Hate and Toxic Speech Detection in Slovak: Comparing BPE and Morphology-Based Tokenization Approaches

Dávid Držík and Lívia Kelebercová

Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1,
SK-949 01 Nitra, Slovakia
lkelebercova@ukf.sk
ddrzik@ukf.sk

Abstract. Hate and toxic speech on the internet pose serious social and ethical challenges, yet automatic detection for morphologically rich and low-resource languages such as Slovak remains underexplored. Traditional subword tokenization methods often fail to capture inflectional and derivational nuances essential for understanding harmful language. This study compares two Slovak RoBERTa-based models differing only in tokenization strategy: SK_BPE_BLM using Byte-Pair Encoding and SK_Morph_BLM, a morphology-aware model preserving root morphemes. Both models were fine-tuned and evaluated on the Slovak Hate Speech and ToxicSK datasets using a stratified 10-fold cross-validation protocol across standard classification metrics. The morphology-aware model achieved slightly higher and more stable results in toxic speech detection (+0.5 pp F1-weighted), while both models performed similarly on hate speech. The gains suggest that morpheme-level segmentation can modestly enhance lexical understanding and classification robustness. Morphology-aware tokenization offers limited but consistent benefits for Slovak toxic speech detection. However, its potential remains constrained by informal online language lacking diacritics and standard morphology.

Keywords: Hate speech; Toxic speech; Morphological tokenization; Slovak; Low-resource languages

1 Introduction

The rise of hate ranging from toxic speech to cyberbullying on the internet has created a need for automated detection systems capable of recognizing harmful language [1]. While considerable research [2,3,4] has focused on high-resource languages, progress in smaller and morphologically rich languages like Slovak remains limited. The hate and toxic speech detection could be framed as a classification task, but it naturally requires high quality data [5] which rises the challenge for low-resource languages such as Slovak. The additional challenge for Slovak is its complex inflectional system, relatively free word order, and scarcity of large annotated datasets. These characteristics make it difficult

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 145–154, 2025.

for traditional subword-based language models to fully capture semantic and grammatical nuances relevant to hate speech detection.

In this study, we explore how tokenization strategies influence the performance of language models for hate and toxic speech detection in Slovak. We compare two approaches: a subword-based model using Byte-Pair Encoding (BPE) [6], and a morphology-aware model utilizing a Slovak Morphological Tokenizer (SKMT) [6]. The former represents a widely used, data-driven tokenization technique, while the latter employs a tokenization algorithm that preserves the integrity of the root morpheme within each token. We hypothesize that morphology-aware tokenization could enhance the model's understanding of syntactic and semantic structures relevant to hate expression in Slovak.

To test this hypothesis, we compare two Slovak RoBERTa-based language models that differ only in their tokenization strategy: SK_BPE_BLM (Slovak Byte-Pair Encoding Baby Language Model), which uses standard Byte-Pair Encoding, and SK_Morph_BLM (Slovak Morphological Baby Language Model), which integrates morphology-aware tokenization that preserves root morphemes. Both models were trained and evaluated using a k-fold cross-validation setup on a Slovak dataset annotated for hate and toxic speech. Their performance was assessed using accuracy, precision, recall, and F1-score. The morphology-aware SK_Morph_BLM achieved slightly higher and more stable results than the BPE baseline, particularly in toxic speech detection. This indicates that linguistically informed tokenization can modestly enhance semantic representation in low-resource, morphologically rich languages such as Slovak. These findings contribute to ongoing efforts toward improving language model fairness and robustness beyond high-resource languages.

2 Related Works

Research on hate and toxic speech detection has made substantial progress, especially in high-resource languages such as English. Studies in these contexts typically explore improvements in model robustness, tokenization, and embedding design, offering valuable but not directly transferable insights to morphologically rich and low-resource languages such as Slovak.

Zhao et al. [3] proposed SWE^2, a hybrid framework for English hate speech detection that integrates word-level semantics, character-level, and phonetic subword embeddings to improve robustness against adversarial spelling variations. Their experiments, conducted on several well-known English datasets, demonstrated that BERT-based embeddings substantially outperform FastText representations. While SWE^2 achieved high accuracy and strong resistance to adversarial perturbations, its applicability to low-resource or morphologically rich languages was not examined, leaving questions about cross-lingual generalization and tokenization sensitivity open.

