# Nozza@LT-EDI-ACL2022: Ensemble Modeling for Homophobia and Transphobia Detection

**Debora Nozza**
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

## Abstract

In this paper, we describe our approach for the task of homophobia and transphobia detection in English social media comments. The dataset consists of YouTube comments, and it has been released for the shared task on Homophobia/Transphobia Detection in social media comments. Given the high class imbalance, we propose a solution based on data augmentation and ensemble modeling. We fine-tuned different large language models (BERT, RoBERTa, and HateBERT) and used the weighted majority vote on their predictions. Our proposed model obtained 0.48 and 0.94 for macro and weighted F1-score, respectively, ranking at the third position.

## 1 Introduction

Despite the progress on LGBT+ rights, Internet still remains a hostile environment for LGBT+ people. The growing number, intensity, and complexity of online hate cases is also reflected in the real world: Anti-LGBT+ hate crimes increased dramatically in the last three years.[1] In 2020, the UK's LGBT+ anti-violence charity (Galop) presented a report about online hate crimes regarding homophobia, biphobia, and transphobia.[2] They surveyed 700 LGBT+ people distributed through online community networks of LGBT+ activists and individuals. The results are worrisome: 8 out of 10 people experienced online hate speech in the last five years, and 1 out of 5 said they had been victims of online abuse at least 100 times. Transgender people experience online harassment at a higher rate (93%) than cisgender ones (70%). It is also alarming that 18% of people claimed that online abuse was linked with offline incidents. These statistics show a worrying picture of the everyday experience that LGBT+ people are living.

Natural language processing (NLP) has emerged as a significant field of research for combating online hate speech because of its ability to automate the process at scale while, at the same time, decreasing the labor and emotional stress on online moderators (Chaudhary et al., 2021). Despite the interest of the NLP community in creating datasets and models for the task of hate speech detection, no research effort has been made to cover homophobia and transphobia specifically. This is a problem because Nozza (2021) has demonstrated that hate speech detection models do not transfer to different hate speech target types.

The shared task of Homophobia and Transphobia Detection (Chakravarthi et al., 2022) enabled researchers to investigate solutions for this problem with the introduction of a novel dataset. The dataset comprises around 5k YouTube comments manually annotated with respect to the presence of homophobia and transphobia. The corpus shows a high imbalance with respect to the non-hateful class, which covers 95% of the dataset. In this paper, we propose an approach designed to overcome the problem of class imbalance. We use ensemble modeling to combine different fine-tuned large language models. We also perform data augmentation from an external dataset to include more homophobic and transphobic instances. However, data augmentation results in lower performance, and we did not use it for the submission.

Our system ranked third for the English track with a macro F1-score of 0.48 and a weighted F1-score of 0.94.

## 2 Data

The shared task on homophobia and transphobia detection in social comments released three different datasets in English, Tamil, and code-mixed Tamil-English (Chakravarthi et al., 2021). The dataset

[1] https://www.theguardian.com/world/2021/dec/03/recorded-homophobic-hate-crimes-soared-in-pandemic-figures-show

[2] https://www.report-it.org.uk/files/online-crime-2020_0.pdf



*Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258 - 264

| | Train | Dev | Test |
|---|---|---|---|
| Size | 3,164 | 792 | 990 |
| # Non-anti-LGBT+ content | 3,001 | 732 | 924 |
| # Homophobic | 157 | 58 | 61 |
| # Transphobic | 6 | 2 | 5 |

Table 1: Statistics of the English dataset.

| | Train |
|---|---|
| Size | 3,678 |
| # Non-anti-LGBT+ content | 3,043 |
| # Homophobic | 626 |
| # Transphobic | 9 |

Table 2: Statistics of the augmented dataset.

comprises YouTube comments of videos from popular YouTubers that talk about LGBT+ topics. The comments have been labelled according to three classes: Non-anti-LGBT+ content (N), Homophobic (H), Transphobic (T). In Table 1 we show the distribution of the English dataset, which is the portion we investigate in this paper.

The numbers clearly show a strong imbalance of the dataset distribution. On average, the class Non-anti-LGBT+ content covers 94% of the instances, while there are only 6% of homophobic instances and 0.3% of transphobic ones.

