MIND at SemEval-2023 Task 11: From Uncertain Predictions to Subjective Disagreement

Giulia Rizzi[1,2], **Alessandro Astorino**[1], **Daniel Scalena**[1], **Paolo Rosso**[2], **Elisabetta Fersini**[1]
[1]University of Milano-Bicocca, Viale Sarca, 336 - Milan, Italy
[2]Universitat Politècnica de València, Camino de Vera, Valencia, Spain
{g.rizzi10, a.astorino2, d.scalena} @campus.unimib.it
prosso@dsic.upv.es, elisabetta.fersini@unimib.it

Abstract

This paper describes the participation of the research laboratory MIND, at the University of Milano-Bicocca, in the SemEval 2023 task related to Learning With Disagreements (Le-Wi-Di). The main goal is to identify the level of agreement/disagreement from a collection of textual datasets with different characteristics in terms of style, language, and task. The proposed approach is grounded on the hypothesis that the disagreement between annotators could be grasped by the uncertainty that a model, based on several linguistic characteristics, could have on the prediction of a given *gold label*.

1 Introduction

Nowadays, natural language processing models are consistently adopted to address several tasks, such as Abusive and Offensive Language Detection (Pradhan et al., 2020; Kaur et al., 2021) and Hate Speech Detection (Alkomah and Ma, 2022). The state-of-the-art works address both general and specific targets (e.g., women, LGBTQ, immigrants) from a unimodal and multimodal perspective, also considering several languages ((Fersini et al., 2022; Magnossao de Paula et al., 2021; Chakravarthi et al., 2022; Lee et al., 2022; Muaad et al., 2022). However, although these tasks are subjective, the majority of the proposed works are based on the assumption that a unique perception and interpretation exists for each instance. However, this assumption does not reflect what happens in the real world, where multiple readers could have different points of view and understanding, resulting in a sort of disagreement. Such disagreement, which typically reflects the subjectivity of the task, can be due to linguistic ambiguity or different beliefs originated from different socio-cultural aspects.

To this purpose, many researchers have proposed corpora that aim at capturing all the distinctive interpretations of a given text by preserving the annotations disagreement instead of just aggregating them in a single *gold label*. The information brought by the annotators' disagreement has been mainly exploited in three different ways: (1) to improve the quality of the dataset by removing those instances characterized by a disagreement between annotators (Beigman Klebanov and Beigman, 2009), (2) to weight the instances during the training phase (Dumitrache et al., 2019) or (3) to directly train a machine learning model from disagreement, without considering any aggregated label (Uma et al., 2021b; Fornaciari et al., 2021). An important contribution in this field has been given by the SemEval-2021 Task 12 - Learning with Disagreements (Le-Wi-Di) (Uma et al., 2021a), whose aim was to provide a unified testing framework from disagreement for interpreting language and classifying images.

In this paper, we address the Learning with Disagreements (Le-Wi-Di) task at SemEval-2023 Task 11 (Leonardellli et al., 2023), where the main goal is to model the disagreement between annotators on different types of textual messages. In particular, we proposed a straightforward strategy that given the probability of the hard-label prediction, it creates an optimal soft-label mapping.

The paper is organized as follows. Detail about the shared task and the related datasets are reported in Section 2. In Section 3 an overview of the state of the art is provided. In Section 4 the proposed approach is detailed focusing on the prediction model, linguistic characteristics, preprocessing and post-processing operations. In Section 5 the results achieved by the proposed models are reported. Finally, conclusions and future research directions are summarized in Section 7.