A more linguistically grounded study by Dang et al. [7] investigated how tokenization strategies influence morphological representation in multilingual transformer-based models. They compared subword-based and character-level

models across 17 languages. The findings showed that morphological information is primarily encoded in middle and late layers of the models and that character-level tokenization captures fine-grained morphology more effectively in inflectional languages. The authors suggest that, for languages such as Slovak, morphology-aware tokenization may outperform subword-based approaches like BPE, which tend to obscure grammatical information through arbitrary subword segmentation. However, their analysis focused on probing tasks rather than downstream classification, leaving its practical implications for hate speech detection untested.

Speaking of inflectional low-resource languages, Mikaberidze et al. [2] performed a systematic comparison of tokenization strategies for Georgian, another morphologically complex, low-resource language. The authors tested WordPiece, BPE, SentencePiece-Unigram, and a token-free approach (ByT5). Their results demonstrated that tokenization granularity significantly influences model performance: finer-grained tokenization with smaller vocabularies performed better on toxicity detection tasks, likely because it captures morpheme-level cues linked to offensive word formation. For morphologically rich, low-resource languages, language-specific subword tokenization is therefore optimal.

Similarly, in the Slavic language domain, Jokić et al. [8] investigated abusive speech detection in Serbian using both traditional machine learning and transformer-based models. The authors evaluated algorithms ranging from Logistic Regression and SVM to deep learning architectures in binary classification (abusive vs. non-abusive). Taking a closer look at language models, the authors used several pretrained transformers such as BERTić, BERTić FRANK Hate, XLM-T, Multilingual BERT (cased), Multilingual DistilBERT, XLM-R, Jerteh-81, and SRoBERTa-F. Although tokenization comparison was not the focus of their study, among the tested models, WordPiece tokenization (used by BERTić) achieved the best results for Serbian hate-speech detection with an F1-score of 0.827. BPE tokenization (SRoBERTa-F) performed slightly worse, while SentencePiece (XLM-T, XLM-R) lagged behind due to weaker alignment with Serbian morphology.

Kocijan et al. [9] presented a rule-based approach for Croatian hate-speech detection using NooJ grammars and a linguistically enriched dictionary tailored to pejorative verbs, nouns, and adjectives. Their results show that linguistically informed rule systems can yield precise detections in a low-resource Slavic language.

Crucially for the Slovak context, Papcunová et al. [10] established a foundational framework for hate speech operationalization in Central Europe, providing a detailed, multi-disciplinary list of indicators and their structural relationships. Building on this operational definition, subsequent computational work explored the semi-automatic recognition of hate speech in Slovak texts [11]. While this work provided initial performance benchmarks for the task, it relied on standard subword tokenization, leaving the specific impact of dedicated morphology-aware tokenization unexplored.

Collectively, these works emphasize that tokenization strategy plays a critical role in hate and toxic speech detection, particularly for morphologically rich, low-resource languages. While English-based frameworks have achieved strong results through advanced embedding and hybrid modeling, studies in Slavic and related languages reveal that linguistically informed tokenization can substantially improve model understanding of inflectional and derivational patterns. Building on these insights, our work focuses on Slovak, a highly inflectional West Slavic language, and compares a standard BPE-based model with a morphology-aware Slovak Morphological Tokenizer (SKMT) [6]. This comparison directly addresses a gap in the current literature, where morphological segmentation remains underexplored in the context of hate and toxic speech detection for low-resource Slavic languages.

3 Methodology

The aim of this study is to compare two Slovak language models that differ solely in their tokenization approach, in order to assess whether morphological tokenization enhances classification performance in hate speech and toxic speech detection tasks. Specifically, we compare the SK_BPE_BLM model, which employs subword Byte Pair Encoding (BPE) tokenization, with SK_Morph_BLM, which integrates BPE with morphological tokenization that preserves root morphemes as indivisible units within tokens [6]. This design allows the model to retain essential lexical semantics. The practical part of the methodology consists of four main stages:

1. **Dataset preparation:** Exploratory analysis, cleaning, and preprocessing of Slovak datasets containing hate and toxic speech.
2. **Text tokenization and input sequence generation:** Preparing textual inputs for the SK_BPE_BLM and SK_Morph_BLM models.
3. **Models fine-tuning:** Adapting both models for classification using 10-fold stratified cross-validation.
4. **Performance evaluation:** Comparing model results on validation sets using standard metrics (accuracy, precision, recall, and F1-score).

3.1 Datasets and Preprocessing

Two publicly available Slovak datasets were used for the experiments: the Slovak Hate Speech and Offensive Language Database [12] and ToxicSK [13], both created at the Technical University of Košice. These represent the most comprehensive Slovak datasets for hate and toxic speech detection, providing valuable resources for a low-resource language.

