HATE-ITA: Hate Speech Detection in Italian Social Media Text

Debora Nozza, Federico Bianchi, Giuseppe Attanasio
Bocconi University
Via Sarfatti 25
Milan, Italy
{debora.nozza,f.bianchi,giuseppe.attanasio3}@unibocconi.it

Abstract

Warning: This paper contains examples of language that some people may find offensive.

Online hate speech is a dangerous phenomenon that can (and should) be promptly counteracted properly. While Natural Language Processing has been successfully used for the purpose, many of the research efforts are directed toward the English language. This choice severely limits the classification power in non-English languages. In this paper, we test several learning frameworks for identifying hate speech in Italian text. We release HATE-ITA, a set of multi-language models trained on a large set of English data and available Italian datasets. HATE-ITA performs better than mono-lingual models and seems to adapt well also on language-specific slurs. We believe our findings will encourage research in other mid-to-low resource communities and provide a valuable benchmarking tool for the Italian community.

1 Introduction

Online hate speech is a dangerous phenomenon that can (and should) be promptly counteracted properly. While Natural Language Processing supplies algorithms to achieve that, most research efforts are directed toward the English language. Indeed, there is now a plethora of approaches and corpora (Indurthi et al., 2019; Kennedy et al., 2020b; D'Sa et al., 2020; Mollas et al., 2022; Kiela et al., 2021, inter alia), that can be adopted for addressing English hate speech detection.

However, this choice strongly limits the classification power in other languages where fewer resources are available, like Italian. Researchers have put a great effort into improving Italian models (Fersini et al., 2018; Bosco et al., 2018; Sanguinetti et al., 2018, 2020). However, previous work does not address the task systematically, resulting in no clear evidence of the performance of these models. Consider also that a competitive baseline for hate speech detection in Italian does not yet exist. Current datasets are not broad enough to cover all the protected categories and are generally based on a few thousand samples. Data annotation is a costly process, and annotating hate speech requires tremendous care.

Multi-lingual models give a possible way out of this issue. Nozza (2021) shows that combining multiple languages in training can help overcome the apparent limitations of hate speech detection models. We start from those conclusions to build up our work by collecting a large dataset of English hate speech data that we combine with some data in Italian. We use this new collection to train multi-lingual models and show the performance and examples across different Italian datasets.

The contribution of this short workshop paper is thus straightforward: we thoroughly evaluate and release to the community a set of models for Italian hate speech detection obtained through fine-tuning of multi-lingual models (HATE-ITA).[1] These models are wrapped in high-level API that will allow the community to access and use these models for future research easily. These models set a new baseline on two state-of-the-art hate speech detection datasets in Italian. To the best of our knowledge, this is the first paper that showcases the use of a large English dataset in combination with a small portion of Italian to create a robust resource for hate speech detection in Italian.

Contribution **1)** our experiments show that multi-lingual models can effectively be used to cover missing ground in some mid-to-low resource languages; **2)** while providing researchers with strong baselines, our models can also be used to study which areas and targets are still not yet covered, thus guiding directions for future research (see Section 4.4). We release HATE-ITA as an open-source Python library[2].

[1] https://huggingface.co/MilaNLProc

[2] https://github.com/MilaNLProc/hate-ita

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 252 - 260
July 14, 2022 ©2022 Association for Computational Linguistics

2 Datasets

2.1 Background

In this work, we consider the task of hate speech as binary (*hate/non-hate*). To control the number of samples for each protected group in the training data, we consider the target of the hateful messages. We select six target attributes based on the type of discrimination, namely origin, gender identity, sexual orientation, religious affiliation, and disability. We consider these targets as the superset of classes able to cover the majority of dataset-specific labels. We discarded the *other* and *none* class from all the datasets because they might represent other classes.

2.2 State-of-the-art Corpora

We describe the datasets we included in the training set in this work. The English corpora have been selected by filtering the ones covering our desired targets from a public list[3].

Italian For Italian, we consider two different corpora proposed for Evalita shared tasks (Caselli et al., 2018): the automatic misogyny identification challenge (AMI18) (Fersini et al., 2018) for hate speech towards women and the hate speech detection shared task (HaSpeeDe18) (Bosco et al., 2018) for the part related to hate speech towards immigrants proposed in (Sanguinetti et al., 2018). Both datasets comprise 2,500 instances for training, 500 for validation, and 1,000 for testing.