### 2.1 Data Augmentation

The low number of instances associated with the hateful classes (homophobic and transphobic categories) may prevent the model from distinguishing them. In order to overcome this issue, we decide to test data augmentation techniques. Including additional hateful instances can increase model performance, even if the definition of hate speech or targets does not match exactly. We perform data augmentation by sampling additional data from the Multilingual and Multi-Aspect Hate Speech (MLMA) (Ousidhoum et al., 2019) corpus. This dataset consists of tweets with various hate speech targets. In order to perform data augmentation, we selected hateful English tweets and *sexual orientation* as the target attribute based on which it discriminates against people. This process allows us to obtain 514 tweets. We proceed by mapping every non-hateful tweet to the Non-anti-LGBT+ content class and every hateful tweet to the Homophobic one. Then, we filtered all the homophobic tweets containing the word "trans", and we associated them with the label Transphobic. Table 2 shows the statistics of the augmented dataset. Note that the MLMA dataset comprises tweets and not YouTube comments.

| Param | Value |
|---|---|
| Batch Size | 128 |
| Warm Up Steps | 50 |
| Learning Rate | 1e-3 |
| Learning Epochs | 10 |
| Optimizer | AdamW |
| Betas | 0.9 and 0.999 |
| Max Length | 200 |

Table 3: Main models' parameters.

### 2.2 Data Preprocessing

Social media textual data strongly differ from formal text, such as newspaper articles (Nozza et al., 2017). They contain slang, emojis, hashtags, URLs, and misspellings. In order to improve the quality of the data, we apply preprocessing techniques. First, we convert the text to lowercase and remove characters that are not words (e.g., numbers and punctuation). Then, we replace URLs, mentions, and emoticons with placeholder tags. Finally, we replace emojis with their textual description (e.g., *rolling on the floor laughing*) following (Corazza et al., 2020).

## 3 Experimental Settings

### 3.1 Fine-tuned Models

We use different large language models (LLMs) exploiting the HuggingFace library (Wolf et al., 2020). We selected two popular LLMs (BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)). We choose these models based on their performance and their low hurtful sentence completion (HONEST) score (Nozza et al., 2021, 2022b). We also selected HateBERT (Caselli et al., 2021), a re-trained BERT model for abusive language detection in English. Caselli et al. (2021) demonstrate that HateBERT has superior abilities for tasks of abusive detection, yielding much better results than BERT.

Each model has been fine-tuned for the task of homophobia and transphobia detection. We train each model with the same parameters (Table 3).

### 3.2 Ensemble Modeling

Ensemble modeling consists in creating a meta-classifier that treats the predicted label of distinct machine learning classifiers as a vote towards the final label that is to be predicted. This paper in-

vestigates two frameworks for ensemble: majority voting and weighted voting. Moreover, we focus only on *hard voting*, i.e., we consider only the predicted class as a vote and not its probability value (which is known as *soft voting*).

**Majority voting** Majority voting is the simplest case of ensemble learning. We consider the prediction of each classifier $C_j$ as a vote, and then we take the predicted class with the highest votes. The predicted class label $\hat{y}$ can be defined as:

$$\hat{y} = \text{mode}\left\{C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_m(\mathbf{x})\right\}$$

where $\mathbf{x}$ is the data instance.

**Weighted Voting** We use the weighted majority vote by associating a weight $w_j$ with classifier $C_j$ to predict the class label $\hat{y}$:

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j \chi_A \left(C_j(\mathbf{x}) = i\right)$$

where $\chi_A$ is the characteristic function $[C_j(\mathbf{x}) = i \in A]$, and $A$ is the set of unique class labels.

Here, as weight we use the *recall* metric for the homophobic class for each classifier. The recall metric represents the percentage of homophobic posts correctly classified by our algorithm.

## 4 Experimental Results

Table 4 shows the precision, recall, and F1-score on the test set disaggregated by class: Non-anti-LGBT+ content (N), Homophobic (H), Transphobic (T). We report the results for each fine-tuned LLMs tested (BERT, RoBERTa, and HateBERT) and the respective version fine-tuned on preprocessed data (*prep*). Finally, we provide the results of our ensemble classifiers using majority and weighted voting on the previous 6 models. From the scores, it is possible to observe that behavior regarding the non-hateful and the transphobic classes are stable for each metric and model. This is due to the class imbalance. Indeed, the Non-anti-LGBT+ content reaches high F1-scores, with a stable 0.97. In contrast, no posts have been predicted as transphobic in the test set, resulting in 0 F1-score. We argue that this is a direct consequence of the limited number of training examples (0.19%), which prevents the models from learning the phenomena. The homophobic class shows more variable performance, with an average of 0.43 and a maximum of

