Exposing the limits of Zero-shot Cross-lingual Hate Speech Detection

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

Abstract

Reducing and counter-acting hate speech on Social Media is a significant concern. Most of the proposed automatic methods are conducted exclusively on English and very few consistently labeled, non-English resources have been proposed. Learning to detect hate speech on English and transferring to unseen languages seems an immediate solution. This work is the first to shed light on the limits of this zero-shot, cross-lingual transfer learning framework for hate speech detection. We use benchmark data sets in English, Italian, and Spanish to detect hate speech towards immigrants and women. Investigating post-hoc explanations of the model, we discover that non-hateful, language-specific taboo interjections are misinterpreted as signals of hate speech. Our findings demonstrate that zero-shot, cross-lingual models cannot be used as they are, but need to be carefully designed.

1 Introduction

An increasing propagation of hate speech has been detected on social media platforms (e.g., Twitter) where (pseudo-) anonymity enables people to target others without being recognized or easily traced. While this societal issue has attracted many studies in the NLP community, it comes with three important challenges. First, "hate speech" covers a **wide range of target types**, including misogyny, racism, and various other forms. While they often intersect, these types require different approaches.

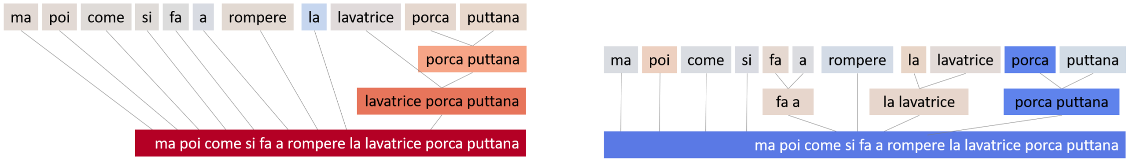
Second, available labeled corpora refer to different definitions of hate speech, collection strategies, and annotation frameworks (Fortuna and Nunes, 2018). This lack of consistency strongly limits research on hate speech, which ultimately needs to apply cross-domain or transfer learning approaches for using different corpora.

Third, most of the research on hate speech detection **consider only English** and only a **limited number of labeled corpora** are available (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2021; Poletto et al., 2020). However, hate speech is not specific to any one language, and approaches proposed for English may not fit other languages. Each language exhibits different complexities in dealing with gender or reflecting cultural ideas around it.

The lack of models and labeled corpora for non-English languages seems a perfect application for zero-shot, cross-lingual learning (Lamprinidis et al., 2021; Bianchi et al., 2021). But is it? In this paper, we investigate the limitations of zero-shot, cross-lingual solutions based on mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) on benchmark data sets of hate speech against immigrants and women in English, Italian, and Spanish.

Our analysis demonstrates that these approaches have significant limitations: (1) they are not able to capture common (taboo) language-specific expressions, and (2) they do not transfer to different hate speech target types. We show that the reasons for these limitations are due to the high presence of language- and target-specific taboo interjections in non-hateful contexts, like *porca puttana* or *puta*.[1] **While derogatory for women, these terms are often used as intensifiers in non-hateful context, blurring the lines for detection**. Since English does not use equivalent words in the same way, zero-shot, cross-lingual models will not observe them in the training data. Consequently, these models consider the literal meaning of these terms as individual words, treating them as misogynous hate speech. These findings demonstrate that, at the current moment, cross-lingual, zero-shot transfer learning is not a solution for solving the lack of models and labeled corpora in non-English languages for hate speech detection.

[1] We report the uncensored words to ensure non-native speaker understanding.



(a) Misclassified prediction by zero-shot, cross-lingual model trained on English and Spanish and tested on Italian data.

(b) Correct prediction by monolingual model trained on Italian and tested on Italian data.

Figure 1: Hierarchical explanations of predictions of a non-hateful Italian tweet. Literal English translation: "how the hell can you break the washing machine".

Contributions 1) We investigate different learning frameworks on benchmark corpora for the detection of hate speech targeting women and immigrants 2) We expose the limits of zero-shot, cross-lingual solutions using the multilingual BERT model (mBERT) 3) We show interpretable results through post-hoc explanation.

