Team_Tamil at HODI: Few-Shot Learning for Detecting Homotransphobia in Italian Language

Rahul Ponnusamy[1], Prasanna Kumar Kumaresan[1], Kishore Kumar Ponnusamy[2], Charmathi Rajkumar[3], Ruba Priyadharshini[4] and Bharathi Raja Chakravarthi[1]

[1]*Insight SFI Research Centre for Data Analytics, University of Galway, Ireland*
[2]*Guru Nanak College, Tamil Nadu, India*
[3]*The American College, Tamil Nadu, India*
[4]*Gandhigram Rural Institute-Deemed to be University, Tamil Nadu, India*

Abstract

This paper presents the novel solution to HODI (Homotranphobia Detection in Italian; [1] at EVALITA 2023 [2]. The task is structured into two subtasks: Task A, homophobic message detection, and Task B, identification of rationales of homophobic messages. We participated in Task A, a binary class classification problem. The main aim of this task is to identify the homotransphobia in Italian tweets. To determine the homotransphobia, we choose the selective models that are available in the huggingface[1] related to our task. We use the zero-shot technique to select the models that are working well for the homotransphobia classification task. With those models, we performed a series of few-shot learning experiments. Our best approach achieves a macro F1-score of 0.673, higher than the baseline, and ranking number 6.

Keywords

Homotranphobia identification, Zero-shot, Few-shot learning, Text classification,

1. Introduction

Hate speech refers to any type of offensive material, including verbal, nonverbal, symbolic, or communicative actions that are intentionally used to demean members of a specific social group based on their membership [3]. Hate speech on social media is a pervasive phenomenon affecting diverse categories of users targeted because of their race, ethnicity, gender, religion, sexual orientation, political views, or other characteristics [4, 5, 6, 7]. Significant effort has been expended to combat harmful and abusive content in the general domain, primarily via reactive measures [8, 9, 10, 11, 12, 13].

One such form of hate is hate speech towards vulnerable LGBTQ+ individuals. Homophobia, a term frequently used to characterize hostile responses to lesbians and gay men, connotes a one-dimensional conception of attitudes as manifestations of irrational fears [14]. Parents, instructors, school administrators, and government are gravely concerned about homophobic bullying against gay, lesbian, bisexual, transgender, and "queer" (LGBTQ+) minorities [15]. Bullying is a form of aggression that involves a victim, an aggressor, and bully victims. Homophobia, the underlying attitude influencing bullying against LGBTQ+ vulnerable minorities, is defined as the negative beliefs, attitudes, stereotypes, and behaviors directed toward sexual minorities [16, 17].

Cross-lingual zero and few-shot learning, in addition to other detection approaches that work well with limited or nonexistent training data sets in the target language, have not been extensively studied in the current literature on homophobia [18, 19]. Neither have other detection methods that work successfully with limited or nonexistent training data sets in the target language [20]. The researchers [21] describe zero-shot learning as an extreme form of transfer learning. When this concept is used in natural language processing (NLP), a model that has been trained on one language or domain can learn to predict samples from an unseen language or domain by making use of the latent structures of a pre-trained language model that is aligned across several languages [22, 23]. In the method of cross-lingual few-shot learning, training on the source language is supplemented with samples of the target language. This helps to strengthen both cross-lingual and task-specific alignment.

In this paper, we describe a novel approach for detecting homotransphobia in Italian-language tweets shared task conducted by [1]. We have plenty of pretrained and fine-tuned (trained for a specific task) models available for the text classification tasks. We selected twelve

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT

rahul.ponnusamy@insight-centre.org (R. Ponnusamy); prasanna.kumaresan@insight-centre.org (P. K. Kumaresan); kishorep161002@gmail.com (K. K. Ponnusamy); charmathirajkumar@gmail.com (C. Rajkumar); rubapriyadharshini.a@gmail.com (R. Priyadharshini); bharathi.raja@insight-centre.org (B. R. Chakravarthi)
0000-0001-8023-7742 (R. Ponnusamy); 0000-0003-2244-246X (P. K. Kumaresan); 0000-0001-9621-668X (K. K. Ponnusamy); 0000-0003-2323-1701 (R. Priyadharshini); 0000-0002-4575-7934 (B. R. Chakravarthi)

models which are related to our task. We performed the zero-shot technique to find the model's capability on our task. It is a technique used to infer a model by only specifying the labels of our task. We selected the top three models that gave high macro F1 scores. We performed a few-shot learning technique using SetFit (Sentence Transformer Fine-tuning) [24] framework. It is a technique to train a model with few samples. Using our approach, the twitter-xlm-roberta-base-sentiment-finetunned model got the top macro F1 of 0.74 on the validation set among the model we selected from the huggingface. Our model achieves macro F1 of 0.67345 on the test set, which is higher than the baseline.