The Slovak Hate Speech and Offensive Language Database contains manually annotated posts from online discussions on sports, politics, and general topics, labeled as 1 (hateful) or 0 (neutral). ToxicSK includes 8,840 annotated Facebook comments, evenly balanced between toxic (4,420) and non-toxic (4,420)

classes. The comments are short and informal, averaging 9.4 words, reflecting typical social media discourse.

Prior to modeling, both datasets were merged (train and test splits combined) and subjected to preprocessing aimed at removing noise and ensuring typographic and linguistic consistency. In this context, noise refers to non-linguistic artefacts or irregularities that do not contribute to semantic interpretation—for example, automatically generated user identifiers (e.g., `user123`), hyperlinks, repeated or malformed character sequences, and duplicated comments with inconsistent labels. The preprocessing pipeline included removal of such identifiers, excessive punctuation and emoticon sequences used outside a syntactic context (e.g., "!!!!!"), normalization of diacritics and quotation marks, unification of typographic variants (e.g., hyphens, ellipses), lowercasing, and deletion of extremely short or non-Slovak entries detected by automatic language identification.

After preprocessing, the hate speech dataset contained 9,401 records (29% hateful), and the toxic speech dataset contained 6,879 records (46% toxic). Both corpora exhibit mild class imbalance, typical for hate speech detection tasks, but the use of multiple evaluation metrics mitigates its effect on comparability and objectivity.

3.2 Used Language Models

Two Slovak language models based on the RoBERTa architecture were used: SK_BPE_BLM and SK_Morph_BLM [6]. Previous research demonstrated that the morphologically informed tokenization of SK_Morph_BLM preserves word-structure information more effectively than standard BPE, yielding superior performance in sentiment-analysis and semantic-similarity tasks. In this study, both models are evaluated on a different task – hate and toxic speech classification – to determine whether this advantage extends to socially sensitive text domains.

Both models share identical architecture and training settings, differing only in tokenization strategy. SK_BPE_BLM applies standard BPE tokenization, while SK_Morph_BLM combines BPE with a morphological lexicon of roughly one million annotated Slovak word forms, ensuring that frequent root morphemes remain indivisible units within tokens. This lexicon, developed as part of the Slovak Morphological Tokenizer [6], integrates lemmas and inflectional patterns derived from publicly available lexical resources and corpus-based annotations. Such morphology-aware tokenization preserves the internal structure of Slovak words and allows the model to maintain semantic and grammatical relationships across inflected forms. However, informal or abusive language in online comments often diverges from standard morphology – users frequently omit diacritics, merge or shorten words, or use creative spelling. These non-standard variants are not always decomposed correctly with respect to their root morphemes, which limits the effectiveness of morphological segmentation in such contexts.

Both tokenizers were trained on the cleaned OSCAR 2019 Slovak corpus (4.2 GB) with a 50,264 vocabulary and minimum frequency 2. The models were further pre-trained on a 455 MB subset. Each model uses 6 encoder layers, 12 attention heads, 576-dimensional vectors, and a maximum sequence length of 256 tokens. Pre-training employed AdamW with a learning rate of 1×10^{-4}, weight decay 0.01, dropout 0.1, and lasted 30 epochs on an NVIDIA A100 (40 GB) GPU. Each model contains approximately 58 million parameters.

3.3 Models Fine-tuning

For fine-tuning, two separate binary classification tasks were conducted—one for hate speech and one for toxic speech. Preprocessed texts were tokenized with the respective tokenizers and converted into input IDs, attention masks, and labels.

To ensure reliable evaluation, we applied stratified 10-fold cross-validation as the experimental evaluation protocol, not as a hyperparameter tuning stage. In each of the ten iterations, the model was re-initialized and fine-tuned from the pretrained checkpoint on nine folds (training set) and evaluated on the remaining fold (validation set). This approach allowed us to assess the stability and generalization ability of the models across different validation subsets.

Each model was fine-tuned for 10 epochs per fold using the AdamW optimizer (learning rate $= 5 \times 10^{-6}$, weight decay $= 0.1$, dropout $= 0.1$) and a batch size of 32. Random seeds were fixed across all libraries (seed $= 42$) to ensure full reproducibility of all training and evaluation runs. Model performance was monitored after every epoch using accuracy, precision, recall, and F1-score (weighted and per class). The epoch achieving the highest weighted F1-score within a fold was retained as that fold's representative result. Final metrics were averaged across all ten folds to obtain the overall model performance. All experiments were executed on a single NVIDIA A100 (40 GB) GPU.

