Annales Mathematicae et Informaticae
61 (2025) pp. 108–117
DOI: 10.33039/ami.2025.10.007
URL: https://ami.uni-eszterhazy.hu

Automated detection of toxic comments in Hungarian

Péter Hatvani[ab], **Zijian Győző Yang**[a]

[a]ELTE Research Centre for Linguistics
yang.zijian.gyozo@nytud.elte.hu

[b]Pázmány Péter Catholic University
Doctoral School of Linguistics
hatvani.peter@hallgato.ppke.hu

Abstract. Moderating toxic online comments in Hungarian remains a challenging NLP task. We introduce the first openly available Hungarian corpus for toxic comment classification, though limited in size (n = 655), sourced from social media and political news forums. We fine-tuned three BERT-based classifiers (huBERT, multilingual BERT, and huBERT-SetFit) and applied data augmentation techniques to expand the training dataset. The best-performing model, huBERT-SetFit, achieved an F1 score of 93.7%. Our results demonstrate the effectiveness of transformer-based models for toxicity detection in low-resource, linguistically complex settings.

Keywords: toxicity, online hate, nlp, classification, logistic regression

AMS Subject Classification: 68T07, 68T30, 68T50, 91F20

1. Introduction

In the digital era, online platforms have become central to social interaction, yet they are frequently plagued by toxic discourse that undermines constructive engagement. Toxic comments can include hate speech, threats, personal insults, and discriminatory remarks, posing challenges to both platform moderation and user safety. Although large-scale datasets and models exist for high-resource languages such as English, many low-resource languages such as Hungarian remain underrepresented in this domain.

This paper addresses the critical need for effective toxicity detection in Hungarian by introducing a novel, publicly available corpus containing both toxic and

Accepted: October 8, 2025
Published online: October 28, 2025

neutral Hungarian comments. Additionally, we evaluate the performance of three fine-tuned transformer-based classifiers – HuBERT, multilingual BERT (mBERT), and huBERT-SetFit – on this dataset.

Drawing inspiration from previous efforts such as detoxify [3] and the Jigsaw Multilingual Toxicity Classification competition [5], this research aims to bring similar capabilities to the Hungarian language. Our experiments show that sentence embedding-based methods like SetFit can outperform traditional fine-tuning in low-data regimes, providing a viable solution for toxicity detection in under-resourced languages.

2. Related works

The detection of toxic comments has gained increasing attention with the growth of user-generated content. One of the most influential initiatives in this space was led by the Jigsaw/Conversation AI team, which introduced large-scale English-language datasets for toxicity classification on platforms such as Wikipedia talk pages [5]. These datasets provided multilabel annotations for categories such as: Toxic, Severe toxic, Obscene, Threat, Insult, Identity hate.

Several benchmark models such as Toxic-BERT [3], based on pretrained transformer architectures, have since been developed and are widely used in high-resource English and multilingual contexts. However, these models often perform inadequately in underrepresented languages due to data scarcity and linguistic differences.

Multilingual BERT [2] and XLM-R [1] offer some generalization to low-resource languages, but studies show that language-specific models like huBERT [6] can outperform them in Hungarian-specific tasks.

SetFit [8] introduced a new paradigm by combining sentence embeddings with lightweight classification heads, enabling few-shot learning with minimal labeled data. Its efficiency and performance in low-resource settings make it particularly suited for tasks such as toxicity detection in Hungarian.

Despite these advances, there remains a lack of open-domain Hungarian datasets and benchmark models for toxic comment classification. Our work contributes to filling this gap by releasing a small but diverse Hungarian dataset and evaluating three transformer-based models on it, including a SetFit variant that requires no data augmentation.

3. Method

To develop an automatic classifier for toxic comment detection, a manually annotated dataset was first created. The training corpus comprises comments from two distinct sources: (i) offensive social media comments from *Reddit* and *napiszar.com*, and (ii) politically charged discussions from Hungarian news sites *mandiner.hu* and

kuruc.info. The final dataset comprises 655 annotated comments, labeled according to the jigsaw competition categories.