English Ousidhoum et al. (2019) present MlMa, a multi-lingual multi-aspect hate speech analysis dataset in Arabic, English, and French. The dataset consists of tweets collected by querying language-specific keywords.

Mollas et al. (2022) propose ETHOS, a multi-label English hate speech detection dataset of Reddit posts. They employ an automatic pre-annotation process where the posts are first labeled with a machine learning classifier. Only the uncertain ones (within the [.4, .6] probability range) are manually labeled using a crowdsourcing platform. Following the authors, we binarise the values of each label (if value $\geq 0.5 \rightarrow 1$ else value $\rightarrow 0$). The targets are identified only when the post is hateful, so we discard the non-hateful ones. Here, we map the targets *national_origin* and *race* to *origin*.

Kennedy et al. (2020c) collected a large set of comments from different social media sources (YouTube, Twitter, and Reddit). The annotation process has been performed via a crowdsourcing platform where each comment receives four ratings. The authors further ensured that every annotator received comments across all the hate speech scale. Since the dataset is annotated with a continuous hate score, we used a threshold set to binarise the problem: if value $< -1 \rightarrow 0$ and if value $> 0.5 \rightarrow 1$. We merged *origin* and *race* classes into the *origin* class.

Mathew et al. (2021) collected English posts from the social media platforms Twitter and Gab. Then, they used a crowdsourcing platform for annotating each post as hate, offensive, or normal speech; annotators also have to select the target communities mentioned in the posts. Labels are aggregated, and the final one is obtained through majority voting. We discard the instance when there is no majority (i.e., the three annotators have assigned a different label). Here, we binarise the targets as suggested by the authors into toxic (*hate-speech* or *offensive*) and non-toxic (*normal*). We also map the targets based on the grouping made in the paper (see Table 3 in (Mathew et al., 2021)), with the only exception of *Indigenous* and *Refugee* that we assign to *origin* class.

Kennedy et al. (2020a) presented the Gab Hate Corpus (GHC), a multi-label English corpus of posts from the social network `gab.com`. Comments were annotated by at least three trained annotators with the following classes: *Call for Violence*, *Assault on Human Dignity*, or *Not Hateful*. Following Kennedy et al. (2020b), we aggregate the first two for obtaining the hateful class. We selected only the targets used in our study (removing *political*) and merged *nationality/regionalism* and *race or ethnicity* classes into the *origin* class.

Kiela et al. (2021) introduced a novel framework for dynamically creating benchmark corpora. The annotators are asked to find adversarial examples, i.e., hard examples that a target model would misclassify. The obtained dataset also provides the target group.[4] Here, we mapped their targets to ours, removing the ones not covered.

Table 1 shows the size of the dataset created by combining all the afore-mentioned English corpora.

[3] `https://hatespeechdata.com/`

[4] `https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset`

	Hate	Non-hate	Total
Disability	3,128	1,488	4,526
Gender	22,655	24,182	46,829
Origin	44,047	31,211	75,327
Religion	17,010	10,840	27,864
Sex. Orientation	9,980	12,312	22,313
Total	97,014	80,729	177,749

Table 1: Statistics of the English dataset.

Model	Mono	Multi	Zero
XLM-Large	59.25	**81.23**	57.27
XLM-Base	52.36	**80.74**	54.47
XLM-Twitter	63.52	**83.34**	56.45
mBERT	66.93	**80.48**	51.87
ITA-Base-XXL	61.20	-	-
ITA-Base	40.45	-	-

Table 2: Macro-F1 results. The most frequent class classifier has a Macro-F1 of 36.85.

3 Experimental Methodology

Our experimental setup illustrates three aspects: 1) the performance of the different models on a train, validation, and test setup that we construct on our data, 2) the performance on different datasets (also considering two new additional datasets that we take as *out-of-domain*) and 3) a qualitative evaluation section in which we use explainability methods to assess which words are contributing more to the prediction.

