Zero-shot Cross-lingual Content Filtering: Offensive Language and Hate Speech Detection

Andraž Pelicon[1], **Ravi Shekhar**[3], **Matej Martinc**[1,2]
Blaž Škrlj[1,2], **Matthew Purver**[2,3], **Senja Pollak**[2]
[1]Jožef Stefan International Postgraduate School
[2]Jožef Stefan Institute, Ljubljana, Slovenia
[3]Computational Linguistics Lab, Queen Mary University of London, UK
andraz.pelicon@ijs.si, r.shekhar@qmul.ac.uk, matej.martinc@ijs.si
blaz.skrlj@ijs.si, m.purver@qmul.ac.uk, senja.pollak@ijs.si

Abstract

We present a system for zero-shot cross-lingual offensive language and hate speech classification. The system was trained on English datasets and tested on a task of detecting hate speech and offensive social media content in a number of languages without any additional training. Experiments show an impressive ability of both models to generalize from English to other languages. There is however an expected gap in performance between the tested cross-lingual models and the monolingual models. The best performing model (offensive content classifier) is available online as a REST API.

1 Introduction

Recent years have seen a dramatic improvement in natural language processing, with machine learning systems outperforming human performance on a number of benchmark language understanding tasks (Wang et al., 2019). This impressive achievement is somewhat tempered by the fact that a large majority of these systems work only for English, while other less-resourced languages are neglected due to a lack of training resources. On the other hand, another recent development is the introduction of systems capable of zero-shot cross-lingual transfer learning by leveraging multilingual embeddings (Artetxe and Schwenk, 2019). These systems can be trained on a language with available resources and employed on a less-resourced language without any additional language specific training.

In this study we present an offensive language classifier available through a REST API which leverages the cross-lingual capabilities of these systems. Due to the exponential growth of social media content, the amount of offensive language and hate speech has seen a steep increase and its identification and removal is no longer manageable by traditional manual inspection of the content (Schmidt and Wiegand, 2017). As a consequence, there is a need for a general model that could be used in content filtering systems to automatically detect such discourse.

Since the majority of research in the area of offensive language and hate speech detection is currently done in monolingual settings, we performed a preliminary study to assess the feasibility of the proposed zero-shot cross-lingual transfer for this task. Two approaches are tested in this study. The first uses multilingual Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019). The second uses Language-Agnostic SEntence Representations (LASER, Artetxe and Schwenk, 2019), a system built specifically for zero-shot cross-lingual transfer using multilingual sentence embeddings. Our best performing model is available online and can be used for detecting offensive content in less-resourced languages with no available training data.

2 Related work

The large majority of research on hate speech is monolingual, with English still the most popular language due to data availability (Wulczyn et al., 2017; Davidson et al., 2017), and a number of English-only shared tasks organized on the topic of hate or offensive speech (e.g., OffenseEval, Zampieri et al., 2019b). Lately, the focus has been shifting to other languages, with several shared tasks organized that cover other languages besides English, e.g. OffenseEval 2020 (Zampieri et al., 2020), EVALITA 2018 (Bai et al., 2018) and GermEval 2018 (Wiegand et al., 2018).

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 30–34

For example, the EVALITA 2018 shared task (Bai et al., 2018) covered hate speech in Italian social media, the GermEval 2018 (Wiegand et al., 2018) shared tasks explored automatic identification of offensive German Tweets, and Semeval 2019 task 5 (Basile et al., 2019) covered detection of hate speech against immigrants and women in Spanish and English Twitter. Schmidt and Wiegand (2017); Poletto et al. (2020); Vidgen and Derczynski (2020) provide excellent surveys of recent hate speech related datasets.

Ousidhoum et al. (2019) conduct multilingual hate speech studies by testing a number of traditional bag-of-words and neural models on a multilingual dataset containing English, French and Arabic tweets that were manually labeled with six class hostility labels (abusive, hateful, offensive, disrespectful, fearful, normal). They report that multilingual models outperform monolingual models on some of the tasks. Shekhar et al. (2020) study multilingual comment filtering for newspaper comments in Croatian and Estonian.

Another multilingual approach was proposed by Schneider et al. (2018), who used multilingual MUSE embeddings (Lample et al., 2018) in order to extend the GermEval 2018 German train set with more English data. They report that no improvements in accuracy were achieved with this approach.

Cross-lingual hate speech identification is even less researched than the multilingual task. The so-called *bleaching* approach (van der Goot et al., 2018) was used by Basile and Rubagotti (2018) to conduct cross-lingual experiments between Italian and English at EVALITA 2018 misogyny identification task. The only other study we are aware of is a very recent study by Pamungkas and Patti (2019) proposing an LSTM joint-learning model with multilingual MUSE embeddings. Google Translate is used for translation in order to create a bilingual train and test input data. Bassignana et al. (2018) report that the use of a multilingual lexicon of hate words, HurtLex, slightly improves the performance of misogyny identification systems. Closest to our work is that of Glavaš et al. (2020), who propose a dataset called XHATE-999 to evaluate abusive language detection in a multi-domain and multilingual setting.