| | | precision | recall | F1-score |
|---|---|---|---|---|
| BERT | N | 0.95 | 0.99 | 0.97 |
| | H | 0.70 | 0.34 | 0.46 |
| | T | 0.00 | 0.00 | 0.00 |
| BERT+*prep* | N | 0.95 | 0.99 | 0.97 |
| | H | 0.72 | 0.21 | 0.33 |
| | T | 0.00 | 0.00 | 0.00 |
| RoBERTa | N | 0.95 | 0.99 | 0.97 |
| | H | 0.60 | 0.25 | 0.35 |
| | T | 0.00 | 0.00 | 0.00 |
| RoBERTa+*prep* | N | 0.95 | 0.99 | 0.97 |
| | H | 0.71 | **0.36** | 0.48 |
| | T | 0.00 | 0.00 | 0.00 |
| HateBERT | N | 0.95 | 0.98 | 0.97 |
| | H | 0.61 | **0.36** | 0.45 |
| | T | 0.00 | 0.00 | 0.00 |
| HateBERT+*prep* | N | 0.95 | 1.00 | 0.97 |
| | H | **0.79** | 0.25 | 0.38 |
| | T | 0.00 | 0.00 | 0.00 |
| Majority Voting | N | 0.95 | 0.99 | 0.97 |
| | H | 0.76 | **0.36** | **0.49** |
| | T | 0.00 | 0.00 | 0.00 |
| Weighted Voting | N | 0.95 | 0.99 | 0.97 |
| | H | 0.78 | 0.34 | 0.48 |
| | T | 0.00 | 0.00 | 0.00 |

Table 4: Results of the different fine-tuned LLMs predictions on the test set for the classes Non-anti-LGBT+ content (N), Homophobic (H), Transphobic (T). Preprocessing is denoted with *prep*.

| | macro F1-score | weighted F1-score |
|---|---|---|
| BERT | 0.48 | 0.94 |
| BERT+*prep* | 0.43 | 0.94 |
| RoBERTa | 0.44 | 0.92 |
| RoBERTa+*prep* | 0.48 | **0.95** |
| HateBERT | 0.47 | 0.93 |
| HateBERT+*prep* | 0.45 | 0.94 |
| Majority Voting | **0.49** | 0.94 |
| Weighted Voting | 0.48 | 0.94 |

Table 5: Macro and weighted F1-score on test set.

| | macro F1-score | weighted F1-score |
|---|---|---|
| BERT+*data* | 0.42 | 0.92 |
| BERT+*data*+*prep* | 0.46 | **0.93** |
| RoBERTa+*data* | 0.45 | **0.93** |
| RoBERTa+*data*+*prep* | 0.45 | **0.93** |
| HateBERT+*data* | 0.42 | 0.92 |
| HateBERT+*data*+*prep* | **0.47** | **0.93** |
| Majority Voting+*data* | 0.46 | **0.93** |
| Weighted Voting+*data* | 0.46 | **0.93** |

Table 6: Macro and weighted F1-score on test set with data augmentation approach.

0.49 obtained by majority voting. Highest scores for this class are highlighted in bold in Table 4.

Concerning the different LLMs, the best results are obtained by RoBERTa+*prep* and HateBERT. We did not observe a consistent effect regarding preprocessing, which has decreased the performance for BERT and HateBERT and has improved the one of RoBERTa. Results also demonstrate the superiority of ensembling methods, in particular, majority voting.

Table 5 reports macro and weighted F1-score. The model obtaining the highest macro F1-score (the score considered by the shared task) is majority voting. Note that we submit to the shared task the weighted voting run cause of its best performance in the dev set.

Finally, we tested the performance of the data augmentation approach (Table 6). Differently from our expectations, we notice a slight decrease in the performance. This is probably due to the different nature of the social media considered in the studies (i.e., Twitter vs. YouTube), resulting in shorter texts comprising emojis, URLs, and user mentions.