3.1. Annotation categories and examples

We have adopted the multi-label approach from Jigsaw with the 5 scale Likert scale to the same categories as mentioned in the Related Works section. Three annotators made judgements with a "slight-agreement" according to Randolph's $\kappa_{\text{free}} = 0.525$ [7]. The annotators were not in the same age group neither were they all of the same gender. Given the subjective nature of toxicity perception, we expected substantial variation in judgments across items when using annotators from diverse backgrounds. Although judgements were not unanimous, no item exhibited extreme disagreement (with one annotator giving the highest rating and another giving the lowest), so no items were removed from the initial collection.

While the moderate inter-annotator agreement reflects the inherent subjectivity in toxicity perception, this represents a limitation of our dataset that may affect model reliability. Future work should explore consensus-building techniques or expert adjudication to improve annotation consistency.

3.2. Analysis of examples from the corpus

Toxicity can show many forms. In this section, we introduce a few examples from the corpus and analyse the toxic content in each.

(1) *Ezt* *az* *arcot* *láttam* *már* *valahol,* *mintha* *egy* *híg*
this.ACC the face-ACC see-1SG.PAST already somewhere as.if a watery
agyú *propaganda-troll* *lenne.*
brain-POSS propaganda-troll be.COND.3SG

'I've seen this face somewhere before, like some dumb-brained propaganda troll.'

Toxicity Example 1 conveys toxicity through indirect derogation. The speaker insinuates that the person resembles a "dumb-brained propaganda troll," a phrase that implies intellectual inferiority and political manipulation. The use of "mintha" (as if) introduces the insult in a covert, sarcastic way, enhancing its rhetorical impact while allowing for plausible deniability.

(2) *Kárpótol* *addig* *hatalmas* *masztika,* *míg* *nem* *jut* *pina.*
compensate until huge jerk.off while not get pussy

'Just jerk off until you can get some pussy.'

Obscenity This utterance in Example 2 is obscene and explicitly sexual, using vulgar slang for masturbation and female genitalia. It expresses objectification and dehumanization, especially through the reduction of sexual partners to means of "compensation." The aggressive tone and crude lexicon contribute to its toxic nature, targeting both sexual frustration and women.

(3) *Írjon neked a boncmester, jegyzőkönyvet.*
write-IMP.3SG you.DAT the autopsy-master report-ACC

'Let the coroner write your report.'

Threat This sentence in Example 3 implies death in a sarcastic manner by suggesting that a coroner should write the recipient's report. While indirect, it is a menacing wish for harm cloaked in formal-sounding language. The choice of a professional figure related to death underscores the implied violent intent without making an overt threat.

(4) *Attól hogy blokkolod, még mindig igaza van, és te*
from.that that block-DERIV-2SG still always true-POSS.3SG is and you
továbbra is egy idióta vagy.
still a idiot are

'Blocking them doesn't change the fact they're right, and you're still an idiot.'

Insult This comment in Example 4 employs direct personal insult by calling the recipient an "idiot." It also undermines their actions (blocking someone) as ineffective, intensifying the disparagement. The phrasing suggests moral superiority while dismissing the recipient's perspective, reinforcing the toxic tone through condescension and name-calling.

(5) *Bárcsak minden buzit kirúgnának!!!!*
if.only all faggot-ACC out-kick-COND.3SG-PL

'If only all the faggots got fired!!!!'

Identity Hate This example in Example 5 is a clear instance of hate speech targeting a marginalized identity group. The slur "buzit" (faggot) is used with a wish for collective punishment ("get fired"), which constitutes discriminatory rhetoric. The multiple exclamation marks amplify the emotional aggression, highlighting the severity of the hate expressed.

(6) *Úgy-úgy, kis hülye! Csússzál térden előttük. A kurva*
so-so little idiot crawl-IMP.2SG knee-SUP before-3PL the whore
anyádat te agyhalott pondró.
mother-POSS.2SG-ACC you brain-dead maggot

'Yeah right, you little idiot! Crawl on your knees in front of them. Fuck your mother, you brain-dead maggot.'

Severe Toxicity This utterance in Example 6 combines multiple forms of toxicity: insult ("idiot," "brain-dead maggot"), obscenity ("fuck your mother"), and verbal domination ("crawl on your knees"). It escalates through commands and extreme invective. The layered abuse represents severe toxicity, intended to intimidate, degrade, and humiliate the addressee completely.

3.3. Training data preparation

As previously discussed, our manually collected toxic corpus helps capture the nuances of Hungarian cultural context in toxic language. However, given the small size of our original corpus, we supplemented it with the Hungarian Twitter corpus[1] as a source of neutral examples.