3 Dataset Description

As an English (EN) training set for *offensive language* classification, we used the training subset of the OLID dataset (Zampieri et al., 2019a). The trained models were evaluated on the test subset of the OLID dataset using their official gold labels and on the test subset of the GermEval 2018 dataset (Wiegand et al., 2018), which also contains manually labeled tweets. Both datasets use hierarchical annotation schemes for annotating hate speech content. For our purposes, we employed only the annotations on the first level which classify tweets into two classes, offensive and not offensive.

We trained the *hate speech* classifiers on the English training set from the HatEval dataset (Basile et al., 2019). For evaluation, we used the English and Spanish (ES) test sets from the HatEval competition, the German (DE) IGW hate speech dataset (Ross et al., 2016), an Indonesian (ID) hate speech dataset (Ibrohim and Budi, 2019) and the Arabic (AR) hate speech dataset LHSAB (Mulki et al., 2019). Each of the test datasets had binary labels that denoted the presence or absence of hate speech, except for the Arabic test set, which modeled hate speech as a three-class task, with labels denoting absence of hate speech, abusive language and hateful language. Since the authors themselves acknowledge there is a fine line between abusive and hateful language, we felt confident to join them into one class that denotes the presence of hate speech in a tweet. Tweets in the German IGW dataset included hate speech labels from two annotators and no common label, so we decided to evaluate only on those tweets where the two annotators agreed. The statistics of the datasets that were used in this study are reported in Table 1.

4 Classification models and methodology

Our models were trained and evaluated on two distinct albeit similar tasks, namely offensive language classification and hate speech detection, using two different approaches.

In the first approach, we tested the multilingual version of BERT to which we attached a classification layer with a softmax activation function. The model was fine-tuned on the chosen training datasets for 20 epochs. We limited the input sequence to 256 tokens and used a batch size of 32 and a learning rate of 2e-5. No additional hyperparameter tuning was performed.

Our second approach was using the pre-trained

	OLID (EN)	GermEval (DE)	HatEval (EN)	HatEval (ES)	IGW (DE)	ID	L-HSAB (AR)
# documents	14,100	8,541	13,000	6,600	541	13,169	5,846
Majority class	67%	66%	60%	60%	85%	57.77%	62.43%
Minority class	33%	34%	40%	40%	15%	43.23%	37.55%

Table 1: Dataset statistics.

LASER model and training a multilayer perceptron classifier with RELU activation function on top of that. To train the models we used the batch size of 32 and a learning rate of 0.001.

5 Results

The results for both tasks together with the majority baselines and the results reported in the literature are presented in Table 2. In the offensive language classification task, our best model (BERT) achieved an F1 score of 82.63 on the English test set, which is on par with the reported results achieved by monolingual classifiers (Zampieri et al., 2019b). When evaluated on the German dataset, we observe a considerable drop in performance compared to the reported results (Wiegand et al., 2018), however, it still achieves a solid F1 score of 70.67, which indicates its ability to generalize to languages it has not seen during training.

In the hate speech classification task, the two models are comparable, with LASER outperforming BERT on the Arabic and Spanish datasets. Overall, the scores for the hate speech classification task proved to be considerably lower for both models as well as lower than the reported results in the monolingual experiments (Basile et al., 2019; Ibrohim and Budi, 2019). Nevertheless, the results again indicate the ability of both models to generalize from English to other languages, as our models perform better than the majority baseline classifiers in terms of macro-averaged F1 score on all the datasets. It should be noted that the performance between our models and the reported performance on the Indonesian and Arabic datasets are not directly comparable as the original training and testing splits from the literature are not available. Therefore, our models were tested on different test splits.

6 Web API design

The best performing cross-lingual model, multilingual BERT for offensive language classification, was implemented as a REST web service in the Flask framework. The design of the web service allows us to easily update the current model with a new version trained on additional data in the future. The web service can be reached programmatically through the endpoint at `http://classify.ijs.si/ml_hate_speech/ml_bert` or through a demo browser-based interface at the URL `http://classify.ijs.si/embeddia/offensive_language_classifier`. The interface is designed for mobile devices and supports most popular screen sizes. It consists of an input area where users can input their sentence and submit it for classification. The classification results as well as the confidence score of the classifier are then displayed under the input area.

