

SMISHI - Ne Nasedaj ("Don't Fall For It")

BHS-Language SMS Phishing (Smishing) Detector for Low-Resource Morphologically Rich Languages

Submission — Apart Research Global South AI Safety Hackathon, June 19–21, 2026

Track: Open / Technical AI Safety

Author: Danijela Stankovic /*metlalchemistspex*/HF: Utaem

Abstract

Phishing detection systems are overwhelmingly trained and evaluated on English-language text, leaving morphologically rich, low-resource languages such as Bosnian/Croatian/Serbian (BHS) structurally underserved. This gap is not cosmetic: BHS morphology produces dozens of inflected surface forms for a single lexical root (e.g. *nagrada/nagradu/nagradi*), which standard tokenizers and English-centric phishing classifiers fail to generalize across — creating an exploitable blind spot for region-specific smishing (SMS phishing) campaigns. We present Ne Nasedaj, a BHS-language SMS phishing detector combining TF-IDF features with a BERTiC-based transformer ensemble, trained on a curated 1,529-message dataset of legitimate and phishing SMS text - in BHS. The system achieves 96.96% accuracy and 96.3% F1 on held-out data, demonstrating that targeted, morphology-aware approaches can close a real-world detection gap for an estimated tens of millions of BHS speakers across the Western Balkans.

1. Problem & Motivation

Most production-grade phishing/smishing detectors are built and benchmarked on English corpora. This creates two compounding risks for Global South and low-resource-language populations:

- Detection blind spots: classifiers trained on English orthography and tokenization do not generalize to BHS inflectional morphology, so even simple paraphrase/inflection changes by attackers can evade detection.
- Attack-surface asymmetry: threat actors increasingly localize phishing content (including via LLM-assisted translation), while defensive tooling for these languages lags years behind.

This mirrors a broader pattern documented in multilingual LLM safety research: refusal and safety behaviors trained predominantly on English do not transfer reliably to other languages, leaving systematic gaps for non-English speakers.

This is not a hypothetical risk. Large-scale smishing campaigns impersonating institutions such as Pošta Srbije (the Serbian Post) have actively targeted citizens for financial fraud and credential theft, and phishing volumes have surged globally since 2022 — a trend amplified by AI tools that make convincing, localized fake messages trivial to produce at scale. A 2026 UNDP/PwC Serbia labor-market study reports a 1,265% increase in phishing attacks since 2022, driven specifically by generative AI's ability to automate convincing, large-scale social engineering.

Despite this, no publicly available smishing detection tool previously supported Serbian, Bosnian, Croatian, or Montenegrin.

2. Methodology

2.1 Dataset

- 1,529 labeled SMS messages covering Serbian (Cyrillic & Latin), Bosnian, Croatian, and Montenegrin (SBCM) — phishing and legitimate, built from scratch due to the absence of any prior public SBCM smishing dataset.
- A separate, independently hand-curated 105-message adversarial stress-test set (smishi_test_cases.csv) is used for targeted evaluation. Composition: 84 phishing / 21 legitimate (80%/20%, mirroring realistic smishing prevalence); 76 Serbian, 11 Bosnian, 10 Croatian, 8 Montenegrin messages; 92 Latin-script, 12 mixed-script, 1 pure-Cyrillic; 72 messages contain a link, 33 do not; only 15 use explicit urgency language, deliberately keeping the set from being dominated by the easiest urgency-keyword cases.
- The set spans 30+ distinct evasion techniques in small, targeted clusters (typosquatting, homographs, Unicode/emoji edge cases, no-link IBAN/phone/email scams, diacritic stripping, mixed casing) rather than broad random sampling, so that each known evasion vector gets explicit, individual coverage.
- A dedicated morphological_case cluster directly tests the core adversarial-robustness claim of this project — that BHS case inflection of a single attack keyword (nagrada / nagradu / nagradi — "prize" in nominative/accusative/dative) is sufficient to potentially evade a classifier trained only on one surface form. All six morphological-variant test messages in this cluster were correctly classified as phishing, providing direct empirical support for treating BHS inflection as a measurable, defensible adversarial surface rather than a hypothetical concern.

2.2 Model Architecture

Three-layer ensemble combining a lexical baseline, a contextual transformer, and rule-based heuristics:

- Model A — TF-IDF (character n-grams, 3–5) + Logistic Regression. Character-level n-grams were chosen specifically because they degrade more gracefully than word-level tokenization under BHS case inflection (nagrada/nagradu/nagradi share overlapping character n-grams even though they are distinct word-level tokens).
- Model B — fine-tuned BERTić transformer (110M parameters), a South-Slavic-pretrained model, providing contextual sub-word representations that generalize across inflected forms and script (Cyrillic/Latin).
- Heuristic layer — suspicious/typosquatted domain detection, message-length analysis, and threat-vector (urgency language) flags, run alongside the two ML models.
- Both model predictions and confidence scores are surfaced side by side in the interface, along with the flagged heuristic indicators, rather than collapsed into a single opaque score — supporting interpretability for end users and reviewers.

2.3 Evaluation Protocol

Two complementary evaluation sets were used: (1) a standard held-out split of the 1,529-message training corpus, and (2) an independently hand-curated 105-message adversarial

stress-test set (smishi_test_cases.csv), purpose-built to probe specific evasion techniques — homographs, typosquatting, IBAN-only scams, diacritic stripping, and full Cyrillic-script messages — rather than random sampling. This second set is the more meaningful evaluation for adversarial robustness claims, since it directly targets the morphological and script-based evasion vectors this project is designed to defend against.

