University *of Ljubljana*
Faculty *of Computer and Information Science*

Cross-lingual offensive language identification

Andrej Hafner, Urša Zrimšek

Abstract

Hateful content has become very common with the development and rapid adoption of different social platforms, hence the need for automatic detection of offensive language has grown as well. Since most of the content generated on the web is in English language, there already exist powerful systems for detection. But for languages with sparse data available, training such system is a challenge. Out of this came the idea of transferring the knowledge learnt on English to other languages. In this work we show three methods for transferring the knowledge inter-language. Models that perform well on English data show poor performance on Slovenian language, which we assume is due to the different domains of training and testing sets and to the inability of transferring the knowledge.

Keywords

offensive language classification, cross-lingual, vector space alignment, context aware models, word embeddings

Advisors: Slavko Žitnik

Introduction

Hateful and toxic content generated on different social platforms is a phenomenon on the rise, which many need to tackle. Since the amount of generated content is growing day by day, moderation by user intervention is becoming harder, therefore the need for automatic detection of hate speech is increasing. For the most widely used language on World Wide Web - English, different kind of detectors are being utilized for some time. Lately, through the rise of deep learning in natural language processing, novel methods are being created for hateful speech classification. The abundance of English data allows these kinds of methods to achieve very decent performance. Because of the lack of data from other languages and the exceptionally long training times of models, ideas started to appear on how to transfer the knowledge learned on English to other languages as well. In this work, we show multiple approaches on how to use models trained to classify hateful speech on English for classification in Slovene language.

Related work

Offensive language classification is a growing field, therefore many approaches have already been researched. There exist multiple benchmark competitions and tasks [1, 2], whose goal is to promote the offensive language classification and offer a common platform to compare state-of-the-art methods. Authors of [3] use three-level classification for detecting hateful language. In the first level they select the most discriminative features obtained from the text, which are further refined on the second and third level. In [4] authors use bi-directional LSTM and CNN for predicting the type of offensive content - hate speech, cyberbullying or cyber-aggression. Furthermore, authors of [5] use context-aware methods that utilize attention mechanism to classify the binary dataset obtained from the Fox news comment section.

An important spread for wide-spread use of these systems is their availability in different languages. Authors of [6] perform multilingual offensive language identification with transfer learning. Initially, they train the XLM-R model on English data, then transfer the weights to the same model for another language. They show that this approach works well for different inter-language transfers as well as domain transfer. Authors of [7] first introduce methods for learning cross-lingual word embeddings that rely on bilingual dictionaries. They show, that it is possible to build a dictionary in an unsupervised way, without any parallel data. It can be even used on distant language pairs and experiment on English-Esperanto pair that is very limited in parallel data. In article [8] they use multilingual word embeddings for cross-lingual hate speech detection. They achieve the best results with custom domain specific embeddings.

Datasets

In this section we present the datasets used for training and testing our approaches. We use dataset with English content for monolingual offensive language classification and datasets in Slovene for testing the cross-lingual knowledge transfer performance of our approaches.

English language

Gab dataset

The dataset [9] of posts collected from the online social platform Gab. The platform is known for having a far-right user base, often attracting users and groups who were banned from other social platforms due to their extreme opinions and disrespectful behaviour. The dataset consists of 32,471 posts, which have binary labels - hateful or non-hateful. There are 14,601 posts labeled as hateful and 17,870 as non-hateful. The authors collected the dataset through searching for posts with hate keywords, in order to identify potentially hateful posts. Collected posts were then manually labeled by the workers on the crowd sourcing platform Mechanical Turk. The dataset is balanced, with the posts labeled as hateful containing many swear words.

Reddit dataset

Reddit dataset is a collection of conversational data, collected from the Reddit platform. The authors [9] collected the hottest submissions from Reddit's whiniest and most toxic subreddits. They also reconstruct the whole conversation around a potentially hateful post, enabling contextual analysis of hateful speech. Labeling was done in the same way as described in description of Gab dataset. The dataset is binary and heavily imbalanced, since out of 22,210 entries 16,959 don't contain hateful language and 5,252 do. Due to the moderation on the platform, very few posts contain swear words.

