Hybrid Emoji-Based Masked Language Models for Zero-Shot Abusive Language Detection

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, Serena Villata

Hybrid Emoji-Based Masked Language Models for Zero-Shot Abusive Language Detection

Michele Corazza[†], **Stefano Menini**[‡],
Elena Cabrio[§], **Sara Tonelli**[‡], **Serena Villata**[§]
[†]University of Bologna, Italy
[‡]Fondazione Bruno Kessler, Trento, Italy
[§]Université Côte d'Azur, CNRS, Inria, I3S, France
michele.corazza2@unibo.it
{menini,satonelli}@fbk.eu
{elena.cabrio,serena.villata}@univ-cotedazur.fr

Abstract

Recent studies have demonstrated the effectiveness of cross-lingual language model pre-training on different NLP tasks, such as natural language inference and machine translation. In our work, we test this approach on social media data, which are particularly challenging to process within this framework, since the limited length of the textual messages and the irregularity of the language make it harder to learn meaningful encodings. More specifically, we propose a hybrid emoji-based Masked Language Model (MLM) to leverage the common information conveyed by emojis across different languages and improve the learned cross-lingual representation of short text messages, with the goal to perform zero-shot abusive language detection. We compare the results obtained with the original MLM to the ones obtained by our method, showing improved performance on German, Italian and Spanish.

1 Introduction

The extensive use of large-scale self-supervised pre-training has greatly contributed to recent progress in many Natural Language Processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019; Conneau and Lample, 2019). In this context, masked language modelling objectives represent one of the main novelties of these approaches, where some tokens of an input sequence are randomly masked, and the objective is to predict these masked positions taking the corrupted sequence as input. Still, little attention has been devoted to the adaptation of these techniques to tasks dealing with social media data, probably because they are characterized by a very domain-specific language, with high variability and instability. Nevertheless, all these challenges make social media data an interesting testbed for novel deep-learning architectures, around the research question: *how could the masking mechanism be adapted to target social media language?*

In this paper, we address the above issue by adapting a novel architecture for cross-lingual models called XLM (Conneau and Lample, 2019) to zero-shot abusive language detection, a task that has gained increasing importance given the recent surge in abusive online behavior and the need to develop reliable and efficient methods to detect it. In particular, we evaluate two methods to pre-train bilingual language models, one similar to the original XLM masked model, and the other based on a novel hybrid emoji-based masked model. We then evaluate them on zero-shot abusive language detection for Italian, German and Spanish, showing that, although our results are below the state-of-the-art in a monolingual setting, the proposed solutions to adapt XLM to social media data are beneficial and can be effectively extended to other languages.

In the following, Section 2 discusses the related work. Section 3 describes our approach to train cross-lingual models for social media data classification, while Section 4 presents the experimental setup. Section 5 reports on the evaluation results, while Section 6 summarizes our findings.

2 Related work

The focus of this paper is the abusive language detection task, which has been widely explored in the last years thanks to numerous datasets, approaches and shared tasks (Waseem et al., 2017; Fišer et al., 2018; Carmona et al., 2018; Wiegand et al., 2018; Bosco et al., 2018; Zampieri et al., 2019b; Roberts et al., 2019) covering different languages. An increasing number of approaches has been proposed to detect this kind of messages (for a survey on the task, see (Schmidt and Wiegand, 2017) and (Fortuna and Nunes, 2018)).

Abusive language detection is usually framed as a supervised learning problem, built using a combination of manually crafted features such as n-grams (Wulczyn et al., 2017), syntactic features (Nobata et al., 2016), and linguistic features (Yin et al., 2009), to more recent neural networks (Park and Fung, 2017; Zhang and Tepper, 2018; Agrawal and Awekar, 2018; Corazza et al., 2018). (Lee et al., 2018) address a comparative study of various learning models on the *Hate and Abusive Speech on Twitter* dataset (Founta et al., 2018), while (Zampieri et al., 2019a) build the *Offensive Language Identification Dataset* and experiment with SVMs, BiLSTM and CNN both on the binary abusive language classification and on a more fine-grained categorization. Our work deals with the same task, addressed from a cross-lingual perspective.