7 Conclusion and future work

In the course of this study, we tested the performance of two multilingual models, BERT and LASER, in zero-shot offensive language and hate speech detection. The results for the offensive language classification task show that even in the multilingual setting the BERT-based classifier achieves results comparable to the monolingual classifiers on English language data and solid performance on the German dataset. On the other hand, hate speech classification still proves to be a hard task for the multilingual classifiers as they achieve considerably lower scores on all languages compared to reported results. Nevertheless, both models show an impressive ability to generalize over languages they have not seen during fine-tuning. We implemented the best performing model, multilingual BERT for offensive language classification, as a REST web service. In the future, we plan to perform similar experiments with other multilingual language models, namely the XLM-R models (Conneau et al., 2019), which show increased performance in standard benchmark tasks compared to multilingual BERT, and the recently released CroSloEngual-BERT (Ulčar and Robnik-Šikonja, 2020).

While all datasets used in this study contain social media posts labeled for hate speech or of-

	Cross-lingual hate speech classification									
	Accuracy					F1-macro				
Model	EN	ES	DE	ID	AR	EN	ES	DE	ID	AR
LASER	**0.5241**	**0.6562**	0.5041	0.5755	**0.7013**	0.4994	0.6538	0.4630	**0.5172**	0.5500
BERT	0.5091	0.6313	**0.6369**	0.5823	0.6264	0.4341	0.5839	**0.6886**	0.4603	0.5033
Reported	/	/	/	**0.7353***	**0.9060***	**0.6510**	**0.7300**	/	/	**0.8930***
Majority	0.6000	0.6000	0.8500	0.5800	0.6200	0.3600	0.3700	0.4600	0.3700	0.3800
	Cross-lingual offensive language classification									
LASER	0.7500	/	0.7129	/	/	0.6823	/	0.6508	/	/
BERT	**0.8279**	/	**0.7148**	/	/	0.8263	/	0.7067	/	/
Reported	/	/	/	/	/	**0.829**	/	**0.7677**	/	/
Majority	0.6700	/	0.6600	/	/	0.4200	/	0.4000	/	/

Table 2: Results of the hate speech classification task (models trained on the English hatEval dataset) and offensive language classification task (models trained on the English OLID dataset) in comparison to the monolingual results as reported in the literature. The forward slash ('/') denotes results which are not reported in the literature. Figures marked with * denote results obtained on a different test split.

fensive language, there are still some differences in the way the data was labeled and collected, as each dataset was collected by a different research team. Therefore, some compromises had to be made in the course of this study to consolidate the datasets as best as possible. In order to better control for such variables, we would like to perform our experiment on the recently released XHate-999 dataset which contains instances in six diverse languages that were collected and annotated by the same research team using a unified annotation process. Given the fact we are working with relatively well-resourced languages, another future endeavour would be to also inspect the differences in cross-lingual model performance between zero-shot and few-shot testing scenarios. Finally, we plan on improving the performance of the model specifically on the task of hate speech classification, and update the existing web service.

8 Acknowledgements

This research is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The work of AP was funded also by the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263). The results of this publication reflect only the authors' views and the Commission is not responsible for any use that may be made of the information it contains. MP was also funded by the UK EPSRC under grant EP/S033564/1. We acknowledge also the funding by the Slovenian Research Agency (ARRS) core research programme Knowledge Technologies (P2-0103).

References

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

X. Bai, F. Merenda, C. Zaghi, T. Caselli, and M. Nissim. 2018. RuG@EVALITA 2018: Hate speech detection in Italian social media. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:245.

A. Basile and C. Rubagotti. 2018. CrotoneMilano for AMI at Evalita2018. a performant, cross-lingual misogyny detection system. In *EVALITA@ CLiC-it*.

V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proc. SemEval*.

E. Bassignana, V. Basile, and V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*.

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

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. ICWSM*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*.

G. Glavaš, M. Karan, and I. Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*.

R. van der Goot, N. Ljubešić, I. Matroos, M. Nissim, and B. Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proc. ACL*.

M. O. Ibrohim and I. Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proc. 3rd Workshop on Abusive Language Online*.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proc. ICLR*.

H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proc. 3rd Workshop on Abusive Language Online*.

N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proc. EMNLP-IJCNLP*.

E. W. Pamungkas and V. Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proc. ACL Student Research Workshop*.

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

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proc. NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.

A. Schmidt and M. Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. 5th International Workshop on Natural Language Processing for Social Media*.

J. M. Schneider, R. Roller, P. Bourgonje, S. Hegele, and G. Rehm. 2018. Towards the automatic classification of offensive language and related phenomena in German tweets. In *14th Conference on Natural Language Processing KONVENS 2018*.

R. Shekhar, M. Pranjić, S. Pollak, A. Pelicon, and M. Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *International Conference on Text, Speech, and Dialogue*.

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

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. ICLR*.

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

E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proc. WWW*.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proc. NAACL-HLT*.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proc. SemEval*.

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