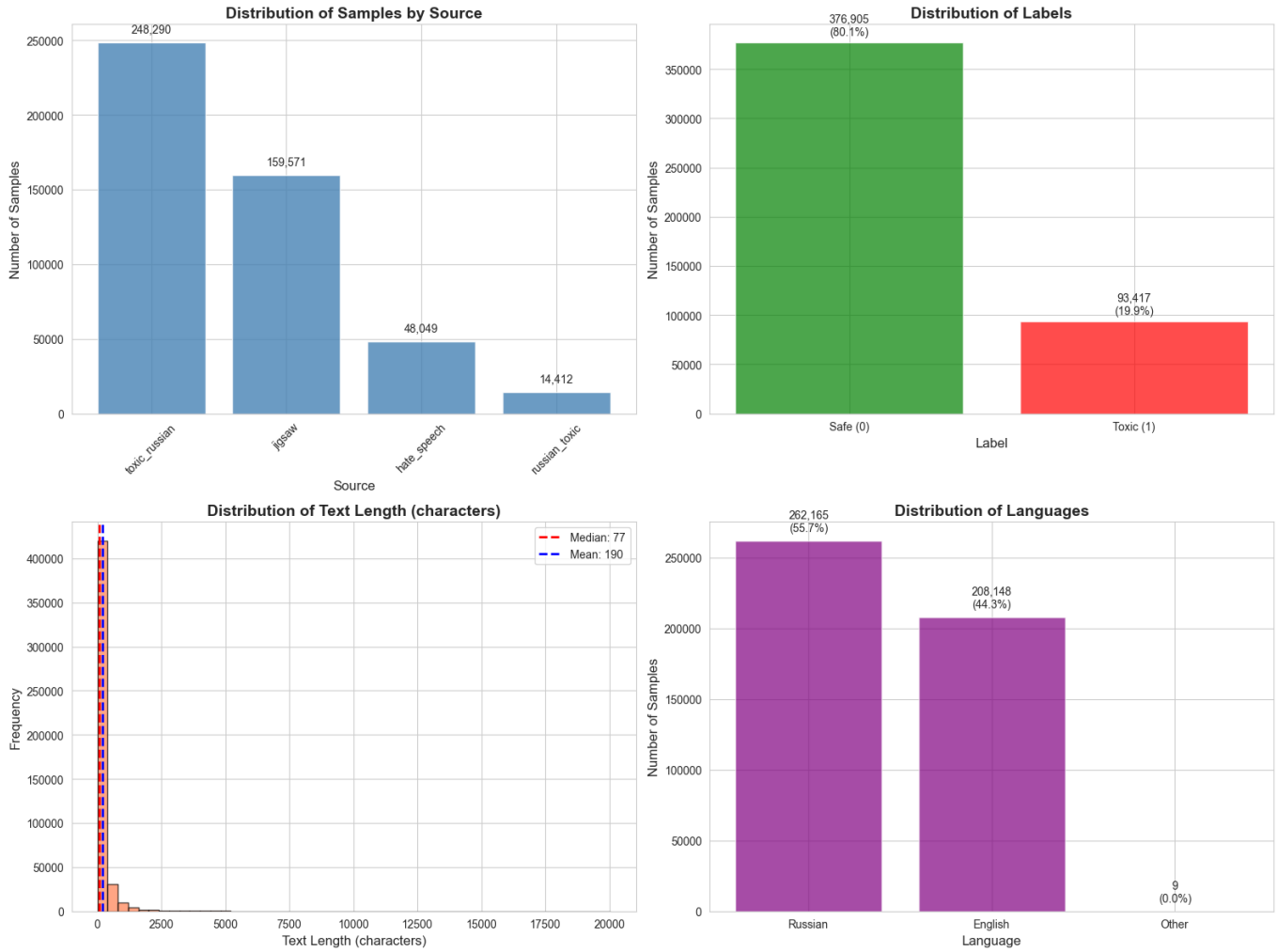
Baseline Report

Date: February 28, 2026 | **Authors:** Arthur Babkin, Alexander Malyy | **Course:** Generative AI, Spring 2026

We compare three toxic text classifiers on quality, speed, and cost. Dataset: 470K multilingual samples (56% Russian, 44% English, 20% toxic), 80/20 stratified split, `random_state=42`.