Wikipedia toxic comment dataset

The dataset is a collection of Wikipedia comments, labeled by human raters for different types of toxic behaviour. It was made publicly available on Kaggle for a toxic comment classification challenge. Every comment is labeled with presence of 7 types of toxicity: *toxic*, *severe_toxic*, *obscene*, *threat*, *insult*, *identity_hate*. In figure 1 we can see the number of comments labeled by each label. In total there are 159,571 comments in the dataset, out of which 143,346 don't contain any type of toxic language.

TRAC-2 aggression classification dataset

TRAC-2 dataset [10] is a collection of comments from YouTube created for the workshop on trolling, aggression and cyberbullying. It contains annotations on two levels - aggression (non-aggressive, covertly aggressive and overtly aggressive) and misogyny (gendered and non-gendered). Data was collected in three languages. For our purpose we only use the comments and aggression labels in the English part of the dataset. Combined *train* and *dev* subset of the dataset contain 5,329 comments, out of which 570 are labeled as covertly aggressive, 548 as overtly aggressive and the rest as non-aggressive.

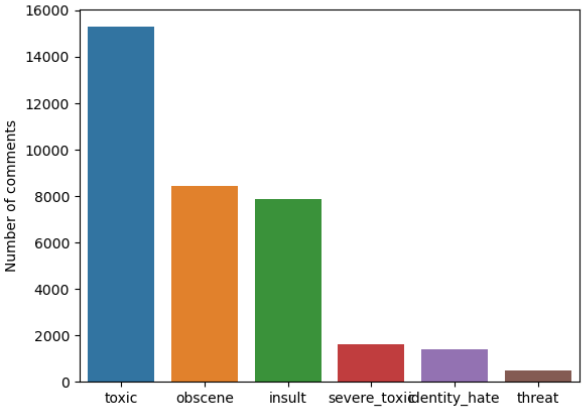

Figure 1. Number of comments labeled with each toxicity category. The toxic types like *toxic*, *obscene* and *insult* heavily outweigh the more strongly toxic types like *severe_toxic*, *identity_hate* and *threat*, introducing even more imbalance to the dataset.

Slovene language

IMSyPP-sl - Twitter hate speech dataset

The dataset [11] is a collection of hand-labeled tweets acquired from Slovenian Twitter. It contains annotations on for the hate speech type (appropriate, inappropriate, offensive, violent) and hate speech target (racism, migrants, Islamophobia, antisemitism, religion, homophobia, sexism, ideology, media, politics, individual and other). Each tweet was annotated twice, most of the times by two different annotators. It contains a disjunctive training and evaluation sets, which we combine four our purpose and get 60,000 labeled tweets.

Data preprocessing

All the datasets we use are cleaned and transformed into a common format. We remove emojies, links and newline characters from the text contents. In Twitter datasets the references to other users are removed from tweets, since they provide no additional information about the text being hateful or not.

Each English dataset is split into training and testing set, with ratio of 8 : 2, respectively. These sets stay the same in for every approach in order to enable fair comparison of our methods. We also unify the hate speech types for binary sets, setting a 0 for non-hateful content and 1 for hateful.

The Wikipedia toxic comment dataset is modified in order to enable the cross-lingual transfer of knowledge to Slovenian language. This is needed because the latter contains 7 types of labels and the Slovenian dataset contains 4 types. We use the following if-else rules for dataset relabeling: entry is labeled as violent, if contains a *threat* label. If the threat label is not present, but it has an *insult* or *identity_hate* label, the entry is

marked as offensive. When none of the already mentioned rules apply, but at least one label is present on the entry, we mark it as inappropriate. In case none of the labels are present, the entry is marked as appropriate. This matches the labels on this multi-class dataset with the Slovenian IMSyPP-sl dataset, enabling cross-lingual multi-class classification. We then re-balance the dataset, since its non-hateful labeled entries represent the majority of the dataset. Out of 143,346 non-hateful entries we randomly sample 7,500 entries, which is almost the same as the number of entries labeled as offensive or inappropriate. The number of entries labeled as violent is still very low, for which we don't account. For our project we only use the training subset obtained from Kaggle. For binary classification the dataset is re-labeled, combining all the labels except non-hateful into hateful.