2 Task Description

The main goal of the SemEval-2023 Task 11 - Learning With Disagreements (Le-Wi-Di) (Leonardellli et al., 2023), is to develop methods

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 556–564
July 13-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Language	Task	Annotators	Split	Instances	Hard label	
						1	0
HS-Brexit	EN	Hate Speech	6	Train	784	72	712
				Dev	168	19	149
				Test	168	18	150
ConvAbuse	EN	Abusive Language detection	2-7	Train	2398	389	2009
				Dev	812	133	679
				Test	840	150	690
ArMis	AR	Misogyny and sexism detection	3	Train	657	270	387
				Dev	141	57	84
				Test	145	62	83
MD_Agreement	EN	Offensiveness detection	5	Train	6592	1962	4630
				Dev	1104	388	716
				Test	3057	1018	2039

Table 1: Datasets characteristics.

able to capture the agreement/disagreement among the annotators toward a specific sentence label. The organizers encourage the usage of shared features between the available datasets for the definition of a model that is able to generalize over potentially different domains. Given the subjectivity of the proposed tasks, and therefore the absence of a *golden label*, the evaluation is based on two different strategies: (i) *hard evaluation* based on the F1 measure estimated on the prediction capabilities on the hard label and (ii) *soft evaluation* that considers how well the model's probabilities reflect the level of agreement among annotators using the Cross-Entropy as a performance metric.

2.1 Dataset

The organizers of Task 11 provided 4 benchmark datasets with different characteristics, in terms of types (social media posts and conversations), languages (English and Arabic), goals (misogyny, hate-speech, offensiveness detection) and annotation methods (experts, specific demographics groups and general crowd). The datasets available for the challenge (Akhtar et al., 2021; Almanea and Poesio, 2022; Curry et al., 2021; Leonardelli et al., 2021a) are summarized in Table 1.

All datasets relate to hate speech detection problems and are characterized by a different number of annotators ranging from 2 to 7. Most of the datasets are composed of tweets, except ConvAbuse which reports dialogues between a user and two conversational agents. Regarding the hard label, there is a significant imbalance over all the datasets. For what concerns the soft label, the ArMIS dataset shows a balanced distribution while we can still observe a disproportion for the remaining datasets. A summary of the available instances of the datasets is reported in Table 2, where dataset-specific attributes are available such as information about the annotators and the task for which the text has been collected.

3 Related Work

Machine learning models require to be trained with a significant amount of data, which usually needs to be labeled by human annotators. In order to collect such data in an efficient and cost-effective way, many researchers have relayed on crowd-sourcing platforms. Since the collected labels typically show disagreement among the annotators, in many cases, the *gold label* is obtained by a majority voting mechanism, that also involve the intervention of a domain expert or more complex approaches based on annotators commonalities or sample weights (Akhtar et al., 2021). Moreover, disagreement can also derive from overlapping labels, subjectivity, or annotator error (Uma et al., 2022). Nowadays, datasets with multiple annotations have become increasingly common and disagreement information have been used as additional information for training predictive models (Basile et al., 2021; Leonardelli et al., 2021b). In (Wich et al., 2020) the authors model the annotators behavior in a graph structure in order to represent their frequency of disagreement with each other and take advantage of information collected through community detection algorithms to train classifiers. In (Fornaciari et al., 2021) the authors implement a multitask learning

Dataset	Text	task	Annotation	Annotators	Hard label	Soft label
HS-Brexit	*UK is the 5th largest economy– wanted sovereignty back. Tired of stupid immigration and paying for others to ride on their success. #Brexit*	Hate speech detection	0,0,0,1,0,0	Ann1,Ann2,Ann3, Ann4,Ann5,Ann6	0	0: 0.83 1: 0.17

Table 2: Example of instance in the HS-Brexit dataset.

model to predict agreement as an auxiliary task in addition to the standard classification task, which improves the performance even in less subjective tasks such as part-of-speech tagging. The approach proposed by (Davani et al., 2022) models each annotator's labels individually: the authors propose a multi-annotator architecture that uses a multi-task approach to separately model each annotator's perspectives. The authors model uncertainty in predictions while maintaining high performances in terms of accuracy and efficiency. Finally, an important contribution to the field is represented by the shared task organized at SemEval-2021 (Uma et al., 2021a). The approaches proposed by the participants (Osei-Brefo et al., 2021) were based on a Sharpness-Aware Minimization (SAM), i.e., an optimization technique that is robust with respect to noisy labels, and on a neural network layer called softmax-Crowdlayer, which was specifically designed to learn from crowd-sourced annotations. Both approaches were able to improve the performance of the state-of-the-art Wide Residual Network and Multi-layer Perception models.

