Spoiler Detection in User Reviews

Magdalena Jeczeń, Piotr Rowicki, Krzysztof Wolny

Natural Language Processing (Winter 2025)

Problem Definition and Motivation

Problem

- Online reviews often contain spoilers that degrade user experience.
- Automatic spoiler detection is a real-world NLP classification task.

Motivation

- Many approaches exist: from simple TF-IDF models to large Transformer-based models.
- Practical question: when is a heavy model really worth it?

Key challenge

- Trade-off between predictive performance and computational efficiency.

Research Goals and Hypotheses

Scientific Goals

- Compare classical ML models with Transformer-based models for spoiler detection.
- Evaluate models not only by accuracy-related metrics, but also by inference time.

Hypothesis

- Transformer models outperform classical approaches in predictive quality,
- but at the cost of significantly higher inference time.

Datasets & EDA

Datasets

- IMDB Spoiler Dataset (movie & TV reviews)
- Goodreads Balanced Dataset (book reviews)

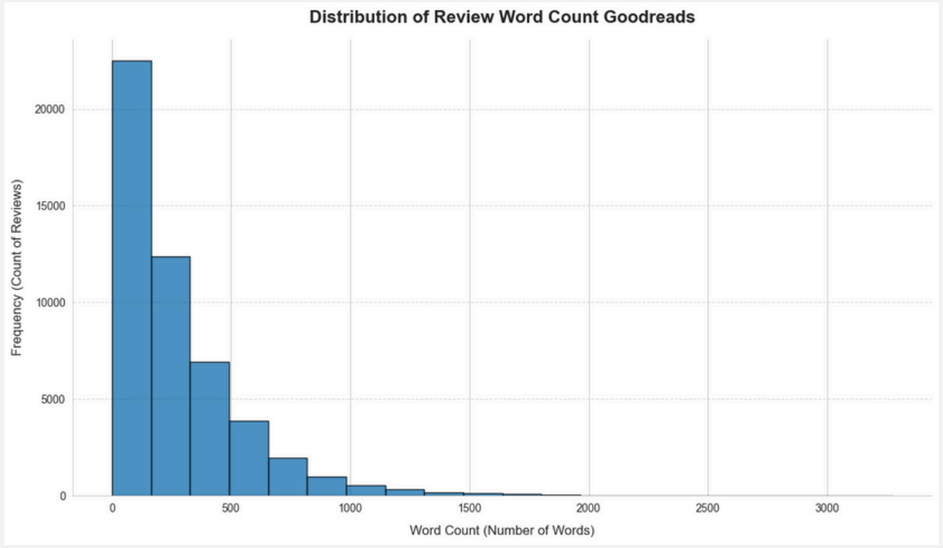
Key EDA Takeaways

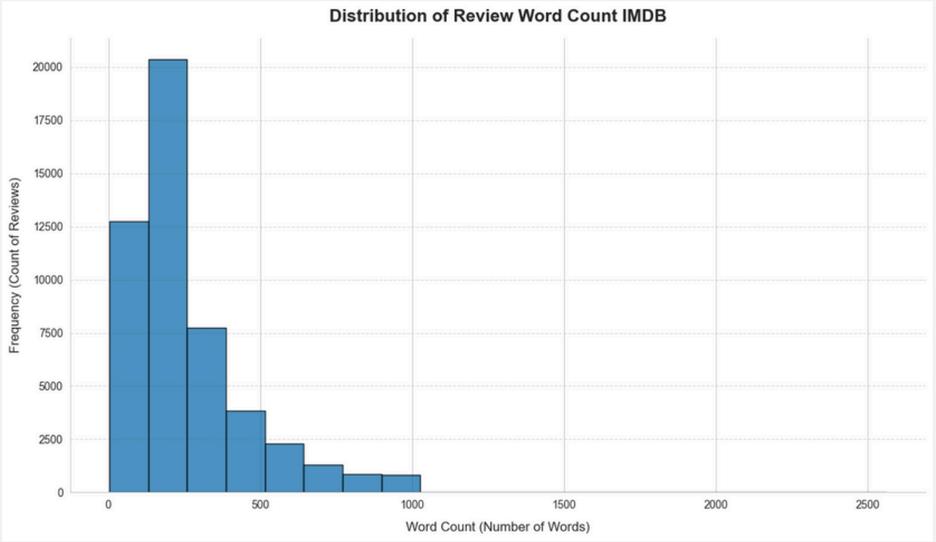
- IMDB reviews: longer, movie-focused vocabulary
- Goodreads reviews: shorter, book-centric language
- Different length distributions → impacts model behavior