2. Related Work

The modern forms of social media are frequently exploited in ways that promote the dissemination of violent messages and remarks as well as hate speech [25]. By analyzing the people's interaction on these issues through posts, videos, and comments, a number of works have been done with the purpose of determining whether or not aggressiveness [26], misogyny [27], racism [28], harassment, and violence are present in social media [29]. On the other hand, the amount of study that has been conducted to identify homophobic and transphobic speech online has been rather limited to few research [18].

Chakravarthi et al. [30] introduced a newly developed hierarchical taxonomy for online homophobia and transphobia, as well as a dataset that has been classified by subject matter experts, that will make it possible for homophobic and transphobic content to be automatically identified. The annotators are provided with thorough annotation criteria because this is a delicate issue. The dataset includes 15,141 comments written in English, Tamil, and Tamil-English, each of which has been annotated. From this data, Chakravarthi et al. [19] organized a shared task to increase research in homophobia/transphobia identification. It garnered 10 systems for the Tamil language, 13 systems for the English language, and 11 systems for the Tamil-English language combination. The best systems for Tamil, English, and Tamil-English each received an average macro F1-score of 0.570, 0.870, and 0.610, respectively.

A similar shared task for Dravidian languages homophobia/transphobia identification was conducted by 10.1145/3574318.3574347. It is conducted in 4 language settings (Tamil, English, Malayalam, Tamil-English). It obtained 8 systems for the Tamil language, 8 systems for the English language, 9 systems for the Malayalam language, and 8 systems for Tamil-English codemixed.

Wenpeng Yin and Roth [31] examined the limitations of prior research on zero-shot text classification (0SHOT-TC), the inadequacy of getting the problem and the significance of the labels, and the chaos of datasets and assessment setups. So, they benchmark 0SHOT-TC by standardizing the datasets and evaluations. They also provided a textual entailment framework that can function with or without the annotated data of observed labels to address the more broadly defined 0SHOT-TC. This motivated us to include the zero-shot technique in our task.

3. Dataset Description

The dataset given by the EVALITA 2023 HODI [2] shared task on Homotransphobia in Italian tweets consists of labeled data categorized into two classes: homotransphobic and not-homotransphobic. The dataset is divided into three sets: training, development, and testing, and each set is carefully constructed to maintain a comparable distribution. The distribution of the dataset is visualized in Figure 1, which provides a graphical representation of the data distribution. Additionally, the class-wise distribution of the training, development, and test sets is presented in Table 3, as provided by the HODI shared task organizers. These distributions offer valuable insights into the composition and balance of the dataset, enabling researchers to analyze and develop models to address homotransphobia in Italian tweets effectively.

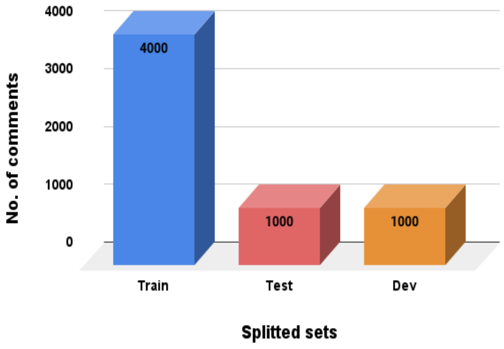

Figure 1: Visualization of splitted dataset

Further, we select 400 random samples from the training set, and we have split the dataset for the few-shot learning into five sets (80, 160, 240, 320, and 400). These sets are divided based on 20, 40, 60, 80, and 100 percent of the data taken from the training data. These datasets are visualized in Figure 2.

4. Methodology

We thoroughly describe our experimental setting in this section. All the experiments that we did adhere to the

	Labels	Comments	Total
Training	0	2379	4000
	1	1621	
Development	0	613	1000
	1	387	
Test	0	489	1000
	1	511	
Total			**6000**

Table 1
This table shows the Data Distributions of the training, development, and test set. 1 denotes homotransphobic and 0 denotes not-homophobic

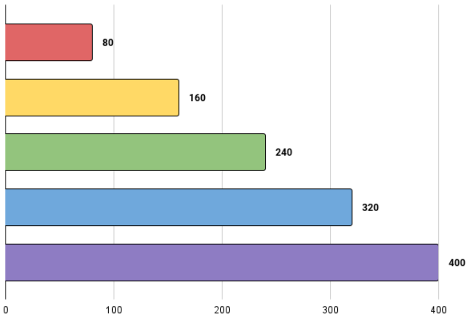

Figure 2: Split the dataset for Few-Shot learning

two phases listed below.