3.1 Models

In this paper, we tested different pretrained language models. As multi-lingual models: the XLM Roberta base and large models from (Conneau et al., 2020) (XLM-Base, XLM-Large), multilingualBERT[5] (mBERT), and a model pre-trained on multi-lingual twitter data (XLM-Twitter) (Barbieri et al., 2021). As mono-lingual models for Italian: *dbmdz/bert-base-italian-xxl-cased* (ITA-Base-XXL) and *dbmdz/bert-base-italian-cased* (ITA-Base).[6] In addition, we used DeHateBert (Aluru et al., 2020), a fine-tuned mBERT model trained on (Sanguinetti et al., 2018).

For the models we train, we run three different experimental frameworks: 1) *mono-lingual* (Mono), in which we train our models only on Italian data; 2) *multi-lingual* (Multi), in which we combine the Italian and the English data for training; 3) *zero shot, cross-lingual* (Zero), in which we train a model only with English data. All the models are tested on the Italian test data (Fersini et al., 2018; Sanguinetti et al., 2018).

3.2 Data Setup

We used the splits provided by the associated shared tasks for the Italian dataset. This setup ensures performance comparability. For Sanguinetti et al. (2018), we isolated 500 instances from the training to be used as the validation set. For the combined English data, we isolate 20% with stratified sampling to be used as the validation set. The details of the parameters used to fine-tune the models can be found in the Appendix A. Models are trained for 5 epochs and evaluated every 50 steps, and we select the best checkpoint considering the validation loss.

4 Results

4.1 Overall Results

Table 2 shows the results only for the models that we trained by testing on the official splits of each Italian dataset (see Section 2.2). We have found two crucial takeaways. First, the best multi-lingual model (XLM-Large) performs sensibly better than the best model trained only on mono-lingual data (mBERT). Second, models subject to multi-lingual training **always** outperforms mono-lingual ones. Recent research (Nozza et al., 2020) has shown that language-specific datasets are more effective when used to fine-tune language-specific models; this research suggests that training only on the small set of Italian data is not enough even when using a language-specific model: joint fine-tuning with larger datasets is an effective way of obtaining more accurate hate speech classifiers. This is a very interesting result: considering the small amount of Italian data used by the multi-lingual model, this opens future applications of multi-lingual pipelines to low-resource languages. Finally, the increase in performance of the multi-lingual framework comes directly from the Italian data we added to the training since the performance of the purely zero-shot cross-lingual models is much worse than the mono-lingual one.

[5] https://github.com/google-research/bert/blob/master/multilingual.md

[6] https://huggingface.co/dbmdz

Model	AMI18	AMI20	Sanguinetti et al. (2018)	HaSpeeDe18	HaSpeeDe20
XLM-Twitter	**82.10**	72.73	78.53	74.59	72.68
XLM-Base	79.88	66.47	79.64	76.40	72.57
XLM-Large	80.37	**73.75**	**79.96**	**78.13**	**75.86**
DeHateBert	42.66	53.97	-	-	70.79

Table 3: Results on different benchmark datasets for the multi-lingual models.

4.2 Results by Dataset

This section shows the results split by datasets for our multi-lingual best models and for DeHateBert. We show the results on the test sets of Sanguinetti et al. (2018) and AMI18 (Fersini et al., 2018). Moreover, we also test on the complete test set of HaSpeeDe18, Bosco et al. (2018) and the shared task re-runs HaSpeeDe20 (Sanguinetti et al., 2020) and AMI20 (Fersini et al., 2020b). Unfortunately, DeHateBert was not fine-tuned following the guidelines described in (Bosco et al., 2018) as the authors used different splits. For this reason, we cannot evaluate the performance of this model on HaSpeeDe18 and (Sanguinetti et al., 2018) (some examples of the examples in the test sets are used for training).

Table 3 shows the results for each dataset. We do not show results for Italian models as they perform much worse (see Table 2). These results show that our models have consistent performance over most categories. Indeed, XLM-Twitter, beats DeHateBert by 39 and 19 points in F1 on AMI18 and AMI20 respectively. This outcome further demonstrates the need for protected group coverage in the training set.