Dataset	Number of rows	Percent of spoilers
IMDB	50 000	26.20%
Goodreads	50 000	50.00%
Combined dataset	100 000	38.10%

Final Dataset

- 100,000 samples (50k IMDB + 50k Goodreads)
- Spoiler ratio: ~38% spoilers / ~62% non-spoilers
- Two different domains → cross-domain robustness





Models Compared

Classical Models

- TF-IDF + SVM
- Bag-of-Words + Logistic Regression

Transformer-based

- BERT (bert-base-uncased, partial fine-tuning: last 2 layers + classifier)
- RoBERTa (roberta-base, full fine-tuning)

Experimental Setup and Testing Suite

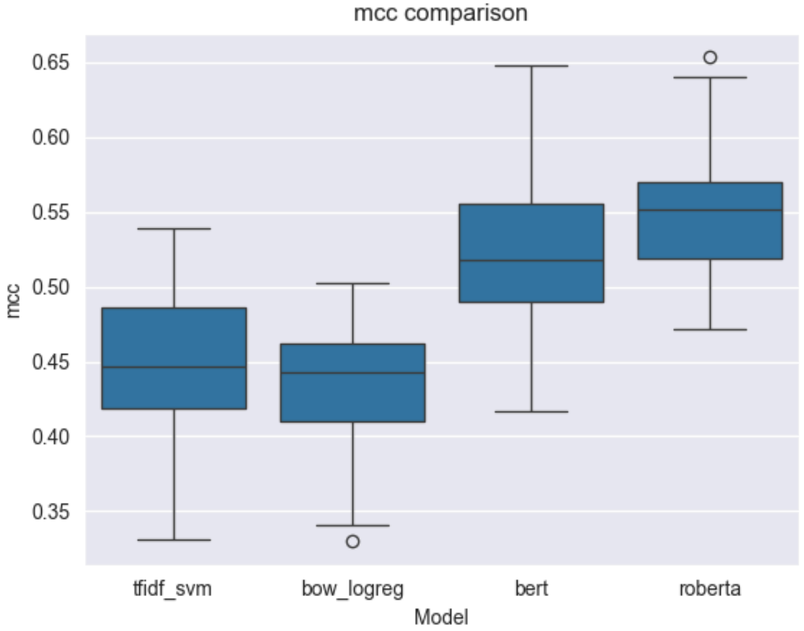
Unified Experimental Setup

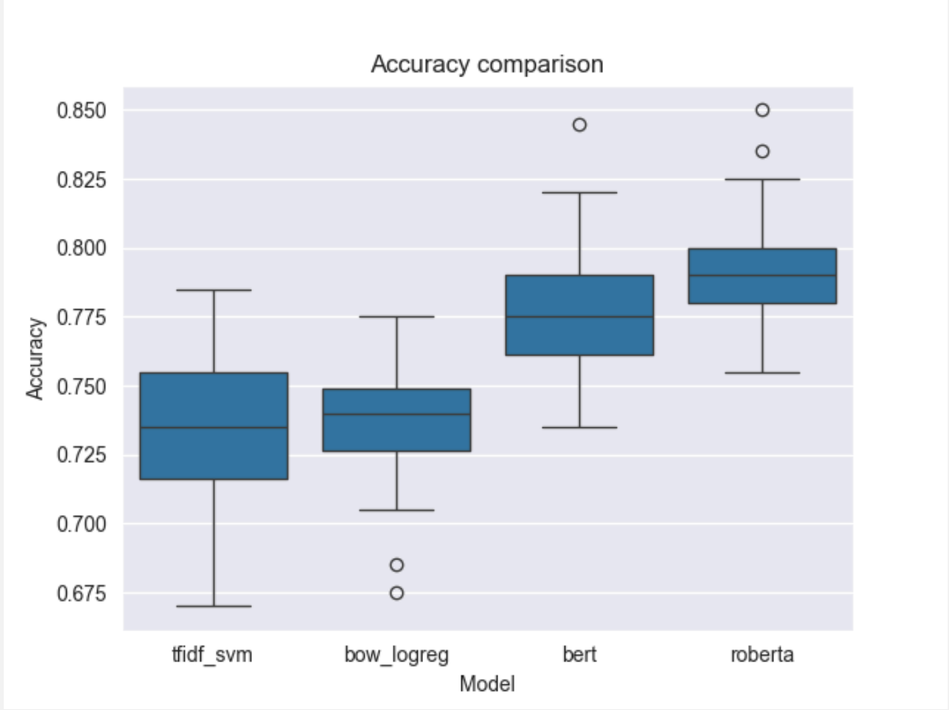
- Identical train/test splits
- Same datasets and preprocessing
- Same evaluation pipeline for all models