Zero-shot learning: In zero-shot learning, the model receives solely natural language instructions describing the task, and demonstrations are not allowed. The model is expected to generalize and perform the task based on the provided instructions without seeing explicit examples during training [32]. We took 12 models in total, which include natural language inference models and the model that trained for different tasks (sentiment, hate, emotions, toxicity) for our study. The list of models can be viewed in Table 4. These models are chosen based on topics related to the homotransphobic detection task.

Few-shot learning: It is a technique in which a limited number of task demonstrations with samples are provided to the model as training, but no weight updates are permitted [33]. With the top three models performing well in zero-shot, we applied a few-shot technique on them with the few sets of data from the training data, shown in Figure 2. We used the SetFit framework to perform few-shot text classification. Before proceeding with few-shot learning, we divided the dataset into five subsets, each containing a different number of samples: 80, 160, 240, 320, and 400. This division allowed us to examine how the models performed with increasing amounts of labeled data. We fine-tuned the models for 8 epochs, trained on GeForce GTX 1080 Ti GPU with a learning rate of 2×10^{-5} and batch size of 2. We used the same setting for all the models used for fine-tuning.

5. Results and Discussion

In this section, we show our results of the experiment and the techniques aimed to address the task given by the HODI shared task at EVALITA 2023.

Our study employed a diverse set of pre-trained models for zero-shot learning. The models used in this phase to detect the homotransphobic content in Italian encompassed a wide range of architectures and pre-trained representations, including the models that are presented in Table 4.

The evaluation of our models' performance was based on the macro F1 score, a widely used metric that considers the precision and recall across all classes. By employing the macro averaging technique, we ensured that each class contributed equally to the overall score, providing a comprehensive assessment of the models. After evaluating the models' performance, we observed notable variations in their effectiveness. Among the models, twitter-xlm-roberta-base-sentiment-finetunned, xlm-roberta-large-it-mnli[1], and mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 emerged as the top performers. These models demonstrated superior performance when compared to the others, achieving macro F1 scores of 0.52, 0.51, and 0.48, respectively, in the validation set. The higher F1 scores indicate that these models effectively captured the homotransphobic information for the given task, showcasing the advantages of leveraging transfer learning techniques in zero-shot learning scenarios.

We performed the few-shot learning experiment with 5 different numbers of few samples. We observed significant performance improvements across the tested models following the few-shot learning experiments. Notably, the twitter-xlm-roberta-base-sentiment-finetunned model outperformed the others, achieving a remarkable macro F1 score of 0.74 on the validation set when trained on the 400-sample subset. The results for the few-shot learning are tabulated in Table 4. This model performed well and got a macro F1 of 0.6735 in the test set, which is higher than the baseline result of 0.6691. This finding suggests that the model effectively learned from the limited labeled examples in the few-shot setting, showcasing its ability to generalize and adapt to novel instances to detect the homotransphobic for the Italian.

It is important to note that while the top-performing models showed promising results in both zero-shot and few-shot learning scenarios, further evaluation and comparison with other state-of-the-art approaches are necessary to establish their performance in a broader context.

[1] https://huggingface.co/Jiva/xlm-roberta-large-it-mnli

Models	ACC	M_R	M_F1	W_P	W_R	W_F1	M_P
twitter-xlm-roberta-base-sentiment-finetunned [34]	0.52	0.53	**0.52**	0.55	0.52	0.53	0.52
xlm-roberta-large-it-mnli	0.52	0.52	**0.51**	0.54	0.52	0.52	0.52
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7 [35]	0.58	0.51	**0.48**	0.53	0.58	0.53	0.51
dehatebert-mono-italian [36]	0.45	0.49	0.45	0.51	0.44	0.44	0.49
mdeberta-v3-base-tasksource-nli [37]	0.44	0.49	0.44	0.52	0.44	0.42	0.49
feel-it-italian-emotion [38]	0.59	0.50	0.43	0.52	0.59	0.50	0.50
mmarco-mMiniLMv2-L12-H384-v1 [39]	0.42	0.44	0.42	0.46	0.42	0.42	0.44
hate_speech_it	0.59	0.51	0.42	0.54	0.59	0.52	0.52
distilbert-base-multilingual-cased-toxicity	0.61	0.50	0.39	0.59	0.61	0.48	0.58
hate-ita-xlm-r-base [40]	0.41	0.49	0.39	0.50	0.41	0.36	0.48
feel-it-italian-sentiment [38]	0.40	0.49	0.34	0.49	0.40	0.29	0.47
setfit-italian-hate-speech [24]	0.39	0.50	0.30	0.51	0.39	0.24	0.49

