and source intention (Tasks 1 and 2) in this dataset. The most significant improvement is observed with the Encoding<sub>annotations</sub> model, where tweets are replicated six times for the six annotators, with the annotator's features concatenated at the end. Although the Encoding<sub>features</sub> model, which adds annotators' features as additional inputs directly to the classifier, enhances the performance of both tasks compared to Encoding<sub>base</sub>, it is outperformed by this Encoding<sub>annotations</sub> model.

Due to the absence of publicly available test data from the shared task organizers, our results are based solely on the publicly available development set. The state-of-the-art model, which won the 2023 shared task, achieved performance scores of 0.81 for the first task and 0.57 for the second task on the test set but did not report performance on the development set.

We experimented with different representations of demographic features, as shown in Examples 1-6. The best performance for both tasks was for representations similar to Examples 4-6, where demographics were added as tokens to the end of instances. For the first and second tasks, using this representation and incorporating *age*, *gender*, *study level*, and *ethnicity* led to a 5% and 7% performance increase, respectively, compared to Encoding<sub>base</sub>. In the preliminary experiments, we included other demographic features from the dataset. However, they did not lead to further improvement and we excluded them from the analysis.

Since each tweet has two labels—one being the gold label derived from majority voting, and the other provided by each annotator—we tested in-corporating the annotators' labels into the input. Along with the tweet and annotators' features for each instance, we also included the label given by that annotator (6). This configuration led to a 1% improvement in performance for the first task, and 4% improvement was observed for the second task.

<sup>&</sup>lt;sup>4</sup>We model Task 2 as a three-way classification after first predicting non-sexist tweets using our Task 1 classifier and classifying the remaining samples as *direct*, *reported*, or *judg-mental*.

Model	Features	Task 1	Task 2
Encoding <sub>base</sub>	-	0.78	0.48
Encoding <sub>features</sub>	age+gender+study+ethnicity	0.80	0.50
<b>Encoding</b> <sub>annotations</sub>	age+gender	0.81	0.48
<b>Encoding</b> <sub>annotations</sub>	age+gender+study+ethnicity	0.83	0.55
Encoding <sub>annotations</sub> ,sep	age+gender+study+ethnicity	0.83	0.53
<b>Encoding</b> <sub>annotations</sub>	age+gender+study+ethnicity+labels	0.84	0.59

Table 1: Experiment results are based on the development set. For Task 1, we used the RoBERTa model fine-tuned on a sentiment dataset, while for Task 2, we used mBERT.

### 7 Analysis

#### 7.1 Annotator demographics as Features

The EXIST dataset is among the recent datasets that adopted the learning with disagreement approach for sexism detection. Our experiments demonstrated that integrating additional demographic information directly into language models significantly enhances performance, proving more effective than incorporating these features separately into classifiers. This underscores the capabilities of the BERT family models, which can achieve impressive results with proper token representations.

Our findings show that adding two demographic features to the model improved performance for the first task, and the inclusion of two more features further increased performance. When annotators' labels were incorporated as an additional feature to the input, there was a slight additional boost in performance. Our experiments also show the critical role of feature representation. Among the various representations tested, adding demographic features as tokens yielded the best results, capitalizing on the inherent power of the corresponding word embeddings. Even when adding annotator's labels as an additional feature, representing them as word tokens (for instance in Example 6 adding a label sexist instead of 1) resulted in slightly better performance.

Including annotator features in the model allows for training on a diverse set of representations for each sample, rather than relying on a single perspective. This leads to better training outcomes and a less biased model, emphasizing the importance of considering annotator demographics in developing more inclusive models.

#### 7.2 Annotator Categories

Studies have shown that annotators' background influences their labeling processes, leading to bias in datasets that may not adequately represent minority voices (Wan et al., 2023; Sap et al., 2022). We examined the distribution of annotations across various demographic features and found that for most features there was about a 10% difference between the lowest and highest groups.

