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

Cross-lingual offensive language identification

Andrej Hafner, Urša Zrimšek

Abstract

Keywords

Advisors: Slavko Žitnik

Introduction

Datasets

1. SemEval2020 - multiple classes (hate/no hate, type of hate) - most promising
2. Aggresion identification - agressive/covertly aggresive/not aggresive [1]
3. FoxNews comments - 1528 annotated comments, 435 of them are labeled as hateful and 1093 of them as non-hateful.
4. Reddit and Gab comments - hate/not hate, together with responses
5. Slovene - 24ur, (also Nova24tv?) comments. We will scrape the comments of news from archive, together with their titles, up/downvotes and responses.

We would ideally need more datasets that are annotated with multiple classes of offensive language.

Related work

- Identifying and Categorizing Offensive Language in Social Media (OffensEval)
- Identification of Offensive Language in Social Media
- A Cross-Lingual Augmentation Approach for Multilingual Offensive Language Identification
- Predicting the Type and Target of Offensive Posts in Social Media
- Detecting Online Hate Speech Using Context Aware Models
- Medjezikovni pristop k abstraktivnemu povzemanju: [2]
- Medjezikovni prenos napovednih modelov za sovražni govor

Methods

Machine translation and classification using fine-tuned BERT model

The idea is to finetune BERT model (pretrained on English data) to classify paragraphs into multiple classes of offensive language. The Slovenian dataset would be translated into English using the Google Translate API and we would test our trained model on the results.

Finetuning on English data and transfer learning into Slovene language with multi-lingual pre-trained BERT

We would finetune the CroSloEn BERT model on English datasets to classify paragraphs into multiple classes of offensive language. Then we would test the model on Slovenian dataset.

Finetuning on English data using BERT and vector space alignment with BERT for Slovenian language

The idea is to use two monolingual BERT models for English and Slovenian language. The English model would be fine-tuned on English datasets for classification and the Slovenian one would be used to create embeddings. Then we would try to create mappings between the embedding spaces of both models. This exploits the similarities between the vector spaces of multiple monolingual models, as the words in different languages have similar relations between them. We would try two methods for creating the mappings. The first one, created by Xing et al. [3], is as supervised learning method

which enforces an orthogonality constraint on the mapping matrix. The second one, created by Conneau et al. [4], utilizes domain-adversarial approach, which is an unsupervised learning method. Paragraphs would be embedded using the Slovenian models, mapped to the space of the English model and then classified.

Clustering comparison of two aligned monolingual models

We would extend the space-alignment approaches proposed in the previous subsection, this time employing different clustering algorithms. Labeled data of the English dataset would be clustered and ideally clusters of each class would be recognized. We could visualize them using a high dimensional visualization method like t-SNE [5]. We would map the embeddings of Slovenian paragraphs into the space of English model, then check to which cluster the aligned embedding is closest to. It would be interesting to see the visualization of labeled Slovenian data, aligned and classified in the space of the English model.

Additional ideas

After quick examination of some articles on 24ur, we saw that hate speech comments are upvoted, and positive ones downvoted. So it could be interesting to add (after determining the type of offensive language on Slovene data), an analysis of how people react to different types of comments by looking at the up/downvotes. We could also check if people respond more to offensive language, and if there is more offensive replies to already offensive original comment, or to neutral ones.

References

[1] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France, May 2020. European Language Resources Association (ELRA).

[2] Aleš Žagar and Marko Robnik-Šikonja. Cross-lingual approach to abstractive summarization. *arXiv preprint arXiv:2012.04307*, 2020.

[3] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, 2015.

[4] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.