4.3 Results on Multi-Lingual HateCheck

We also use the recently introduced Multi-Lingual HateCheck (MHC) (Röttger et al., 2022). MHC is a suite of functional tests for multi-lingual hate speech detection models that extend the original English HateCheck (Röttger et al., 2021). MHC tests several functionalities that can affect hate prediction (e.g., counterspeech, spelling variations, use of slurs). Here, we used only the Italian subset. MHC should serve as an external testbed to validate our models.

Results in Table 4 show the consistent performance of our models. XLM-Twitter and XLM-Large strongly outperform the results of the original baseline proposed by Röttger et al. (2022).

Model	F1-h	F1-nh	Macro-F1
XLM-Twitter	84.74	61.17	72.96
XLM-Base	82.71	55.10	68.90
XLM-Large	**88.63**	**65.88**	**77.26**
Röttger et al.	81.50	57.80	69.60

Table 4: Results on different MULTILINGUAL HATECHECK. We report F1 score the for **h**ateful and **n**on-**h**ateful cases, and the overall macro-F1 score.

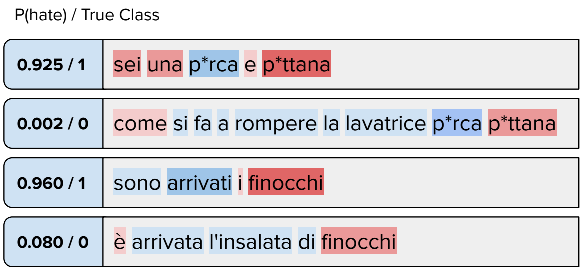

Figure 1: Examples of predictions with SHAP (Lundberg and Lee, 2017) contributions on a color scale; color scale: blue (not-hate), red (hate). Translation available in Appendix B.

4.4 Qualitative Evaluation

Figure 1 reports token contribution explanations of four correct predictions from our multi-lingual XLM-Large. The texts are complex examples in Italian, as standard models usually misclassify them (Nozza, 2021). We extracted token contributions using the interpretability suite provided in Attanasio et al. (2022b). The first two examples regard the taboo Italian expression *p*rca p*ttana* (literally *p*rca* (*pig*) + *p*ttana* (*sl*t*)). When used separately (*porca e puttana* (*pig and slut*)), they should be considered literally; when used together, the two words form taboo expressions that do not have a misogynistic connotation. The latter two examples regard the ambiguous Italian term *finocchi*. The word means *fennels* in a food-related context, but can also be translated to *f*ggots* when refereed to individuals.

5 Related Work

National evaluation campaigns and shared tasks played a significant role in releasing non-English corpora for hate speech detection (Wiegand et al., 2018; Mulki and Ghanem, 2021; Basile et al., 2019; Ptaszynski et al., 2019). Indeed, the research of hate speech detection in Italian in mono-lingual settings mainly revolves around the datasets (Fersini et al., 2018; Bosco et al., 2018; Sanguinetti et al., 2020; Fersini et al., 2020b) released for shared tasks (Bakarov, 2018; Cimino et al., 2018; Attanasio and Pastor, 2020; Lees et al., 2020; Lavergne et al., 2020; Fersini et al., 2020a; Attanasio et al., 2022a, inter alia).

In NLP, the scarcity of data in languages beyond English has generated an interest in zero-shot learning (Srivastava et al., 2018; Ponti et al., 2019; Pfeiffer et al., 2020; Wu et al., 2020; Bianchi et al., 2021, 2022, inter alia) and the application of this to hate speech detection methods (Corazza et al., 2020; Stappen et al., 2020; Aluru et al., 2020; Leite et al., 2020; Rodríguez et al., 2021; Feng et al., 2020; Pelicon et al., 2021). In particular, Aluru et al. (2020) exploited several deep learning models and multi-lingual embeddings for performing an extensive analysis on 16 datasets in 9 different languages in few- and zero-shot learning settings. Rodríguez et al. (2021) use the pre-trained Language Agnostic BERT Sentence Embeddings (Feng et al., 2020) obtaining good results. Other research efforts focused on translating English data to enrich data availability in other languages with mixed results: Ibrohim and Budi (2019) shows that translations do not bring good results using traditional machine learning classifiers. However, more sophisticated pipelines of translation and pre-training can indeed provide some improvement over standard benchmarks (Pamungkas et al., 2021; Wang and Banko, 2021).

