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

Cross-lingual offensive language identification

Robert Košir, Edi Čebokli, and Žiga Kleine

Abstract

This project focuses on Cross-lingual offensive language classification. Our goal is explore the options of constructing models to solve the mentioned task, from traditional models like for example logistic regression, naive Bayes and random forest classifiers using tf-idf and custom features and more modern, neural network based models, like ELMo, mBERT and XLM-RoBERTa where some of them allow us to use transfer learning and use primarily English data to train the models and classify on Slovene. We evaluate and compare the models with numerous experiments such as binary and multi-label classification on English and Slovene datasets. We also prepare a translated English dataset in hopes of improving the performance. For testing on the Slovene language we prepared our own Slovene hate speech dataset. We conclude that modern approaches have better performance on the given data and reach decent results on the Slovene dataset.

Keywords

Hate speech classification, BERT, Cross-lingual language classification

Advisors: Slavko Žitnik

Introduction

With the ongoing rise in popularity of social medias, consequently the amount of offensive language used is rising as well. In this assignment we are provided with multiple datasets containing annotations for different types of offensive language. Our task is to use natural language processing techniques to predict whether a sentence contains offensive language and it's type in Slovene. Since there is a lack of annotated datasets in Slovene, we train our models on English datasets and then transfer the learned knowledge to Slovene. In the assignment we explore different techniques to achieve this goal.

1. Related work

Cross-lingual transfer is an active research field in natural language processing. There are different techniques for solving cross-lingual offensive language identification. Success in identifying offensive language was found by using neural networks like LSTM and bidirectional LSTM [1] and transformer models like BERT [2] and ELMO [3].

For Slovene language, the cross-lingual transfer applied to the topic of hate-speech classification, which was focused on anti-immigrant and anti-LGBT hate-speech classification based on BERT models, has been explored in [4].

The article [5] approaches the task of hate speech detection by utilising transfer learning on pre-trained BERT models. The authors of the article start by initialising a BERT model pre-trained on English Wikipedia and BookCorpus, then slightly modifying the models to achieve a structure more appropriate for hate speech detection. The base BERT model takes in tokenized text input, and has an output layer of 768 dimensions, that is later tweaked to better suit the purpose of the task. The results of different configurations are evaluated with metrics of precision, recall, and F1-score.

For the purpose of identification in a less-resourced language, in [6] cross-lingual word embeddings are used which represent lexical items from different languages with the same vector space. Classifiers trained on one language achieve good results since the words of the other language appear close to the words of the trained language in vector space. In the article the authors build cross-lingual word embeddings for several languages instead of the common pair of English and another less-resourced language.

In article [7], authors tackled a more relevant problem to ours. The goal was to identify offensive language using multilingual models trained on English data with two classes (offensive/non-offensive) which were then used to build models for Bengali, Hindi and Spanish. They compared the classification performance of cross-lingual contextual embeddings and transfer learning. Two models were used - XLM-R and BERT. The authors achieved slightly better results using trans-

fer learning on all three datasets.

Similar approaches with cross-language learning for different tasks were also used in [8] for argumentative relation and complex word identification in [9].

2. Existing solutions and datasets

2.1 Multilingual and Multi-Aspect Hate Speech Analysis

The article [10] presents a dataset containing tweets in three languages, English, French and Arabic. The number of English tweets amounts to 5,647. The tweets are labeled according to different attributes, like directness, hostility, target and group and multiple labels within the attributes. For our experiments we extracted tweets containing sexism, which amounted to 876 tweets.

2.2 Hate speech dataset from a white supremacist forum

In the paper [11] a dataset is presented consisting of 10,568 sentences. The data was extracted from a white supremacy forum - Stormfront using a web crawler. Sentences were classified as conveying hate speech or not and as a relation, where individual sentences do not necessarily convey hate speech but a combination of several sentences do. The authors inspected the dataset by making 3 models that would predict whether a sentence conveys hate speech or not. The three models were based on Support Vector Machines, Convolutional Neural Networks and Recurrent Neural Networks with Long Short-term Memories (LSTM). Out of the three models, the model based on LSTM performed the best with an accuracy of 0.78. Authors note the difficulty of annotating hate speech as it is of subjective nature and related to topics as free speech and tolerance. We used the 1196 comments that conveyed hate speech and labeled them as comments containing racism.

