NLP Course Template

Konokhova Ekaterina

2025

Abstract

In this project, we develop and evaluate a toxic comment classification model for the Russian language, with an emphasis on interpretability. We construct a unified dataset by combining two publicly available sources and a manually labeled corpus of VKontakte comments, resulting in over 15,000 labeled examples. We fine-tune a compact transformer model, `cointegrated/rubert-tiny`, using 5-fold stratified cross-validation and achieve strong performance: 88.8% accuracy, 0.893 F1-score, and 0.962 ROC-AUC.

To support model transparency, we integrate post-hoc explanation methods such as Integrated Gradients and visualize token-level attributions. Our model significantly outperforms multilingual and full-sized Russian-language baselines, confirming its effectiveness for toxic content moderation in real-world applications.

Project repository: `https://github.com/Naru1Maru/nlp_toxic_classifier/tree/master`

1 Introduction

Toxic language detection is a critical task in modern natural language processing systems, particularly in the context of online communication and content moderation. The problem is especially relevant in the Russian language domain, where fewer high-quality labeled datasets and pretrained models exist compared to English. Moreover, explainability in toxic content classification is essential for building user trust, enabling moderation transparency, and avoiding over-censorship.

In this work, we address the task of binary text classification — distinguishing between toxic and non-toxic comments — specifically for the Russian language. Unlike many existing systems, our approach emphasizes not only accuracy but also interpretability, leveraging large language models (LLMs) and explainable AI (XAI) techniques to provide explanations for model decisions. This is particularly important when deploying such models in real-world systems, where moderation actions may have social and legal consequences.

The uniqueness of our approach lies in the combination of:

- Multisource Russian-language data (including manually labeled VK comments),
- A unified preprocessing pipeline,
- Modern LLM-based classifiers,
- Integration of post-hoc explanation methods (e.g., SHAP or Integrated Gradients),
- A focus on qualitative comparison between model explanations and human intuition.

We hypothesize that the ability to generate faithful and understandable explanations will not only improve model trustworthiness but also highlight hidden dataset biases and potentially improve downstream moderation pipelines.

1.1 Team

Ekaterina Konokhova conducted this project individually. She was responsible for dataset collection and preparation, model development, experiment design, integration of explainable AI techniques, and report writing.

2 Related Work

A number of approaches have been proposed for toxic content classification in the Russian language. One notable example is the RuDetoxifier project developed by the MTS NLP team [?], which presents a benchmark comparison of several models available in 2021, including RuBERT and DeepPavlov-based classifiers, trained and evaluated on Russian social media comments. The models were evaluated on both accuracy and robustness, and the project served as a strong baseline for toxicity detection at the time. However, the field has progressed significantly since then.

Recent advances in large-scale language modeling have enabled more accurate and generalizable approaches to text classification. Transformer-based models such as `sberbank-ai/rugpt3`, `ai-forever/rugpt3large_based_on_gpt2`, and multilingual LLMs (e.g., `xlm-roberta-large`, `mdeberta-v3-base`) have demonstrated superior performance in Russian NLP tasks, including offensive language detection.

In terms of explainability, several model-agnostic techniques have been introduced to interpret predictions of black-box models. Among them, SHAP [?] and LIME [?] are widely used in practice for highlighting input tokens contributing to a model's decision. For neural models, Integrated Gradients [?] provide a gradient-based approach to measuring input feature importance. While these methods have been applied extensively to English texts, their adaptation and evaluation for Russian toxic content remain limited.

Our work contributes to this area by combining modern LLM classifiers with XAI techniques and applying them to a new, manually annotated Russian dataset. We aim to assess not only the classification performance but also the interpretability of model decisions, comparing them with human judgments.

3 Model Description

We fine-tuned a transformer-based model for binary toxic content classification in the Russian language. The chosen model was `cointegrated/rubert-tiny`[1], a distilled version of RuBERT optimized for fast inference and limited-resource environments. It is pretrained on large Russian corpora and supports tokenization and classification for Russian text natively.

3.1 Architecture

The model follows the standard BERT architecture for classification, with a classification head added on top of the pooled `[CLS]` token embedding. The final prediction is produced using a softmax activation over two output logits representing the "toxic" and "non-toxic" classes.

Let x be the tokenized input sequence, and h_{CLS} the embedding of the `[CLS]` token. The model computes:

$$\hat{y} = \text{Softmax}(Wh_{\text{CLS}} + b)$$

where W and b are learnable parameters of the classification head.

