Toxic comment classification with BERT-based models in Hungarian

Péter Hatvani

(PPKE Doctoral School of Linguistics, HUN-REN Hungarian Research Center for Linguistics)

January 2024

Keywords: Toxicity Detection, BERT Classifiers, Natural Language Processing (NLP), Online Moderation, Language Model,

1 Introduction

In the digital era, online platforms have become central to social interaction, but they also face significant challenges due to toxic discourse as much as their users. This study addresses the urgent need for effective toxicity detection in Hungarian online spaces, drawing inspiration from the capabilities of Toxic-BERT [2]. Since the Multilingual Jigsaw Competition [3] toxicity classification halted, as the most used languages were covered. This research focuses on developing classifiers that identify toxic content on par with Toxic-BERT for a low resource language, such as Hungarian.

2 Corpus

Firstly, a small corpus $N_{\text{items}} = 655, \quad N_{\text{tokens}} = 14,207$ of toxic comments were created based on the categories of the competition. The only difference is the omission of the category 'severely toxic' for simplicity's sake. Three annotators made judgements with a 'slight-agreement' according to Randolph's $\kappa_{\text{free}} = 0.525$ [6] which means that deciding on the toxicity of a text is a difficult task. The comments were collected from news sites' (mandiner.hu, kuruc.info) comment sections and social media interactions from (reddit.com, napiszar.com) to balance out the political and personal topics.

Model	Step	Training Loss	Validation Loss	F1 Score
HuBERT	140	0.317000	0.325885	0.873582
mBERT	380	0.593200	0.490828	0.790007
BERT-60k	500	0.501400	0.466688	0.788360

Table 1: Training and Validation Losses and F1 Scores of Different Models

3 Modelling

The classification was performed via 3 contextual language model trained on Hungarian data. On the collected corpus and the Hungarian twitter sentiment corpus[1] three BERT classifiers were trained: huBERT[5], mBERT[1] and a BERT-like model trained from scratch - denoted as BERT-60k meaning the steps of pretraining. The three models were trained to equilibrium with different training steps (cf. Table 1). The corpus will be available on github and the models on huggingface. To meet the recommended minimum of 10'000 items for fine tuning I have used the neutral and positive annotated items from the twitter sentiment corpus and those items were also augmented with nlpaug[4] as the toxic comments. The augmentation prepared the models for the constant typographical errors found in comments. I have employed the masking replacement to further harden the models' judgement and make them more robust with limited training data. For the training I have used 80/20 splits of training / validation. The fine tuning dataset consisted of 27'676 items. The twitter sentiment corpus was only augmented 5 times with 10% chance of any word being misspelled (17'226) all together. The toxic part was augmented ten folds with typography (6550) and five time with masking (3275).

After training, a random set of 40 items were selected from the validation set to perform a pairwise confirmation with Cohen's kappa as can be seen on Table 2. The guesses of the models are very close to the averaged annotator judgement.

[1] https://opendata.hu/dataset/hungarian-twitter-sentiment-corpus - property of Precognox Ltd.

Model	Cohen's Kappa
mBERT	0.7930
BERT-60k	0.6891
huBERT	0.6875

Table 2: Cohen's Kappa for Model Validation with Annotator Averaged Results

4 Conclusion

The models show promise for application as they are the only Hungarian toxicity classifier and they present high accuracy, (0.87358) the fine tuned huBERT's f1 score. With more training data and better preprocessing an even higher accuracy can be achieved.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[2] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify. 2020.

[3] Ian Kivlichan et al. Jigsaw Multilingual Toxic Comment Classification. 2020. URL: https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification.

[4] Edward Ma. NLP Augmentation. https://github.com/makcedward/nlpaug. 2019.

[5] Dávid Márk Nemeskey. "Introducing huBERT". In: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021). Szeged, 2021, TBA.

[6] Justus J. Randolph. "Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa". In: Joensuu Learning and Instruction Symposium. Vol. 2005. 2005. URL: https://eric.ed.gov/?id=ED490661.