2.3 Detecting Online Hate Speech Using Context Aware Models

The article [12] tackles content aware hate speech detection on comment sections of the Fox News website. The dataset provided with the article contains 1528 annotated comments, with 435 of them labeled as hateful and 1093 of them labeled as non-hateful. Contrary to many other hate speech datasets, the dataset provided with the article also contains context for each comment included, in the form of the title of the article and a short description of the article under which the comment was added. Each comment also contains pointers to possible previous and next comments in a comment thread. Each comment is tagged with either containing hate speech or not. The authors used two types of models for context aware hate speech detection, feature-based logistic regression models and LSTM neural network models. While both types of models achieved good results, an improvement in accuracy was made by combining both types of models into an ensemble model. We manually annotated the comments with 5 labels that we would use in our experiments. Since the comments had a headline from a news article that they belong, we could more easily annotate, based on the headline. Not every comment was suitable and in total we extracted 65 racist and 51 homophobic comments.

2.4 CONAN COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech

The authors of the article [13] formed a dataset consisting of English, Italian and French hate speech/counter-narrative pairs. The pairs were obtained artificially with the help of Non-government organizations focused on preventing hate speech. The dataset focuses on annotating sentences that convey Islamophobia with sub-types like terrorism, islamization, generic, crimes and other. Authors augmented the English part of the dataset by paraphrasing sentences and translating Italian and French parts of the dataset amounting to 6654 pairs in English. The dataset features also suitable responses to such hate speech. The total amount of hate speech comments from the English part amounted to 408 comments that contained Islamophobia, which we used in our merged dataset.

2.5 A Benchmark Dataset for Learning to Intervene in Online Hate Speech

The main concern for the authors in the article [14] was to intervene with a friendly response in any kind of online hate speech. Their modal was learned on two different datasets. The first one was obtained from Reddit, where they have collected approx. 22000 comments. 10% of the comments were labelled as hate speech. The second one was from Gab which is similar to Reddit. They have collected almost 34000 posts where almost half of them were recognized as hate speech. The authors introduced four different strategies to intervene in online hate speech: hate words identification, hate speech classification, friendly response and suggestion of proper action. We used the comments that contained no hate speech to balance the hate speech comments in our merged dataset. The small portion of comments were randomly chosen.

2.6 Slovenian Twitter hate speech dataset IMSyPP-sl

In this article [15] a dataset containing tweets in Slovene language is presented. It contains 120,000 tweets, however not all were available, although the majority were. The tweets were annotated with 12 labels, out of which we select the tweets that are annotated with the 5 labels we chose for our multi-label classification experiment. Tweets that were not labeled as hate speech represent around 69% of the whole dataset. For our experiments, with our 5 chosen labels, we constructed a smaller dataset with better balance of the data, however there is some imbalance within tweets conveying hate speech as seen in Figure 1, especially with the lack of tweets conveying homophobic remarks.

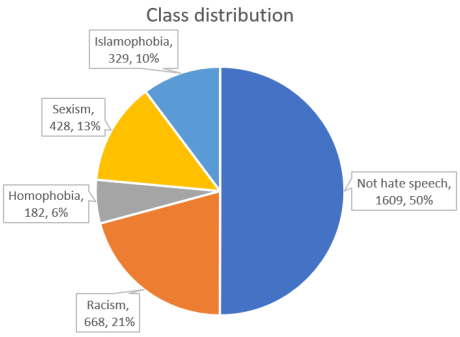

Figure 1. Pie chart showing the distribution of classes in the modified Slovenian Twitter hate speech dataset.

3. Methods

3.1 Preparing datasets

In total we prepared 4 datasets for our experiments. The first was created by combining multiple datasets mentioned in the previous chapter, creating a large training and testing dataset from which we created a dataset for binary classification, so that label 0 represents no hate speech, while 1 represents hate speech and a dataset for multi label classification, where data is labeled with subtypes of hate speech - racism, homophobia, sexism and Islamophobia. The distribution of the dataset is seen in Figure 2. Figure 2.