In recent years, some proposals have been made to tackle abusive language detection in a cross-lingual framework (Sohn and Lee, 2019; Pamungkas and Patti, 2019; Casula et al., 2020), with some attempts at zero-shot learning (Stappen et al., 2020). Most systems, however, rely on pretrained models and do not investigate the potential of in-domain data for pretraining. Additionally, as regards masked language models, we are not aware of any work in the literature modifying masking mechanisms for this task.

3 Cross-Lingual Language Models

3.1 MLM and HE-MLM training objectives

Our basic architecture relies on the XLM approach described in (Conneau and Lample, 2019), specifically developed to learn joint multilingual representations enabling knowledge transfer across languages. In particular, we borrow from XLM the method developed for unsupervised machine translation, that relies on the Masked Language Model (MLM) objective (Devlin et al., 2019) applied to multiple monolingual datasets as pretraining. We choose to adopt the unsupervised approach because the alternative (i.e., the supervised one based on the Translation Language Modeling) would need to be trained on parallel data, which are not available at scale for social media. As in XLM, we use Byte Pair Encoding (BPE) (Sennrich et al., 2016) to learn a shared vocabulary of common subwords between the languages. This technique has proven beneficial to the alignment of embeddings from different languages, when used on languages that share some common traits, such as alphabet and digits.

Following the original approach to MLM, 15% of the tokens in a sentence get selected, which get masked 80% of the times, replaced by a random token in 10% of the cases and kept unchanged 10% of the times. In order to reduce the impact of relatively frequent words on the model, tokens are sampled according to a multinomial distribution that is proportional to the square root of their inverted frequency. While the original XLM operates on streams of text, split by sentence separators, we split the stream of tweets, so that each example contains only one tweet.

Since using a standard pre-trained language model to classify irregular data obtained from social networks would prove very challenging, we try to adapt our cross-lingual model to social media data as much as possible. Specifically, we rely on two main intuitions: emojis are linked to emotion expressions, correlated in turn with various forms of online harassment (Arslan et al., 2019). Besides, emojis could be seen as common traits that are present in tweets across different languages, maintaining a similar meaning at least when comparing Indo-European languages (Lu et al., 2016). If we consider the data used in this paper, we can find a good coverage of emojis, with 16.82% of the tweets containing at least one emoji for English, 16.15% for German, 7.68% for Italian, and 18.39% for Spanish. Furthermore, in these datasets the most frequent emojis are shared among all the four languages, with 'red heart', 'face with tears of joy', 'thinking face' and 'smiling face with heart-eyes' among the top ten emojis in each dataset. We therefore compare a standard masked language model with one that targets emoji prediction instead of the *cloze* task (Taylor, 1953). However, since emojis are not always present in each tweet, we adopt a hybrid approach: when emojis are not present, the previously described MLM objective is trained. When emojis are found, we select them as candidates to be masked 80% of the time, replaced by a random token 10% of the time or kept unchanged 10% of the time as in MLM. With this technique, which we call Hybrid Emoji-based Masked Language Model (HE-MLM) we can use all the available data, while also leveraging the common information conveyed by emojis.

We test also a variant of MLM and HE-MLM, in which we put special tokens "<emoji>" and

"</emoji>" around all emojis in the dataset, given that we are effectively performing two different tasks with the same model. This approach allows the model to distinguish between normal words and emojis in the text while training masked language models.

3.2 Fine-tuning for abusive language detection

In order to assess how invariant our tweet embeddings are with respect to the language provided as input to the encoder, we create a zero-shot framework, where the system is only trained on English tweets and is evaluated on multiple languages. In particular, we first load the pretrained transformer and attach to it a single feed-forward layer on top of the encoder with a single, sigmoid activated output neuron. The entire model is then fine-tuned on the English hate speech detection dataset using a binary cross-entropy loss function. The system uses early stopping with the minimum F1 score between the two classes as a stopping criterion, relying on a balanced dataset that contains all languages as validation set. Finally, the performance is evaluated on the German, Italian and Spanish test sets to assess how our classifier performs on the different languages using the bilingual models.

