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

Andraž Pelicon¹, Ravi Shekhar³, Matej Martinc¹,²
Blaz Škrlj¹,², Matthew Purver²,³, Senja Pollak²
¹Jožef Stefan International Postgraduate School
²Jožef Stefan Institute, Ljubljana, Slovenia
³Computational 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).
work is that of words, HurtLex, slightly improves the performance with more English data. They report that no im-
Arabic tweets that were manually labeled with six
setting.
language detection in a multi-domain and multilingual
dataset called XHATE-999 to evaluate abusive lan-
dation task. The only other study we are aware of is
conduct cross-lingual experiments between Italian
bleaching
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 identifi-
cation task. The only other study we are aware of is
a very recent study by Ousidhoum et al. (2019)
proposing an LSTM joint-learning model with mul-
tilingual MUSE embeddings (Lample et al., 2018) in or-
der to extend the GermEval 2018 German train set
with more English data. They report that no im-
provements in accuracy were achieved with this
approach.

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

Ousidhoum et al. (2019) conduct multilingual
hate speech studies by testing a number of tradi-
tional bag-of-words and neural models on a mul-
tilingual 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 mod-
els on some of the tasks. Shekhar et al. (2020)
study multilingual comment filtering for newspaper
comments in Croatian and Estonian.

3 Dataset Description
As an English (EN) training set for offensive lan-
guage 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 man-
nually labeled tweets. Both datasets use hierarchi-
cal 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 En-
glish training set from the HatEval dataset (Basile
et al., 2019). For evaluation, we used the English
and Spanish (ES) test sets from the HatEval compe-
tition, 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 la-

defined that denoted the presence or absence of hate
speech, except for the Arabic test set, which mod-
eled 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 an-
notators and no common label, so we decided to
evaluate only on those tweets where the two an-
notators 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 dis-
tinct 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 classi-
fication layer with a softmax activation function.
The model was fine-tuned on the chosen training
datasets for 20 epochs. We limited the input se-
cquence to 256 tokens and used a batch size of 32
and a learning rate of 2e-5. No additional hyperpa-
parameter tuning was performed.

Our second approach was using the pre-trained
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 CroSloEngualBERT (Ulčar and Robnik-Šikonja, 2020).

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

<table>
<thead>
<tr>
<th>OLID</th>
<th>GermEval</th>
<th>HatEval</th>
<th>HatEval</th>
<th>IGW</th>
<th>ID</th>
<th>L-HSAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(EN)</td>
<td>(DE)</td>
<td>(EN)</td>
<td>(ES)</td>
<td>(DE)</td>
<td></td>
<td>(AR)</td>
</tr>
<tr>
<td># documents</td>
<td>14,100</td>
<td>8,541</td>
<td>13,000</td>
<td>6,600</td>
<td>541</td>
<td>13,169</td>
</tr>
<tr>
<td>Majority class</td>
<td>67%</td>
<td>66%</td>
<td>60%</td>
<td>60%</td>
<td>85%</td>
<td>57.77%</td>
</tr>
<tr>
<td>Minority class</td>
<td>33%</td>
<td>34%</td>
<td>40%</td>
<td>40%</td>
<td>15%</td>
<td>43.23%</td>
</tr>
</tbody>
</table>

Table 1: Dataset statistics.
Cross-lingual hate speech classification

<table>
<thead>
<tr>
<th>Model</th>
<th>EN</th>
<th>ES</th>
<th>DE</th>
<th>ID</th>
<th>AR</th>
<th>EN</th>
<th>ES</th>
<th>DE</th>
<th>ID</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASER</td>
<td>0.5241</td>
<td>0.6562</td>
<td>0.5041</td>
<td>0.5755</td>
<td>0.7013</td>
<td>0.4994</td>
<td>0.6538</td>
<td>0.4630</td>
<td>0.5172</td>
<td>0.5500</td>
</tr>
<tr>
<td>BERT</td>
<td>0.5091</td>
<td>0.6313</td>
<td>0.6369</td>
<td>0.5823</td>
<td>0.6264</td>
<td>0.4341</td>
<td>0.5839</td>
<td>0.6886</td>
<td>0.4603</td>
<td>0.5033</td>
</tr>
<tr>
<td>Reported</td>
<td>/</td>
<td>/</td>
<td>0.7353*</td>
<td>/</td>
<td>/</td>
<td>0.6510</td>
<td>0.7300</td>
<td>/</td>
<td>/</td>
<td>0.8930*</td>
</tr>
<tr>
<td>Majority</td>
<td>0.6000</td>
<td>0.6000</td>
<td>0.8500</td>
<td>0.5800</td>
<td>0.6200</td>
<td>0.3600</td>
<td>0.3700</td>
<td>0.4600</td>
<td>0.3700</td>
<td>0.3800</td>
</tr>
</tbody>
</table>

Cross-lingual offensive language classification

<table>
<thead>
<tr>
<th>Model</th>
<th>EN</th>
<th>ES</th>
<th>DE</th>
<th>ID</th>
<th>AR</th>
<th>EN</th>
<th>ES</th>
<th>DE</th>
<th>ID</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASER</td>
<td>0.7500</td>
<td>/</td>
<td>0.7129</td>
<td>/</td>
<td>/</td>
<td>0.6823</td>
<td>/</td>
<td>0.6508</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>BERT</td>
<td>0.8279</td>
<td>/</td>
<td>0.7148</td>
<td>/</td>
<td>/</td>
<td>0.8263</td>
<td>/</td>
<td>0.7067</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Reported</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>0.829</td>
<td>/</td>
<td>0.7677</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Majority</td>
<td>0.6700</td>
<td>/</td>
<td>0.6600</td>
<td>/</td>
<td>/</td>
<td>0.4200</td>
<td>/</td>
<td>0.4000</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

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.

In offensive 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 EM-BEDDIA (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


