UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 2. stopnja

Tina Vovk

Projekt pri predmetu matematika z računalnikom

Ljubljana, 2024

1. Introduction

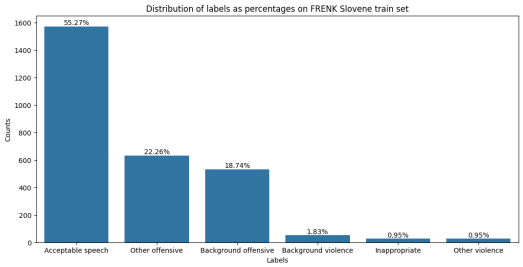
The aim of the project is to investigate whether we can improve NLP methods for hate speech detection, by exploiting the fact that hate speech usually has some sexist and/or racist basis; so by using multi-task learning or intermediate training of a model on datasets annotated for sexism/racism we should improve performance on hate speech detection. There are many existing datasets for these phenomena, but not many in Slovene; so we want to approach the problem using multi-lingual or cross-lingual techniques, e.g. training on datasets annotated in English but testing on Slovene or by training the model based on one hate speech target and testing it on a different one. The was done using Jupyter notebook in Google Colab. It was agreed that all the datasets shared with me can not be shared externally.

2. Setting up the datasets

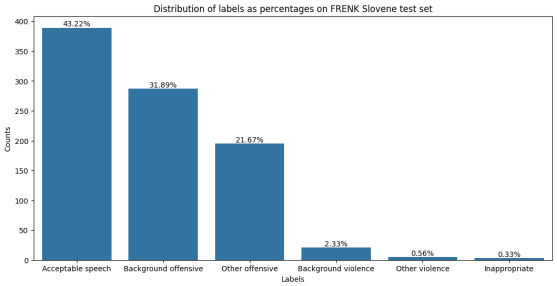
At the beginning of the project I had to investigate the structure of our datasets. I was given two sets of datasets.
First ones are so called *FRENK Datasets of Socially Unacceptable Discourse in Slovene and English*[1], where we had various comments in English, Slovene and Croatian language.
In each of those languages we have seperate datasets where some comments are targeting migrant groups and some lgbt community. The columns in the datasets consist of text column containing the possibly offensive comments, offensiveness label column where it says whether the comment is deemed offensive or not, and target Column describing who the target of the hate speech is. Then in each set of targets the data is split into two datasets: train and test sets. The main advantages of these datasets are that there are identical sampling procedures, producing comparable data across languages and an annotation schema that takes into account six types of SUD and five targets at which SUD is directed.
Then we have second set of Datasets called *IMSyPP Datasets* [2]. It includes four annotated datasets, one for each target language: English, Italian, Slovenian and Dutch. The Slovenian dataset consists of Twitter posts and was drawn from an exhaustive set of all Slovenian Twitter posts of the last three years. The dataset is not focused on any specific topic.
I first had to perform some statistics tests on the dataset for example which labels appear, distribution of labels, the lengths of datasets, distribution of targets etc.
I was focusing mainly on the Slovene datasets. The code is available here.

2.1. **FRENK Datasets.**

The labels in FRENK Datasets were labeled as "Acceptable speech", "Other offensive", "Background offensive", "Background violence", "Inappropriate"and "Other violence"in both Datasates with targets of hatespeech being migrants or LGBT. Following images are the distributions of train and test sets of Datasets where the target is LGBT.

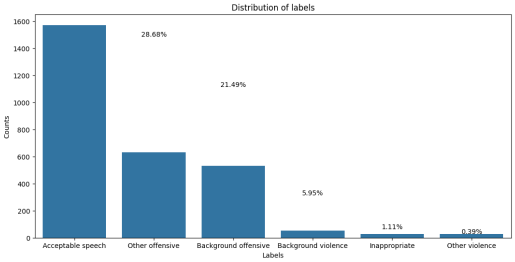


Slika 1. Distribution of LGBT train set

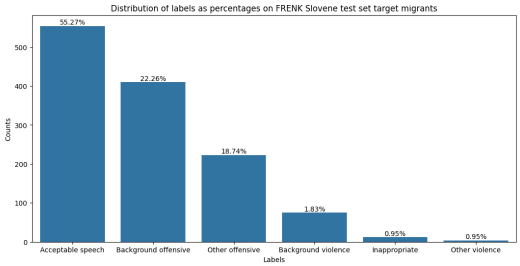


Slika 2. Distribution of LGBT test set

And bellow are the distributions of train and test sets of Datasets where the targets are migrants. We can observe that the distributions for both datasets are similar if test and train sets are compared.

