Improving Zero-Shot Cross-Lingual Hate Speech Detection with Pseudo-Label Fine-Tuning of Transformer Language Models

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, Gareth Tyson
Queen Mary University of London
h.b.zia@qmul.ac.uk, i.castro@qmul.ac.uk, a.zubiaga@qmul.ac.uk, g.tyson@qmul.ac.uk

Abstract

Hate speech has proliferated on social media platforms in recent years. While this has been the focus of many studies, most works have exclusively focused on a single language, generally English. Low-resourced languages have been neglected due to the dearth of labeled resources. These languages, however, represent an important portion of the data due to the multilingual nature of social media. This work presents a novel zero-shot, cross-lingual transfer learning pipeline based on pseudo-label fine-tuning of Transformer Language Models for automatic hate speech detection. We employ our pipeline on benchmark datasets covering English (source) and 6 different non-English (target) languages written in 3 different scripts. Our pipeline achieves an average improvement of 7.6% (in terms of macro-F1) over previous zero-shot, cross-lingual models. This demonstrates the feasibility of high accuracy automatic hate speech detection for low-resource languages. We release our code and models at https://github.com/harisbinzia/ZeroshotCrosslingualHateSpeech.

Introduction

Despite its benefits, social media has also been used to disseminate hateful material at an unprecedented scale (Müller and Schwarz 2018). The sheer volume of hateful content poses a challenge for timely and effective human moderation. Hence, automatic hate speech detection has received significant attention from the Natural Language Processing (NLP) research community (Schmidt and Wiegand 2017). Still, much of this research focuses on a single language, generally English (Fortuna and Nunes 2018; Poletto et al. 2020; Vidgen and Derczynski 2020), and lesser-resourced languages have been rarely studied. To tackle this limitation, we study zero-shot, cross-lingual transfer learning (Artetxe and Schwenk 2019) *i.e.* training on a high-resource (source) language and testing on a low-resource (target) language.

Existing methods for zero-shot, cross-lingual hate speech detection leverage language-agnostic sentence representations to embed training data in the high-resource language (mostly English) or multilingual transformer language models fine-tuned on a high-resource language (again mostly English) (Pelicon et al. 2021). However, recent research suggests that these methods struggle with taboo interjections in the target language (Nozza 2021), *i.e.* common language-specific hateful expressions *e.g. puta* (meaning *bitch*) in Spanish. Since English does not necessarily use equivalent words in the same way, zero-shot, cross-lingual models trained/ fine-tuned on English fail to capture the context of these expressions. Consequently, they consider them hate speech regardless of their use in context. To address this limitation, we propose a novel pipeline based on pseudo-label fine-tuning of transformer language models for zero-shot, cross-lingual hate speech detection. We experiment with benchmark datasets in 6 languages written in 3 different scripts, outperforming previous zero-shot, cross-lingual results, with preliminary proof of improvement on taboo expressions. Our contributions include:

1. We propose a novel model to use pseudo-labeled in-domain data in the target language, in addition to gold-labeled data in English to fine-tune transformer language models for zero-shot, cross-lingual hate speech detection.
2. Our method consistently outperforms the previous state-of-the-art zero-shot, cross-lingual models and improves the comprehension of taboo interjections by 11.1% (average macro-F1 improvement over different datasets).

Related Work

The vast majority of research in hate speech detection is monolingual, with English being the most prevalent language due to the availability of resources (Waseem and Hovy 2016; Wulczyn, Thain, and Dixon 2017; Davidson et al. 2017; Zampieri et al. 2019). There has been limited research on non-English hate speech detection too, *e.g.* for Italian (Sanguinetti et al. 2018), and French (Chiril et al. 2020). Lately, several shared tasks have helped increase the coverage of non-English hate speech datasets, *e.g.* AMIEvalita 2018 (Fersini, Nozza, and Rosso 2018), AMIIberEval 2018 (Fersini, Rosso, and Anzovino 2018) and AMIEvalita 2020 (Fersini, Nozza, and Rosso 2020). These covered misogyny detection in Italian, Spanish and English. GermEval 2018 (Wiegand, Siegel, and Ruppenhofer 2018) explored identification of offensiveness in German tweets, HatEval 2019 (Basile et al. 2019) covered hate speech against immigrants and women in Spanish and English. HASOC 2019 (Mandl et al. 2019) and HASOC 2020 (Mandl et al. 2020) introduced resources for hate speech detection in Hindi, German and English. OffensEval 2020 (Zampieri et al. 2020) featured offensive language identification datasets in Arabic, Danish, Greek, Turkish and English. We direct interested readers to relevant surveys for further information (Schmidt and Wiegand 2017; Fortuna and Nunes 2018; Poletto et al. 2020; Vidgen and Derczynski 2020; Pamungkas, Basile, and Patti 2021b).