6 Conclusion

This paper presents a novel resource for Italian hate speech detection on social media text, HATE-ITA. Researchers can use this new set of models to assess the quality of new systems by providing a more reliable benchmark. However, this is just the first step. Indeed, we do not claim to have released the final model for Italian hate speech detection; HATE-ITA requires careful benchmarking to understand if it can accurately capture hate speech on other targets.

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, Federico Bianchi, and Giuseppe Attanasio are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

Ethical Statement

While promising, the results in this work should not be interpreted as a definitive assessment of the performance of hate speech detection in Italian. We are unsure if our model can maintain a stable and fair precision across the different targets and categories. HATE-ITA might overlook some sensible details, which practitioners should treat with care.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, page 423–439, Berlin, Heidelberg. Springer-Verlag.

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 (Forthcoming)*. 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.

Amir Bakarov. 2018. Vector space models for automatic misogyny identification (short paper). In *Proceedings of the Sixth Evaluation Campaign of Natural*

Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

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

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. 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.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task learning in deep neural networks at EVALITA 2018. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. Hybrid Emoji-Based Masked Language Models for Zero-Shot Abusive Language Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 943–949, Online. Association for Computational Linguistics.

Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2020. Label propagation-based semi-supervised learning for hate speech classification. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 54–59, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

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.

Muhammad Okky Ibrohim and Indra Budi. 2019. Translated vs non-translated method for multilingual hate speech identification in twitter. *International Journal on Advanced Science, Engineering and Information Technology*, 9(4):1116–1123.

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, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian*, Gabriel Cardenas*, Alyzeh Hussain*, Austin Lara*, Adam Omary*, Christina Park*, Xin Wang*, Clarisa Wijaya*, Yong Zhang*, Beth Meyerowitz, and Morteza Dehghani. 2020a. The Gab Hate Corpus: A Collection of 27k Posts Annotated for Hate Speech.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020b. 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.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020c. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Eric Lavergne, Rajkumar Saini, György Kovács, and Killian Murphy. 2020. Thenorth @ haspeede 2: Bert-based language model fine-tuning for italian hate speech detection (short paper). In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and haspeede2: Fine-tuning a pre-trained comment-domain BERT model. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

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. 2020. What the [MASK]? Making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

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.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559. Publisher: PeerJ Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019. Towards zero-shot language modeling. 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 2900–2910, Hong Kong, China. Association for Computational Linguistics.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the PolEval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter. *Proceedings of the PolEval 2019 Workshop*, page 89.

Sebastián E. Rodríguez, Héctor Allende-Cid, and Héctor Allende. 2021. Detecting Hate Speech in Cross-Lingual and Multi-lingual Settings Using Language Agnostic Representations. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 77–87, Cham. Springer International Publishing.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022)*. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection 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 41–58, Online. Association for Computational Linguistics.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Lukas Stappen, Fabian Brunn, and Björn W. Schuller. 2020. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850.

Cindy Wang and Michele Banko. 2021. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, 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)*.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

A Model Training

All the models are trained with the same pipeline. We report the shared parameters in Table 5. As mentioned in the table we fix the maximum number of epochs, but we evaluate the models during training and eventually select the model with the best validation loss. All the experiments are run 5 times and we take the average of the results.

Param	**Value**
Batch Size	64
Warm Up Steps	50
Learning Rate	1e-3
Learning Epochs*	5
Optimizer	AdamW
Betas	0.9 and 0.999
Max Length	100

Table 5: The main parameters we used to run the models. *While epochs are 5, we remark that we are running a step-wise evaluation.

B Examples Translation

We provide as literal as possible translations.

- IT: *Sei una p*rca e p*ttana*
- EN: *you are pig and sl*t*

- IT: *Come si fa a rompere la lavatrice p*rca p*ttana*
- EN: *How the hell can you break the washing machine*

- IT: *Sono arrivati i finocchi*
- EN: *Here come the f*ggots*

- IT: *È arrivata l'insalata di finocchi*
- EN: *Here it comes the fennel salad*