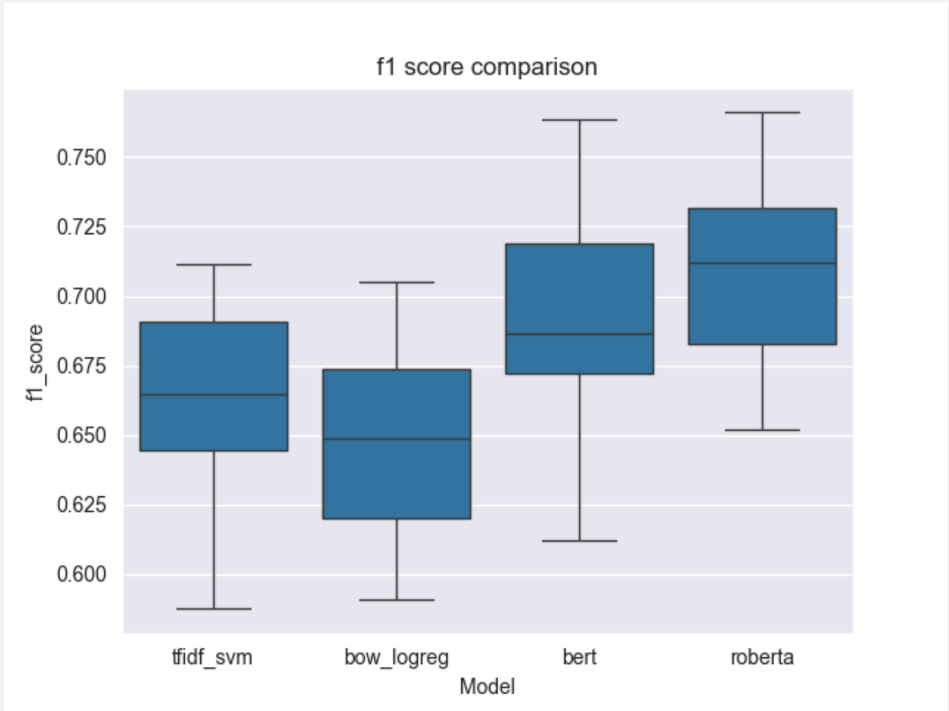
Testing Suite

- Standardized model interface
- Batch-based evaluation
- Supports multiple runs and latency measurement

