Using Deep Learning and Visual Analytics to Investigate Hate speech Patterns in Lithuanian Politics

VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

AIM OF THE RESEARCH

The aim of this research is to detect and analyse hate speech of the Lithuanian politicians. For this a hate speech detection in Lithuanian language model was trained and used to classify the stenograms from the Lithuanian parliament meetings.

TRAINING DATASET

- Deep learning models were trained on the comments collected from various Lithuanian social media pages *(Facebook)* and news websites *(lrytas, 15min, alkas, delfi)*.
- Overall, 25 219 comments were collected and annotated by four annotators.
- The comments were annotated into three classes: neutral, offensive and hate speech.

Class	No. of comments
Neutral	15 316
Offensive	7 821
Hate	2 082

FINE-TUNING MODELS

Three different deep learning base models were chosen for the fine-tuning stage: Multilingual-BERT, LitLat-BERT and Electra transformer. Two BERT models were already pretrained for the Lithuanian language, and we trained the Electra base transformer from scratch.

The models fine-tuning pipeline was the following:

- The dataset was divided into training, validations and testing sets by the ratio 0.6:0.2:0.2.
- The comments for the hateful and abusive language classes were replicated (duplicated) in each of the samples to get the balanced datasets.
- Since the generated embeddings were vectors of length 512, any comments with more tokens were discarded.
- The model was trained with three different seeds and the one having the best results was chosen for further comparison.

The trained models were evaluated using various classification metrics. We present here an overall F1-score metric for each model.

Model	Overall F1-score
Electra	0,55
Multilingual-BERT	0,63
LitLat-BERT	0,72

Since LitLat-BERT model had the best F1-score, we decided to use only this model for the hate speech detection in Lithuanian politics.

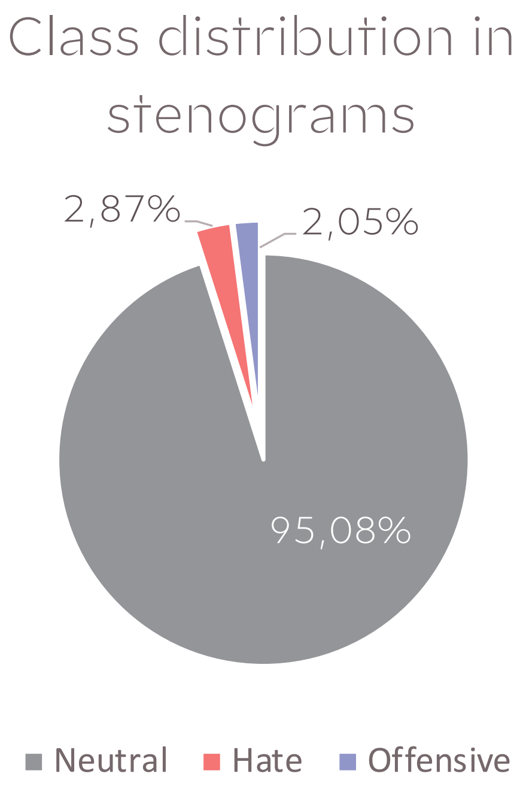
CONCLUSIONS

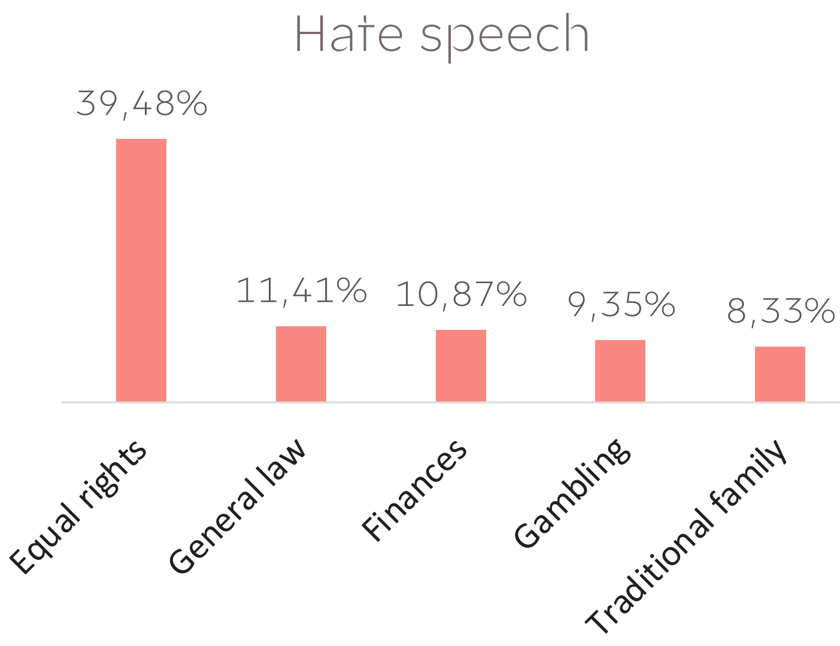
- Three transformer-based models were tested for the hate speech detection. The best performing model was LitLat-BERT.
- Overall, 2.87% of hate and 2.05% of offensive speeches were found in the stenograms from the Lithuanian parliament meetings.
- The opposition factions had more hate and offensive speeches than the government factions.