Only a handful of studies have investigated zero-shot cross-lingual transfer learning for hate speech detection. Both Pamungkas and Patti (2019) and Jiang and Zubiaga (2021) proposed hybrid approaches with neural models and a multilingual lexicon to cross-domain and cross-lingual detection of abusive content. A novel attention-based classification block for zero-shot, cross-lingual learning was proposed by Stappen, Brunn, and Schuller (2020). They demonstrated highly competitive results on the Spanish and English subsets of HatEval 2019 (Basile et al. 2019). Bigoulaeva, Hangya, and Fraser (2021) used bilingual word embedding-based classifiers to transfer learn hate speech detection for German from English. Pamungkas, Basile, and Patti (2021a) experimented with traditional and recent neural architectures, and proposed two joint-learning hate speech detection models, using different multilingual language representations to transfer knowledge between pairs of languages.

Closest to our work is Pelicon et al. (2021) and Nozza (2021). The former use a multilingual Bidirectional Encoder Representations from Transformers (mBERT) (Devlin et al. 2018) and Language Agnostic SEntence Representations (LASER) (Artetxe and Schwenk 2019) with multilayer perceptron classifier to generalize hate speech detection from English to other languages. In contrast, the latter only fine-tuned mBERT on English for hate speech detection and demonstrated that zero-shot, cross-lingual models were not able to capture target language specific taboo interjections. Our work is different in that we exploit pseudo-labeled in-domain data in target language along with gold-labeled data in English to fine-tune transformer language models that overcome the limitation of taboo expressions highlighted by Nozza (2021) and significantly enhances their performance in zero-shot, cross-lingual settings.

Methodology

We propose to use the cross-lingual hate speech classifier trained on high-resource (source) language data as a teacher to obtain pseudo-labels for training monolingual hate speech model on the low-resource (target) language. The pipeline is shown in Figure 1 and consists of three main steps:

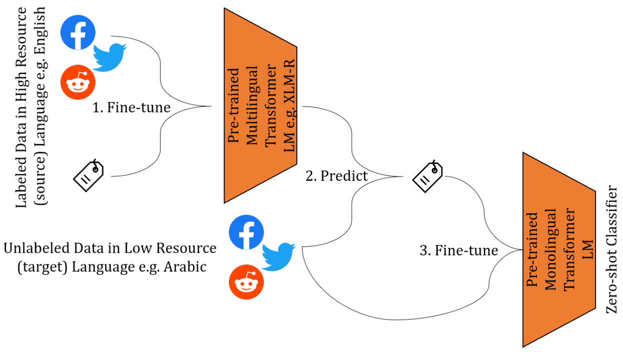

Labeled Data in High Resource (source) Language e.g. English
1. Fine-tune
Pre-trained Multilingual Transformer LM e.g. XLM-R
2. Predict
Unlabeled Data in Low Resource (target) Language e.g. Arabic
3. Fine-tune
Pre-trained Monolingual Transformer LM
Zero-shot Classifier

Figure 1: Our proposed zero-shot, cross-lingual transfer learning pipeline for hate speech detection.

1. First, we fine-tune a pre-trained multilingual transformer language model on gold-labeled source language data. This gives us our zero-shot, cross-lingual teacher.
2. Next, we perform inference with the zero-shot, cross-lingual teacher on in-domain target language data to predict labels. This generates a new psuedo-labeled dataset in the target language.
3. Finally, we use the in-domain target language data and its predicted (pseudo) labels to fine-tune a monolingual transformer language model pre-trained on the target language. This gives us our final classifier (zero-shot, monolingual student).

Our approach provides two main benefits: (*i*) it circumvents the need for gold-labels in the target language; (*ii*) it allows the final zero-shot classifier to better capture the hateful expressions (including taboo interjections) in target language.