When examining ethnicity, 54% of annotators who identified as multi-racial labeled tweets as sexist, compared to only 41% of annotators who identified as Black or African American. Given the diverse backgrounds of annotators from 45 different countries, we focused on the top nine countries with more than 1,000 annotators each. Notably, 39% of annotators from Poland found tweets to be sexist, while 50% and 51% from the US and Italy identified tweets as sexist. Language also played a role in the labeling process. Spanish-speaking annotators labeled tweets as sexist at a higher rate (49%) compared to English-speaking annotators (41%). Our analysis confirms the importance of considering annotators' demographics in the labeling process to ensure a more representative and less biased dataset.

#### 8 Conclusion and Limitations

The aim of this study is not to reach SOTA in detecting sexism and source intention, but rather to examine the potential advantage of adding annotator features in representing more diverse judgments in the EXIST dataset. We showed that incorporating annotators' demographic information as inputs into the BERT models can enhance the performance of models in detecting sexism and understanding the author's intention in the EXIST dataset. Additionally, our findings indicate significant variations in how tweets are ranked as sexist, influenced by factors such as the annotators' spoken language and country of residence.

However, including demographic features in the training process may introduce new biases into the models. Future research will focus on further exploring the effects of these demographic features on trained models. We anticipate that the strict conditions for corpus construction used in corpus linguistics can inform addressing some of these issues. Additionally, we plan to investigate methodologies that assess the degree of disagreement among annotators, building on existing studies in this area (Rodrigues and Pereira, 2018; Gordon et al., 2022). This will help us better understand and address potential biases in the labeling process, aiming for more robust models.

## 9 Statement of Bias

The EXIST dataset captures a wide range of biases against women, spanning from hateful or offensive sentences to humorous or friendly ones. At its core, sexism encompasses any prejudice or discrimination directed towards women based solely on their gender (Plaza et al., 2023). As a result, statements like *As usual, the woman was the one quitting her job for the family's welfare*, while not explicitly sexist, are classified as such due to their description of stereotypical situations of gender bias.

However, it is important to recognize that interpretations of sexism can vary among annotators, influenced by their language backgrounds and cultural perspectives. While efforts are made to mitigate bias in the training process, incorporating diverse viewpoints and background information from annotators, rather than relying on a single label, can result in a fairer and more robust model. Additionally, our work examined the impact of annotators' demographics on the labeling process, highlighting how these factors can lead to biased datasets.

In the annotation guidelines of the EXIST dataset, tweets that express gendered stereotypes, hatred and violence toward women, or tweets that reject inequality between men and women are both labelled sexist.

## References

- Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-Abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for automated identification of online gender-based violence: A systematic review. In *Proceedigs of the 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186.
- 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.

- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64. Springer International Publishing.
- Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.
- Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066.
- Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the 2nd CEUR Workshop on Perspectivist Approaches to NLP*.
- Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective *Isms*? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, page 275–282.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 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, pages 2591–2597.
- Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.
- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336– 1350.
- Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):1–22.

- Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, page 1637–1648. Association for Computing Machinery.
- Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation* (*SemEval-2023*), pages 2304–2318.
- Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland.
- Moeen Mostafavi and Michael D. Porter. 2021. How emoji and word embedding helps to unveil emotional transitions during online messaging. In 2021 IEEE International Systems Conference (SysCon), pages 1–8.
- 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.
- Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017– 1029.
- Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing* & *Management*, 57(6):102360.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.
- Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of exist 2023 – learning with disagreement for sexism identification and characterization. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 316– 342, Cham. Springer Nature Switzerland.
- Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

- Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings* of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 175–190.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906.
- Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622. Association for Computing Machinery.
- Narjes Tahaei, Harsh Verma, Parsa Bagherzadeh, Farhood Farahnak, Nadia Sheikh, and Sabine Bergler. 2022. Identifying Author Profiles Containing Irony or Spreading Stereotypes with SBERT and Emojis. In CLEF 2022 Labs and Workshops, Notebook Papers. CEUR-WS.org.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 338– 347, Online.
- Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. J. Artif. Int. Res., 72:1385–1470.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: quantifying annotation disagreement using demographic information. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence AAAI-23*, AAAI Special Track on AI for Social Impact.
- Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93.
- Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22.