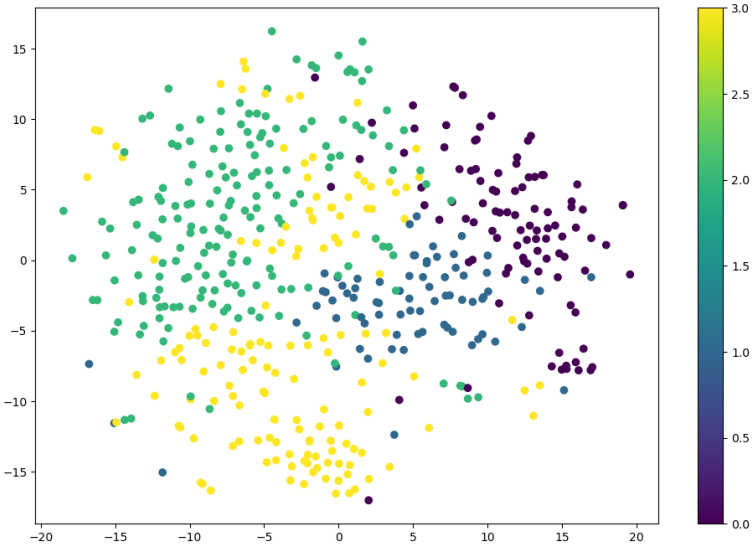

Figure 1. K-means clustering of the comments.

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

Clustering of the toxic corpus To better understand the structure of our toxic comment corpus, we visualized huBERT-generated sentence embeddings using t-SNE dimensionality reduction (Figure 1). We then applied K-means clustering with k=4, motivated by the four distinct data sources: comments from *kuruc.info*, *mandiner.hu*, the *napiszar.com* shock site, and Reddit posts. As shown in Figure 1, the resulting clusters reveal meaningful groupings in the embedded space. Although some overlap exists (common in language data), the clusters are generally well-formed and reflect source-specific patterns in toxic language use. In particular, the Reddit and *napiszar.com* content forms broader, more dispersed clusters, likely due to their conversational and informal nature, while politically charged comments from news sites appear more tightly grouped.

Hungarian Twitter Sentiment Corpus The *Hungarian Twitter Sentiment Corpus* (HTS) is a publicly available collection of approximately 4,000 Hungarian-language tweets annotated for sentiment polarity. The corpus was created by Precognox and made accessible through `opendata.hu`. It includes five sentiment labels on a Likert-like scale, ranging from 1 (very negative) to 5 (very positive), which form the HTS5 variant. A binary version (HTS2) was also derived by grouping positive and negative classes and excluding neutral tweets. In our work, we used HTS5 labels and treated tweets rated 3–5 as *neutral* training data, following a pragmatic interpretation where mid-scale ratings reflect low-intensity sentiment or ambiguous tone. This decision enabled a clearer separation between offensive and non-offensive language in our toxicity classification task. The corpus provides a valuable resource for sentiment modeling in Hungarian, and our adaptation aligns with common practices in low-resource language scenarios where neutral and ambiguous categories are often merged to improve class balance and model performance.

Models trained and evaluated There are a plethora of models available for moderation. Most of them are products of companies that are available for a fee per request or token. To establish a baseline, we have evaluated the models finetuned by us with OpenAI's Moderation API. The API in question is omni-moderation-latest that was available on 2025.07.02. We have evaluated four models collected in the list 3.3 from which we fine-tuned three of the models. The OpenAI endpoint was the baseline for the evaluation for it is widely used moderation tool for AI models currently. The toxic-hubert model was fine-tuned from the HuBERT model [6], same method as for the multilingual BERT. The hubert-embedding-setfit-toxic model was fine-tuned from an embedding model [4] with the SetFit toolset.

- Openai Moderation endpoint (omni-moderation-latest)
- RabidUmarell/toxic-hubert
- RabidUmarell/toxic-mbert
- RabidUmarell/hubert-embedding-setfit-toxic

Models were evaluated using standard classification metrics including precision, recall, F1-score, and accuracy. Statistical significance was assessed using McNemar's test with $\alpha = 0.05$.

Training Data Augmentation Before training, we have applied two data augmentation techniques: Typographical error simulation and masked token replacement to increase the training set size, resulting in 10,185 toxic and 17,266 neutral instances.