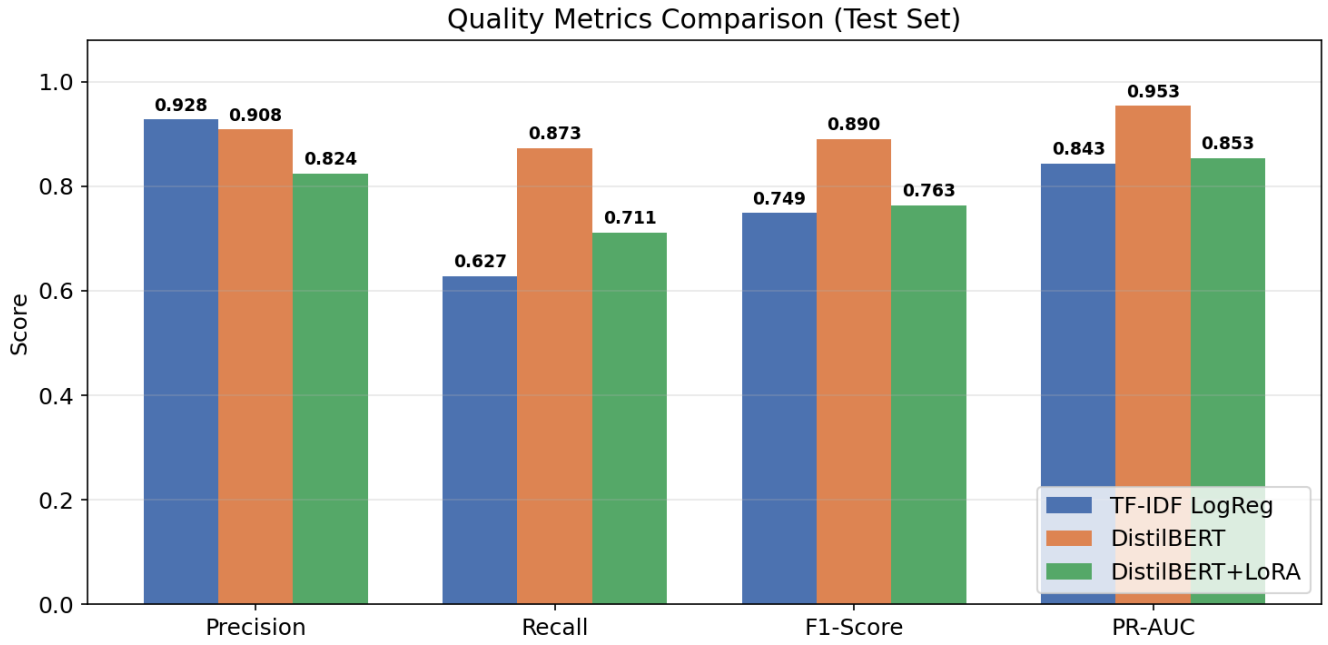
Model	F1	Precision	Recall	PR-AUC	Latency (ms)	Throughput	Disk
TF-IDF + LogReg	0.75	**0.93**	0.63	0.84	**0.018**	**58K/s**	0.5 MB
DistilBERT (fine-tuned)	**0.89**	0.91	**0.87**	**0.95**	3.08	334/s	255 MB
DistilBERT + LoRA	0.76	0.82	0.71	0.85	3.15	322/s	255 MB