4 Proposed Approach

The proposed approach is grounded on the hypothesis that the disagreement between annotators could be grasped by the uncertainty that a model, based on several linguistic characteristics, could have on the prediction of a given *gold (hard) label*. This hypothesis reflects the uncertainty that a human annotator might have when annotating a text with disagreement. To validate this hypothesis, the proposed approach maps the probability values returned by a classifier trained to predict the hard label into agreement/disagreement values. The main two steps of the proposed approach are synthesized in Figure 1 and described as follows:

- **Hard label (HL) training.** A classifier is trained to predict if a given message has to be classified as positive or negative as hard label. The probability distributions over all the samples in the training dataset are used to compute the optimal classification threshold according to the Youden's J statistics (Youden, 1950). The Youden's J statistics, which is a linear combination of sensitivity and specificity, is maximized evaluating several cut off of the decision threshold of the classifier to obtain the optimal value. The proposed strategy, which basically optimizes the Area Under the Curve (AUC), allowed us to select the best classification threshold for the hard label and partially overcome the imbalance of the dataset labels.
- **Soft label (SL) representative estimation.** The model designed to predict the hard label is used to determine representative probability values that can be associated to each soft label. In particular, focusing on the positive label, the sub-samples of the training data sharing the same soft label (e.g. SL = 0.8) are used to obtain the corresponding probability distributions from the hard label model. Such probability distributions are then used to compute, for the positive class, the corresponding mean value subsequantly used as representatives for determining the corresponding SL labels.

Once the model for the hard label prediction has been trained and the soft label representative identification has been performed, we can move to the inference phase. In particular, when an unseen sample has to be classified, the hard label model is exploited. For the hard label prediction, the most probable class is selected according to the comparison of the probability distribution and the optimal cut-off obtained via the Youden's statistics. For the soft label prediction, the probability of the positive class is compared to the representative mean values estimated in the previous step, and assigned to the closest one. This allows us to automatically map the probability of the hard label to the corresponding soft label.