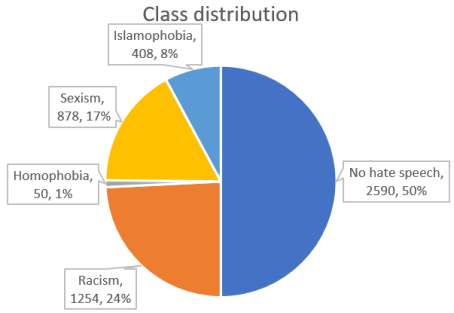

Figure 2. Pie chart showing the distribution of classes in the combined hate speech dataset.

The second dataset is the modified Slovenian Twitter hate speech dataset described in the previous chapter, where we extracted the labels we will be using for our experiments. The distribution is seen in Figure 1.

We also prepared a small Slovene hate-speech dataset, which we manually gathered from the comment sections of the nova24tv news website https://nova24tv.si/. We created a small dataset in two parts, meant mainly for evaluating our cross-lingual models. The first part contains multi-labeled data and consists of 100 comments, from which 20 are classified as neutral (labeled 0), 20 of them are racist (labeled with 1), 20 are homophobic (labeled 2), 20 are sexist (labeled 3), and 20 of them are Islamophobic (labeled 4). The second part we prepared is binary, and it consists of 160 comments, from which 80 are marked as hate speech (labeled 1), and 80 are neutral (labeled with 0).

The last dataset that we prepared was by taking the English combined dataset and translating it with a machine translation program called DeepL.

3.2 Traditional methods

Because of the abundance of hate-speech datasets in English, we first tackled the problem of hate speech classification on the English language, using some basic machine learning algorithms that were talked about in this course. We then also experimented with hate-speech classification on the Slovene language, as well as trying to apply some custom features to the data.

3.2.1 Hate-speech classification on English data

With the two datasets, one multilabel and one binary from the combined dataset, we first preprocessed the data by removing all non-alphabet characters, removing unnecessary spaces, removing stopwords and stemming the data. After that, we ran the data trough the tf-idf vectorizer, which generates feature vectors of tf-idf similarity between tokens in the dataset. After that, the data was split into training and test sets.

Then we ran our training data through 5 different models: a logistic regression model, a gaussian naive bayes model, a random forest classifier model, a multinomial naive bayes model and a bernoulli naive bayes model. We used our training data to train and the test data to evaluate the trained models. We calculated the accuracy, precision, recall and f-scores of the models. In the case of multilabel classification models, we used weighted metrics for evaluation.

3.2.2 Hate-speech classification on Slovene data

We also performed binary and multilabel classifications on the before mentioned Slovene Twitter dataset. The dataset contains tweets labeled with 5 different labels. Neutral tweets are labeled 0, racist tweets are labeled with 1, homophobic with 2, sexist with 3, and Islamophobic with 4. Here, we were using similar preprocessing as well as the same tf-idf vectorizer as for the multilabel classification on the English language mentioned before.

We then split the data into train and test sets, and ran them trough the same 5 machine learning models we talked about earlier. We evaluated the models by calculating accuracy, precision, recall and f-scores of the models. We used weighted metrics for evaluating multilabel data.

3.2.3 Hate-speech classification using custom features

Instead of using the tf-idf vectorizer to create a vector of features, we have also tried to create our feature vector. We have chosen text features that might be useful in offensive language detection. Such as the number of upper case characters in the text, exclamation mark count, word count and many more. After that, the data was split into training and test sets that were then used on the same models as above.

3.3 Neural network based approach

3.3.1 ELMo

To test the performance of a neural network based approach we perform transfer learning by fine-tuning a pre-trained ELMo model trained on the 1 Billion Word Benchmark [16]. We modify the architecture by adding a dense layer and a binary classification layer. We then train the model on our merged English dataset.