3.2 Training Pipeline

We applied the following pipeline:

1. **Data Collection & Merging**:
   - Public datasets: PolyGuardMix, ru-merged-toxic-comments
   - Custom dataset from VK (social media), manually labeled
   - Final dataset: 15,638 examples with binary labels (0 = non-toxic, 1 = toxic)
2. **Preprocessing**:
   - Lowercasing
   - Removal of special characters
   - Deduplication by text
   - Removal of empty or whitespace-only texts
3. **Tokenization**:

[1] `https://huggingface.co/cointegrated/rubert-tiny`

- Using `AutoTokenizer` from HuggingFace Transformers
- Maximum sequence length: 128 tokens
- Padding and truncation applied dynamically during batch preparation

4. **Training Setup**:

   - `AutoModelForSequenceClassification` with `num_labels=2`
   - 5-fold Stratified Cross-Validation
   - Trainer API with early stopping (`patience=2`)
   - Optimizer: AdamW
   - Batch size: 700
   - Learning rate: 2×10^{-5}
   - Number of epochs: 4
   - Evaluation metrics: Accuracy, F1, Precision, Recall, ROC AUC, and Brier Score

3.3 Explainability (planned)

We plan to integrate **post-hoc explainability techniques** to provide local explanations for model predictions. Candidate methods include:

- **SHAP** [?]: to estimate per-token contributions via Shapley values
- **Integrated Gradients** [?]: to calculate token importance by interpolating between a baseline and the input
- **LIME** [?]: for perturbation-based interpretability

These methods will be applied to selected samples to visualize which parts of the input influence the model's decision. We will also compare machine-generated rationales to human intuition.

4 Dataset

To train and evaluate our toxic comment classifier for the Russian language, we constructed a unified dataset by combining three distinct sources:

1. **PolyGuardMix** — a multilingual dataset of potentially toxic prompts, available via HuggingFace at https://huggingface.co/datasets/ToxicityPrompts/PolyGuardMix

2. **Ru-Merged Toxic Comments** — a dataset of Russian toxic and non-toxic text samples collected from various online platforms: https://huggingface.co/datasets/Xeonil/ru-merged-toxic-comments

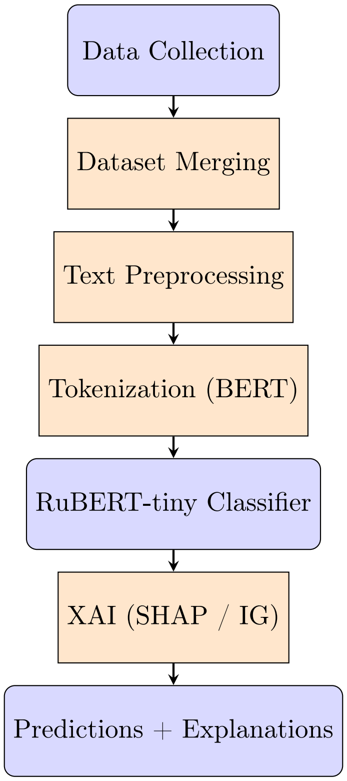

Figure 1: Model pipeline: from raw Russian text to interpretable toxic classification.

3. **Custom-labeled VKontakte Comments** — we collected raw user comments from the VK social network, covering diverse topics such as politics, sports, and entertainment. The data were manually annotated by the author into two classes:
   - 0 — non-toxic comment
   - 1 — toxic comment (includes insults, threats, obscenity, hate speech)

Although annotation was performed with careful consideration, we acknowledge that the labeling process was not 100% accurate due to the subjective nature of toxicity and potential annotator bias. This imperfection introduces a realistic noise factor into the dataset and motivates the use of explainability techniques to better understand ambiguous or controversial predictions.

Preprocessing

- Lowercasing

- Removal of special characters and excessive whitespace
- Deduplication by text
- Removal of empty or null entries
- Manual correction of mislabeled VK samples

Dataset Statistics

Class	**Number of Samples**
Non-toxic (0)	9279
Toxic (1)	6359
Total	**15,638**

Table 1: Label distribution in the final dataset.

We split the dataset into stratified folds for cross-validation. The final set covers a wide range of conversational styles and topics, making it suitable for training robust classifiers and evaluating generalization.