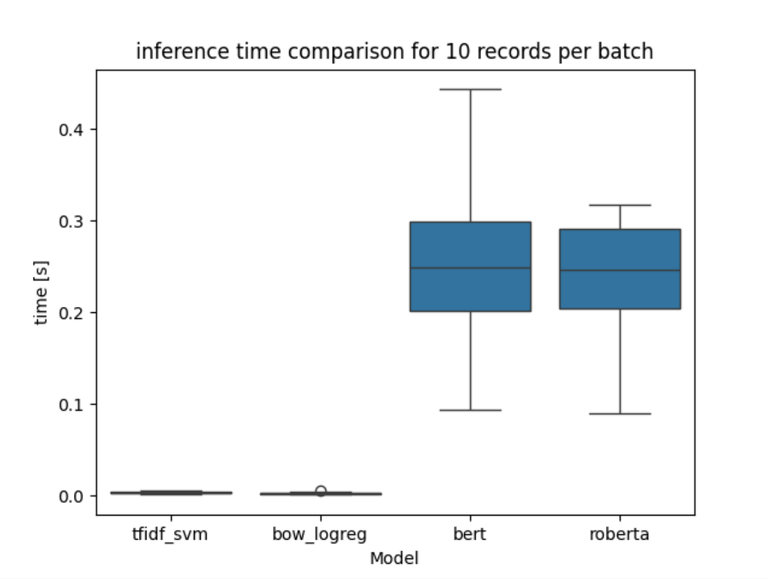
Results: Predictive Performance

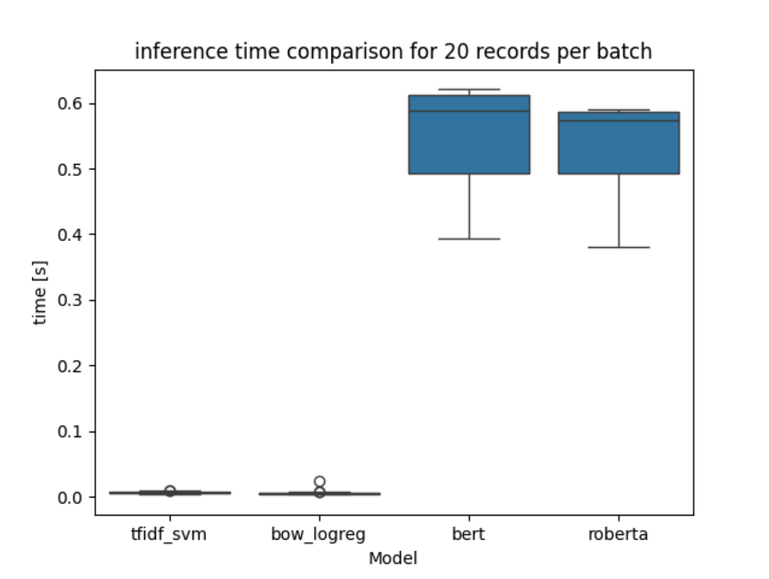


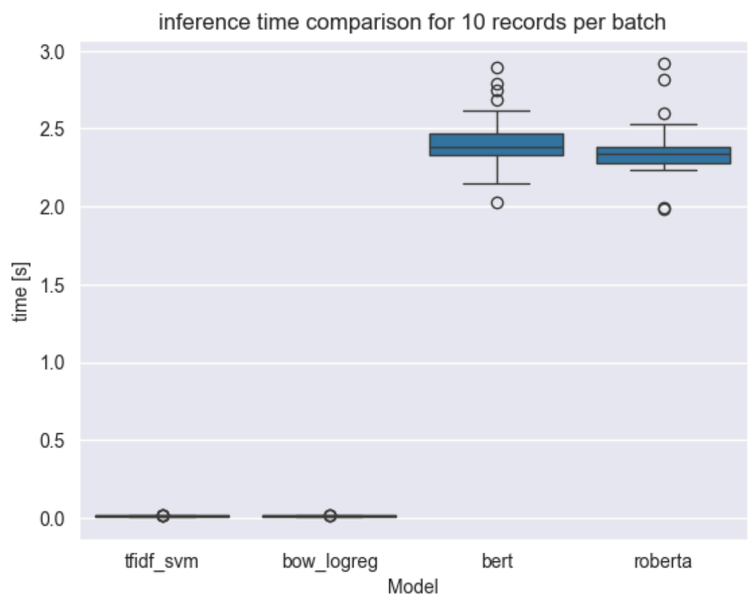


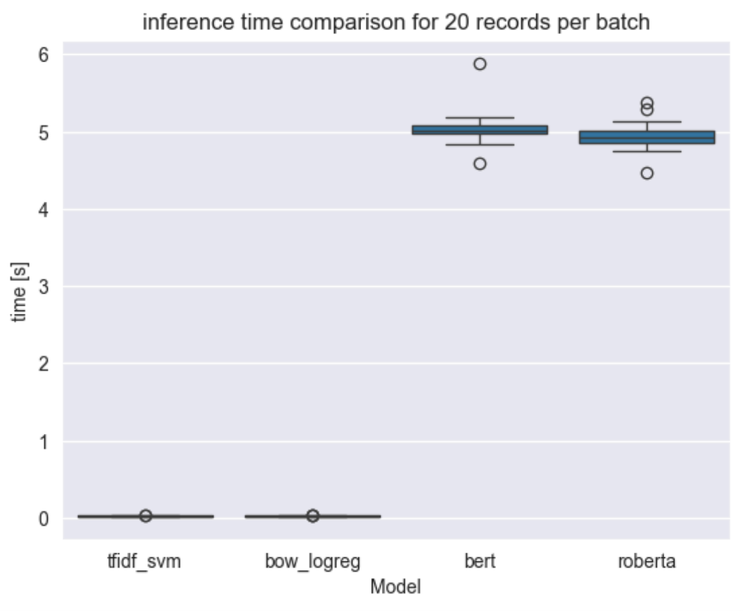


Results: Inference Time









Discussion and Takeaways

Main Takeaways

- RoBERTa > BERT > Classical (predictive quality)
- Classical models ≫ Transformers (speed)
- No universally best model

Practical Implications

- Real-time, low-resource systems → classical ML
- High-stakes moderation → Transformers justified

Limitations

- Text-only models
- Limited hyperparameter tuning
- No metadata or multimodal features

Conclusions and Future Work

Conclusions

- We delivered a reproducible, benchmark-style comparison.
- Explicit performance vs efficiency trade-offs were demonstrated.
- Transformer improvements are measurable but costly.

Future Work

- Extended hyperparameter tuning
- Exploring deeper fine-tuning strategies
- Metadata and multimodal models
- Larger-scale evaluation

Thank You!