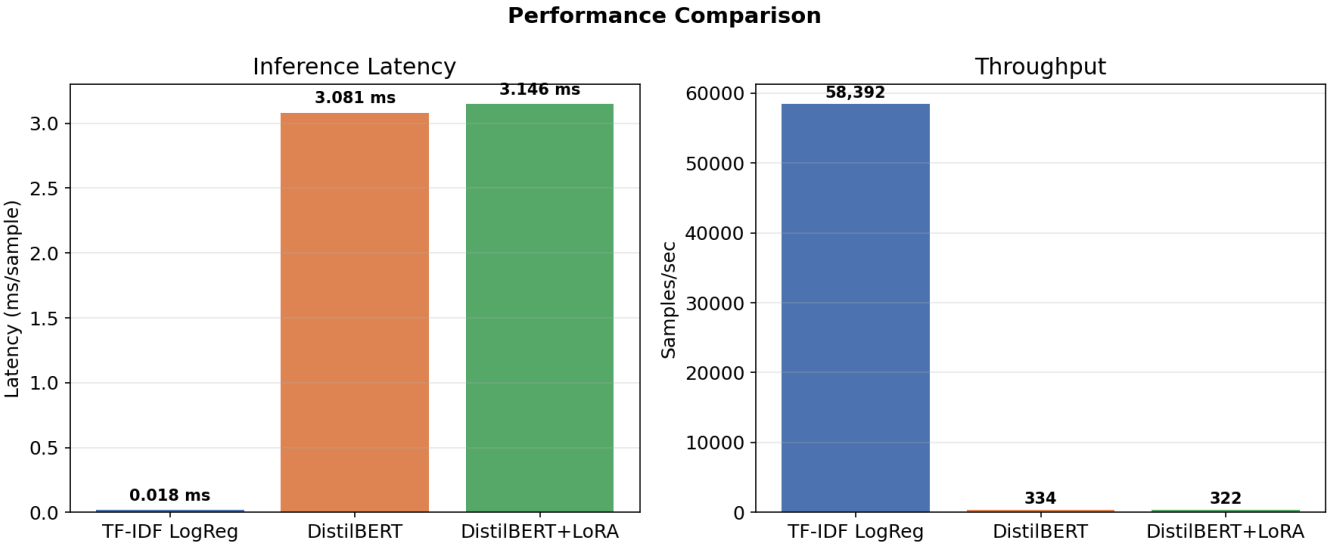


TF-IDF + LogReg — 10K features (unigrams+bigrams), C=1.0. Extremely fast, high precision, but misses 36% of toxic content due to no semantic understanding.

DistilBERT — `distilbert-base-uncased` (67.6M params), 3 epochs, lr=2e-5 on full 376K training set. Best quality, catches implicit and multilingual toxicity. ~170x slower than TF-IDF.

DistilBERT + LoRA — same base model, LoRA r=4/alpha=16 on `q_lin`/`v_lin`. Only 665K trainable params (0.98%). Trained on 100K subset, 2 epochs — 6x faster than full fine-tuning, reaches 86% of its F1.

