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

Cross-lingual offensive language identification

Maj Šavli, Blaž Rupnik and Leon Premk

Abstract

The abstract goes here.

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

Since the outburst of social media, freedom of speech has allowed anyone to share their opinion on internet. While that allows people to make a change, it can also have negative consequences. Offensive language or hate speech has become a constant on online forums [1]. Best definition of online hate speech we can use are hateful messages (posts on social platforms, comments on news articles) directed against an individual or a group of individuals based on their identity. Because of these messages the group can be viewed as undesirable which warrants hostility towards them.

That's why automatic offensive language detection is highly required task. Some solutions for hate speech detection already exist but most are in english. We wish to train state-of-the-art models such as mBERT and XLM-R on english datasets ([2, 3, 4, 5, 6]) and later transfer our models to Slovenian language.

Related work

With rapid growth of information on internet, automatic tools for detecting hate speech are in huge demand. Earlier implementations of offensive language detection were based on basic machine learning classifiers such as naive bayes and SVM. By increasing hardware capabilities in recent years deep learning methods became became the new state-of-the-art outperforming previous methods by large margin.

Pitenis et al. [7] used deep learning methods to detect offensive language in Greek Tweeter posts. In another work Rizwan et al. [8] proposed their Convolutional Neural Network n-gram to detect hate speech on dataset containing Roman Urdu tweets. Ranasinghe et al. [9] used different state-of-the-art natural language processing methods such as BERT and XLM to detect offensive language in Bengali, Hindi and Spanish social media posts. In OffensEval 2020 [10] competitors were detecting offensive language, categorizing it based on offense type and identifying toward whom offense was targeted. Datasets were in English, Arabic, Danish, Greek and Turkish language. Most teams used pre-trained Transformers such as BERT [11] and it's variations like RoBERTa [12], or AL-BERT [13]. Other Transformers, most notably GPT-2 [14], were also used for classification. Word embeddings were mostly done by BERT or RoBERTa and BERT's multilingual variant mBERT [11].

Methods

Data

English dataset consist of five datasets from different sources. They mainly consist of social media posts on Twitter [5], Reddit and Gab [3] aswell as Wikipedia posts [2], news articles [4] and forum Stormfront posts [6]. Lots of acquired data is politicaly oriented, which can be benefitial since majority of Slovenian offensive language has political base.

One part of the Slovene dataset was acquired from scraping a slovenian news platform, 24UR. We extracted user comments of various articles. To achieve better variance in data, we made sure the scraped comments belonged to articles of various themes. There were a lot of emoticons present in the comments. Since they could negatively impact on the learning of the algorithms, we removed them.

We also included data from social platforms. Our main goal was to use extract comments from posts on Facebook where a lot of hate speech can be found. For this purpose we create a facebook developer app with which we can use their Graph API. It turns out that they recently made privacy changes in latest versions so for acquiring such data from public pages through their API we would require read access for

the page given by manual approval. Because of this obstacle we decided to include a simple posts scraper for retrieving data from Facebook.

References

[1] Konrad Rudnicki and Stefan Steiger. Online hate speech - introduction into motivational causes, effects and regulatory contexts. 08 2020.

[2] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[3] 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.

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

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[6] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

[7] Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*, 2020.

[8] Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. Hate-speech and offensive language detection in roman urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, 2020.

[9] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*, 2020.

[10] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.