Training Each model was trained with standard BERT hyperparameters: learning rate $\eta = 2 \cdot 10^{-5}$, batch size of 8, weight decay of 0.01, and mixed precision training using `fp16`. The SetFit model, on the contrary, did not require data augmentation; it used the huBERT sentence transformer [4] to generate embeddings and used a lightweight logistic regression head for classification. This approach aligns with SetFit's promise of achieving high performance with minimal computational overhead, particularly in low-resource scenarios. A detailed summary of training performance can be seen in Table 1. The HuBERT model was only trained for one epoch because this model reached equilibrium quickly and additional epochs only degraded the performance.

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

4. Results

We evaluated the trained models on a small test dataset (available[2])

Toxicity Classification Accuracy Comparison We evaluated four models (`SetFit Toxic-HuBERT`, `Toxic-HuBERT`, `Toxic-mBERT`, and `OpenAI Moderations API`) across eight manually annotated toxicity categories plus a neutral control condition. The `SetFit Toxic-HuBERT` model achieved consistently strong performance, reaching perfect or near-perfect accuracy in *Hate Speech*, *Threat*, *Obscenity / Profanity*, and *Harassment / Bullying* (100% in all these categories except for a minor drop in *Toxic Generalization*, 75%).

`Toxic-mBERT` closely followed, also achieving 100% accuracy in four toxicity categories and demonstrating solid generalization. Meanwhile, `Toxic-HuBERT` showed

[2] https://huggingface.co/datasets/RabidUmarell/hu-toxic-test-set

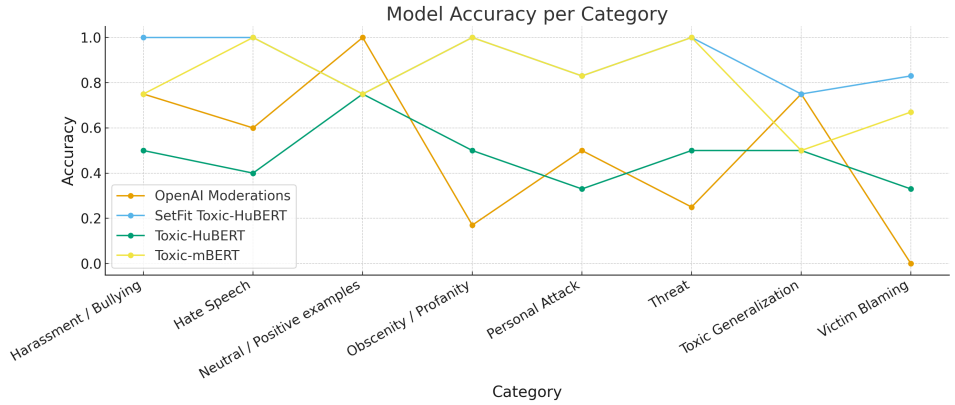

Figure 2. Model accuracy on the test set.