Slovenian Twitter dataset IMSyPP-sl is mostly annotated by two different annotators. Upon analysis of the dataset, we discovered that in a large part of the entries the annotators didn't agree on the type of hate speech present in the tweet. This lead to the decision to narrow down the dataset only to those entries, in which the annotators have reached consensus. Resulting dataset contains 18,459 entries, which is more than enough for a representative test set.

Methods

Machine translation and classification using fine-tuned BERT model

In this approach we use the Google Translate API to translate Slovenian dataset into English, then use a BERT model finetuned on English data for classification. Machine translation today is very accurate and should convert most of the hate speech elements from Slovenian to English. We use the Translate API to translate the modified IMSyPP-sl dataset and then use it for testing. We finetune the *bert-base-uncased* version of BERT, pre-trained on English data. We use all the described datasets for training and testing of monolingual offensive language classification.

The original training set is split into a training and validation set with ratio 8:2. We set the batch size to 32, learning rate to 2×10^{-5} and train the models for 3 epochs. Maximum length of tokenized text is set to 128. The finetuned models are then evaluated on their respective testing datasets and on the translated IMSyPP-sl dataset. The model trained on toxic comment dataset is evaluated on IMSyPP-sl dataset with all of their classes, while for other models the labels of IMSyPP-sl are converted into a binary form (hate/non-hate).

Multilingual pre-trained BERT finetuned on English data used for classification on Slovene datasets

The idea behind this approach is to use a model that was trained on both Slovenian and English data. Model creates a shared space for both languages, where in theory the similar words in both languages are close together, or the relations between different words and contexts is similar. Based on this concept, we could train the model to classify hate speech in English and it could work as well on Slovenian language. We use a version of the original BERT transformer model [12], called CroSloEngual BERT [13]. The model was trained on Croatian, Slovenian and English data for the purpose of various NLP tasks.

For training we set the batch size to 32, learning rate to 5×10^{-5}, max length of tokenized text to 128 and train the models for 3 epochs. The testing protocol is the same as for the machine translation approach, but in this case we use the non-translated version of IMSyPP-sl dataset.

Vector space alignment of word embeddings and predictions based on occurrences in offensive comments.

This approach is based on words that appear in the hateful comments, and tries to determine which words are distinctive for hate speech. The advantage of using word embeddings instead of sentences is the availability of word dictionaries, that help us to easily find linear mapping between vector embeddings of multiple languages. For alignment of spaces of sentence embeddings (that would tell us more about the whole comment), we would need sentence dictionary, that would probably need to be a lot bigger, since we would then be in the space of sentences not words. So for this cross lingual approach, we decided to use word embeddings.

Alignment of Slovene and English word embeddings

For alignment of Slovene and English word spaces we used the approach described in [14].

We are using two FastText [15] models, pre-trained on Slovene and English Wikipedia data. With these models, we extracted embeddings, for pairs of words $\{x_i, y_i\}$ that appeared in dictionary taken from the repository of [7]. Our goal is to find a linear mapping W between English and Slovene word space, such that we minimize $\|WX - Y\|_F$, where X is the matrix with columns x_i and Y consists of columns y_i, for $i \in 1, n$. Both matrices are of size $d \times n$, with $n = 7111$ being the length of dictionary, and $d = 300$ being the dimension of embeddings. In [14] they showed, that we get better results, if we constrain W to be orthogonal. In that case, the solution is obtained with singular value decomposition of YX^T:

$$W^* = \arg\min \|WX - Y\|_F = UV^T, \text{ where } U\Sigma V^T = YX^T.$$

Prediction of offensive language on English data

To generate predictions of word offensiveness, we broke the sentences into words, and for each word w, that appeared in our train dataset, calculated the number of occurrences of in offensive comments, n_o, and not offensive comments, n_n. We can then express the probability the word w is offensive by:

$$P(w \text{ is offensive}) = \frac{n_o}{n_o + n_n}.$$

The probabilities we calculated are only available for the words that are contained in our train dataset. To find the probabilities of new words, we will again use word embeddings

from the same FastText models that we used in word space alignment. We generated matrix X, in which are the embeddings for all words that appeared in train data. Our target variables are their probabilities.

The main reason for using word embeddings is the hypothesis, that the embeddings of similar words will be close together in the word vector space. So for offensiveness prediction, we will use support vector regression, that will find a hyperplane in the word space. As we suspect that we will have multiple clusters of hate words in the space, we will use radial basis function kernel, and we will also try quadratic polynomial kernel.

We also need to determine the right function f to combine the probabilities of words in sentence into classification if the sentence is offensive. We tried different approaches:

1. mean of word probabilities,
2. calculating `logit` of probabilities, dividing it by some coefficient, taking the mean of new values and then calculating the `inverse logit`,
3. dividing the probabilities by average probability of all word occurrences in the train set, to normalize the mean value to 1, then calculating their product,
4. inverting the probabilities and then doing step 3.

For each approach we determined a threshold, how high should the value be, to classify it as offensive. The results differed depending on the dataset, but generally, the second approach was the best.

So the function used to combine words probabilities into sentence probability is

$$\texttt{inv\_logit}(\frac{1}{n}\sum_i \texttt{logit}(w_i)),$$

where w_i are the words contained in a sentence of length n.

Cross lingual prediction of offensive language on Slovene data

The main goal of this method is to combine the techniques described above, to be able to classify hate speech in Slovene language, that was not used in training. We approached it by again splitting sentences into words and generating their embeddings with Slovene FastText model. We then mapped them into English word space with matrix W, and used SVM regressor that was trained on English data, for defining the probability they are offensive. Lastly we combined these probabilities into sentence probability with use of `logit` function, and classified it using a threshold.

This method is only implemented and tested on binary data. It can be extended to multi-class method by calculating and predicting probabilities for each class separately.

Results

In this section we present the results of methods used.

Machine translation and classification using finetuned BERT model

In this section we present the results of our first method - machine translation and classification using finetuned BERT model. We first show the results of testing models on their respective test datasets, then we present the results of models trained on English and tested on Slovenian dataset.

For binary datasets we use the accuracy, f_1-score, precision and recall metrics. Multi-class classification is evaluated with the same metrics, but we report the macro version of f_1-score, precision and recall. We use the macro average because it gives the same importance for every class. This means that its value will be low for models that perform well on common classes, but perform poorly on the rare classes. In our case we find this important, since most of our datasets are imbalanced.

Results of monolingual method

In table 1 we can see the results of testing our models on their respective testing sets. We can see that the best performing binary classification on the Gab dataset, with Reddit coming up close. We assume that the best results were acquired from Gab because it contains more explicit language, which is often a good indicator of hateful speech and an useful feature for the model to learn. In multi-class classification, reasonable performance was achieved on the toxic comment dataset.

Table 1. Evaluation on English datasets. Datasets marked with * are multi-class and were evaluated using macro f_1-score, precision and recall.

	f_1-score	*precision*	*recall*	*accuracy*
Gab	0.92	0.92	0.92	0.92
Reddit	0.88	0.88	0.89	0.91
TRAC-2*	0.56	0.62	0.54	0.81
Toxic*	0.72	0.72	0.73	0.79

Results of multilingual method

In table 2 we can see the results of models trained on different English datasets and evaluated on the translated Slovenian IMSyPP-sl dataset. Both binary models performed very poorly on the Slovenian dataset, achieving high precision, but almost zero recall. The reason for this is that both of the models predicted most of the entries as non-hateful, probably because they were trained on much different datasets that contained many swear words, which are not as present in the Slovenian dataset.

Multilingual pretrained BERT finetuning

In this section we show the results of of our CroSloEn BERT model trained on English data and then tested on English and Slovenian. We use the same metrics as described in the previous section, since we follow the same testing protocol.

Results of monolingual method

In table 3 we present the results of testing the finetuned CroSloEn BERT on the respective testing sets. We can see

Table 2. Evaluation on Slovenian dataset. Datasets marked with * are multi-class and were evaluated using macro f_1-score, precision and recall. Each row represents the results of testing with a different model, the dataset is the same for all tests.

	f_1-score	*precision*	*recall*	*accuracy*
BERT Gab	0.04	0.77	0.02	0.5
BERT Reddit	0.03	0.81	0.01	0.5
BERT Toxic*	0.23	0.31	0.33	0.39

that results on binary datasets and the TRAC-2 dataset are very similar to the English BERT model used in the machine translation approach, while the performance is not as good on the toxic comment dataset.