3.4 Research Assumptions

Two main hypotheses were formulated for the experiment:

1. **The morphologically informed model will achieve higher classification performance.** It is expected that SK_Morph_BLM will outperform SK_BPE_BLM, as preserving root morphemes allows more accurate modeling of semantic and grammatical relations in Slovak.
2. **Higher performance will be obtained on the toxic speech dataset.** Both models are anticipated to perform better on ToxicSK than on the Slovak Hate Speech dataset, due to its more balanced class distribution and lower data noise, which enable more stable model learning.

4 Results

Hypothesis 1: On the Hate Speech dataset, both models achieve nearly identical performance across all evaluated metrics, with differences below 0.1%. The

Table 1: Average performance of the models across the Hate Speech and ToxicSK datasets.

Model	Dataset	Accuracy	Prec_w	Rec_w	F1_w	Prec(1)	Rec(1)	F1(1)
SK_BPE	Hate Speech	0.760	0.751	0.760	0.753	0.611	0.507	0.553
SK_Morph	Hate Speech	0.760	0.751	0.760	0.753	0.610	0.510	0.554
SK_BPE	Toxic Speech	0.891	0.892	0.891	0.891	0.876	0.890	0.883
SK_Morph	Toxic Speech	0.896	0.896	0.896	0.896	0.885	0.891	0.888

SK_Morph_BLM model (hereafter referred to as the *Morph* model) shows a marginal improvement in recall and F1-score for the hateful class (class 1), indicating a slightly higher sensitivity to hate expressions. However, both models find this class challenging, achieving recall ≈ 0.51 and F1 ≈ 0.55. These results suggest that the effects of tokenization are largely dominated by dataset characteristics – particularly class imbalance ($\approx$ 29% hateful instances) and high linguistic variability within hateful expressions. While transformer-based architectures such as BERT are generally robust to moderate imbalance, the combination of relatively few hateful samples and their lexical diversity likely reduced the model's ability to generalize effectively.

On the Toxic Speech dataset, the Morph model consistently outperforms the SK_BPE_BLM baseline (hereafter referred to as the *BPE* model) across all metrics, though the improvements are modest. Gains include +0.5 pp in accuracy and weighted F1, +0.9 pp in precision (class 1), and +0.6 pp in F1 (class 1). This consistent pattern implies that morphology-aware tokenization provides a small but measurable advantage by preserving root-based lexical information and inflectional patterns relevant to the formation of offensive words. The Morph model's ability to retain morphemic integrity appears to slightly enhance robustness in recognizing toxic language.

Hypothesis 2: As anticipated, both models perform substantially better on the ToxicSK dataset than on the Slovak Hate Speech corpus. Accuracy and weighted F1 increase by more than 13 percentage points on average across models (from $\approx$ 0.76 to $\approx$ 0.89). This improvement is primarily due to the balanced class distribution (50/50) and the shorter, more homogeneous social media comments in ToxicSK, which facilitate more stable model training. In contrast, the Hate Speech corpus is considerably noisier and more context-dependent. It contains a high proportion of anonymized user identifiers (e.g., user1234), long thread-like political discussions, mixed-topic comments, and irregular use of diacritics or punctuation. Many hateful expressions are implicit, relying on irony or allusion rather than explicit slurs, which makes them harder to detect automatically. These factors jointly introduce lexical and contextual noise that complicates the model's ability to form consistent representations of hate speech.

The averaged cross-validation results (aggregated across all ten fine-tuning runs) for both models are presented in Table 1.

These results partially support Hypothesis 1, showing that morphological tokenization has a minor yet consistent effect only in the toxic speech domain, while offering no clear advantage for hate speech classification. Hypothesis 2 is fully confirmed, as both models exhibit substantially higher performance on the balanced ToxicSK dataset.

Overall, the results demonstrate that morphological tokenization yields a small but consistent improvement in toxic speech detection, while its impact on hate speech classification remains minimal due to data-specific limitations. The potential of the Morph model is not yet fully realized, as hateful and toxic online comments are often written in non-standard Slovak, frequently omitting diacritics or using irregular spellings. Such orthographic variation limits the effectiveness of morphological segmentation, since words without diacritics cannot always be decomposed correctly with respect to their root morphemes.

5 Discussion

Since both models were fine-tuned independently in each fold under the same cross-validation protocol, the reported averages reliably reflect model behavior across multiple data partitions rather than results from a single training run.