2 Zero-shot, Cross-lingual Hate Speech Detection

We investigate different learning settings: 1) *zero-shot, cross-lingual*, i.e., training on one language and testing on unseen languages; 2) *monolingual*, i.e., training and testing on the same language; 3) *few-shot, cross-lingual*, i.e., training on one language and a small percentage of samples from the test language and testing on the test language; 4) *augmented cross-lingual*, i.e., training on several languages and testing on a language included in the training.

Multilingual BERT Recently, *contextual embeddings* pretrained on large corpora substantially advanced research for several major Natural Language Processing (NLP) tasks (Nozza et al., 2020). In particular, multilingual BERT (mBERT) (Devlin et al., 2019), a model pretrained on monolingual Wikipedia dumps in 104 languages, has shown surprisingly good abilities for zero-shot, cross-lingual model transfer for different NLP tasks (Pires et al., 2019). In this paper, we fine-tune the mBERT model on the task of hate speech detection considering data from one or multiple languages.

Post-hoc Explanation One of the biggest limitations of using complex black-box models, such as BERT, is the lack of interpretability. Following Kennedy et al. (2020), we use the Sampling and Occlusion (SOC) algorithm (Jin et al., 2020) to generate hierarchical explanations of predictions. SOC assigns an importance score to show how much a given word or sequence of words contributes to classifying a sentence as hate speech. Then, it combines this score hierarchically following semantic compositions. Visual representation examples are given in Figures 1 and 2. The hierarchy reflects how the model captures compositional semantics (e.g., stress or negation) in making predictions. Color intensity represents how much each phrase contributes to classifying the sentence as hate speech. The label prediction is encoded in the color: blue for non-misogynous and red for misogynous.

3 Data

	Immigrants			Women		
	EN	**IT**	**ES**	**EN**	**IT**	**ES**
Train	4500	2000	1618	4500	2500	2882
Dev	500	500	173	500	500	327
Test	1499	1000	800	1472	1000	799

Table 1: Corpora splits # of instances by target type.

To assess the cross-lingual evaluation framework, we use hate speech benchmark data sets with consistent definitions, annotation schema, and collection strategies (see Appendix C). For English and Spanish, we adopt the data sets proposed in the shared task of hate speech against immigrants and women on Twitter (HatEval) (Basile et al., 2019). For Italian, we consider two different corpora proposed for Evalita shared tasks (Caselli et al., 2018): the automatic misogyny identification challenge (AMI) (Fersini et al., 2018) for hate speech towards women, and the hate speech detection shared task on Facebook and Twitter (HaSpeeDe) (Bosco et al., 2018) for hate speech towards immigrants. Table 1 reports data distributions across languages and targets.

		Immigrants		
	Test	**IT**	**EN**	**ES**
Train	**IT**	*0.777*	*0.635***	*0.666*
	EN	*0.590***	*0.368*	*0.633*
	ES	*0.683***	*0.596***	0.630
	EN+ES	*0.706*$*$	0.353	0.676$*$
	ES+IT	*0.757*	*0.538***	0.686$*$
	EN+IT	0.771	*0.340*	*0.657*
	Baseline	0.799	-	-

(a)

		Women		
	Test	**IT**	**EN**	**ES**
Train	**IT**	0.808	*0.545*	*0.463***
	EN	*0.449***	0.559	*0.546***
	ES	*0.337***	*0.558*	0.839
	EN+ES	*0.440*	0.449**	0.873*
	ES+IT	0.820	*0.502*	0.878*
	EN+IT	0.798	0.469**	*0.603***
	Baseline	0.844	-	-

(b)

Table 2: Macro-F1 results for the two hate speech targets. Monolingual results are underlined. Zero-shot cross-lingual results are highlighted in *italic*. * = differs significantly from monolingual at p ≤ 0.05. ** = significant difference at p ≤ 0.01.

4 Experimental Results

Table 2 shows the macro-averaged F1 score for hate speech detection on different training and test languages (in rows and columns, respectively). Underlined numbers refer to the monolingual setting results, while zero-shot, cross-lingual results are italicized. We report as *baselines* the best performing model for each of the considered data set released in conjunction with shared tasks.[2] Since the aim of this paper is to investigate classification abilities of cross-lingual, zero-shot models, we do not aim to overcome the baselines but to provide comparable results.