In order to accomplish the above-mentioned

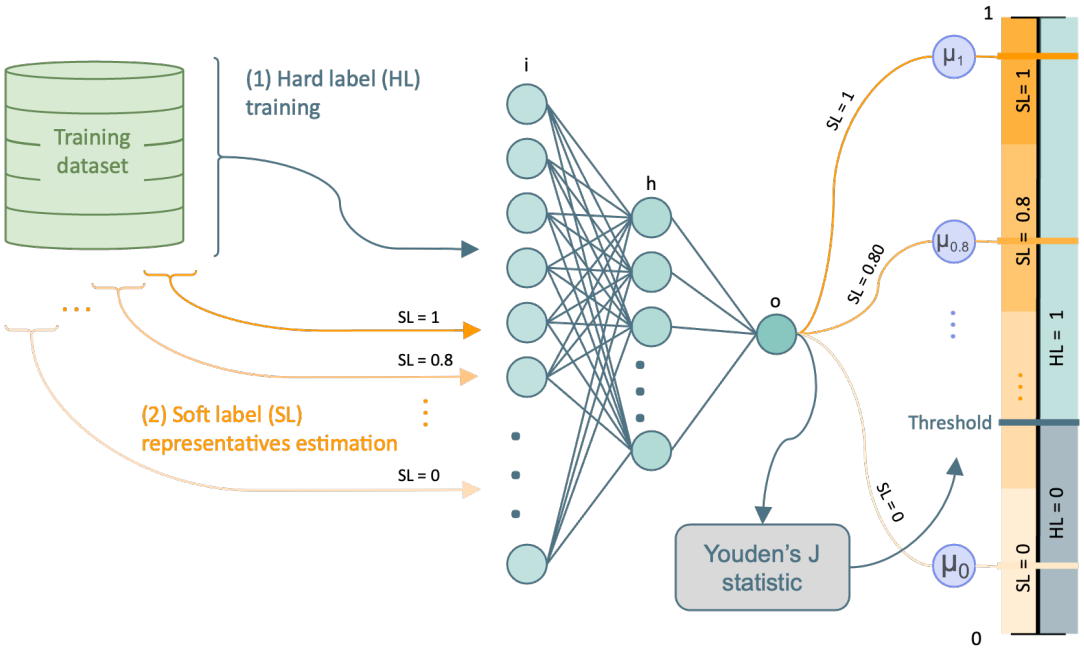

Figure 1: Schematic representation of the proposed approach. The first step refers to the training phase of the hard label prediction model. The second step consists in the estimation of the representative probability values for the soft label attribution.

steps, we developed a system based on four main constituents subsequently described.

4.1 Hard Label Classifier

The aim of the proposed approach is to investigate a potential relationship between classifier predictions on hard labels and annotator agreement. To this purpose, a simple cost-effective neural network has been adopted. The model, trained to predict the hard label, is composed of the following layers:

- *Input:* layer with length that varies according to the combination of language characteristics described in subsection 4.2;
- *Hidden:* dense layer with 256 internal neurons, with LeakyReLU activation function and dropout equal to 0.2;
- *Output:* dense layer with 1 output neuron, with sigmoid activation function able to predict the hard label.

4.2 Language Characteristics

The input of the proposed system has been defined to consider multiple aspects of the language of each dataset.

The first descriptor of each text is represented by the **sentence encoding** of the text itself. In order to guarantee a representation that could model multiple tasks, we adopted the Universal Sentence Encoder (USE) (Cer et al., 2018). The objective function, based on the binary cross-entropy loss, is minimized using the Adam (Kingma and Ba, 2014) optimization technique. Due to limited computational resources, we selected a training phase of 10 epochs with a batch size of 20 instances. The optimal classification threshold has been selected through Youden's J statistic (Youden, 1950), which allows selecting the best tradeoff between specificity and sensitivity.

Additional metadata has been considered in order to enhance the neural network discrimination capabilities. The selected metadata are strictly related to the nature of the proposed tasks (i.e., hate speech, misogyny, abusive and offensive language detection). In particular, such metadata are based on the assumption that hateful messages are characterized by specific emotions, such as anger and disgust, and a peculiar writing style, i.e., long sentences with a consistent number of special characters and uppercase. Moreover, since the majority of the proposed datasets are a collection of tweets, metadata representing online users' pragmatics have been considered (e.g. related to the use of hashtags). The selected metadata are summarized as follows:

- **Emotion**: emotion-related features have been derived through the NRC Emotion Lexicon (Mohammad and Turney, 2013). A value for each of the eight emotions (anger, fear, expectation, trust, surprise, sadness, joy, and disgust) has been computed and stored into an 8-dimensional vector.

- **Hashtag Disagreement**: for each textual message, a score representing a measure of disagreement related to the contained hashtags is computed. In particular, given a text s and the corresponding hashtag set $H_s = \{h_{s1}, ...h_{sn}\}$, the Hashtag Disagreement (HD) score is computed as follows:

$$HD(s) = \frac{1}{|H_s|} \sum_{i=1}^{|H_s|} \frac{1}{N_i} \sum_{j=1}^{|T(h_{si})|} \frac{A_j^+ - A_j^-}{A_j} N_{ij}$$

  where:

  - $T(h_{si})$ dentoes the set of training messages containing the hashtag h_{si}
  - A_j represents the number of annotators that labelled the training sample j containing h_{si}
  - A_j^+ is the number of annotators that labelled the training sample j containing h_{si} as positive
  - A_j^- denotes the number of annotators that labelled the training sample j containing h_{si} as negative
  - N_{ij} represents the number of occurrences of hashtag h_{si} in the training sample j
  - N_i is the total number of occurrences of hashtag h_{si} in the training dataset.