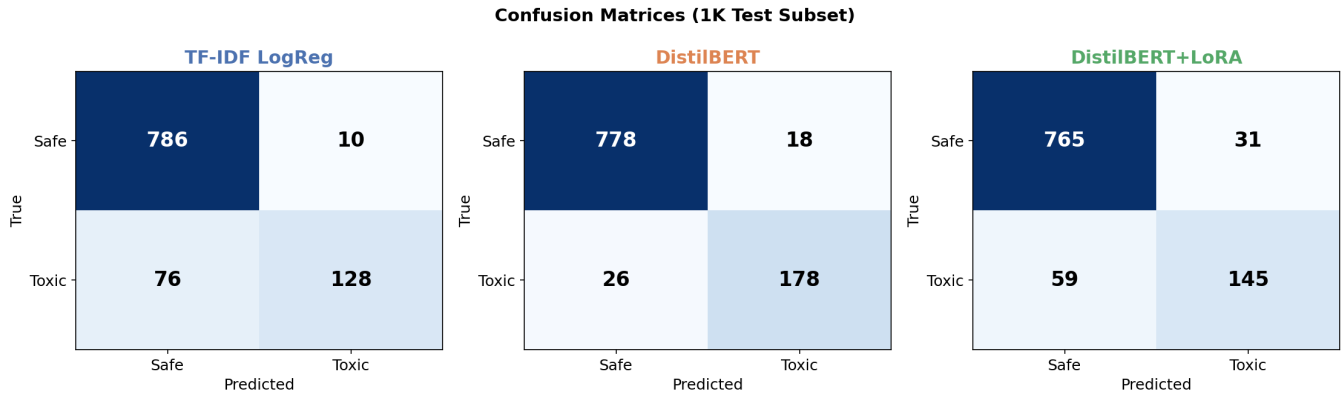




Confusion matrices

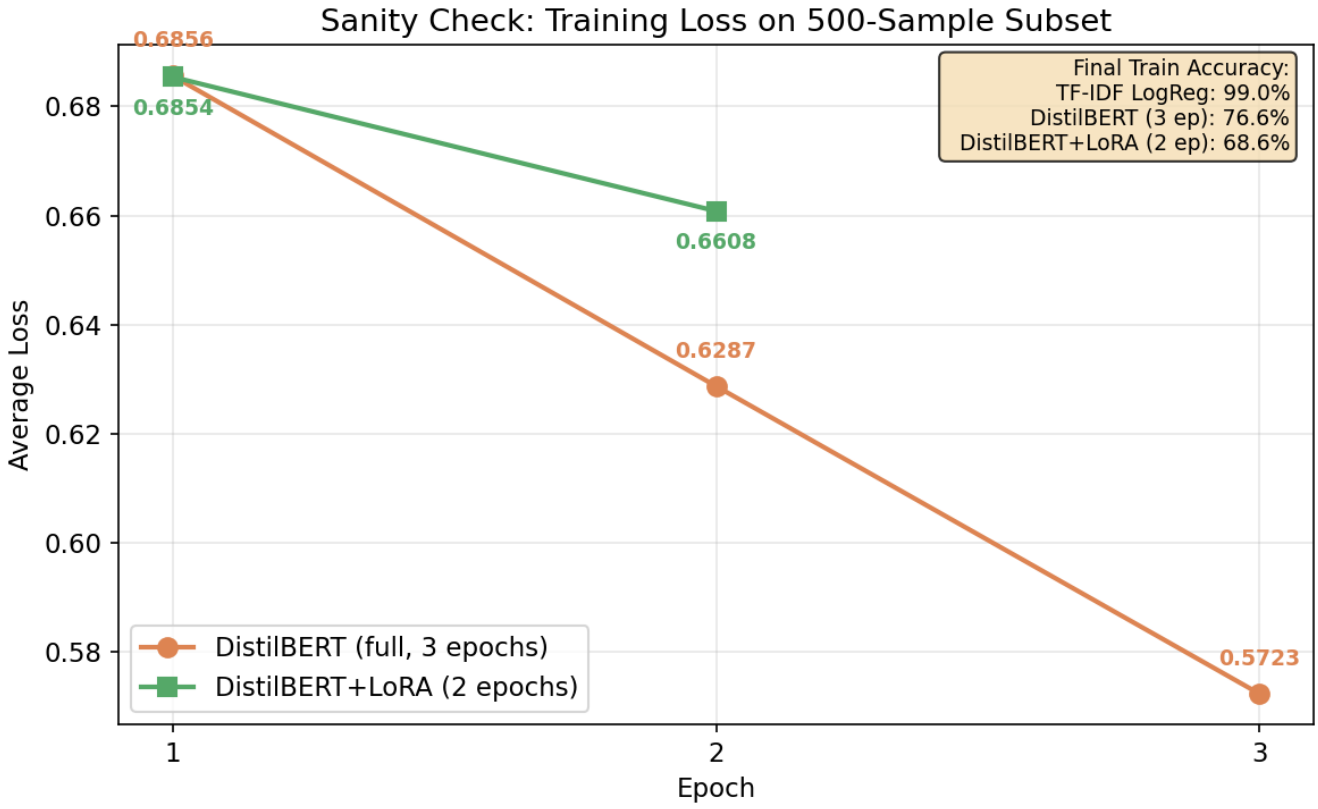
Model	Test size	FP rate	FN rate
TF-IDF LogReg	94K	1.6%	36.2%
DistilBERT	1K subset	2.3%	12.7%
DistilBERT+LoRA	1K subset	3.9%	28.9%

TF-IDF has the lowest false positive rate but misses over a third of toxic content. DistilBERT catches 87% of toxic samples. LoRA sits in between.



Sanity check

Overfitting test on 500 balanced samples to verify training pipelines work. TF-IDF reaches 99% train accuracy. DistilBERT: 76.6% (loss 0.69→0.57, 3 epochs). LoRA: 68.6% (loss 0.69→0.66, 2 epochs). All pipelines functional — loss decreases consistently.