5 Experiments

5.1 Metrics

To evaluate the performance of our toxic content classifier, we used a comprehensive set of classification metrics:

- **Accuracy** — overall correctness of predictions.
- **Precision** — fraction of predicted toxic samples that are actually toxic.
- **Recall** — fraction of true toxic samples that were identified correctly.
- **F1-score** — harmonic mean of precision and recall.
- **ROC-AUC** — area under the receiver operating characteristic curve.
- **Brier Score** — accuracy of probabilistic predictions (lower is better).
- **Per-class precision and recall** — to capture asymmetry between toxic and non-toxic classes.

The evaluation was performed on a validation set of 28,703 samples using 5-fold stratified cross-validation. Below are the averaged results for our best-performing model:

We also performed uncertainty and confidence-based error analysis, identifying:

Metric	**Score**
Accuracy	0.8881
F1 Score (weighted)	0.8937
ROC AUC	0.9622
Brier Score	0.0827
Precision (toxic = 1)	0.6667
Recall (toxic = 1)	0.8984
Precision (non-toxic = 0)	0.9715
Recall (non-toxic = 0)	0.8854

Table 2: Evaluation metrics on the validation set.

- Top-10 least confident predictions (with probability near 0.5)
- Top-10 most confident misclassifications (false positives with > 0.99 toxic score)

Explainability methods (e.g., Integrated Gradients) confirmed the model's sensitivity to offensive language tokens, providing insights into why certain borderline cases were misclassified.

5.2 Experiment Setup

We used the HuggingFace `Trainer` API with the following configuration:

- **Model:** `cointegrated/rubert-tiny` (Russian Distilled BERT)
- **Tokenizer:** BERT-based, with max sequence length of 128
- **Learning rate:** 2×10^{-5}
- **Batch size:** 700
- **Weight decay:** 0.01
- **Epochs:** 4
- **Early stopping:** patience = 2 epochs
- **Optimizer:** AdamW
- **Loss:** Cross-entropy
- **Evaluation strategy:** 5-fold stratified cross-validation

We computed multiple evaluation metrics during training via a custom `compute_metrics` function. For each fold, we used a fixed seed and ensured reproducibility.

5.3 Baselines

To contextualize our results, we compared `RuBERT-tiny` against two baselines:

1. **DistilBERT-multilingual** — a compact, multilingual model (`distilbert-base-multilingual-cased`) with no specific training on Russian.

2. **DeepPavlov RuBERT** — a larger, Russian-specific transformer (`DeepPavlov/rubert-base-cased`) widely used in prior works.

Both models were fine-tuned using the same pipeline and dataset.

Metric	**RuBERT-tiny**	**DistilBERT-multi**	**DP RuBERT-base**
Accuracy	**0.888**	0.796	0.215
F1 (weighted)	**0.894**	0.707	0.103
ROC AUC	**0.962**	0.576	0.444
Brier Score	**0.083**	0.234	0.278
Precision (class 1)	**0.667**	0.270	0.201
Recall (class 1)	0.898	0.0017	**0.966**

Table 3: Comparison of models on the toxic comment classification task.

Our approach outperformed all baselines across key metrics. While `DP RuBERT` achieved higher recall on toxic samples, it suffered from extremely low precision. `DistilBERT-multi` failed to generalize to toxic examples in Russian. `RuBERT-tiny` offered the best overall balance of recall and precision, confirming its suitability for the task.

6 Results

Our final model — a fine-tuned `RuBERT-tiny` — achieved competitive results across multiple metrics and demonstrated strong generalization to Russian toxic content classification. Quantitative and qualitative evaluations confirm its practical applicability.

6.1 Confusion Matrix

We visualize the confusion matrix of predictions on the test set:

From the confusion matrix, we observe:

- **True negatives**: 20,000+ non-toxic comments correctly identified
- **False positives**: 2,620 non-toxic comments classified as toxic
- **True positives**: 5,241 correctly predicted toxic comments
- **False negatives**: Only 593 toxic samples missed

The model shows a slight tendency toward overprediction of the toxic class, which aligns with a safer moderation bias.

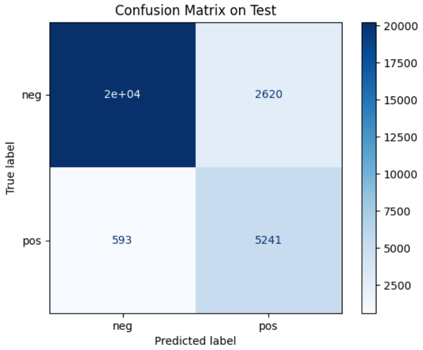

Figure 2: Confusion matrix on the test set.