Table 2
Results for Zero-Shot Learning in the validation set

Models	Shots	Acc	M_P	M_R	M_F1	W_P	W_R	W_F1
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	**80**	0.65	0.63	0.62	0.63	0.65	0.65	0.65
	160	0.66	0.64	0.64	0.64	0.66	0.66	0.66
	240	0.73	0.71	0.71	0.71	0.73	0.73	0.73
	320	0.68	0.66	0.65	0.65	0.67	0.68	0.68
	400	0.72	0.71	0.69	0.70	0.72	0.72	0.72
twitter-xlm-roberta-base-sentiment-finetunned	**80**	0.68	0.66	0.63	0.63	0.67	0.68	0.66
	160	0.69	0.67	0.67	0.67	0.69	0.69	0.69
	240	0.72	0.71	0.70	0.70	0.72	0.72	0.72
	320	0.75	0.75	0.72	0.72	0.75	0.75	0.74
	400	0.76	0.74	0.75	**0.74**	0.76	0.76	0.76
xlm-roberta-large-it-mnli	**80**	0.64	0.61	0.59	0.59	0.62	0.64	0.62
	160	0.70	0.55	0.52	0.51	0.63	0.70	0.64
	240	0.61	0.31	0.50	0.38	0.38	0.61	0.47
	320	0.61	0.31	0.50	0.38	0.35	0.61	0.47
	400	0.61	0.31	0.50	0.38	0.38	0.61	0.47

Table 3
Results for Few-Shot Learning in the validation set

Additionally, future research should explore the behavior of these models with larger datasets and investigate their robustness and generalization capabilities in different domains. These additional investigations will help to provide a more comprehensive understanding of the model's strengths and limitations.

6. Conclusion

In this study, our team, TEAM_TAMIL, presented our solutions for the Homotransphobia Detection in Italian (HODI) Shared Task at EVALITA 2023. Our approach focused on utilizing pre-trained models for zero-shot learning, aiming to detect homotransphobia in Italian text. We experimented with multiple pre-trained models and evaluated their performance to identify the most effective one. Additionally, we employed a few-shot learning technique by utilizing a dataset consisting of five sets, as depicted in the dataset section diagram. Our findings revealed that the selected model from zero-shot learning performed well in the few-shot learning scenario. As a result, our model achieved the 6th position in the ranking for subtask A with macro F1 of 0.6735 in the test samples. From this, we conclude that our novel approach works better than the baseline results and can be used to improve homotransphobia identification in the future.

References

[1] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th

evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] O. Ștefăniță, D.-M. Buf, Hate speech in social media and its effects on the LGBT community: A review of the current research, Romanian Journal of Communication and Public Relations 23 (2021) 47–55.

[4] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 19–26. URL: https://aclanthology.org/W12-2103.

[5] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[7] N. Chetty, S. Alathur, Hate speech review in the context of online social networks, Aggression and violent behavior 40 (2018) 108–118.

[8] A. Founta, L. Specia, A survey of online hate speech through the causal lens, in: Proceedings of the First Workshop on Causal Inference and NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 74–82. URL: https://aclanthology.org/2021.cinlp-1.6. doi:10.18653/v1/2021.cinlp-1.6.

[9] P. K. Kumaresan, Premjith, R. Sakuntharaj, S. Thavareesan, S. Navaneethakrishnan, A. K. Madasamy, B. R. Chakravarthi, J. P. McCrae, Findings of Shared Task on Offensive Language Identification in Tamil and Malayalam, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 16–18. URL: https://doi.org/10.1145/3503162.3503179. doi:10.1145/3503162.3503179.

[10] R. Priyadharshini, B. R. Chakravarthi, S. Cn, T. Durairaj, M. Subramanian, K. Shanmugavadivel, S. U Hegde, P. Kumaresan, Overview of abusive comment detection in Tamil-ACL 2022, in: Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 292–298. URL: https://aclanthology.org/2022.dravidianlangtech-1.44. doi:10.18653/v1/2022.dravidianlangtech-1.44.

[11] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from Bengali memes and texts, in: Speech and Language Technologies for Low-Resource Languages: First International Conference, SPELLL 2022, Kalavakkam, India, November 23–25, 2022, Proceedings, Springer, 2023, pp. 293–308.