3. Results

Two evaluation sets are reported. Training/held-out split: 96.96% accuracy, 96.3% F1 (full dataset). Targeted edge-case stress test (smishi_test_cases.csv, 105 hand-curated examples specifically designed to probe homographs, typosquatting, IBAN-only scams, Cyrillic/Latin script-switching, and diacritic stripping): 93.3% accuracy.

Metric	Score
Accuracy (training/held-out split)	96.96%
F1 Score (training/held-out split)	96.3%
Accuracy (105-example adversarial edge-case set)	93.3% (97/105 correct, 7 incorrect, 1 model-disagreement)

Error analysis (105-example edge-case set):

- IBAN-only phishing without urgency language or a URL was the dominant failure mode: all three no_link_iban test cases (bank-impersonation messages requesting a refund/payout via IBAN, with no link and no time-pressure phrasing) were either misclassified as legitimate or flagged uncertain. This is a 0% reliable-detection rate on this specific technique cluster within the test set, empirically confirming the system's documented weak spot around no-URL, IBAN-manipulation scams as the single most concrete area for future work.
- One very long, multi-sentence phishing message was misclassified as legitimate, suggesting message length and sentence structure can dilute the signal the ensemble relies on, independent of vocabulary.
- One very short legitimate message (a single word) was misclassified as phishing, a false positive likely caused by low-signal text giving the classifier too little context — flagged as urgency-related despite containing no actual urgency cues.
- One example produced model disagreement (TF-IDF model leaning legitimate at 58.8% vs. BERTiC leaning phishing at 69.0%), both well below confident-prediction thresholds — the ensemble correctly flagged this as "uncertain" rather than forcing a wrong high-confidence call, which is itself a useful safety property.
- Critically, every Cyrillic-script and homograph/Unicode-substitution test case in the 105-example set was correctly classified — direct empirical evidence that the character n-gram approach generalizes across script and morphological surface-form variation, the core adversarial-robustness claim of this project.

4. Why This Matters for AI Safety in the Global South

- Demonstrates that BHS morphological variation is a measurable, exploitable adversarial attack surface — not just a translation inconvenience — directly relevant to multilingual LLM safety evaluation.

- Provides a reusable methodology (TF-IDF + morphology-aware transformer ensemble) that other low-resource, morphologically rich language communities (e.g. other South Slavic, Baltic, or Bantu languages) could adapt.
- Closes a concrete real-world harm gap: SMS phishing is a high-volume, low-sophistication attack vector that disproportionately affects populations without access to enterprise-grade email/SMS filtering.
- A relevant industry parallel: in June 2026, a Mythos-class frontier model (Anthropic's Claude Fable 5) was jailbroken in part through Cyrillic/Unicode homoglyph character substitution — replacing Latin characters with visually identical characters from other scripts to evade keyword-based safety classifiers, contributing to a US government export-control directive suspending the model. Ne Nasedaj was built to defend against functionally the same evasion technique from the opposite direction: scammers using Cyrillic/homoglyph substitution to dodge phishing filters, rather than attackers using it to dodge safety filters. Because the system uses character n-gram representations rather than exact keyword matching, it is structurally resistant to this evasion class — the 105-example adversarial test set (Section 3) confirms every Cyrillic-script and homograph test case was correctly classified. This suggests character/sub-word-level representation, rather than keyword-list matching, is a generalizable hardening principle for any classifier — phishing detection or AI safety guardrail alike — that needs to resist script-substitution evasion.
- This project is one concrete output of a broader, ongoing independent research direction treating BHS morphological variation as a measurable adversarial attack surface for multilingual NLP safety — the same throughline underlying an active application to the Apart Research / Atlas Computing SPS Fellowship (Track 4, Adversarial Robustness). Ne Nasedaj's morphological_case test results (Section 2.1) and Cyrillic/homograph results (Section 3) serve as early empirical validation for that larger research agenda.

5. Limitations & Future Work

- Dataset size (1,529 messages) is modest; broader, continuously updated data collection would improve robustness against novel campaigns, especially for Bosnian and Montenegrin regional variants.
- No-URL phishing is a known, documented weak spot: IBAN-manipulation scams, vishing scripts, and pure social-pressure messages without a link currently evade the system more often than URL-based scams.
- Novel attack patterns: AI-generated smishing may increasingly use phrasing outside the current training distribution, requiring continuous dataset refresh.
- Future work: extend the morphological-variation methodology into a formal adversarial robustness benchmark (planned as part of ongoing independent research into BHS as an adversarial attack surface).

6. Repository & Reproducibility

Live demo (Hugging Face Space): huggingface.co/spaces/Utaem/ne-nasedaj-sms-phishing

TF-IDF baseline model: huggingface.co/metalalchemistspex/Utaem | BERTiC fine-tuned model: huggingface.co/ravi2505/ne-nasedaj-sms-phishing

License: MIT. Test set (105 labeled edge-case examples) included in repo as `smishi_test_cases.csv` for independent verification.

Acknowledgements

Built during the Build Small Hackathon, June 2026, by Danijela Stankovic metalalchemistspex/Utaem — system design, Gradio UI, TF-IDF baseline training, dataset creation/preparation, rule-based heuristics, and full application integration — and Jnana Durga Ravi Chand Mutthina (ravi2505) — BERTić model fine-tuning and Spaces deployment support.

References

- UNDP / PwC Serbia, "Tržište rada u oblasti sajber bezbednosti u Srbiji," Belgrade, June 2026.
- National CERT Serbia, Phishing campaign alert, August 2024.
- M. Drolet, "AI Is Amping Up Phishing, Smishing and Vishing Attacks," Forbes Technology Council, May 2025.
- RATEL, Annual Report 2024.
- Nikola Ljubešić and Davor Lauc. 2021. BERTić — The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, pages 37–42, Kyiv, Ukraine. Association for Computational Linguistics.