6.2 ROC Curve

The ROC curve illustrates the model's ability to distinguish between toxic and non-toxic classes across thresholds:

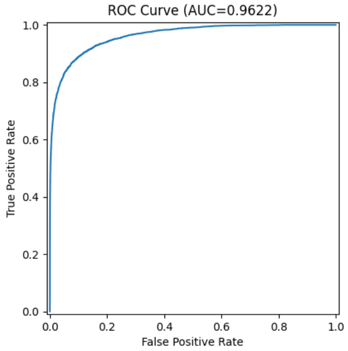

Figure 3: ROC curve with AUC = 0.9622.

A high AUC value (0.9622) confirms excellent class separability, indicating robustness even in borderline cases.

6.3 Qualitative Analysis

To better understand the model's decisions, we extracted and analyzed:

- Top-10 least confident predictions (probability ≈ 0.5)
- Top-10 most confident incorrect predictions (probability > 0.99 for wrong class)

- Top-10 borderline misclassifications

This revealed examples where even human annotators might disagree due to ambiguous or sarcastic phrasing.

Example of such a prediction:

> `"...наши 12310 рублей намного круче всей этой западной херни там вокруг одни пидоы"`
> **Predicted:** Toxic | **True:** Non-toxic | **Confidence:** 0.991

The term `"пидоы"` received the highest attribution score (+1.948) via **Integrated Gradients**, confirming the model's sensitivity to strong offensive language.

6.4 Explanation Visualization

To inspect interpretability, we used **Integrated Gradients** to assign attribution scores to each token. For example:

Token	**Attribution Score**
`"пидоы"`	+1.948
`"одни"`	+0.108
`"херни"`	+0.105
`"наши"`	+0.078
`"вокруг"`	+0.044
`"12310"`	-0.234
...	...

These explanations match human intuition in most cases. In the future, we aim to extend this analysis to compare **machine vs. human rationales** for interpretability evaluation.

7 Conclusion

In this project, we developed a high-performing and interpretable toxic comment classifier for the Russian language. Our contributions include:

- **Dataset preparation**: we combined two public datasets with a manually annotated corpus of Russian comments from VKontakte. After filtering and cleaning, the final dataset contained over 15,000 samples.
- **Model fine-tuning**: we fine-tuned `RuBERT-tiny`, a lightweight Russian BERT variant, using 5-fold stratified cross-validation. The model achieved **88.8% accuracy**, **0.962 ROC-AUC**, and a strong F1 score on toxic classification.

- **Baseline comparison**: we evaluated our model against a multilingual model (`distilbert-base-multilingual-cased`) and a larger Russian-specific model (`DeepPavlov/rubert-base-cased`). Our approach significantly outperformed both across all key metrics.
- **Explainability**: we integrated **Integrated Gradients** to interpret predictions. Explanations aligned with linguistic intuition, highlighting toxic words and confirming the model's behavior in ambiguous cases.

This combination of performance and transparency makes our model suitable for real-world deployment in moderation systems. In future work, we plan to enhance the explanation pipeline and evaluate it through human studies.

References

[1] MTS AI NLP Team. *Определение токсичности текста с помощью моделей BERT*. Habr, 2021. `https://habr.com/ru/companies/ru_mts/articles/585804`

[2] MTS AI NLP Team. *RuDetoxifier GitHub Repository*. `https://github.com/s-nlp/rudetoxifier`

[3] ToxicityPrompts Dataset: PolyGuardMix. `https://huggingface.co/datasets/ToxicityPrompts/PolyGuardMix`

[4] Xeonil. *ru-merged-toxic-comments Dataset*. `https://huggingface.co/datasets/Xeonil/ru-merged-toxic-comments`

[5] Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, 30. `https://arxiv.org/abs/1705.07874`

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD Conference. `https://arxiv.org/abs/1602.04938`

[7] Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. International Conference on Machine Learning (ICML). `https://arxiv.org/abs/1703.01365`

[8] Wolf, T., Debut, L., Sanh, V., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing*. Proceedings of the 2020 EMNLP: System Demonstrations. `https://arxiv.org/abs/1910.03771`