Experimental Setup

Datasets

We use English as the source and 6 other target languages: Spanish, Italian, German, Arabic, Greek and Turkish. We use HatEval 2019 (Basile et al. 2019), AMIEvalita 2018 (Fersini, Nozza, and Rosso 2018), GermEval 2018 (Wiegand, Siegel, and Ruppenhofer 2018) and OffensEval 2020 (Zampieri et al. 2020) datasets (see Table 1). To ensure consistency in experiments, we use source and target language data pertaining to the same shared task.[1] For each dataset, we treat the English subset as the source language and non-English as target language. We keep the original splits[2] for comparability with previous work.

Pre-trained Language Models

For pre-trained language models, we rely on community-driven Hugging Face's Model Hub.[3] Specifically, we use XLM-R$_{large}$ (Conneau et al. 2019) as zero-shot, cross-lingual teacher, RoBERTa for Spanish (Gutiérrez-Fandiño et al. 2021) & Greek (Papaevagelou 2021) and BERT for Italian (Polignano et al. 2019), German (Chan, Schweter, and Möller 2020), Arabic (Antoun, Baly, and Hajj 2020) & Turkish (Schweter 2020). All models are fine-tuned using Hugging Face's Transformers (Wolf et al. 2020) with input sequence length of 128, batch size of 32 and learning rate of 2e-5 for 3 epochs.

[1] Except for GermEval 2018 dataset used in combination with OffensEval 2020 in line with previous studies.

[2] Note: we do not use gold-labels from non-English train and valid splits (so as to ensure zero-shot settings) but use their text to obtain pseudo-labels.

[3] https://huggingface.co/models

Dataset	Task	Labels	Lang.	Train	Valid	Test	PIR
HatEval	Hate Speech Detection	1 - hateful 0 - non hateful	EN	9000	1000	3000	0.42
			ES	4500	500	1600	0.41
AMIEvalita	Automatic Misogyny Identification	1 - misogynous 0 - non misogynous	EN	3600	400	1000	0.44
			IT	3600	400	1000	0.46
GermEval	Offensive Language Identification	1 - offensive 0 - non offensive	DE	4508	501	3532	0.33
OffensEval	Offensive Language Identification	1 - offensive 0 - non offensive	EN	11916	1324	860	0.32
			AR	7055	784	2000	0.19
			EL	7868	875	1544	0.26
			TR	28149	3128	3528	0.19

Table 1: Dataset Statistics. PIR is Positive Instance Rate.

Baselines

We compare three strong baselines, following Pelicon et al. (2021): (*i*) Pre-trained XLM-R$_{large}$ fine-tuned only on labeled data in source (English) language (*i.e.* our zero-shot, cross-lingual teacher); (*ii*) Logistic Regression trained on language agnostic sentence embeddings LASER (Artetxe and Schwenk 2019) and LaBSE (Feng et al. 2020) using labeled data in source (English) language; and (*iii*) the previous best zero-shot, cross-lingual results for each dataset.

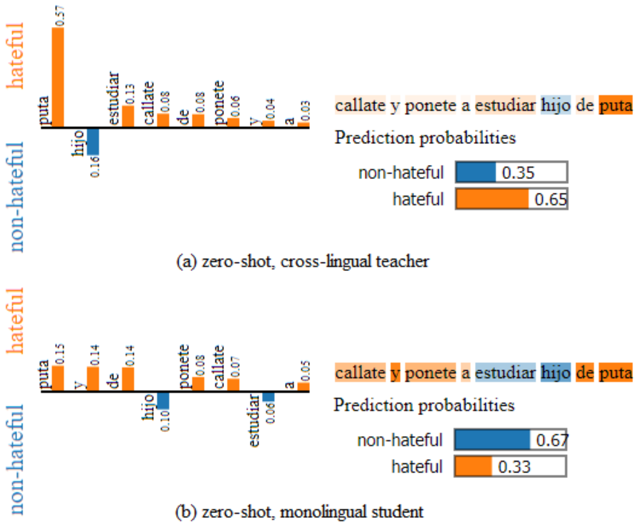
Results

The results (macro-F1 scores) are given in Table 2. The pseudo-label fine-tuned model outperforms its teacher in all settings. The largest improvement is for OffensEval-EL, where the macro-F1 increases from 0.67 to 0.72 (7.46%). The smallest improvement (2.08%) is observed in AMIEvalita-IT. This may be attributed to the uniqueness of the task for which the Italian model was trained (*i.e.* automatic misogyny identification) as language models are less likely to capture such features during pre-training. On average, the proposed method achieves 4.5% macro-F1 improvement over its teacher, 20.5% over logistic regression classifiers and 7.6% over previous state-of-the-art models.