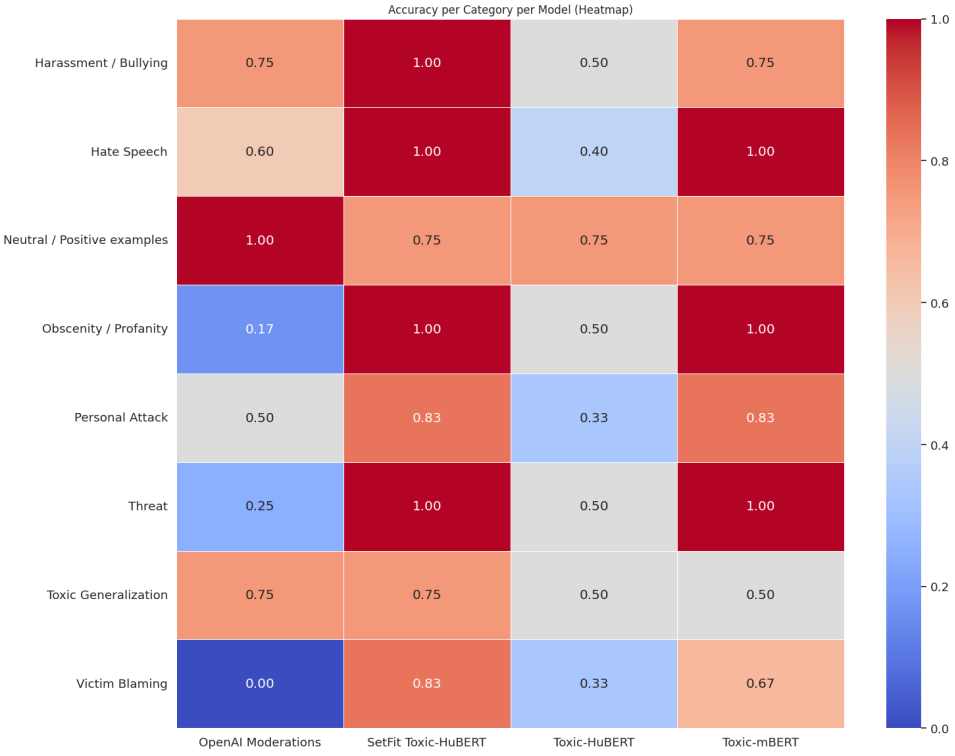

Figure 3. Heatmap of the models' performance on the test set.

weaker overall performance, with accuracy between 33%–75% across all categories, suggesting lower confidence or overfitting. In contrast, the `OpenAI Moderations`

`API` displayed a more varied performance profile: high accuracy in broader categories like *Personal Attack* (50%), *Hate Speech* (60%), and *Harassment / Bullying* (75%), but weaker results in more nuanced cases such as *Victim Blaming* (0%) and *Threat* (25%).

Neutral examples were handled best by `OpenAI Moderations` and `Toxic-HuBERT` (100%), indicating reliable non-toxic classification. Overall, `SetFit Toxic-HuBERT` and `Toxic-mBERT` emerged as the most balanced models, with robust performance in detecting multiple forms of toxicity across languages and expressions.

Statistical significance was assessed using McNemar's test for paired comparisons. The huBERT-SetFit model significantly outperformed both huBERT ($p < 0.01$) and mBERT ($p < 0.05$), while the difference between huBERT and mBERT was not statistically significant ($p = 0.12$).

5. Conclusion

This paper introduced a novel, manually annotated Hungarian dataset for toxic comment classification and presented a comparative evaluation of three transformer-based models – huBERT, mBERT, and SetFit – on this task. Our experiments demonstrate that sentence embedding-based approaches, particularly SetFit combined with huBERT, offer strong and reliable performance across a wide range of toxicity categories, even in low-data conditions. Notably, SetFit achieved high accuracy without requiring data augmentation, confirming its utility in low-resource scenarios.

Our results indicate that while general-purpose multilingual models like mBERT provide reasonable baseline performance, Hungarian-specific models better handle the morphological complexity and cultural nuances of Hungarian toxic language. The huBERT-based SetFit model consistently outperformed traditional fine-tuned counterparts, especially in categories such as obscenity, threat, and identity hate, where subtle linguistic cues play a key role.

Several limitations should be acknowledged. The relatively small dataset size (655 comments) may limit generalizability, and the moderate inter-annotator agreement ($\kappa_{free} = 0.525$) suggests inherent challenges in toxicity annotation. Additionally, the cultural and platform-specific nature of our data sources may not fully represent the diversity of Hungarian toxic language across all digital contexts.

By releasing both the annotated corpus and the model evaluation results, this work contributes a much-needed resource for Hungarian NLP and opens the door to further research on toxicity detection in underrepresented languages. Beyond the immediate task, the dataset and findings also provide a foundation for developing safer and more context-aware content moderation tools in Hungarian digital spaces. In doing so, our work supports broader efforts toward building inclusive, multilingual language technologies that reflect the full diversity of online communication.

References

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov: *Unsupervised Cross-lingual Representation Learning at Scale*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ed. by D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Online: Association for Computational Linguistics, July 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747, url: https://aclanthology.org/2020.acl-main.747/.

[2] 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.

[3] L. Hanu, Unitary team: *Detoxify*, Github. https://github.com/unitaryai/detoxify, 2020.

[4] 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.

[5] I. Kivlichan, J. Sorensen, J. Elliott, L. Vasserman, M. Görner, P. Culliton: *Jigsaw Multilingual Toxic Comment Classification*, 2020, url: https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification.

[6] D. M. Nemeskey: *Introducing huBERT*, in: XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2021), Szeged, 2021, pp. 3–14.

[7] J. 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.

[8] 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.