VYTAUTO
DIDŽIOJO
UNIVERSITETAS
Informatikos
fakultetas

DEEP LEARNING MODELS FOR HATE SPEECH DETECTION

RELEVANCE

Online hate speech is assumed to be an important factor in political and ethnic violence. Therefore, media platforms are pressured to timely detection and elimination of hate speech. This tendency led to increasing efforts in terms of hate speech detection, and several hate speech detection models have been developed. Hate speech is not only a complex phenomenon that is difficult to detect but even its definitions vary in different studies, therefore comparison of different hate speech detection models not in terms of performance but in terms what is marked as hate speech could contribute to more comprehensive understanding of the phenomenon and its timely identification.

DEFINITIONS

HATE SPEECH: Describes negative attributes or deficiencies to groups of individuals because they are members of a group. Hateful comment occurs toward groups because of race, political opinion, sexual orientation, gender, social status, health condition, or similar.

OFFENSIVE CONTENT: Posts that are degrading, dehumanizing, insulting an individual, threatening with violent acts, fall into this category.

GOAL

The purpose of this experiment is to compare selected hate speech detection models for English from the perspective of inter-annotator agreement.

DATA

For model comparison, we used an English dataset from HASOC 2019 shared task:

- Sources – Twitter & Facebook
- 2 subsets of English dataset:
  - Training subset (*5852 posts*)
  - Test subset (*1153 posts*)
- Classes:
  - NOT – Non-Hate-Offensive: posts do not contain any hate speech or offensive content
  - HATE – Hate speech: posts contain hate speech content
  - OFFN – Offensive: posts contain offensive content

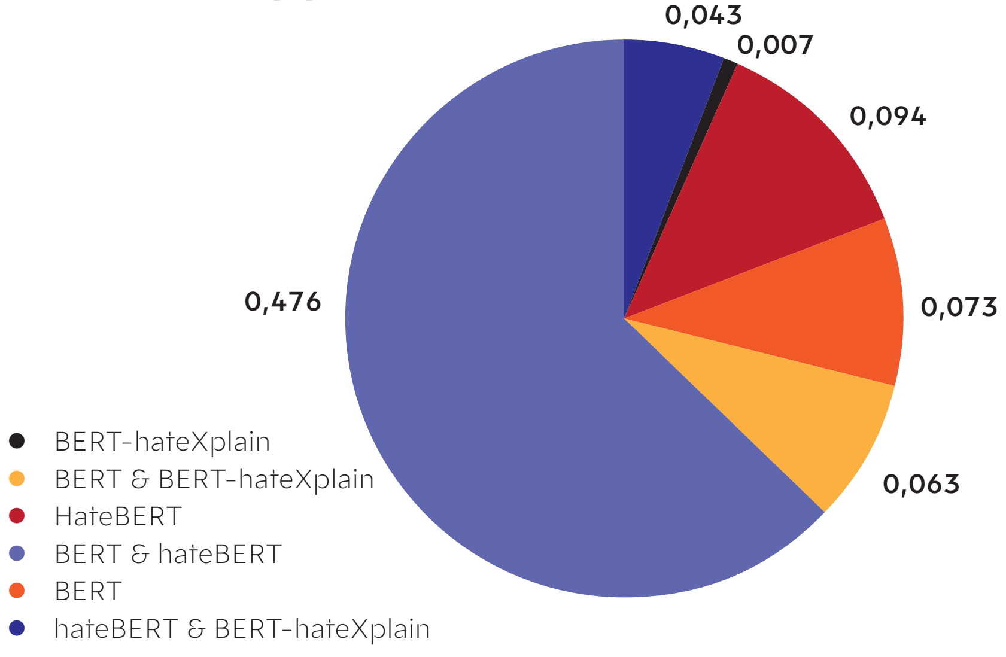
	NOT POSTS	HATE POSTS	OFFN POSTS
English training subset	4042	1443	667
English test subset	958	124	71