  The HD score associated with each text is bounded in the interval [-1; 1], where negative values denote a high correlation between the hashtag and the negative label, while positive values represent a high correlation between the hashtag and the positive label.

- **Language pragmatics**: for each textual message, features representing the presence of a few special characters (i.e., @, #, ! and ") and emoji are measured. In particular, given a textual message s, the number of occurrences of each of the above-mentioned pragmatic elements is estimated and represented through a 5-dimensional vector. The above-mentioned special characters have been selected as proxies of hateful content. In fact, from a preliminary analysis of the datasets it has been observed that users using a lot of exclamation marks strengthen their point of view towards hateful content. Moreover, hateful tweets usually include hashtags and mentions to directly and explicitly address specific groups or gain visibility by mentioning specific trends. Finally, users usually make use of quotes both to report sentences they're willing to comment on or to highlight the usage of sarcasm.

- **Stylometric features**: for each textual message, two stylometric characteristics have been considered. In particular, the number of words contained in each message and the percentage of uppercase characters have been considered as stylometric features, originating a 2-dimensional vector. This choice is motivated by the fact that hateful messages are frequently characterized by a longer text and the presence of uppercase to emphasize the exasperation related to a given topic.

4.3 Soft Label Mapping

The mechanism to predict the level of agreement of a given message is based on the assumption that samples with similar values of agreement are characterized by a similar probability distribution related to the hard label. In other words, the overall assumption is that a model trained on the hard label will predict a similar probability value for samples with the same value of agreement. In particular, we designed a mapping strategy that is able to associate the probability prediction related to the hard label with a given representative value denoting the corresponding soft label. The proposed mapping strategy is based on the following steps:

1. the training samples are grouped in subsets according to their value of agreement, i.e., the corresponding soft label (e.g. SL = 0.8), and given as input to a neural network that is able to predict the probability distribution of the corresponding hard labels. In particular, the training dataset has been split in order to create subsets of samples with the same agreement value (i.e. with the same soft labels). Then for each sample, the probability distribution for the hard label has been computed, while keeping track of the assigned subset;

2. for each of the above-mentioned subsets, the mean of the classifier prediction values are computed considering the positive class label;

3. given a new sample, its prediction probability of being predicted as positive by the hard label model is mapped to the closest mean value among the ones available for each level of agreement estimated at the previous step.

4.4 Prediction Refinement

An additional post-processing operation has been performed in order to improve the predictions given by the soft mapping. In particular, the goal is to adjust the soft label prediction towards the number of annotators that labelled a given sample. To this purpose, the soft label prediction has been shifted to the closest plausible value in accordance with the number of annotators.

This step was particularly effective for the dataset with different numbers of annotators (i.e., ConvAbuse) while resulting in a basic rounding operation for the other datasets. The post-processing steps to refine the soft label prediction allowed us to achieve an improvement of the performance up to 4.28 in terms of cross-entropy.

5 Results

In order to provide the label predictions for the test set, several input configurations have been investigated related to the input layer of the hard label classifier. In particular, the following models have been considered during the training phase:

- *M1*: For each dataset, the classification model has been trained using only the sentence encoder (USE) as input, without considering any additional information. The resulting input used to train the M1 has a size of 512 descriptors given by the USE encoder.

- *M2*: For each dataset, the classification model has been trained using the embeddings given by the sentence encoder, coupled with emotion-related features and hashtag disagreement score. The resulting input used to train M2 consists of 512 descriptors given by the USE, 8 dimensions for capturing the emotions of the message and 1 dimension for denoting the hashtag disagreement.

- *M3*: For each dataset, the classification model has been trained using the embeddings given by the sentence encoder, together with emotion-related features, hashtag disagreement, language pragmatics and stylometric features. The resulting input used to train M3 consists of a 528-dimension vector.