## 5 Related Work

In the last years, many shared tasks have been organized with the aim of detecting hate speech on social media comments (Kumar et al., 2018; Basile et al., 2019; Zampieri et al., 2020, inter alia). While the majority of them focus on English, some efforts have been made to include other languages (e.g., Italian, Arabic) (Bosco et al., 2018; Fersini et al., 2018; Wiegand et al., 2018; Fersini et al., 2020b; Mubarak et al., 2020; Mulki and Ghanem, 2021, inter alia). Chaudhary et al. (2021) proposed a one-of-a-kind shared task for Homophobia and Transphobia detection on social comments for three languages (English, Tamil, and code-mixed Tamil-English).

Several NLP approaches have been proposed for the task of hate speech detection (Qian et al., 2018; Indurthi et al., 2019; Vidgen et al., 2021; Fersini et al., 2020a; Attanasio and Pastor, 2020; Kennedy et al., 2020; Attanasio et al., 2022b, inter alia). While ensemble modeling has been proven to be effective for several tasks in NLP (Garmash and Monz, 2016; Nozza et al., 2016; Fadel et al., 2019; Bashmal and AlZeer, 2021), a limited number of research work have investigated its potentiality for hate speech detection (Plaza-del Arco et al., 2019; Ramakrishnan et al., 2019; Zimmerman et al., 2018).

Only recently, researchers have focused on detecting and measuring harmfulness against LGBTQIA+ community members in NLP. Some research work investigated bias in co-reference resolution (Cao et al., 2020), conversational language models (Barikeri et al., 2021), and LLMs (Nozza et al., 2022b). In a similar spirit, Dev et al. (2021) discussed the harms of treating gender as binary in English language technologies, and pointed to the complexity of gender representation. Focusing on the notion of referential gender, Lauscher et al. (2022) presented an overview on phenomena relating to 3rd person pronouns and discussed how NLP can and should model pronouns.

## 6 Conclusion

This article describes our approach for the shared task of Homophobia and Transphobia on social media comments. We propose to couple ensemble learning and data augmentation to address the problem of class imbalance of the dataset. We found that augmenting the dataset with a corpus from a different domain was ineffective. Our submitted model consists of the weighted majority vote of different fine-tuned LLMs (BERT, RoBERTa, and HateBERT) ranked at the third position out of 13 submissions. In the future, we aim to explore how fine-tuned LLMs are biased towards members of the LGBT+ community and propose a bias mitigation solution following (Nozza et al., 2019, 2022a; Attanasio et al., 2022a).

## Acknowledgements

This project has partially received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). Debora Nozza is member of the MilaNLP group, and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

## References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022a. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL2022*. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Eliana Pastor, and Dirk Hovy. 2022b. Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection. In *Proceedings of the First Workshop on Efficient Benchmarking in NLP*. Association for Computational Linguistics.

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identificationin italian tweets. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 48–54. Accademia University Press.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Laila Bashmal and Daliyah AlZeer. 2021. ArSarcasm shared task: An ensemble BERT model for SarcasmDetection in Arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 323–328, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, volume 2263, pages 1–9, Turin, Italy. CEUR.org.

Qingqing Cao, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020. Towards accurate and reliable energy measurement of NLP models. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 141–148, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An NLP perspective. *arXiv preprint arXiv:2109.02941*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Trans. Internet Technol.*, 20(2).

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

Ali Fadel, Ibraheem Tuffaha, and Mahmoud Al-Ayyoub. 2019. Pretrained ensemble learning for fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 139–142, Hong Kong, China. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. 2020a. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France. European Language Resources Association (ELRA).

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2018)*, 12:59.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020b. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. *arXiv preprint arXiv:2202.11923*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Hala Mulki and Bilal Ghanem. 2021. Working notes of the workshop arabic misogyny identification (armi-2021). In *Forum for Information Retrieval Evaluation*, FIRE 2021, page 7–8, New York, NY, USA. Association for Computing Machinery.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, , and Dirk Hovy. 2022a. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of the First Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022b. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2016. Deep learning and ensemble methods for domain adaptation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 184–189.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, Maite Martin, and L. Alfonso Ureña-López. 2019. SINAI at SemEval-2019 task 5: Ensemble learning to detect hate speech against inmigrants and women in English and Spanish tweets. In *Proceedings of the 13th International Workshop*

*on Semantic Evaluation*, pages 476–479, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium. Association for Computational Linguistics.

Murugesan Ramakrishnan, Wlodek Zadrozny, and Narges Tabari. 2019. UVA wahoos at SemEval-2019 task 6: Hate speech identification using ensemble machine learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 806–811, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).