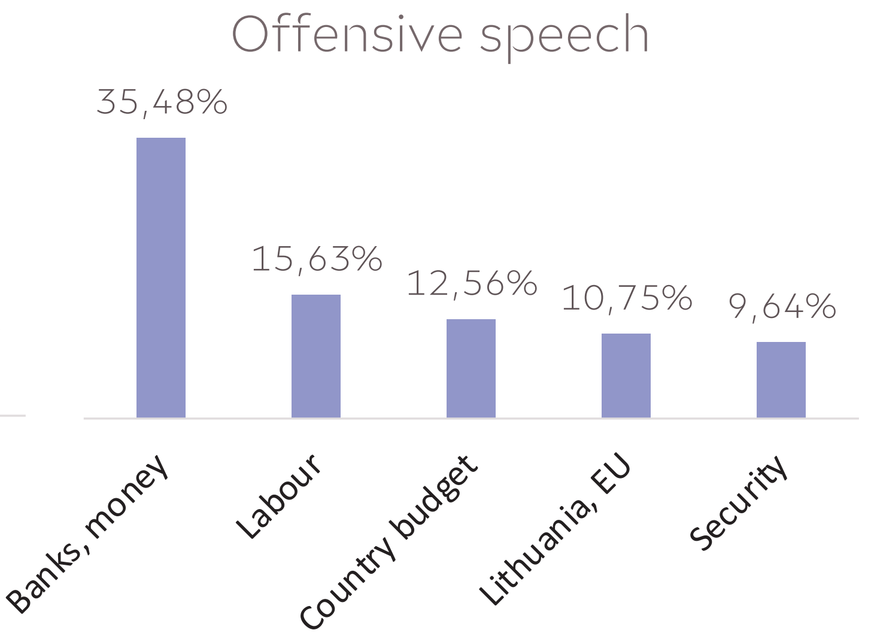
OVERALL ANALYSIS RESULTS

The hate and offensive speech classes were the minority of all the politician speeches.

For the hate and offensive speeches we used topic modelling (LDA method) to see what the most occurring topics are. The texts were modelled into 15 topics, however, only the top 5 of each class are displayed.