Table 3. Evaluation on English datasets. Datasets marked with * are multi-class and were evaluated using macro f_1-score, precision and recall.

	f_1-score	*precision*	*recall*	*accuracy*
Gab	0.92	0.92	0.92	0.91
Reddit	0.87	0.87	0.87	0.90
TRAC-2*	0.61	0.66	0.59	0.70
Toxic*	0.38	0.54	0.39	0.79

Results of multilingual method

In table 4 we show the results of testing the CroSloEn BERT models on the Slovenian dataset. We can see that results are very similar to the ones achieved by the machine translation approach, achieving a bit lower scores on the toxic comment dataset.

Table 4. Evaluation on Slovenian dataset. Datasets marked with * are multi-class and were evaluated using macro f_1-score, precision and recall. Each row represents the results of testing with a different model, the dataset is the same for all tests.

	f_1-score	*precision*	*recall*	*accuracy*
BERT Gab	0.05	0.61	0.02	0.5
BERT Reddit	0.06	0.61	0.06	0.5
BERT Toxic*	0.17	0.25	0.25	0.24

From the results we can see that both finetuned models performed better than the baseline model without training, which shows that finetuning on English data is somewhat effective and improves the classification on Slovenian data. Out the two pre-trained models, the model trained on Gab dataset came out on top, achieving *f_1-score* of 0.357. The performance is not very good, which could be explained by the lack of training data or the inability of the model to use the knowledge learned on the English data in classifying Slovenian data.

Classification based on word occurrences in offensive comments

In this section we will first show the results of the method on the testing part of the English dataset that was used to train it, and then compare it with the results on translated Slovene data, to see if this model is good for cross lingual hate speech detection.

Results of the monolingual method

In table 5 are the results of classification of offensive comments on English datasets, where we got the embeddings with FastText model trained on data from Wikipedia. These results were obtained using SVR with radial basis kernel for predicting the probabilities and `logit` function for combining them. On *Toxic* dataset we showed results with normalized embeddings before training SVR and mapping, as we got better results like this.

Table 5. Evaluation on English datasets.

	f_1-score	*precision*	*recall*	*accuracy*
Gab	0.82	0.88	0.76	0.85
Reddit	0.72	0.84	0.63	0.88
TRAC-2	0.51	0.40	0.73	0.71
Toxic	0.77	0.95	0.65	0.81

In table 5 we can see, that this approach is good on Gab dataset, where there is a lot of explicit language, which helps our model select the words that best define offensive comments. Hate words are much better for out of context evaluation, then datasets that have them filtered out. We also reach the best accuracy on Reddit dataset, but lower F-1 score, the reason for this is that it is highly unbalanced. That is why we shouldn't measure the success by accuracy. TRAC-2 dataset is much smaller than the others, and we can see that the results on it are poor. The reason for this is that we can't accurately calculate word offensiveness probability, if it only appears in the train set a couple of times. Toxic dataset is the biggest, and we also achieve good results on it.

Results of the multilingual method

In table 6 are the results of transferring the knowledge gathered on English datasets to Slovene. We tested the method trained on each English dataset on Slovene Twitter dataset. The methods were trained with support vector regression with radial basis kernel, and the embedding vectors were not normalized in all but *Toxic* dataset.

Table 6. Evaluation on Slovene dataset.

	f_1-score	*precision*	*recall*	*accuracy*
Gab	0.27	0.54	0.18	0.50
Reddit	0.01	0.5	0	0.49
TRAC-2	0.51	0.52	0.51	0.51
Toxic	0.12	0.53	0.07	0.5

Discussion

Overall, we show that transformer models show good performance on monolingual offensive language classification, in both English and mixed variations. When it comes to the cross-lingual transfer of knowledge, none of the approaches work very well. This could be due to the different domains of datasets they were trained on in English language and then tested in Slovenian language. Most of the datasets in English contain many swear words, which serve as a good indicator of hate speech, but are not present in the Slovenian Twitter dataset.