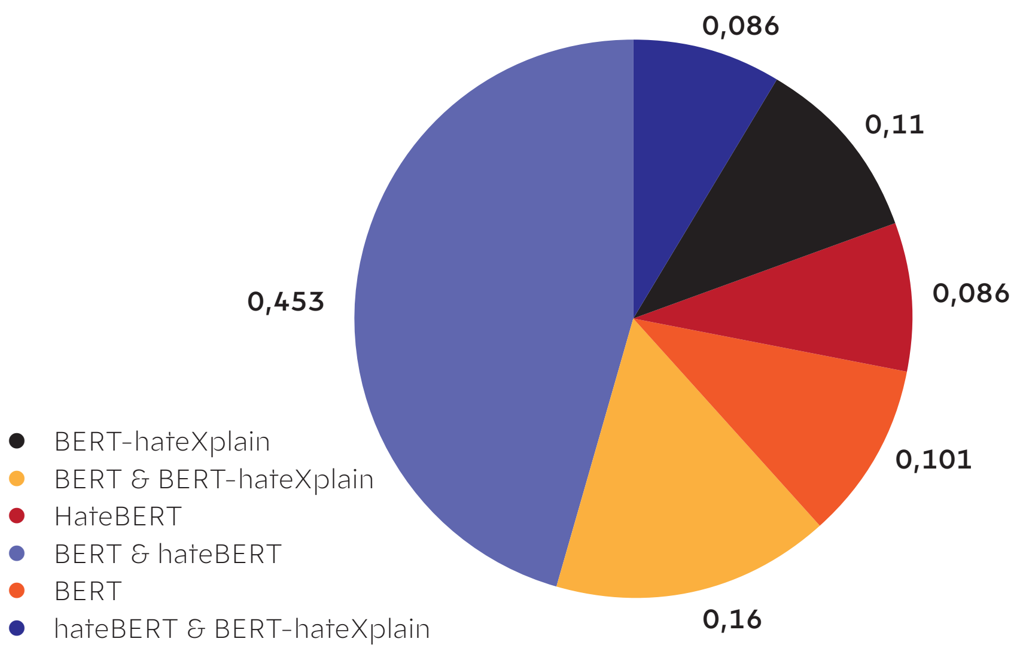
RESULTS:

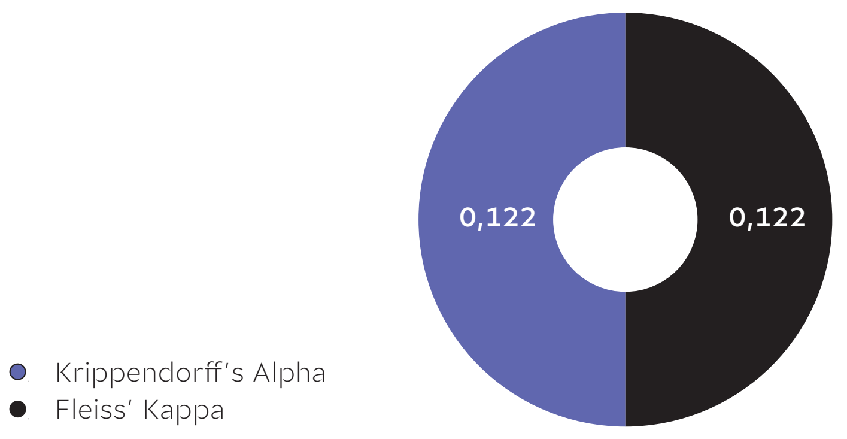
English training subset

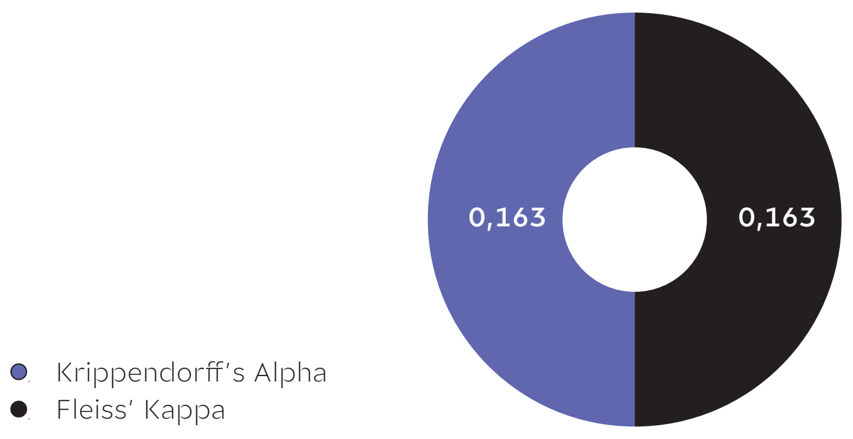
English test subset

Cohen's Kappa

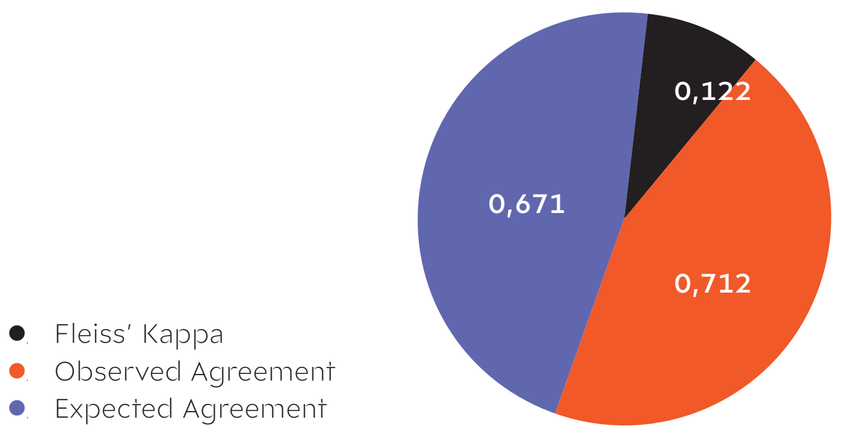


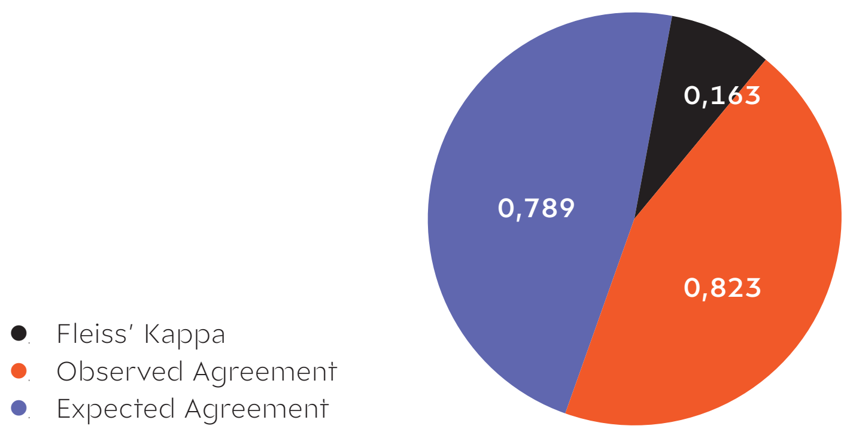




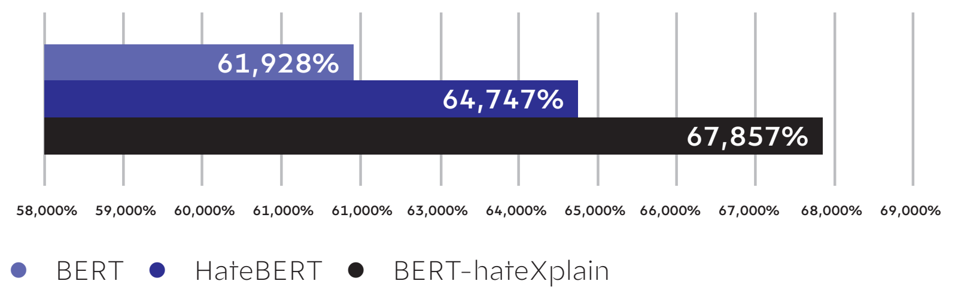


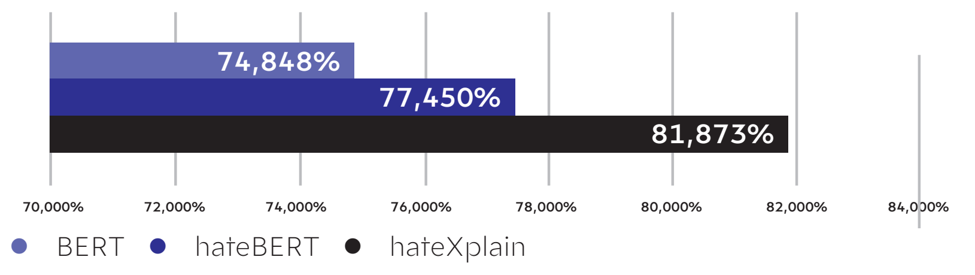
Fleiss' Kappa

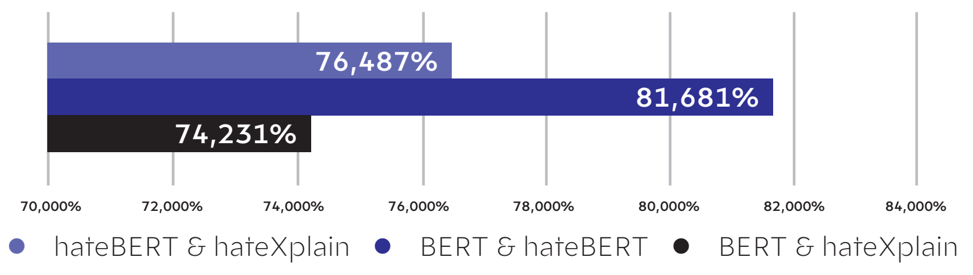


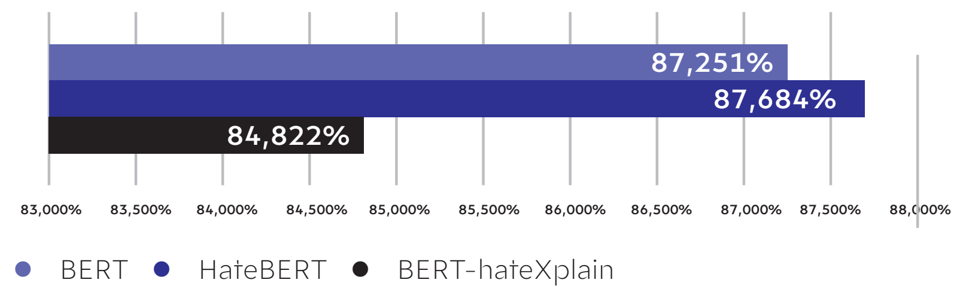


Accuracy









Average Pairwise Percent Agreement

Average Pairwise Percent Agreement	BERT-HateXplain	HateBERT	BERT	BERT & HateXplain	BERT & HateBERT	HateBERT & HateXplain
71.155%	67.857%	64.747%	61.928%	74.231%	81.681%	76.487%

Average Pairwise Percent Agreement	BERT-HateXplain	HateBERT	BERT	BERT & HateXplain	BERT & HateBERT	HateBERT & HateXplain
82.321%	81.873%	77.450%	74.848%	84.822%	87.684%	87.251%

METHODS & EXPERIMENTAL SETUP

Inter-annotator agreement:

- **Linguistics**: To evaluate the reliability of an annotation process
- **Our experiment**: To evaluate how the selected models "agree" in terms of annotation of hate speech instances
- Selected metrics:
  - Pairwise Cohen's Kappa
  - Fleiss' Kappa
  - Krippendorff's Alpha

Selected hate speech detection models for comparison:

- BERT-HateXplain
- HateBERT
- BERT

FUTURE PLANS

Our future plans include:

- **Experiments with different corpora and languages**
- **Experiments with higher variety of hate speech detection models**
- **Additional evaluation methods & metrics**

AUTHORS:

Milita Songailaitė
milita.songailaite@stud.vdu.lt

Eglė Kankevičiūtė
egle.kankeviciute@stud.vdu.lt

Justina Mandravickaitė
justina.mandravickaite@vdu.lt

Danguolė Kalinauskaitė
danguole.kalinauskaite@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

CARD
CENTRE
FOR APPLIED
RESEARCH
AND
DEVELOPMENT