4.1 Hate speech towards immigrants

Observing monolingual results (underlined numbers in Table 2), we see that training and testing in English gives the poorest performance. This behavior is due to an over-sensitivity to specific words/hashtags used during data collection (e.g. *#SendThemBack*, *#StopTheInvasion*), which leads to overfitting. In Appendix A, we report the SOC explanation of a misclassified tweet containing these hashtags. We confirm this finding by training the monolingual English model on data deprived of these hashtags, which lead to higher macro-F1 (from 0.368 to 0.438).

The zero-shot, cross-lingual configuration (italic numbers in Table 2) shows very different results between the two targets. Zero-shot learning obtains good performance for detecting hate speech towards immigrants: when testing Italian and Spanish, results are very similar; when testing on English, training on a different language is better than including English data, resulting in a 22% macro-F1 improvement on average. This is because training sets based on other languages do not contain the above-mentioned specific words and therefore do not suffer from over-sensitization.

4.2 Hate speech towards women

Concerning hate speech towards women, *the zero-shot, cross-lingual model obtains significantly lower performance for Spanish and Italian.* To better understand this substantially different finding, we analyze wrongly labeled instances. We discover that zero-shot, cross-lingual models are strongly influenced by common, language-specific taboo interjections to mislabel non-hateful text as misogynous. In particular, expressions that contain literal insults towards women but are not misogynistic per se. For example in Spanish, beyond its misogynistic meaning, the word *puta* (literally *bitch*) is also used as an exclamation of surprise (e.g., *puta mierda*). The Italian expressions *porca troia* and *porca puttana* (literally *porca* (*pig*) + *troia*/*puttana* (*slut*)) are very generic taboo interjections that do not have a misogynistic connotation. It is important to notice that these interjections are not directly translatable and usually used in combination, e.g. *porca* + *puttana*, *puta* + *mierda*.

To demonstrate this finding, in Table 3 we report the number of times a zero-shot cross-learning model correctly predicts the labels of instances containing taboo interjections for Italian and Spanish (i.e., *porca puttana, porca troia, puta*). The high frequency of instances containing taboo interjections (29% and 78% of the test set), due also to the keyword-driven collection strategy, proves the importance of understanding these linguistic expressions. The following numbers illustrate the

[2] State-of-the-art performance do not exist for every combination, since Hateval (English and Spanish) consider hate speech towards women and immigrant in conjunction.

Test Lang	Frequency	Zero-Shot, Cross-Lingual	Monolingual
IT	294 (29%)	9 (3%)	291 (99%)
ES	627 (78%)	365 (58%)	514 (82%)

Table 3: Correct predictions for instances containing Italian and Spanish taboo interjections.

impact of taboo interjections: all the 276 Italian tweets containing *porca puttana* are labeled as non-misogynous and are consistently misclassified by zero-shot, cross-lingual model; the Spanish expression *hijo de puta* appears in 64 tweets (of which 57 are non-misogynous) for which the zero-shot, cross-lingual model achieves 62% accuracy vs. 90% accuracy of the monolingual model. We confirm this finding by training models on data deprived of these taboo interjections, obtaining improvements: 0.627 for **ES⇒IT**; 0.479 for **IT⇒ES**; 0.662 for **EN⇒IT**; 0.660 for **IT⇒EN**.

Figure 1 shows the SOC explanation of a non-hateful tweet correctly classified by the monolingual Italian model and wrongly classified by the zero-shot, cross-lingual model trained on English and Spanish data. As expected, training and testing on Italian teach the model that *porca puttana* is a very general exclamation that does not imply misogyny (high importance score for non-misogynous prediction). However, when training on other languages, this taboo interjection is not recognized because it is strictly related to the *test* language. We observe that zero-shot, cross-lingual models consider the literal meaning of individual words, and consequently treat terms like *porca puttana* as misogynous regardless of their use in context.

To further validate this major finding, we conduct an additional experiment on the corpus of hate speech towards women: we train *few-shot, cross-lingual* models randomly sampling 1% of training data in the test language. The averaged results on 10 runs in terms of macro-F1 are: 0.660 for **ES+EN⇒IT**; 0.702 for **EN+IT⇒ES**. The significant improvements with respect to zero-shot performances prove that misogyny detection is strongly entangled with common, language-specific taboo interjections that are very frequent in the data set.