We can see that for the classification based on word occurrences in offensive comment the results on Slovene data are a lot worse than on English datasets. The reason for bad results could be that the prediction of hate probabilities is bad, or that the translation from Slovene to English space is bad. Since the predictions on English data is good, it is probably the translation.

To confirm this hypothesis, we visualized the predictions on figure 2. We got the real values of probabilities of word offensiveness on test dataset (we used *Toxic*, as it is largest and the shapes can be best seen) so that we summed the occurrences of these words in train dataset with occurrences in test dataset, and then calculated probabilities in the same way as on train set. In test dataset we don't have enough comments, to get nice probabilities from them. We would get even more vertical lines, than we can see on left plot of figure 2. Wrong predictions of words with true probabilities 0 and 1 can be explained with those being some words that only appear once (for example wrongly written), and do not have special meaning for hatefulness.

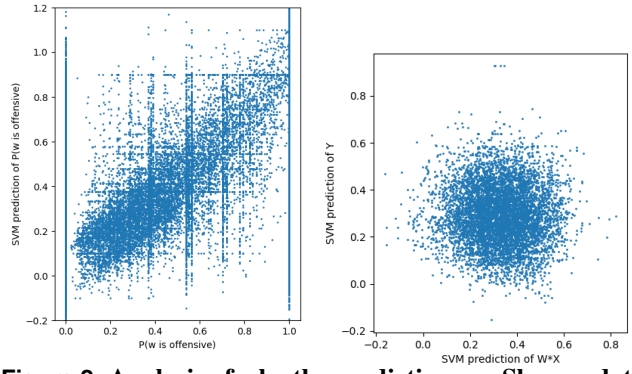

Figure 2. Analysis of why the predictions on Slovene data are bad. On the left we can see a plot of true values of word hate probabilities and of predictions that SVR makes on the embedding of this word. The shape shows that the regression is working. On the right side are the predictions on the train dictionary, used for making the mapping W. On the x axis we have the prediction of SVR on the Slovene word embeddings mapped to English space and on the y axis we have the direct prediction of the English word. We can see that those predictions are not correlated, so we can conclude that this mapping is not right for our case.

To see if the mapping was even done correctly, we calculated average cosine distance between Slovene and English embeddings. Before alignment, it was 0, and after alignment we reached 0.5. So we can see that we did shift the Slovene language space closer to the English one, but not enough, for the embeddings to be so similar, that the regressor would be able to correctly predict translated Slovene embeddings.

We tried to normalize the embeddings before calculating the mapping, but that didn't solve the problem. We can conclude that this way of calculating the mapping can not align the spaces enough, at least not English and Slovene, that are not very similar.

Conclusion

We tested various methods for offensive language classifications, two with contextual awareness and one that only uses word knowledge. With both we managed to achieve good results on monolingual task, but failed when tried to transfer our knowledge. The reason for this could be the distance between Slovene and English language, or wrong approaches. On monolingual classification we achieved best results with monolingual BERT. We can't really compare results on Slovene data, as none were much better as those that majority or random classifier would have. To improve these methods, we would propose to try different mappings for vector space alignment, based more on aligning the part of space that includes hate words. If we would want to improve the approach describe here, we would need to include more hate words into our dictionary. We would also propose that it is important to train the model in the same environment as we would use it. Even if we train it on English data, and use it on Slovene, dataset domain should be related. Similar amount of curse words and maybe even topics discussed, would improve the algorithm, if that kind of data is available.

References

[1] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, 2020.

[2] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*, 2020.

[3] Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer, 2010.

[4] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting

the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.

[5] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*, 2017.

[6] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*, 2020.

[7] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[8] Aymé Arango, Jorge Pérez, and Barbara Poblete. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *arXiv preprint arXiv:2104.14728*, 2021.

[9] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*, 2019.

[10] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France, May 2020. European Language Resources Association (ELRA).

[11] Petra Kralj Novak, Igor Mozetič, and Nikola Ljubešić. Slovenian twitter hate speech dataset imsypp-sl. 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Matej Ulčar and Marko Robnik-Šikonja. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*, 2020.

[14] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[15] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.