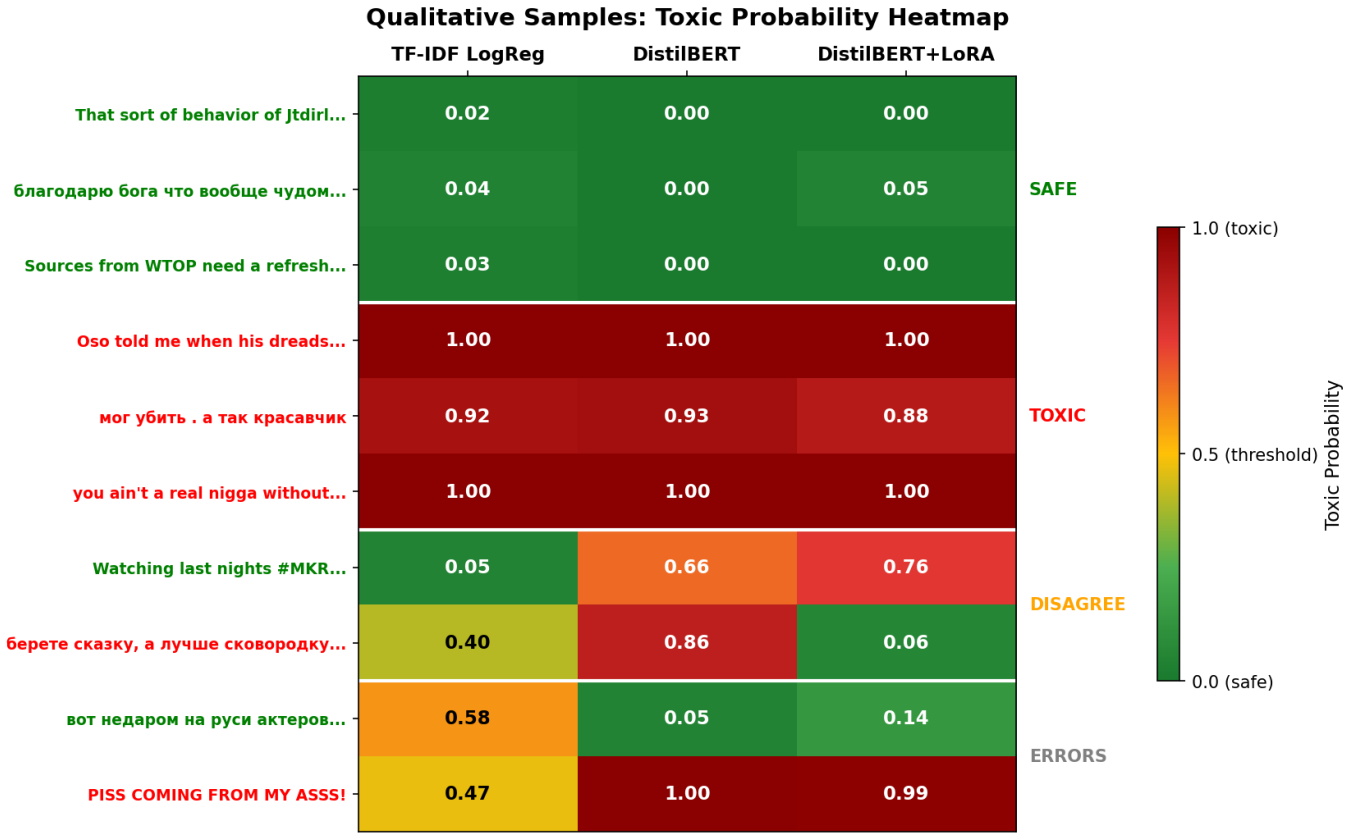


Qualitative analysis

Predictions on 500-sample random test subset. Toxic probability in parentheses. All models agree on clear-cut cases (explicit slurs → toxic, neutral text → safe). Interesting disagreements:

- *"Watching last nights #MKR ... Celine would be an absolute hoot to know"* — safe, but both transformers flag it as toxic (0.66/0.76). TF-IDF gets it right (0.05). Transformers over-trigger on informal slang.
- *"берете сказку, а лучше сковородку потяжелее..."* — toxic (implicit violence in Russian). Only full DistilBERT catches it (0.86). TF-IDF (0.40) and LoRA (0.06) miss the context.