Class distribution in stenograms

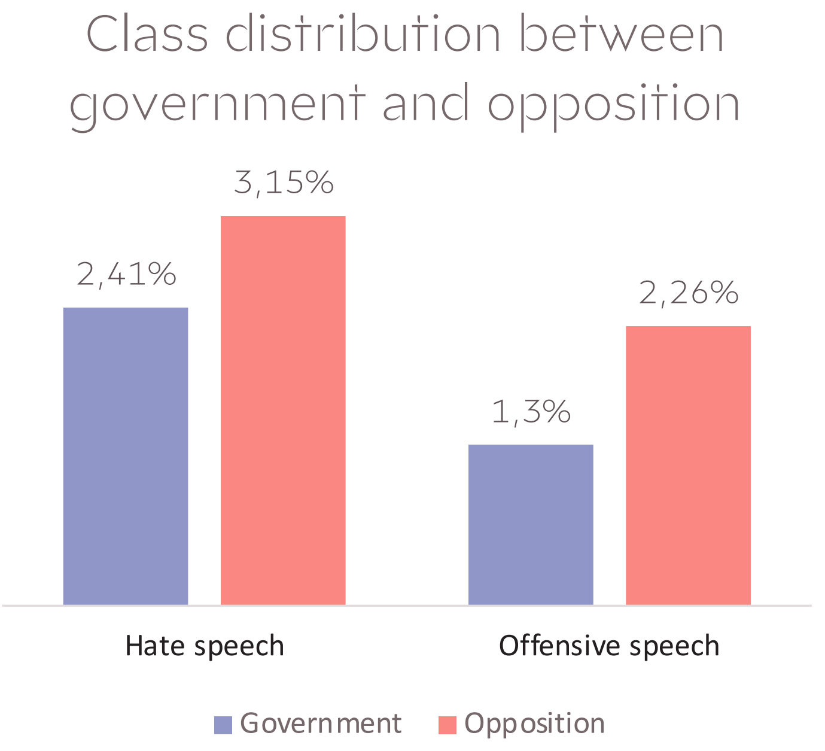

Hate speech


Offensive speech

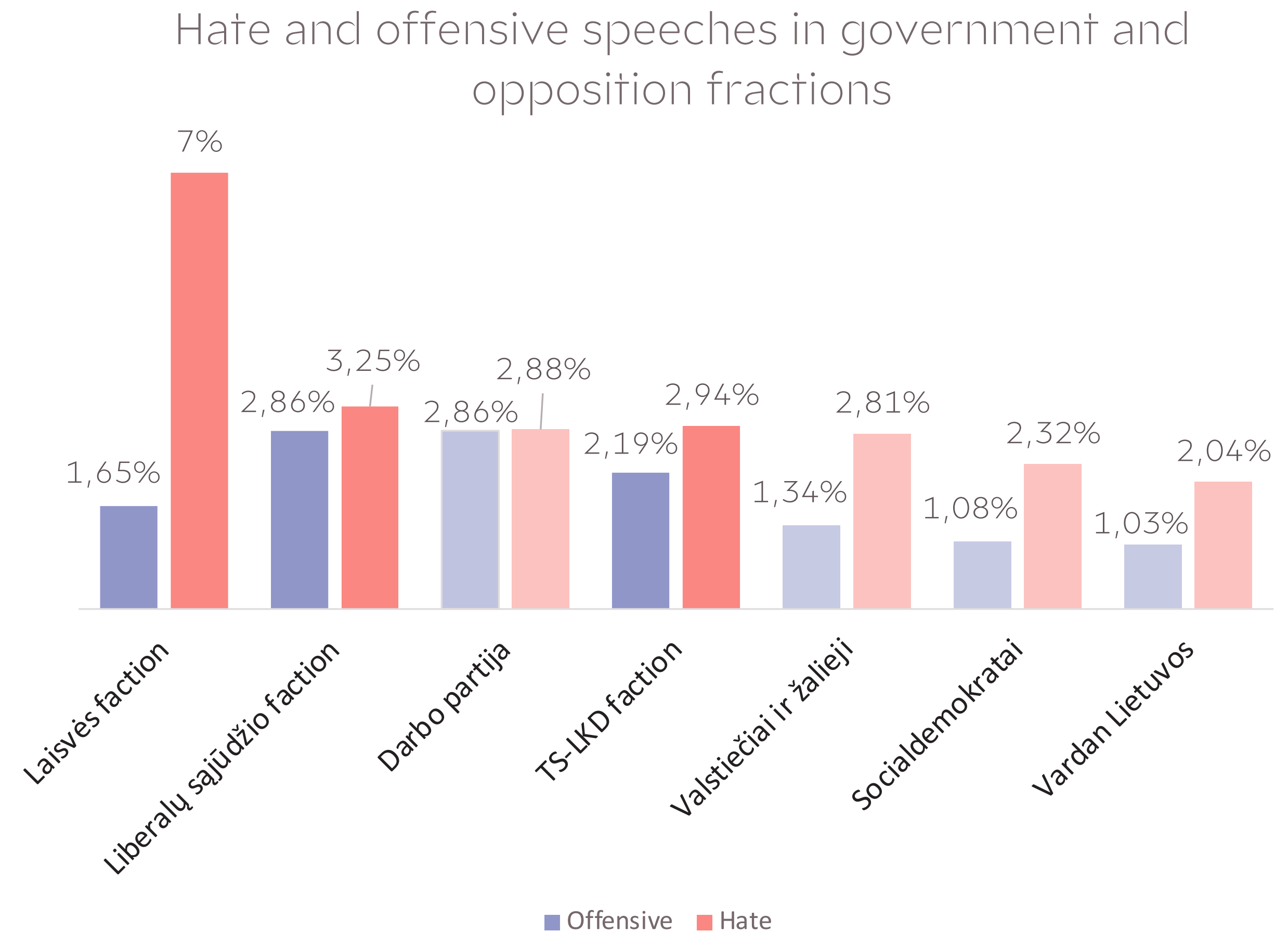
GOVERNMENT AND OPPOSITION COMPARISON

The percentages of hate and offensive speeches were compared between the government and opposition factions.

In general, the opposition factions had higher proportion of hate and offensive speeches than the government factions.


Class distribution between government and opposition

However, by looking at each faction separately, we can see that the two most hate speech using factions are government factions.


Hate and offensive speeches in government and opposition fractions

FUTURE PLANS

- Develop a real-time analysis based dashboard, where Lithuanian citizens could find all the statistics of hate and offensive speeches from the Lithuanian parliament meetings.
- Analyse correlations between how the voters vote and how many hate and offensive speeches appear in the parliament meetings of each political party.

AUTHORS:

Milita Songailaitė
milita.songailaite@vdu.lt

Justina Mandravickaitė
justina.mandravickaite@vdu.lt

Justinas Juozas Dainauskas
justinas.dainauskas@vdu.lt

CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT