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

Cross-lingual offensive language identification

Andrej Hafner, Urša Zrimšek

Abstract

Keywords

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

- Identifying and Categorizing Offensive Language in Social Media (OffensEval)
- Identification of Offensive Language in Social Media
- A Cross-Lingual Augmentation Approach for Multilingual Offensive Language Identification
- Predicting the Type and Target of Offensive Posts in Social Media
- Detecting Online Hate Speech Using Context Aware Models
- Medjezikovni pristop k abstraktivnemu povzemanju: [1]
- Medjezikovni prenos napovednih modelov za sovražni govor

Datasets

In this section we present the datasets used for training and testing our approaches.

English language

1. **SemEval2020 - multiple classes (hate/no hate, type of hate)**
2. **Aggresion identification - agressive/covertly aggresive/not aggresive** [2]
3. **FoxNews comments** Consists of 1528 annotated comments, 435 of them are labeled as hateful and 1093 of them as non-hateful.
4. **Comments from the Reddit and Gab platform.** The Gab dataset contains 32471 entries out which 17870 are non-hateful and 14601 are hateful. Most of the content contains explicit language. The Reddit dataset is rather unbalanced, since out of 22210 entries 16959 don't contain hateful language and 5251 do. Since the platform is moderated, there are not that many explicit words.

Slovene language

1. **Slovenian Twitter hate speech dataset IMSyPP-sl.** The dataset [3] contains Tweets containing different types of hate speech as well as appropriate language. It consist of the training set with 50000 tweets and the evaluation set with 10000 tweets.
2. **News platform 24UR article comments** For this project we created a scrapper to fetch the comments from the

articles of one of the most visited Slovenian news platform 24UR. We fetched all of the comments from 850 articles, then stopped since the amount of content was already enough. Since the platform archives all of the articles ranging all way back to year 1995, there exists a possibility to fetch comments from more than half a million articles. Due to the moderation of the site, comments don't contain a lot of explicit words, but hateful content is still very common. We plan to use this dataset for qualitative evaluation of our methods.

Methods

Machine translation and classification using fine-tuned BERT model

The idea is to finetune BERT model (pretrained on English data) to classify paragraphs into multiple classes of offensive language. The Slovenian dataset would be translated into English using the Google Translate API and we would test our trained model on the results.

Multilingual pretrained 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 [4], called CroSloEngual BERT [5]. The model was trained on Croatian, Slovenian and English data for the purpose of various NLP tasks.

We create two models, the first one finetuned on the dataset from Gab and the second one on the dataset from Reddit. Both dataset contain binary labels on whether the text is hateful or not. We use the Google Colab platform for training, since it offers GPUs with good performance. For training we set the batch size to 96, learning rate to $5 \times 10^{-}5$, max length of tokenized text to 128 and train the models for 3 epochs. In order to test the performance on the English datasets as well, we split both datasets into a train, validation and testing set with ratio 7:2:1, respectively. Both models are then tested on the Slovenian IMSyPP-sl Twitter hate speech dataset (binary hate/not hate).

In future research we plan to test this approach using different datasets and test the performance on the finetuned models on Slovenian data.

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

This approach is based on words that appear in the 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, 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.

Alignment of Slovene and English word embeddings

For alignment of Sloven and English word spaces we used the approach described in [6].

We are using two FastText [7] models, pretrained on Slovene and English data. With these models, we extracted embeddings, for pairs of words $\{x_i, y_i\}$ that appeared in dictionary taken from the repository of [8]. 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 [6] 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 major question here is how 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. In approach 1. and 2. we tried setting the threshold to 0.5 and to average probability of all word occurences, and got better results with the average. We tried these combination only on the train set, because we wanted to choose the right function if we have

good word probabilities. In 3. approach we took threshold 1, and in the last one 0.2. The results differed depending on the dataset, so here we should experiment some more.

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, y, are their probabilities. On this data we trained a support vector machine regressor, which will then output the probability that any new embedding belongs to offensive word.

Further work

This approach needs to be further tested on Slovene dataset: splitting sentences into words, generating their embeddings with Slovene FastText model, translating them into English word space, defining the probability they are offensive with regression model, and combining these probabilities into sentence classification.

Improvements can be made by generalizing the words during probability calculations of training words - we could count similar words as one, and maybe get a better representation. We should further experiment with the function that combines word probabilities into sentence classification, as this is a major part of the model, and we only tried some. Also for classification, SVR was our only choice up to now, so we should also try some other approaches.

Results

In this section we present the results of methods used.

Multilingual pretrained BERT finetuning

We take a look at the performance of the CroSloEngual BERT model in classifying the test sets from the Gab and Reddit datasets. These results don't use our cross-lingual approaches and are meant to show the performance of monolingual hate speech classification. In table (1) we show the results of testing the finetuned CroSloEngual BERT model on their respective datasets. We can see that the model performed

Table 1. Test set evaluation of finetuned CroSloEngual BERT.

Dataset	f_1-*score*	*precision*	*recall*
Gab	**0.900**	**0.895**	**0.905**
Reddit	0.806	0.814	0.799

best on the Gab dataset, achieving f_1-*score* of 0.900. The first reason for the better performance on the Gab dataset is probably that its balanced, containing roughly the same amount of hateful labels as non hateful. The second is that the dataset contains a lot of unfiltered explicit language. Explicit words usually mean that the text is in some way hateful, which allows the model to better classify the content. Although the content on Reddit is moderated and doesn't contain a lot of explicit language, the model still performed quite well, achieving f_1-*score* of 0.806.

The pretrained models were then tested on the Slovenian IMSyPP-sl dataset (evaluation set). In table (2) we present the results of testing, where CSE stands for CroSloEngual. We did three experiments, the first one was a baseline estimation where we used the CroSloEngual BERT for classification of IMSyPP-sl dataset in order to show the performance without finetuning on English data. In the other two we used the models that were finetuned on Gab and Reddit dataset. From

Table 2. Evaluation of finetuned CroSloEngual BERT on IMSyPP-sl dataset.

Finetuned model	f_1-*score*	*precision*	*recall*
CSE BERT - Gab	**0.357**	0.264	**0.550**
CSE BERT - Reddit	0.287	**0.273**	0.302
CSE BERT - no training	0.244	0.219	0.276

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 pretrained 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

Here are the results of classification of offensive comments on English Gab and Reddit datasets, where we got the embeddings with FastText model trained on data from Wikipedia, used SVR for predicting the probabilities and `logit` function for combining them.

Table 3. Evaluation on Gab and Reddit datasets.

	f_1-*score*	*precision*	*recall*
Gab	0.82	0.88	0.76
Reddit	0.72	0.84	0.63

In table 3 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.

References

[1] Aleš Žagar and Marko Robnik-Šikonja. Cross-lingual approach to abstractive summarization. *arXiv preprint arXiv:2012.04307*, 2020.

[2] 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).

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

[4] 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.

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

[6] 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.

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

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