The findings of this study align with and extend previous research on the influence of tokenization strategies in hate and toxic speech detection for morphologically rich, low-resource languages. The comparison between SK_BPE_BLM and SK_Morph_BLM demonstrates that morphology-aware tokenization provides a modest but consistent performance improvement, particularly in toxic speech detection. These results confirm earlier claims that tokenization granularity and linguistic awareness are critical for downstream text classification in complex languages.

The small yet consistent performance gain of SK_Morph_BLM on the ToxicSK dataset supports the observations made by Dang et al. [7] and Mikaberidze et al. [2], who argued that morphology-sensitive segmentation enhances a model's ability to capture inflectional and derivational patterns.

However, the limited improvement on the Hate Speech dataset indicates that tokenization benefits may be constrained by data characteristics. The hate speech corpus used in this study is noisier and exhibits strong class imbalance, leading to performance saturation regardless of tokenization type. This echoes Zhao et al. [3], who noted that dataset quality and class balance often dominate model performance in hate speech detection. The morphological model's advantage thus appears more pronounced in cleaner, balanced datasets, where morphological regularities can be better exploited.

Jokić et al. [8] reported that WordPiece tokenization yielded the best results for Serbian hate speech detection, outperforming BPE and SentencePiece approaches. Similarly, our results suggest that subword strategies preserving meaningful morphemic units—whether through WordPiece or morphology-aware tokenizers—are more effective in morphologically rich languages.

6 Conclusion

This study compared two Slovak language models differing only in tokenization strategy. The morphology-aware SK_Morph_BLM achieved slightly higher and more stable results in toxic speech detection, while both models performed similarly on hate speech. Morphological tokenization offers limited but consistent benefits, though its impact is constrained by informal Slovak orthography. Future research should explore diacritic restoration and cross-lingual generalization to other Slavic languages.

Acknowledgements. The authors gratefully acknowledge the support of the Slovak Research and Development Agency under Contract No. APVV-23-0554, Nitra Competence Center for Cybersecurity within the Recovery and Resilience Plan of the Slovak Republic No. 17R05-04-V01-00003, and University Grants Agency No. VII/4/2025.

References

1. Mazari, A.C., Boudoukhani, N., Djeffal, A.: BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing* **27**, 325–339 (2024).
2. Mikaberidze, B., Sakhinadze, T., Mikaberidze, G., Kalandadze, R., Pkhakadze, K., van Genabith, J., Ostermann, S., van der Plas, L., Muller, P.: A Comparison of Different Tokenization Methods for the Georgian Language. In: *Proc. of ICNLSP 2024*, pp. 199–208 (2024).
3. Zhao, Z., Zhang, Z., Hopfgartner, F.: A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification. In: *WWW Companion*, pp. 500–507. ACM, Ljubljana (2021).
4. Mou, G., Ye, P., Lee, K.: SWE2: SubWord Enriched and Significant Word Emphasized Framework for Hate Speech Detection. In: *CIKM*, pp. 1145–1154 (2020).
5. Madukwe, K.J., Gao, X., Xue, B.: Token replacement-based data augmentation methods for hate speech detection. *World Wide Web* **25**, 1129–1150 (2022).
6. Držík, D., Forgac, F.: Slovak morphological tokenizer using the Byte-Pair Encoding algorithm. *PeerJ Computer Science* **10**, e2465 (2024).
7. Dang, T.A., Raviv, L., Galke, L.: Tokenization and Morphology in Multilingual Language Models: A Comparative Analysis of mT5 and ByT5. arXiv (2024).
8. Jokić, D., Stanković, R., Todorović, B.: Abusive Speech Detection in Serbian using Machine Learning. In: *NLP and AI for Cyber Security*, pp. 153–163. Lancaster (2024).
9. Kocijan, K., Košković, L., Bajac, P.: Detecting Hate Speech Online: A Case of Croatian. In: *Formalizing Natural Languages with NooJ 2019*, pp. 185–197. Springer (2020).
10. Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., Moro, R., Pikuliak, M., Šimko, M., Adamkovič, M.: Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. In: *Complex & Intelligent Systems*, Vol. 9, No. 3, pp. 2827–2842 (2021). DOI: 10.1007/s40747-021-00561-0.
11. Horníková, L.: Hate speech: semi-automatic recognition in Slovak texts. Bachelor's Thesis, Masaryk University, Faculty of Informatics, Brno (2022). `https://is.muni.cz/th/ncomq/`

12. Ferko, V., Hládek, D.: Slovak Hate Speech and Offensive Language Database. https://huggingface.co/datasets/TUKE-KEMT/hate_speech_slovak.
13. Sokolová, Z.: ToxicSK. https://aifasthub.com/datasets/TUKE-KEMT/toxic-sk.