[12] B. R. Chakravarthi, M. B. Jagadeeshan, V. Palanikumar, R. Priyadharshini, Offensive language identification in Dravidian languages using MPNet and CNN, International Journal of Information Management Data Insights 3 (2023) 100151. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000945. doi:https://doi.org/10.1016/j.jjimei.2022.100151.

[13] B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, J. P. McCrae, Detecting abusive comments at a fine-grained level in a low-resource language, Natural Language Processing Journal 3 (2023) 100006. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000031. doi:https://doi.org/10.1016/j.nlp.2023.100006.

[14] G. M. Herek, Beyond" homophobia": A social psychological perspective on attitudes toward lesbians and gay men, Journal of homosexuality 10 (1984) 1–21.

[15] S. S. Horn, J. G. Kosciw, S. T. Russell, Special issue introduction: New research on lesbian, gay, bisexual, and transgender youth: Studying lives in context, 2009.

[16] J. S. Hong, J. Garbarino, Risk and protective factors for homophobic bullying in schools: An application of the social–ecological framework, Educational Psychology Review 24 (2012) 271–285.

[17] B. R. Chakravarthi, Detection of homophobia and transphobia in youtube comments, International Journal of Data Science and Analytics (2023). URL: https://doi.org/10.1007/s41060-023-00400-0. doi:10.1007/s41060-023-00400-0.

[18] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance, International Journal of Information Management Data Insights 2 (2022) 100119. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000623. doi:https://doi.org/10.1016/j.jjimei.2022.100119.

[19] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, 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, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377. URL: https://aclanthology.org/2022.ltedi-1.57. doi:10.18653/v1/2022.ltedi-1.57.

[20] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.

[21] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[22] G. Winata, S. Wu, M. Kulkarni, T. Solorio, D. Preotiuc-Pietro, Cross-lingual few-shot learning on unseen languages, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 777–791. URL: https://aclanthology.org/2022.aacl-main.59.

[23] Y. Gu, X. Han, Z. Liu, M. Huang, PPT: Pre-trained prompt tuning for few-shot learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8410–8423. URL: https://aclanthology.org/2022.acl-long.576. doi:10.18653/v1/2022.acl-long.576.

[24] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, arXiv preprint arXiv:2209.11055 (2022).

[25] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[26] S. T. Aroyehun, A. Gelbukh, Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 90–97. URL: https://aclanthology.org/W18-4411.

[27] P. Zeinert, N. Inie, L. Derczynski, Annotating online misogyny, 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), Association for Computational Linguistics, Online, 2021, pp. 3181–3197. URL: https://aclanthology.org/2021.acl-long.247. doi:10.18653/v1/2021.acl-long.247.

[28] A. Field, S. L. Blodgett, Z. Waseem, Y. Tsvetkov, A survey of race, racism, and anti-racism in NLP, 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), Association for Computational Linguistics, Online, 2021, pp. 1905–1925. URL: https://aclanthology.org/2021.acl-long.149. doi:10.18653/v1/2021.acl-long.149.

[29] K. Glasgow, R. Schouten, Assessing violence risk in threatening communications, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 38–45. URL: https://aclanthology.org/W14-3205. doi:10.3115/v1/W14-3205.

[30] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[31] J. H. Wenpeng Yin, D. Roth, Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, in: EMNLP, 2019. URL: https://arxiv.org/abs/1909.00161.

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[33] D. Nozza, 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), 2021, pp. 907–914.

[34] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[35] M. Laurer, W. v. Atteveldt, A. S. Casas, K. Welbers, Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI, Preprint (2022). URL: https://osf.io/74b8k, publisher: Open Science Framework.

[36] S. S. Aluru, B. Mathew, P. Saha, A. Mukherjee, Deep learning models for multilingual hate speech detection, arXiv preprint arXiv:2004.06465 (2020).

[37] D. Sileo, tasksource: Structured dataset preprocessing annotations for frictionless extreme multi-task learning and evaluation, arXiv preprint arXiv:2301.05948 (2023). URL: https://arxiv.org/abs/2301.05948.

[38] F. Bianchi, D. Nozza, D. Hovy, "FEEL-IT: Emotion and Sentiment Classification for the Italian Language", in: Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2021.

[39] W. Wang, H. Bao, S. Huang, L. Dong, F. Wei, MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2140–2151. URL: https://aclanthology.org/2021.findings-acl.188. doi:10.18653/v1/2021.findings-acl.188.

[40] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in italian social media text, in: Proceedings of the 6th Workshop on Online Abuse and Harms, Association for Computational Linguistics, 2022.