3.3.2 mBERT

To classify hate speech on a Slovene dataset using mBERT we use a pre-trained mBERT model called CroSloEngual BERT [17], which is a trilingual model trained on Croatian, Slovene and English data. By using a pre-trained model we can transfer the learned contextual dependent word embeddings to another problem by fine-tuning the model and saving a huge amount of time and lowering the risk of overfitting. We set a binary classification layer and start fine-tuning the model with two different approaches - by training the entire architecture and by freezing the entire architecture except the classification layer. As necessary for the BERT model, the text from the dataset was tokenized, and necessary tokens were added. We limit the number of tokens in a sentence to 64 for faster training speed and also to limit longer sentences which are although rare.

3.4 XLM-RoBERTa

XLM-RoBERTa is a model based on Facebook's RoBERTa model from 2019 [18]. It generally outperforms mBERT on various cross-lingual tasks. For our experiments we use a pretrained model that was trained on 100 different languages, Slovene and English included called xlm-roberta-base from Huggingface. Similarly to mBERT we also use the fine-tuning technique to relatively quickly train a model for our task.

4. Results

We conducted several experiments for the cross-lingual offensive language identification task, specifically binary and multi-label classification using our 4 prepared datasets on datasets with English comments and datasets. For evaluating performance we calculate the accuracy and weighted precision, recall and f-score.

4.1 Traditional methods

As mentioned, we ran our training data through 5 different machine learning models and evaluated their performance on the testing data.

4.1.1 Hate-speech classification on English data

The results on the combined English multilabel dataset can be seen in table 1, while the results on the combined English binary dataset can be seen in table 2. As we will see in the next chapter, the traditional methods are not significantly far from the neural network based approaches in terms of performance.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.848	0.849	0.848	0.840
Gaussian Naive Bayes	0.522	0.569	0.522	0.523
Random Forest Classifier	0.865	0.863	0.865	0.859
Multinomial Naive Bayes	0.837	0.837	0.837	0.829
Bernoulli Naive Bayes	0.853	0.850	0.853	0.850

Table 1. Performance of traditional classification methods on the multilabel dataset [19].

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.877	0.884	0.877	0.876
Gaussian Naive Bayes	0.811	0.813	0.811	0.811
Random Forest Classifier	0.879	0.880	0.879	0.878
Multinomial Naive Bayes	0.864	0.865	0.864	0.864
Bernoulli Naive Bayes	0.892	0.894	0.892	0.892

Table 2. Performance of traditional binary classification methods on our combined dataset.

4.1.2 Hate-speech classification on Slovene data

The results of the binary hate-speech classification on the Slovene twitter dataset can be seen in table 3, while the results of the multilabel classification can be seen in table 5.

We can see that the traditional models perform a bit worse on the Slovene dataset as they did on the English one, probably because the training set here is significantly smaller. We can also observe that the multilabel Slovene classification performs worse than the binary one almost all across the board, except for the random forest classifier, where the multilabel classification performs a bit better. The multilabel classification performs worse probably again because of an even smaller number of training examples provided for each hate-speech type.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.738	0.740	0.738	0.738
Gaussian Naive Bayes	0.698	0.708	0.698	0.695
Random Forest Classifier	0.831	0.831	0.831	0.831
Multinomial Naive Bayes	0.728	0.728	0.728	0.728
Bernoulli Naive Bayes	0.741	0.744	0.741	0.740

Table 3. Performance of traditional classification methods on the binary Slovene Twitter dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.669	0.739	0.669	0.636
Gaussian Naive Bayes	0.565	0.651	0.565	0.568
Random Forest Classifier	0.843	0.845	0.843	0.841
Multinomial Naive Bayes	0.675	0.725	0.675	0.647
Bernoulli Naive Bayes	0.690	0.711	0.690	0.670

Table 4. Performance of traditional classification methods on the multilabel Slovene Twitter dataset.

4.1.3 Hate-speech classification using custom features

As described in chapter 3.2.3, we have also included some custom features to detect hate speech. It's important to notice that the performance was way below the tf-idf approach. That's probably due to the importance of feature selection. In table, we can observe the performance of custom classifiers using traditional models on multilabel Slovene Twitter dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.501	0.501	0.501	0.489
Gaussian Naive Bayes	0.504	0.540	0.503	0.448
Random Forest Classifier	0.738	0.738	0.738	0.737
Multinomial Naive Bayes	0.520	0.523	0.520	0.520
Bernoulli Naive Bayes	0.507	0.540	0.507	0.458

Table 5. Performance of traditional classification methods with custom classifiers on the multilabel Slovene Twitter dataset.

