FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Cross-lingual offensive language identification

Robert Košir, Edi Čebokli, and Žiga Kleine

Abstract

This project will focus on Cross-lingual offensive language classification. Our goal is to test out different models for the mentioned task, ranging from different multilingual BERT configurations, to cross-lingual embedding. We will evaluate and compare different models accordingly. For evaluation, we will prepare a small Slovene hate speech 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].

The article [4] 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 [5] 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 [6], 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 transfer learning on all three datasets.

Similar approaches with cross-language learning for different tasks were also used in [7] for argumentative relation and complex word identification in [8].

2. Existing solutions and datasets

2.1 Automated Hate Speech Detection and the Problem of Offensive Language

This dataset was presented in the paper [9]. The goal of the paper is to present the way to separate hate-speech from other instances of offensive language. Firstly, the authors created the dataset by searching for tweets, using the Twitter API that contained terms from the online lexicon available at

Hatebase.org. Afterwards, they asked CrowdFlower workers to label each tweet as one of three categories: hate speech, offensive but not hate speech or neither. They were also asked to think about the context in which the ”hate” term is present in a particular sentence. After the dataset was successfully prepared, the authors used different predefined models such as logistic regression, naive Bayes, decision trees and many more. They've found out that the best model has had an overall precision of 0.91 and recall of 0.90. Almost 40% of hate speech were misclassified.

2.2 Hate speech dataset from a white supremacist forum

In the paper [10] 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.

2.3 Detecting Online Hate Speech Using Context A-ware Models

The article [11] 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 1 or 0, 1 annotating hate speech, and 0 no hate speech. 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.

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

The authors of the article [12] 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.

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

The main concern for the authors in the article [13] 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.

3. Methods

3.1 Preparing datasets

We combined 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, such as racism, sexism and islamophobia.

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 two small datasets, meant mainly for evaluating our cross-lingual models. The first dataset is a multilabel one, consisting of 100 comments, from which 20 are classified as xenophobic (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 neutral (labeled 4). The second dataset 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).

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 did multilabel classification on the dataset provided alongside the paper called Automated Hate Speech Detection and the Problem of Offensive Language [9]. The dataset provided consists of tweets labeled with labels 0, 1 or 2, with 0 representing hate speech, 1 representing offensive language and 2 representing neutral language.

With the two datasets, one multilabel and one combined binary, we first preprocessed the data by removing all non-alphabet characters, removing unnecessary spaces, removing stopwords and stemming the data. After that, we vectorized the data and split the datasets 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 evaluate the trained models. We calculated the accuracy, precision, recall and f-scores of the models.

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 [14]. 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 BERT

To classify hate speech on a Slovene dataset using BERT we use a pre-trained BERT model called CroSloEngual BERT [15], 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.

4. Results

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. We calculated the the accuracy, precision, recall and f-score. The results on the multilabel dataset can be seen in table 1, while the results on the combined 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
LogisticRegressionModel	0.8989	0.8862	0.8989	0.8843
GaussianNaiveBayesModel	0.7038	0.7829	0.7038	0.7375
RandomForestClassifierModel	0.9010	0.8849	0.9010	0.8877
MultinomialNaiveBayesModel	0.8480	0.8302	0.8480	0.8132
BernoulliNaiveBayesModel	0.8796	0.8603	0.8796	0.8664

Table 1. Performance of traditional classification methods on the multilabel dataset [9].

Model	Accuracy	Precision	Recall	F1-Score
LogisticRegressionModel	0.8921	0.8920	0.8921	0.8921
GaussianNaiveBayesModel	0.7394	0.8291	0.7394	0.7525
RandomForestClassifierModel	0.8836	0.8878	0.8836	0.8851
MultinomialNaiveBayesModel	0.8490	0.8472	0.8490	0.8398
BernoulliNaiveBayesModel	0.8681	0.8665	0.8680	0.8671

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

4.2 Neural network based approach

We train the ELMo model for 20 epochs, while the BERT model with frozen layers was trained for 10 epochs on frozen layers and 3 epochs on unfrozen, and the BERT model trained on the whole architecture was trained for 5 epochs. The results of binary classification are shown in Table 3 for the English dataset and Table 4 for the Slovene dataset. The ELMo model was tested only on the English dataset. We see that BERT outperforms ELMo by a large margin, while the difference between training with frozen layers and the whole network are slight. Slightly larger differences can be observed on the Slovene dataset, although the amount of data is much smaller, which has a bigger impact on the metrics in case of a few failures. While the model trained on the frozen layers has a larger precision it fails to detect a larger portion of the sentences containing hate speech.

Model	Accuracy	Precision	Recall	F1-Score
ELMo	0.8771	0.8990	0.8497	0.7757
BERT (frozen layers)	0.9401	0.9377	0.9428	0.8872
BERT	0.9413	0.9413	0.9413	0.8891

Table 3. Performance of neural network based approach binary classification methods on our combined dataset.

Model	Accuracy	Precision	Recall	F1-Score
BERT (frozen layers)	0.7938	0.9796	0.6000	0.5925
BERT	0.8250	0.9333	0.7000	0.6666

Table 4. Performance of neural network based approach binary classification methods on our Slovene dataset.

5. Ideas for the final submission

For our future work we will use neural network based approaches for multi label classification on English and Slovene datasets that we already prepared. We will explore different transfer learning training techniques such as freezing different layers or gradual unfreezing to try and achieve better results. Furthermore, we will use mBERT and XLM-R models and compare results. The other approach we would test is cross-lingual mapping of ELMo embeddings to Slovene.

Another idea for improving the performance from there is to use a translation model to translate the English dataset into Slovene, and then use the translated data to perform the fine tuning on the multilingual BERT and see how it affects the performance. Additionally we will try to expand our Slovene dataset by gathering more data from other sources and get a more diverse English dataset to train on in terms of the number and quantity of hate speech subtypes.

6. Conclusion

Our project consists of multiple English and Slovene hate speech classification solutions that are appropriately evaluated and compared. We provided a small Slovene dataset containing hate speech. So far we used traditional models and ELMo for hate speech classification in English and multilingual BERT models for classification in Slovene.

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] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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