Analysis of Taboo Expressions. We next seek to understand *why* our approach outperforms prior baselines. Nozza (2021) demonstrated that zero-shot, cross-lingual models fine-tuned only on source (English) language data fail to capture target language specific taboo expressions such as *puta* (meaning *bitch*) in Spanish and *porca* (meaning *slut*) in Italian. While derogatory for women, these words are often used as intensifiers in non-hateful contexts *e.g. hijo de puta* (meaning *son of a bitch*) and *porca puttana* (meaning *holy shit*).

Following (Nozza 2021), we also analyze the performance of our proposed pseudo-label fine-tuned model on texts containing taboo expressions and compare it with zero-shot, cross-lingual teacher that is fine-tuned on English only. Specifically, we examine expressions *puta* and *porca* in the test subsets of HatEval-ES and AMIEvalita-IT.[4]

Figure 2 shows the LIME (Ribeiro, Singh, and Guestrin 2016) explanation of an example non-hateful Spanish tweet wrongly classified by zero-shot, cross-lingual teacher and correctly classified by our zero-shot, monolingual student model. The teacher model gives high hateful importance to term like *puta* regardless of its context as it considers the literal meaning of individual words. On the contrary, our proposed method teach the model that *puta* is a taboo expression and does not imply hatefulness in this particular context. The explanation also reveal our model's ability to assign correct importance to non-hateful words *e.g.* teacher model considers the word *estudiar* (meaning *study*) as hateful whereas our model correctly identifies it as non-hateful.



(a) zero-shot, cross-lingual teacher

(b) zero-shot, monolingual student

Figure 2: LIME explanations of predictions of a non-hateful Spanish tweet by (a) zero-shot, cross-lingual teacher and (b) zero-shot, monolingual student. English translation: "shut up and study son of a bitch".

To further exemplify this finding, in Table 3 we report the macro-F1 of zero-shot, monolingual student and its teacher in predicting the labels of tweets containing taboo expressions *puta* and *porca* in the test subsets of HatEval-ES and AMIEvalita-IT respectively. On average, the student model increases the macro-F1 by 11.1%. These numbers show that our proposed method understands these general exclamations with much better accuracy and thus limits the number

[4] The reason we restrict our analysis to HatEval-ES and AMIEvalita-IT is because remaining datasets are tagged for offensive language identification task thus any occurrence of taboo expression must be in the positive class irrespective of the context.

Dataset	Previous zero-shot, cross-lingual results	Embeddings		Zero-shot, cross-lingual teacher	Zero-shot, monolingual student
		LaBSE	LASER		
HatEval-ES	0.65 (Nozza 2021)	0.63	0.67	0.70	**0.73**
AMIEvalita-IT	0.48 (Pamungkas and Patti 2019)	0.47	0.32	0.48	**0.49**
GermEval-DE	0.70 (Pelicon et al. 2021)	0.64	0.64	0.72	**0.76**
OffensEval-AR	-	0.68	0.70	0.77	**0.81**
OffensEval-EL	-	0.58	0.56	0.67	**0.72**
OffensEval-TR	-	0.62	0.58	0.70	**0.72**

Table 2: Experiment results (macro-F1) for all proposed and baseline models.

Term	Frequency	Zero-shot teacher	Zero-shot student
puta	565 (35.3%)	0.68	0.75
porca	306 (30.6%)	0.25	0.28

Table 3: Macro-F1 on tweets with taboo expressions.

of false positives.

Conclusion & Future Work

This paper proposes a novel pipeline based on pseudo-label fine-tuning of transformer language models for zero-shot, cross-lingual hate speech detection. Experimenting on benchmark datasets containing English and 6 different non-English languages, our approach not only outperforms previous zero-shot, cross-lingual models but also overcomes their limitation by improving detection of taboo expressions.