4.2 Neural network based approach

For the mBERT model we first explore and find the best way to train the model, whether by freezing the layers except the classification one, or by training without freezing. We test the models on the binary and multi-labeled English combined dataset and get the results shown in Table 6 and Table 7. The model with frozen layers was traind for 5 epochs with frozen layers and 5 epochs without The difference between the two models in binary classification is almost non-existent, however we get slight differences in multi-label classification. For the following experiments we train the model with unfrozen layers.

Model	Accuracy	Precision	Recall	F1-Score
mBERT (frozen layers)	0.943	0.943	0.943	0.943
mBERT	0.943	0.944	0.943	0.943

Table 6. Performance of mBERT models by freezing layers during training and without tested on the binary English combined dataset.

Model	Accuracy	Precision	Recall	F1-Score
mBERT (frozen layers)	0.916	0.928	0.916	0.918
mBERT	0.934	0.935	0.934	0.934

Table 7. Performance of mBERT models by freezing layers during training and without tested on the multi-labeled English combined dataset.

Next we test our models on the combined English dataset, first for binary classification. We train the models on the training part of the dataset that is 80% of the whole dataset. The ELMo model was trained for 10 epochs, while the BERT and XLM models were trained for 5 epochs and 8 epochs respectively. The results are shown in Table 8. mBERT and XLM-RoBERTa significantly outperform ELMo. The main difference between the three models are that mBERT and XLM-RoBERTa use transformers, while ELMo uses LSTM. The two transformer models are deeply bidirectional as apposed to ELMo which uses a concatenation of both directions, which could be the reason we get better performance As reported in [18] XLM-RoBERTa outperforms mBERT, which is also the case in this experiment. The transformer models achieve good results for binary classification on a relatively small dataset.

We also train models for the multi-label classification. Using the same amount of epochs we get results shown in 9. We get similar results to binary classification, however ELMo and XLM-RoBERTa even outperform models for binary classification by a slight margin. It could be just a coincidence or having separate types of hate speech could improve distinguishing between hate speech and not hate speech by more accurately tying some typical words to a type of hate speech.

Model	Accuracy	Precision	Recall	F1-Score
ELMo	0.804	0.810	0.804	0.803
mBERT	0.943	0.943	0.943	0.943
XLM-RoBERTa	0.975	0.975	0.975	0.975

Table 8. Performance of the three models on the binary English combined dataset.

Model	Accuracy	Precision	Recall	F1-Score
ELMo	0.828	0.836	0.828	0.826
mBERT	0.971	0.971	0.971	0.971
XLM-RoBERTa	0.984	0.984	0.984	0.984

Table 9. Performance of the three models on the multi-label English combined dataset.

Next we test the models on the Slovene dataset that we prepared with 160 comments for binary classification and 100 comments for multi-label classification. The results for both binary and multi-label are shown in Table 10. The results did not meet expectations as the models achieved poor performance. The majority class for binary classification is 0.5 and 0.2 for multi-label. The models achieve slightly better results than that, however still not useful. The models mistakenly classify the majority of comments as racist. One reason could be the presence of dialects found in the Slovene dataset paired with poor writing of the comments. Another reason could be that the English dataset contains specific types of hate speech inside the sub-types that would not appear in the Slovene language, which would have that much bigger of an effect when using a small dataset as is the case with our experiments.

Model	Accuracy	Precision	Recall	F1-Score
mBERT (binary)	0.638	0.705	0.638	0.605
XLM-RoBERTa (binary)	0.625	0.679	0.625	0.594
mBERT (multi-label)	0.380	0.453	0.380	0.301
XLM-RoBERTa (multi-label)	0.340	0.329	0.340	0.282

Table 10. Performance of mBERT and XLM-RoBERTa models on the binary and multi-label Slovene dataset.