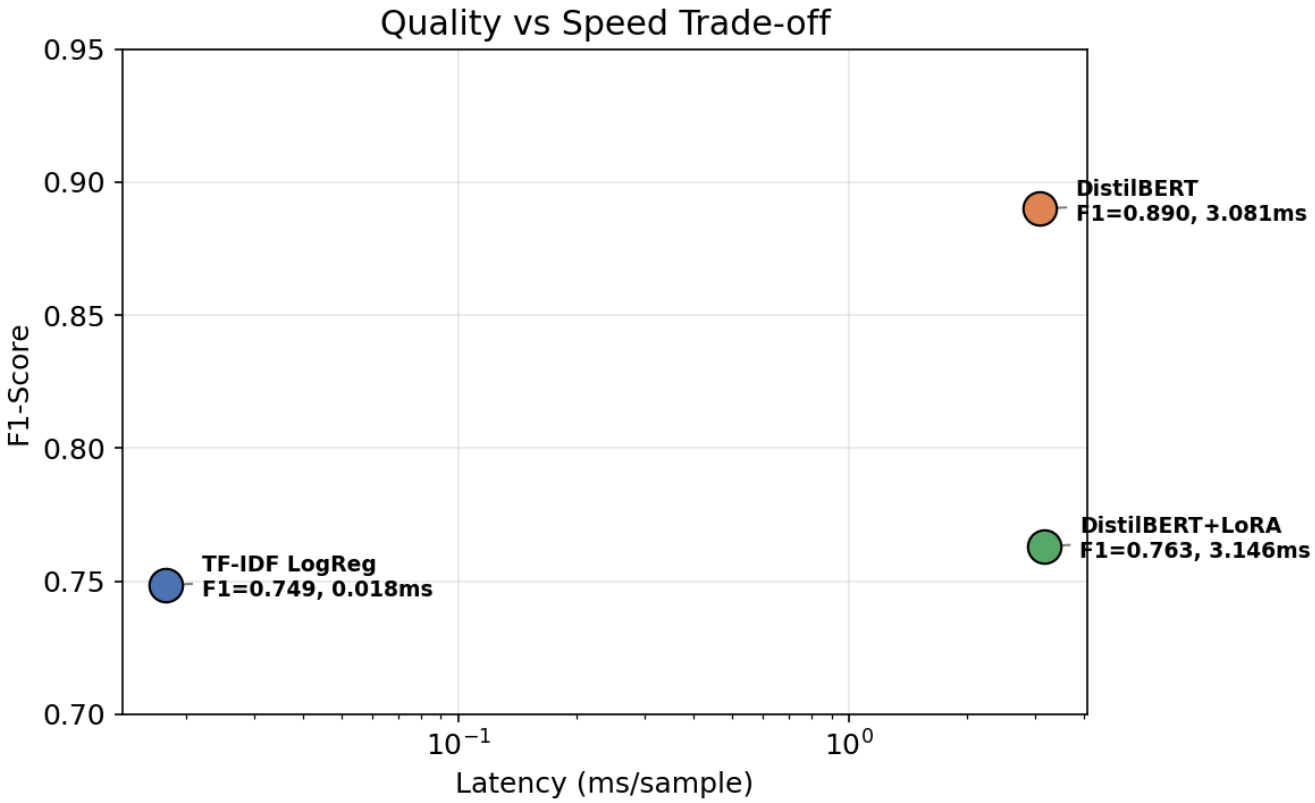
- *"PISS COMING FROM MY ASSS!"* — toxic, but TF-IDF scores it 0.47 (safe). Transformers: 1.00/0.99. Bag-of-words can't detect vulgarity from context alone.



Trade-offs and deployment

~19% F1 improvement (0.75→0.89) costs ~170x latency. LoRA offers a middle path: fast training, decent quality, same inference cost as full fine-tuning.

Use case	Model	Rationale
Real-time moderation	TF-IDF	<0.02ms, CPU-only, 0.5 MB
Batch review	DistilBERT	Best quality
Rapid prototyping	LoRA	20min training



Next steps

- Threshold tuning on validation set
- Deobfuscation preprocessing (spacing, leetspeak)
- Robustness testing on adversarial inputs
- Ensemble (TF-IDF + DistilBERT routing)
- Separate Russian vs English evaluation

Reproducibility

Notebooks: `model/experiments/{tf_idf,transformer,transformer_lora}.ipynb`. Pre-trained models: `data/models/{logreg,transformer,transformer_lora}/`. All seeds fixed at 42.