4.3 Hate speech towards immigrants and women

Finally, to demonstrate the *need for treating target types separately*, we run the zero-shot, cross-lingual model on the merged data sets of hate speech towards immigrants and women. The results in terms of macro-F1 are: 0.572 for **ES+IT⇒EN**; 0.513 for **ES+EN⇒IT**; 0.632 for **EN+IT⇒ES** (see Appendix B).

Following Stappen et al. (2020), these scores suggest a sufficient adaptation by the models. However, they represent a compromise between the high results of zero-shot cross-lingual hate speech detection against immigrants and the low results of hate speech detection against women. By showing the results for the two separate targets, we demonstrated that zero-shot cross-lingual models suffer from limitations when predicting hate speech detection against women and that, in general, zero-shot cross-lingual hate speech detection has yet to be solved.

4.4 Impact of language-specific taboo interjections on XLM-R

In order to understand whether common, language-specific taboo interjections play a role in other language models, we conducted experiments with XLM-R (Conneau et al., 2020). XLM-R is a large cross-lingual language model based on RoBERTa (Liu et al., 2019), trained on 2.5TB of filtered CommonCrawl data, which significantly outperformed mBERT on a variety of cross-lingual benchmarks.

XLM-R achieves high macro-F1 scores in monolingual settings for detecting hate speech towards women in Italian and Spanish (0.806 for **IT⇒IT**; 0.859 for **ES⇒ES**). Similar to the previously presented findings, we observe a significant drop of 36% in macro-F1 when considering the zero-shot cross-lingual settings (0.604 for **EN⇒IT**; 0.511 for **ES⇒IT**; 0.404 for **IT⇒ES**; 0.658 for **EN⇒ES**). This drop in macro-F1 is more evident when considering the performance when training on Spanish and testing on Italian and vice versa. These results on XLM-R bring more evidence about the role that language-specific taboo interjections have in impacting the performance.

5 Related Work

Hate speech detection has attracted great interest in the NLP community. This has led to the proposal of automatic detection approaches based on machine learning (Indurthi et al., 2019; Nozza et al., 2019; Fersini et al., 2020a; Kennedy et al., 2020; D'Sa et al., 2020, inter alia) and the creation of benchmark data sets, usually distributed through shared tasks (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Bosco et al., 2018; Kumar et al., 2018; Wiegand et al., 2018; Basile et al., 2019; Fersini et al., 2018; Zampieri et al., 2020; Fersini et al., 2020b, inter alia).

Only a few studies have investigated hate speech detection across different languages. Steimel et al. (2019) asked which factors affect multilingual settings for German and English, concluding that a shared classification algorithm is not conceivable due to lack of corpora comparability. In Sohn and Lee (2019), the authors proposed a multi-channel model exploiting multilingual BERT and language-specific BERT for Chinese, English, German, and Italian. Finally, Stappen et al. (2020) proposed a novel, attention-based classification block for performing zero- and few-shot, cross-lingual learning on the HatEval data set. While they state that transfer learning is effective for hate speech detection, we argue that there is a need to investigate hate speech targets separately since these models consistently fail misogyny classification.

6 Conclusion

We demonstrate that cross-lingual, zero-shot transfer learning, in its traditional settings, is not a feasible solution for solving the lack of models and labeled corpora for hate speech detection. We argue that hate speech is language specific, and NLP approaches to identifying hate speech must account for that specificity and the adoption of related techniques must be done with care (Bianchi and Hovy, 2021). We plan to expand this evaluation to other languages and to investigate a solution based on bias mitigation (Nozza et al., 2019; Kennedy et al., 2020) and on pragmatic role-aware models (Holgate et al., 2018; Pamungkas et al., 2020) to reduce the impact of this problem on classification. Future work will also focus on modeling language's social factors (Hovy and Spruit, 2016; Hovy, 2018; Hovy and Yang, 2021), such as speaker and receiver characteristics, and study their impact on hate speech detection classifiers.

Ethical Considerations

We are aware that the inherent (gender) biases of sentence and word embeddings are affecting the model's performance on detecting hate speech towards women (Bolukbasi et al., 2016; Sheng et al., 2019; Nangia et al., 2020; Nozza et al., 2021). We believe that this issue plays a role in the classification models. However, in this paper we extensively demonstrate that the presence of taboo interjections is one of the main hurdles that specifically hinder zero-shot, cross-lingual hate speech detection results.

