International Conference on Formal Methods
and Foundations of Artificial Intelligence
Eszterházy Károly Catholic University
Eger, Hungary, June 5–7, 2025

# Automated Detection of Toxic Comments in Hungarian

**Péter Hatvani**[a],[b], **Zijian Győző YANG**[a]

[a]HUN-REN Research Center for Linguistics
(familyname).(givenname)@nytud.hun-ren.hu

[b]PPKE Doctoral School of Linguistics
hatvani.peter@hallgato.ppke.hu

## Abstract

Moderating toxic comments online remains a significant and unresolved challenge. This paper presents the first openly accessible small-scale Hungarian corpus (n=655) for toxic comment classification, together with three fine-tuned Hungarian BERT-based classifiers: huBERT [5], multilingual-BERT [1], and huBERT-Setfit [6]. The corpus comprises social media comments collected from screenshots shared on Reddit and comments from Hungarian news websites, mandiner.hu and kuruc.info. To ensure robustness, training data was augmented using typo simulation and masked token replacement techniques. The best performing classifier, huBERT-SetFit, achieved an F1 score of 93.72%. These classifiers effectively identify offensive content in online discussions in real-world Hungarian.

The corpus was constructed from various sources on the Internet. Social media comments labeled as *social* originated from Reddit's *r/szopjatokle* subreddit and the notorious napiszar.com, featuring interpersonal aggression and general offensive content. Comments in the news category labeled *news* were collected from politically polarized discussion forums of mandiner.hu and kuruc.info, which included ethnically charged remarks and context-dependent offensive content.

The data can be visually analyzed. In this way, it is clear that four groups are present in the toxic data set, as can be observed in 1. Clustering of k-means [3] was achieved with t-SNE [4] with reduced dimensionality.

Data augmentation involved the introduction of typographical errors and masked token substitutions to expand the limited dataset, resulting in 10,185 toxic and 17,266 neutral instances. The classifiers were fine-tuned using standard BERT pa-

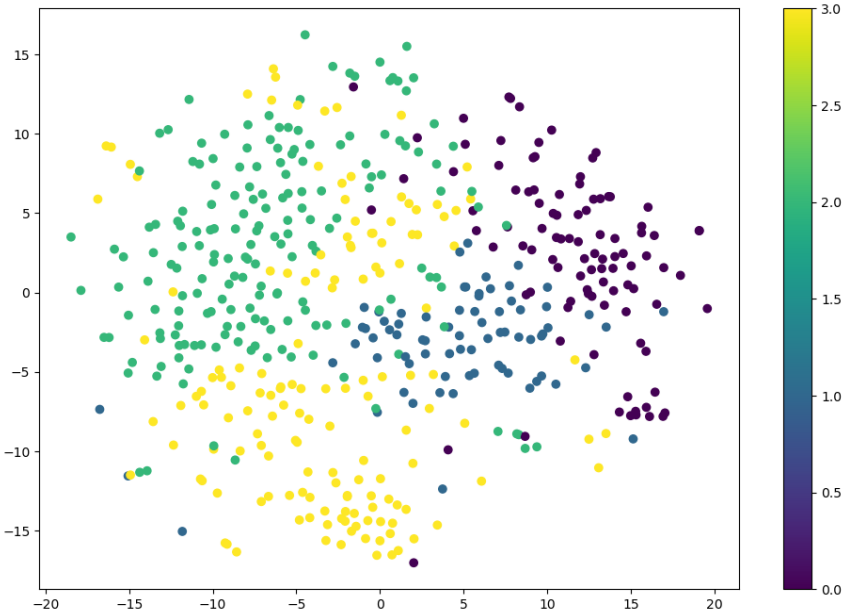


**Figure 1.** K-means clustering of the comments

rameters (learning rate $2 \times 10^{-5}$, batch size 8, weight decay 0.01, precision FP16). The evaluation of the performance of the model demonstrated a robust classification, with huBERT showing superior results in context of the fine-tuned models. The SetFit model, based on the huBERT fine-tuned sentence transformer model [2], is a classifier model that is partly an embedding - from the base model, and partly a logistic regression head. This model achieved comperative results from the data set without any data augmentation. This fact is aligned with the promise of the presenting paper of SetFit, that this method offers results with minimal computational strains.

**Table 1.** Training and Validation Losses and F1 Scores of Different Models

| Model | Epochs | Training Loss | F1 Score |
|---|---|---|---|
| HuBERT | 1 | 0.317000 | 0.873582 |
| mBERT | 3 | 0.593200 | 0.790007 |
| huBERT-embedding-setfit-toxic | 3 | 0.2175 | 0.93725 |

All in all, thee huBERT-Setfit model achieved the highest accuracy, followed by huBERT and mBERT. All classifiers successfully differentiated offensive and neutral content, albeit with minor confidence variations. Future research aims include the expansion of the corpus into other on-line interactions, refined labeling

methodologies, and improving the enhancement of Hungary's specific data.

# References


[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *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 J. Burstein, C. Doran, T. 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] P. Hatvani, Z. G. Yang: *Training Embedding Models for Hungarian*, in: Proceedings of the 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS), Debrecen: University of Debrecen, 2024, pp. 75–80, isbn: 9798350387889.

[3] S. Lloyd: *Least squares quantization in PCM*, IEEE transactions on information theory 28.2 (1982), pp. 129–137.

[4] L. van der Maaten, G. Hinton: *Visualizing High-Dimensional Data Using t-SNE*, Journal of Machine Learning Research 9.Nov (2008), pp. 2579–2605.

[5] D. M. Nemeskey: *Introducing huBERT*, in: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021), Szeged, 2021, TBA.

[6] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg: *Efficient Few-Shot Learning Without Prompts*, 2022, doi: 10.48550/ARXIV.2209.11055, url: https://arxiv.org/abs/2209.11055.