- *M4*: A single model has been trained using all the available datasets to take advantage of hateful content commonalities among them. The classification model M4 has been trained using only the sentence encoder (USE) as input, without considering any additional information.

The results achieved by the proposed approaches on development and test data are summarized in Tables 3 and 4 respectively [1]. The achieved results highlight that the exploitation of the considered metadata, i.e., emotion, hashtag disagreement score, language pragmatics and stylometric features, allows a better representation of disagreement resulting in an improvement of the Cross-Entropy values in all the considered datasets. However, since this representation emphasizes the uncertainty in the model predictions it also results in a decreasing performance in terms of F1 measure on the hard label. It is important to note that while M4 brings an overall improvement (with respect to the other models) when considering the Cross-Entropy measure, we can not observe any difference in terms of F1-Micro measure.

6 Considerations about the evaluation metrics

6.1 Performance evaluation

The state of the art about performance evaluation is characterized by two main types of evaluation metrics (Uma et al., 2021b) in order to evaluate how well the model performs when (i) all items are treated equally and (ii) the items are weighted depending on disagreement. The most frequent hard label measures in the first scenario are the percentage **accuracy** and the class-weighted **F1**. An alternative is based on the crowd-truth weighted f-measure (**CT-F1**). This approach is based on the intuition that items characterized by a large value of disagreement should be weighted less than items characterized by a large value of agreement.
Regarding the soft label predictions, many approaches have been discussed (Uma et al., 2021b)

[1] Only the M1 and the M2 model predictions have been submitted for the challenge partecipation

Model	HS-Brexit		ArMIS		ConvAbuse		MD-Agreement		Average	
	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro
Organizer Baseline	2.71	**0.89**	8.23	0.59	3.38	**0.95**	7.74	0.78	5.52	0.74
M1	1.328	0.845	8.547	0.468	1.084	0.888	2.705	0.755	3.416	0.739
M2	1.227	0.839	6.241	**0.603**	1.050	0.881	2.478	0.764	2.749	**0.772**
M3	**0.995**	0.821	6.434	0.518	1.287	0.842	**2.188**	**0.812**	2.726	0.749
M4	1.114	0.881	**4.794**	0.596	**0.843**	0.839	2.533	0.774	**2.321**	**0.772**

Table 3: Results associated with the proposed models on the dev dataset

Model	HS-Brexit		ArMIS		ConvAbuse		MD-Agreement		Average	
	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro	Cross-Entropy	F1-micro
Organizer Baseline	2.71	**0.89**	8.91	0.57	3.48	0.82	7.38	0.67	5.62	0.74
M1	1.648	0.839	9.160	0.469	1.168	0.861	2.580	0.749	3.639	0.730
M2	**1.040**	0.732	7.845	0.559	1.137	0.857	2.440	0.703	3.116	0.713
M3	1.258	0.869	6.656	**0.572**	1.251	**0.863**	**2.272**	**0.754**	2.859	**0.765**
M4	1.485	0.720	**4.771**	0.524	**1.109**	0.726	2.973	0.715	**2.585**	0.671

Table 4: Results associated with the proposed models on the test dataset

to capture (i) the similarity in the class distributions between the predicted labels and the ones given by the annotators or (ii) the ability of the model to reproduce human uncertainty in the prediction. The first type of evaluation is based on the assumption that the label distribution obtained by the annotation process is representative of the ambiguity of each item. To measure the label distribution similarity two approaches have been proposed:

- **cross-entropy (CE)**: to capture how confident the model is with respect to humans and the reasonableness of the distribution over alternative categories.

- **Jensen-Shannon Divergence (JSD)**: a standard symmetric method to measure the similarity between two probability distributions based on the Jensen-Shannon Divergence.

The second type of evaluation assesses the model's capability to capture the disagreement between annotators via **normalized entropy**. This approach is based on the assumption that the entropy of the annotation distribution represents how confusing the annotators find an item.