As part of future steps, we plan to expand our work to other low-resource languages such as Urdu, Indonesian etc., as well as other task types such as toxicity and racism detection. We would also like to analyze the performance of recently released larger multilingual transformer language models XLM-R$_{XL}$ and XLM-R$_{XXL}$ (Goyal et al. 2021) as a zero-shot, cross-lingual teacher. Finally, we plan to investigate the primacy of English as source language in zero-shot, cross-lingual hate speech detection.

References

Antoun, W.; Baly, F.; and Hajj, H. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Artetxe, M.; and Schwenk, H. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7: 597–610.

Basile, V.; Bosco, C.; Fersini, E.; Debora, N.; Patti, V.; Pardo, F. M. R.; Rosso, P.; Sanguinetti, M.; 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*, 54–63. Association for Computational Linguistics.

Bigoulaeva, I.; Hangya, V.; and Fraser, A. 2021. Cross-Lingual Transfer Learning for Hate Speech Detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, 15–25.

Chan, B.; Schweter, S.; and Möller, T. 2020. German's Next Language Model. *arXiv preprint arXiv:2010.10906*.

Chiril, P.; Moriceau, V.; Benamara, F.; Mari, A.; Origgi, G.; and Coulomb-Gully, M. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1397–1403.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Fersini, E.; Nozza, D.; and Rosso, P. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12: 59.

Fersini, E.; Nozza, D.; and Rosso, P. 2020. AMI@ EVALITA2020: Automatic Misogyny Identification. In *EVALITA*.

Fersini, E.; Rosso, P.; and Anzovino, M. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval@ SEPLN*, 2150: 214–228.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.

Goyal, N.; Du, J.; Ott, M.; Anantharaman, G.; and Conneau, A. 2021. Larger-Scale Transformers for Multilingual Masked Language Modeling. *arXiv preprint arXiv:2105.00572*.

Gutiérrez-Fandiño, A.; Armengol-Estapé, J.; Pàmies, M.; Llop-Palao, J.; Silveira-Ocampo, J.; Carrino, C. P.;

Gonzalez-Agirre, A.; Armentano-Oller, C.; Rodriguez-Penagos, C.; and Villegas, M. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Jiang, A.; and Zubiaga, A. 2021. Cross-lingual Capsule Network for Hate Speech Detection in Social Media. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 217–223.

Mandl, T.; Modha, S.; Kumar M, A.; and Chakravarthi, B. R. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, 29–32.

Mandl, T.; Modha, S.; Majumder, P.; Patel, D.; Dave, M.; Mandlia, C.; and Patel, A. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, 14–17.

Müller, K.; and Schwarz, C. 2018. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*.

Nozza, D. 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*. Online: Association for Computational Linguistics.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2021a. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4): 102544.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2021b. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 1–27.

Pamungkas, E. W.; and Patti, V. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, 363–370.

Papaevagelou, D. 2021. Greek RoBERTa. https://huggingface.co/cvcio/roberta-el-uncased-twitter-v1. Accessed: 2022-04-24.

Pelicon, A.; Shekhar, R.; Martinc, M.; Škrlj, B.; Purver, M.; and Pollak, S. 2021. Zero-shot Cross-lingual Content Filtering: Offensive Language and Hate Speech Detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, 30–34. Online: Association for Computational Linguistics.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 1–47.

Polignano, M.; Basile, P.; De Gemmis, M.; Semeraro, G.; and Basile, V. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, 1–6. CEUR.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; and Stranisci, M. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Schmidt, A.; and Wiegand, M. 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*, 1–10.

Schweter, S. 2020. BERTurk. https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased. Accessed: 2022-04-24.

Stappen, L.; Brunn, F.; and Schuller, B. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850*.

Vidgen, B.; and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12): e0243300.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wiegand, M.; Siegel, M.; and Ruppenhofer, J. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. *Overview of the germeval 2018 shared task on the identification of offensive language*.

Wolf, T.; Chaumond, J.; Debut, L.; Sanh, V.; Delangue, C.; Moi, A.; Cistac, P.; Funtowicz, M.; Davison, J.; Shleifer, S.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Zampieri, M.; Nakov, P.; Rosenthal, S.; Atanasova, P.; Karadzhov, G.; Mubarak, H.; Derczynski, L.; Pitenis, Z.; and Çöltekin, c. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.