Finally, we want to highlight that the presented findings are specifically related to the considered languages and data sets. Hopefully, our work will generate more conscious research about the use of hate speech detection models in zero-shot, cross-lingual frameworks.

Acknowledgments

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). Thanks to Dirk Hovy, Federico Bianchi, Salvatore Alessandro Casa, Tommaso Fornaciari, and Silvia Terragni for their invaluable feedback. Debora Nozza is a member of the Bocconi Institute for Data Science and Analytics (BIDSA) and the Data and Marketing Insights (DMI) unit.

References

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.

Federico Bianchi and Dirk Hovy. 2021. On the gap between adoption and understanding in NLP. In *Findings of the Association for Computational Linguistics: ACL 2021*. 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.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

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.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515. AAAI Press.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

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.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

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.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, 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*.

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, 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.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Hajung Sohn and Hyunju Lee. 2019. MC-BERT4HATE: hate speech detection using multichannel BERT for different languages and translations. In *2019 International Conference on Data Mining Workshops, ICDM Workshops 2019, Beijing, China, November 8-11, 2019*, pages 551–559. IEEE.

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.

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1151–1160, Varna, Bulgaria. INCOMA Ltd.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. 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)*.

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.

A Additional Post-Hoc Explanation

Figure 2 shows the hierarchically clustered explanations from SOC for an example of non-hateful speech wrongly classified as hateful by the monolingual English model. It is evident how the (incorrect) high score of the hashtag eclipses the influence of non-hateful words such as *days*, *kids*, and *school*.

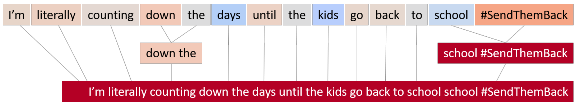


Figure 2: Hierarchical explanations of the incorrect prediction of a non-hateful English tweet by a monolingual model trained on English and tested on English data.

B Additional Results

		Immigrants+Women		
	Test	IT	EN	ES
Train	**IT**	0.804	*0.571***	*0.596***
	EN	*0.564***	0.416	*0.648***
	ES	*0.513***	*0.576***	0.752
	EN+ES	*0.513***	0.335**	0.768
	ES+IT	0.797	*0.572***	0.744
	EN+IT	0.802	0.399	*0.632***
	Baseline	-	0.651	0.730

Table 4: Results in terms of macro-F1 for the merged corpora containing hate speech towards immigrants and women. Monolingual results are underlined. Zero-shot cross-lingual results are highlighted in *italic*.
* = differs significantly from monolingual at p $\leq$ 0.05.
** = significant difference at p $\leq$ 0.01.

C Experimental Configuration

C.1 Consistent Data sets

We use benchmark hate speech data sets with consistent definitions, annotation schema, and collection strategies. All the three data sets (Bosco et al., 2018; Fersini et al., 2018; Basile et al., 2019) refer to the same definitions of hate speech towards immigrant and women.[3] This paper focuses on the common binary classification task (hateful/non-hateful) across all data sets, ensuring the same annotation schema. Finally, all data sets have been collected by following three strategies: (1) monitoring potential victims of hate accounts, (2) downloading the history of identified haters and (3) filtering Twitter streams with keywords, i.e. words, hashtags and stems.

For experimental evaluation, we use the data set splits provided in the associated shared task for comparability with previous work.

C.2 Implementation Details

We implement the proposed work exploiting the public code implementation of the classification model presented by Kennedy et al. (2020)[4]. We use their hyperparameter configuration for training: batch size is set to 32, the learning rate of the Adam optimizer is set to 2×10^{-5}, the loss function is the binary cross entropy.

Computing Infrastructure We independently run the experiments on two machines: the first one is equipped with two NVIDIA RTX 2080TI and has 64GB of RAM. The other one is equipped with four GPUs, NVIDIA GTX 1080TI, and has 32GB of RAM.

[3] https://github.com/msang/hateval/blob/master/annotation_guidelines.md

[4] https://github.com/BrendanKennedy/contextualizing-hate-speech-models-with-explanations