6.2 Comparison of Evaluation Metrics

Despite both F1 and CT-F1 are considered valid to rank systems based on their performances, F1 measure is more suitable with class imbalance, while the accuracy measure results in a rank strictly associated with the majority class.
Regarding the soft labels' evaluation, several authors highlighted a few limitations of the cross-entropy measure (Uma et al., 2021b,a). The first weakness of cross-entropy is related to its unbound nature. Moreover, since the Cross-Entropy measure is asymmetrical, i.e. it tends to have large values when close to the boundaries of the distribution, and it depends on the intrinsic entropy of the ground truth distribution, a few behaviours result to be unintended. Additionally, in a few cases, Cross-Entropy assumes values that do not reflect the quality of the best prediction. For instance, considering a Ground Truth (GT) distribution of [0.8, 0.2], a prediction of [0.95, 0.05] will result in a cross-entropy value of 0.32 while a prediction of [0.6, 0.4], despite more distant from the real value, results in a better cross-entropy value that is equal to 0.2959.
For these reasons, other metrics such as the Jansen-Shannon Divergency of the Wasserstein Distance appear to be more suitable for a soft-label evaluation being positive, symmetric and satisfying the triangle inequality. Moreover, the Wasserstein metric may calculate the distance between two distributions even if they are not in the same probability space, while the Jansen-Shannon Divergency, when the distributions are supported on non-overlapping domains, could fail (Kolouri et al., 2019).

7 Conclusions and Future Work

The proposed approach represents a preliminary attempt to deepen the relation between model predictions and samples' disagreement. The achieved results highlight a correlation between metadata and disagreement predictions: considering the average performances, the inclusion of emotion and lexical-related metadata resulted in an improvement of the performance in terms of disagreement detection. As future research direction, the correlation between annotators' agreement and both emo-

tion and lexical-based features will be deepened in order to achieve a representation more suitable for the model to address the task of disagreement prediction. Despite the poor improvement introduced by the universal model, the data provided by the organizer resulted to be referred to tasks too different to be assimilated into a single classification model. As a future study, it would be interesting to integrate additional datasets collected for the same tasks in order to realize task-related models. Finally, more complex architectures (e.g. transformer-based) will be implemented to improve prediction capabilities.

Acknowledgment

The work of Paolo Rosso was done in the framework of the FairTransNLP-Stereotypes research project on Fairness and Transparency for equitable NLP applications in social media: Identifying stereotypes and prejudices and developing equitable systems (PID2021-124361OB-C31) funded by MCIN / AEI / 10.13039 / 501100011033 and by ERDF, EU A way of making Europe.

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).

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

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.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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

Elisabetta Fersini, Giulia Rizzi, Aurora Saibene, and Francesca Gasparini. 2022. Misogynous meme recognition: A preliminary study. In *AIxIA 2021–Advances in Artificial Intelligence: 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1–3, 2021, Revised Selected Papers*, pages 279–293. Springer.

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

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: a survey. *Procedia Computer Science*, 189:274–281.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. 2019. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*.

Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Fatima El Barakaz, Wajdi Aljedaani, and Imran Ashraf. 2022. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE Access*, 10:9717–9728.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021a. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021b. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Angel Felipe Magnossao de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, pages 356–373. CEUR Workshop.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Abdullah Y Muaad, Hanumanthappa Jayappa Davanagere, JV Benifa, Amerah Alabrah, Mufeed Ahmed Naji Saif, D Pushpa, Mugahed A Al-Antari, and Taha M Alfakih. 2022. Artificial intelligence-based approach for misogyny and sarcasm detection from arabic texts. *Computational Intelligence and Neuroscience*, 2022.

Emmanuel Osei-Brefo, Thanet Markchom, and Huizhi Liang. 2021. Uor at semeval-2021 task 12: On crowd annotations; learning with disagreements to optimise crowd truth. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1303–1309.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. *Advances in Data and Information Sciences: Proceedings of ICDIS 2019*, pages 433–439.

Alexandra Uma, Dina Almanea, and Massimo Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. Semeval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347.

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

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199.

William J Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.