4 Experimental setting

4.1 Datasets

Since we run our classification in a zero-shot scenario, we use English data for training, and tweets in German, Italian and Spanish for validation and test. The datasets we used and the related number of tweets are reported in Table 1. To guarantee a comparable setting for our experiments, we carefully investigated data samples and the annotation schemes adopted for the different languages, concluding that the tweet content as well as the binary annotation tagsets (hate-speech/offensive and other) of the datasets are similar enough to use them in the same classification framework. Also the class distribution is similar, with the abusive class covering around 30% of the tweets in each dataset.

To pre-train our cross-lingual language models with in-domain data, we gather 5 million tweets for each of the targeted languages (i.e., *English*, *German*, *Italian* and *Spanish*). Such tweets have been collected in different time periods spanning from March to August 2019 through the Twitter Streaming API using the stopwords of the target language as filter to query the API, as in (Scheffler, 2014).

ENGLISH (Waseem and Hovy, 2016)		
Train	Validation	Test
9,534	–	
GERMAN (Wiegand et al., 2018)		
Train	Validation	Test
–	1,002 (+ 1,002 EN)	3,532
ITALIAN (Bosco et al., 2018)		
Train	Validation	Test
–	600 (+ 600 EN)	1,000
SPANISH (Basile et al., 2019)		
Train	Validation	Test
–	500 (+ 500 EN)	1,600

Table 1: Number of tweets used for fine-tuning (English), validation and testing (German, Italian, Spanish). For each classification language, the validation set comprises the same amount of language-specific and English tweets.

4.2 Data splitting

Concerning the dataset splits into training and test instances, for the English dataset - since no standardized split is provided - we randomly selected 60% of the dataset for training, 20% for validation and 20% for testing. For the German and Italian datasets, we use the training and test split provided by the Germeval and Evalita task organisers, respectively. In both cases, we use 20% of the training set as validation set. Whenever we split the datasets, we use the *train_test_split* function from scikit-learn (Pedregosa et al., 2011), using 42 as a seed value. Finally, for Spanish, we use the development, test and training set provided by the HatEval task organisers.

For each combination of languages tested in our experiments (i.e., English-German, English-Italian and English-Spanish), the validation test is obtained by keeping the language-specific validation set as is and undersampling the English one to the same size, so that each language has the same weight during the early stopping phase.

Before classification, the text is first lowercased, all accents are removed, then it is tokenized with (Koehn et al., 2007)'s system. Finally, Byte Pair Encoding is applied to all datasets by using the

		pre-trained model			MLM			MLM with <emoji>			HE–MLM			HE–MLM with <emoji>		
Lang.	Category	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN	Non-hate speech	0.682	0.993	0.809	0.700	0.736	0.688	0.698	0.387	0.465	0.690	0.830	0.738	0.685	0.907	0.773
	Hate speech	0.423	0.013	0.023	0.340	0.293	0.257	0.320	0.625	0.412	0.304	0.185	0.175	0.248	0.099	0.101
	macro avg	0.553	0.503	0.417	0.520	0.515	**0.473**	0.509	0.506	0.439	0.497	0.507	0.456	0.466	0.503	0.437
DE	Non-hate speech	0.660	0.998	0.795	0.626	0.371	0.359	0.477	0.114	0.141	0.575	0.319	0.283	0.656	0.821	0.667
	Hate speech	0.142	0.002	0.005	0.294	0.637	0.379	0.342	0.890	0.487	0.286	0.676	0.375	0.112	0.180	0.109
	macro avg	0.401	0.500	**0.400**	0.460	0.504	0.369	0.409	0.502	0.314	0.430	0.497	0.329	0.384	0.500	**0.388**

Table 2: Average performance (10 runs) on English and German, comparing the En-De model from (Conneau and Lample, 2019) pre-trained on Wikipedia, our MLM re-trained on English and German tweets, and our Hybrid Emoji-based MLM (HE–MLM). MLM and HE–MLM are evaluated with and without the use of <emoji> tokens.