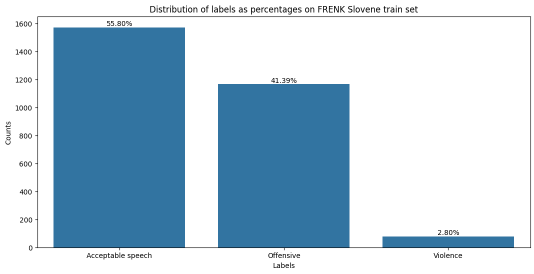


Slika 3. Distribution of Migrants train set

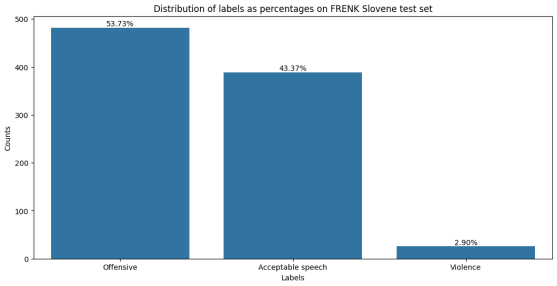


Slika 4. Distribution of Migrants test set

Due to instances of label *inappropriate* being so low in number compared to other labels, it was decided that it's best to remove that label to improve accuracy of the model. It was also decided that the labels of *other* and*bacground violence* and *other* and *bacground offensive* would be mapped to *violence* and *offensive*, respectively (for possible future comparison between the two types of datasets). The training of the models was then later done on the improved labeling. The distibutions of corrected labels are as shown below for LGBT target.



Slika 5. Distribution of LGBT train set



Slika 6. Distribution of LGBT test set

The distributions of corrected train and test sets of Datasets where the targets are migrants are simillar to those where target is LGBT if test and train data are

copared and will not be included due to repetition, but can be found in the provided link to the code.

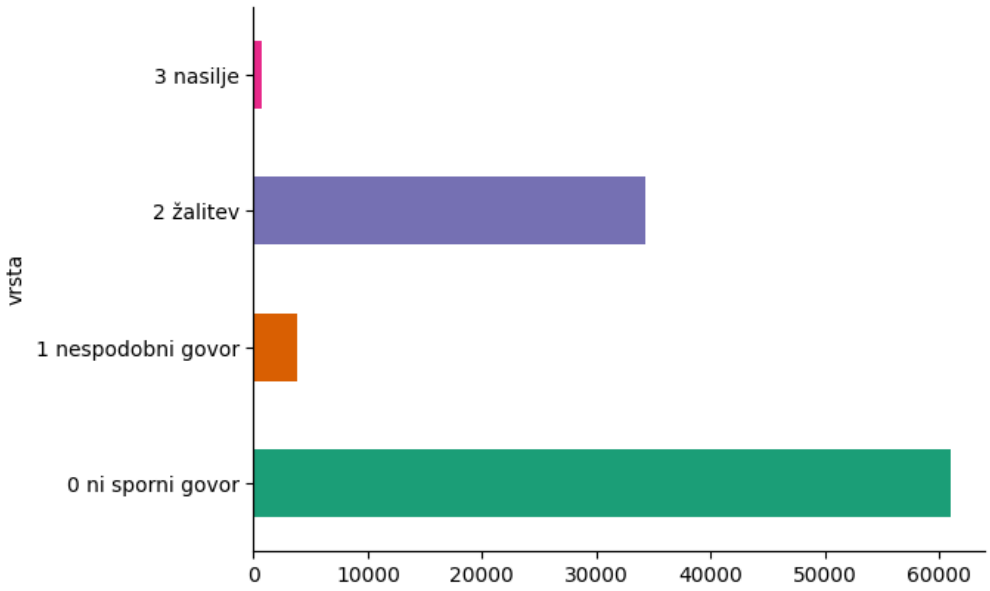
2.2. **IMSyPP Datasets.** IMSyPP Datasets were similar with a couple more columns (ID, Username, text, label, target, language, anotator), but the important ones were once again the text, label and target columns. The comments were again labeled as in Slovene but renamed into english to match the FRENK Datasets.

The mappings are shown in Table 1.

Tabela 1. Label Mapping to Numerical Form

Original Label	**Mapped Label**
2 žalitev	Offensive
3 nasilje	Violence
0 ni sporni govor	Acceptable

Distibution of the labels is shown bellow:



Slika 7. Distribution of IMSyPP dataset

Once again due to instances of label *inappropriate* being so low in number compared to other labels, it was decided that it's best to remove that label.

2.3. **Preparing Datasets.** After mapping the labels, a dictionary is created to assign a unique index to each unique label in the dataset. This process converts the categorical labels into numerical form, which is required for further analysis and model training. Next, we create a dictionary that assigns a unique index to each of these mapped labels, as shown in Table 2.

The texts are then extracted into a list, and their corresponding labels are transformed using the previously defined dictionary to create a list of numerical labels.

Tabela 2. Label Dictionary with Numerical Indices

Mapped Label	**Index**
Offensive	0
Violence	1
Acceptable	2

These lists prepare the dataset for further processing or model input. The code for preparing the datasets can be found here.