We also try to train the mBERT model on the English dataset translated to Slovene and see if we get improved performance. The results for binary and multi-label classification are shown in Table 11. We observe a drop in accuracy and recall but an increase in precision and f1-score. We do not gain any performance by translating the dataset. Since most English comments were not written in formal English, the translations are not as accurate. Some comments also contain derivatives of words that the translation program has no translations of.

To improve the performance we additionally trained the models trained on English data with data from the Slovenian Twitter hate speech dataset that we described in previous chapters. We then tested the models on our Slovene dataset for binary and multi-label classification. The results are shown

Model	Accuracy	Precision	Recall	F1-Score
mBERT	0.380	0.453	0.380	0.301
mBERT (translated)	0.350	0.533	0.350	0.310

Table 11. Performance of mBERT model trained on the English combined dataset and mBERT model trained on the English combined dataset and additionally on the translated English combined dataset to Slovene and tested on the multi-label Slovene dataset.

in Table 12. We can observe a significant increase in performance. While XLM-RoBERTa outperforms mBERT in binary classification it slightly falls behind in multi-label classification. The Slovene dataset has a slightly better distribution of labeled comments which most likely had a larger impact on the performance, especially for identifying homophobic comments. Having comments that include typical words found in hate speech comments and informal Slovene words enables the model to learn such connections.

Model	Accuracy	Precision	Recall	F1-Score
mBERT (binary)	0.900	0.900	0.900	0.900
XLM-RoBERTa (binary)	0.913	0.913	0.913	0.913
mBERT (multi-label)	0.810	0.853	0.81	0.803
XLM-RoBERTa (multi-label)	0.790	0.857	0.790	0.785

Table 12. Performance of mBERT and XLM-RoBERTa models, additionally trained on Slovene data from Twitter, on the binary and multi-label Slovene dataset.

In comparison to traditional methods we achieve better performance in all experiments with neural network based approaches.

5. Conclusion

For our project we used traditional and neural network based approaches to solve cross-lingual offensive language identification. We combined multiple datasets in a larger one with multiple labels we chose and trained the models on the data. Additionally we tried translating the English dataset with machine translation in hopes of achieving better performance, however it decreased. We found out that mBERT and XLM-RoBERTa models perform similarly and achieve quite good results for binary classification of hate speech in Slovene. In comparison to traditional methods, we get significantly better performance with neural network based approaches.

To improve the results we could expand our datasets, especially the English dataset. Possibly choose different labels with more available data, however the labels we chose are important to identify as they are quite common on the internet.

References

[1] Akanksha Bisht, Annapurna Singh, H. Bhadauria, Jitendra Virmani, and Dr Kriti. *Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model*, pages 243–264. 03 2020.

[2] Rajalakshmi Sivanaiah, Angel Suseelan, S Milton Rajendram, and Mirnalinee T.t. TECHSSN at SemEval-2020 task 12: Offensive language detection using BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2190–2196, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[3] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[4] ŽAN PEČOVNIK. Medjezikovni prenos napovednih modelov za sovražni govor. 2020.

[5] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2019.

[6] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 894–904, Valencia, Spain, April 2017. Association for Computational Linguistics.

[7] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online, November 2020. Association for Computational Linguistics.

[8] Gil Rocha, Christian Stab, Henrique Lopes Cardoso, and Iryna Gurevych. Cross-lingual argumentative relation identification: from English to Portuguese. In *Proceedings of the 5th Workshop on Argument Mining*, pages 144–154, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[9] George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. Cross-lingual transfer learning for complex word identification, 2020.

[10] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2019.

[11] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels,

Belgium, October 2018. Association for Computational Linguistics.

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

[13] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] 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.

[15] Petra Kralj Novak, Igor Mozetič, and Nikola Ljubešić. Slovenian twitter hate speech dataset IMSyPP-sl, 2021. Slovenian language resource repository CLARIN.SI.

[16] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling, 2014.

[17] Matej Ulčar and Marko Robnik-Šikonja. CroSloEngual BERT 1.1, 2020. Slovenian language resource repository CLARIN.SI.

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

[19] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Weblogs and Social Media*, ICWSM '17, 2017.