		MLM			MLM with <emoji>			HE–MLM			HE–MLM with <emoji>		
Lang.	Category	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN	Non-hate speech	0.473	0.181	0.220	0.699	0.712	0.610	0.449	0.317	0.321	0.616	0.732	0.635
	Hate speech	0.326	0.837	0.458	0.113	0.293	0.162	0.273	0.689	0.374	0.170	0.270	0.179
	macro avg	0.400	0.509	0.339	0.406	0.503	0.386	0.361	0.503	0.347	0.393	0.501	**0.407**
IT	Non-hate speech	0.688	0.718	0.679	0.680	0.891	0.765	0.698	0.740	0.713	0.664	0.446	0.452
	Hate speech	0.352	0.301	0.262	0.221	0.122	0.137	0.381	0.326	0.326	0.296	0.587	0.349
	macro avg	0.520	0.510	0.470	0.451	0.507	0.451	0.539	0.533	**0.519**	0.480	0.517	0.401

Table 3: Average performance (10 runs) on English and Italian after re-training the Masked Language Model (MLM) on tweets and using Hybrid Emoji-based MLM (HE–MLM).

fastBPE implementation[1]. We evaluate the classifier performance over a maximum of 100 training epochs, and use an early stopping mechanism with a patience of 5. The selected model is then used to evaluate performance on the test set.

4.3 Pretraining methods

Since we want to assess the impact of emojis on the pretraining results, we train four different configurations:

- Using the base MLM training objective;
- Using the base MLM training objective and <emoji> tokens;
- Using the HE-MLM training objective;
- Using the HE-MLM training objective and <emoji> tokens.

For each configuration, we pretrain two models in order to reduce the impact of random initialization on the final results and we fine-tune each model 10 times (20 total). The final results are obtained by averaging the results of these 20 runs.

5 Evaluation

We report the experiment results for each language in Tables 2, 3, 4. For all languages, training is performed using only English data.

[1]https://github.com/glample/fastBPE

Results for German (Table 2) show that using in-domain unlabeled data from Twitter instead of pre-trained models yields an improvement in performance on English, while on German the model is not able to outperform the pre-trained model. In this case, however, the pretrained model is only learning the non-hate class, while the other three models all achieve non zero recall on both classes. Beside the baseline, the HE–MLM model with <emoji> is the best performing one on the German data, while on English the best performance is achieved by using the vanilla MLM model.

We evaluate MLM and HE–MLM also for zero-shot Italian hate speech classification, comparing the configurations with and without <emoji> tokens like in the previous experiments (Table 3). For English, the best performing model is HE–MLM with emoji tokens, while on Italian the HE–MLM model with no tokens is better in terms of macro averaged F1. When comparing configurations, we observe that the MLM model with emoji tokens has better F1 score than the MLM one in the non hate speech class, while the MLM model has improved performance on the hate class. This results in the MLM model having better macro average F1 for Italian, while the MLM model with emoji tokens shows higher average F1 on English. When considering the hybrid emoji-based models, HE–MLM achieves a higher F1 for the hate speech class in English and for the non hate class in Italian. This results in the HE–MLM model having a higher

		MLM			MLM with <emoji>			HE–MLM			HE–MLM with <emoji>		
Lang.	Category	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN	Non-hate speech	0.667	0.927	0.762	0.692	0.847	0.706	0.752	0.308	0.305	0.699	0.722	0.676
	Hate speech	0.072	0.072	0.048	0.118	0.146	0.090	0.316	0.698	0.370	0.296	0.307	0.234
	macro avg	0.369	0.499	0.405	0.405	0.497	0.398	0.534	0.503	0.337	0.498	0.515	**0.455**
ES	Non-hate speech	0.599	0.760	0.655	0.577	0.679	0.599	0.598	0.725	0.648	0.595	0.740	0.643
	Hate speech	0.365	0.267	0.275	0.407	0.307	0.298	0.438	0.303	0.332	0.451	0.275	0.280
	macro avg	0.482	0.513	0.465	0.492	0.493	0.449	0.518	0.514	**0.490**	0.523	0.507	0.461

Table 4: Average performance (10 runs) on English and Spanish after re-training the Masked Language Model (MLM) on tweets and using Hybrid Emoji-based MLM (HE–MLM).

macro averaged F1 in the Italian language, while the HE–MLM model with emoji tokens is better on English.

As a final test, we evaluate the performance of the model trained on English and Spanish (Table 4). Our English–Spanish models show a similar behaviour to the one observed for the English–Italian pair. In terms of macro averages, the HE–MLM model with emoji tokens has a higher average F1 for English, while the HE–MLM model has higher macro F1 for Spanish.