3. Building a model

After preparing and getting familliar with the datasets, I started building, and training a BERT-based classifier for text classification found here, specifically focusing on labeling offensive or violent comments.
For Bert you first have to load and prepare the datasets as described above. Then I had to define custom dataset class for text classification: *TextClassificationDataset* function for training and validation. It handles tokenization and encoding of text data, preparing it for input into the BERT model. Then a BERT-based classifier function is defined: *BERTClassifier*. It consists of a BERT model followed by a dropout layer and a linear layer to adapt BERT's outputs to the specific number of classes (in this case 3 after removing the others).Before training a model tokenizer is initialized. This tokenizer is configured with a pre-trained BERT model. We use-pretrained model obtained from huggingface open repository and the choice of which model to use depends on the language of dataset. Then the training function was defined, that was later enhanced using Automatic Mixed Precision as that speeds up the operations without major impact on performance. And finally functions for training and evaluating the model are defined. These handle batching of data, setting the model to training or evaluation mode, computing loss, and updating model parameters. The evaluation function calculates accuracy and other classification metrics to assess model performance.

3.1. **Comparing the models.**

I then compared performance of models using previously defined evaluate function, after training the models in 20 epochs and using different pre-trained models. The pre-trained models I tried were all trained on mutliple languages including Slovenian and are the following: *Bert-base-multilingual cased*[3], *Bert-base uncased*[4] and *XLM-Roberta*[5]. This was done on FRENK Datasets.
The majority class in the dataset, both in terms of percentage and count, is *Acceptable speech*. The detailed distribution is presented in Table 3.

As shown in Table 3, the class 'Acceptable speech' constitutes 43.22% of the dataset and has 389 instances, making it the predominant class.

Bellow are the evaluation results after training the models in 20 epochs on LGBT targeted datasets:

Tabela 3. Majority Class Distribution

Class	**Percentage (%)**	**Count**
Acceptable speech	43.22	389
Background offensive	31.89	287
Other offensive	21.67	195
Background violence	2.33	21
Other violence	0.56	5
Inappropriate	0.33	3

Tabela 4. Comparison of the models

	BERT-base mulilingual cased	BERT-base uncased	XLM RoBERTa
Accuracy	0.67	0.35	0.34
Macro average f1	0.53	0.27	0.28
Weighed average f1	0.67	0.35	0.34

Here the BERT-base multilingual cased model achieved the highest accuracy and weighted average f1 score, suggesting it is most effective in classifying comments accurately.
Results of training different models for migrant targetted datasets are shown in Table 5

Tabela 5. Comparison of the models for migrants

	BERT-base mulilingual cased	BERT-base uncased	XLM RoBERTa
Accuracy	0.44	0.44	0.44
Macro average f1	0.21	0.32	0.30
Weighed average f1	0.28	0.43	0.42

Here all models tested achieved the same accuracy, but there were variations in the macro and weighted average f1 scores.

4. Tests

We then wanted to test the difference in accuracy of our model if the model is trained based on a specific target and then evaluated on the same target compared to it being trained on one target and then evaluated on a different target. This was done on IMSyPP Datasets.
First, I had to separate the data based on specific targets and then split the data into training and test sets. The model was trained for 20 epochs and saved. The saved model was then loaded and used to evaluate a different dataset with the same or different target, and the results were compared. The results are presented in Table 6.

Tabela 6. Comparison of models based on training them on different targets

	Racism-Racism	Homophobia-Homophobia	Racism-Homophobia	Homophobia-Racism
Accuracy	0.86	0.80	0.90	0.82
Macro average f1	0.58	0.79	0.90	0.55
Weighted average f1	0.85	0.79	0.90	0.80

The results show that the model can generalize well across different targets, particularly when trained on one target and evaluated on another. The highest accuracy and F1 scores were observed when the model trained on racism was evaluated on homophobia. However, this could be due to a difference in training data quantity and data quantity on which evaluation was performed. For example, when training on racism, there were 168 acceptable samples, 159 offensive samples, and 6 violence samples, while the evaluation on homophobia had only 11 offensive and 11 acceptable samples. This imbalance could impact the observed performance.

I will continue the testings on other targets and different datasets and see if I can improve the results. The codes for this analysis can be found here for Racism-Racism and Racism-Homophobia and here for Homophobia-Racism and Homophobia-Homophobia.

5. References

(1) Ljubešić, N., Fišer, D., Erjavec, T. (2019). The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In: Ekštein, K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science(), vol 11697. Springer, Cham. https://doi.org/10.1007/978-3-030-27947-9_9

(2) Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., Zollo, F. (2022). Handling Disagreement in Hate Speech Modelling. In: Ciucci, D., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2022. Communications in Computer and Information Science, vol 1602. Springer, Cham. https://doi.org/10.1007/978-3-031-08974-9_54

(3) https://huggingface.co/google-bert/bert-base-multilingual-cased

(4) https://huggingface.co/google-bert/bert-base-uncased

(5) https://huggingface.co/docs/transformers/model_doc/xlm-roberta