On all the runs, the classifier achieves a lower performance on German than on the other two languages, while the results on Italian and Spanish are comparable. This confirms the findings in (Corazza et al., 2020) suggesting that, even when using the same classification framework, experimental setting and amount of training data, offensive speech detection on German achieves lower performance than on other languages. This may have two possible reasons: on the one hand, German may have inherent characteristics that make it more challenging to classify for abusive language detection, for example the presence of compound words makes hashtag splitting more error-prone. On the other hand, the Germeval dataset was built by sampling data from specific users and avoiding keyword-based queries, so to obtain the highest possible variability in the offensive language. This led to the creation of a very challenging dataset, where lexical overlap between training and test data is limited and where hate speech is not associated with specific topics or keywords, as suggested in (Wiegand et al., 2019).

6 Conclusions

In this paper, we present a novel zero-shot framework for multilingual abusive language detection. We compare two cross-lingual language models, i.e., standard MLM and a hybrid version of MLM based on emojis (HE–MLM), highlighting that the latter shows some advantages over the MLM model when used on social media data: first of all, when using emojis, the pre-training step is aimed at predicting tokens that are inherently more relevant for the final abusive language detection task whenever possible, as opposed to random tokens. Secondly, emojis convey similar meaning in the languages that we consider, serving as a common trait between languages during pre-training. We also use <emoji> tokens around emojis to help the system discriminate between the two training objectives when using HE–MLM.

The proposed methods represent a novel contribution with respect to social media data processing and abusive language detection. Our aim is not to create a system comparable with monolingual state-of-the-art solutions, but to investigate the possibility to use an unsupervised approach for zero-shot cross lingual abusive language detection. As a first step in this direction, we focused on four European languages, for which similar data were available. The only existing work dealing with zero-shot abusive language detection, presented in (Stappen et al., 2020), only focuses on a language pair and, while obtaining promising results, relies on the English and Spanish corpora annotated for HatEval 2019 following the same guidelines and focusing on hate against immigrants and women. Our approach aims to be more robust, comparing datasets annotated for different shared tasks which may adopt slightly different guidelines.

In the near future, we plan to further extend the social media-specific datasets we are collecting to pre-train HE-MLM, since 5 million tweets we used for each language correspond to a small-sized corpus compared to standard pre-trained language models. Then, to investigate whether our results can be generalised also when dealing with typologically different languages, we will test our approach on additional abusive language datasets covering other languages (Ousidhoum et al., 2019; Zampieri et al., 2020).

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *ECIR*, pages 141–153.

Pinar Arslan, Michele Corazza, Elena Cabrio, and Serena Villata. 2019. Overwhelmed by Negative Emotions? Maybe You Are Being Cyber-bullied! In *Proceedings of the 34th ACM/SIGAPP Symposium On Applied Computing (SAC)*, Limassol, Cyprus.

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

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of EVALITA 2018*.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *Proceedings of IberEval 2018*, pages 74–96.

Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. FBK-DH at SemEval-2020 Task 12: Using Multi-channel BERT for Multilingual Offensive Language Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2020)*. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing different supervised approaches to hate speech detection. 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.*

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

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.

Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors. 2018. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Antigoni-Maria 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. In *ICWSM*, pages 491–500.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL: demo and poster*, pages 177–180.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. *CoRR*, abs/1808.10245.

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.

Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, and Qiaozhu Mei. 2016. Learning from the ubiquitous language: An empirical analysis of emoji usage of smartphone users. In *UbiComp*, pages 770–780, New York, NY, USA. ACM.

Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*, pages 145–153.

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 4667–4676, Hong Kong, China. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 363–370. Association for Computational Linguistics.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Workshop on Abusive Language Online*, pages 41–45.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors. 2019. *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy.

Tatjana Scheffler. 2014. A german twitter snapshot. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, Berlin, Germany.

Hajung Sohn and Hyunju Lee. 2019. MC-BERT4HATE: hate speech detection using multi-channel 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.

Wilson L. Taylor. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Bulletin*, 30(4):415–433.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. 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 *GermEval 2018*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of WWW Conference*, pages 1391–1399.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the Web*, pages 1–7.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

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 SemEval*.

Robinson D. Zhang, Z. and J. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *ESWC*, pages 745–760. Springer Verlag.