NLP Fake News Detection

Group 4

Rodrigo Torralba, Hrishikesh Reddy Sanivarapu, Krzysztof Giwojno

Executive summary

Best Model: DistilBERT (fine-tuned) — **98.14% accuracy**, F1: 0.9814

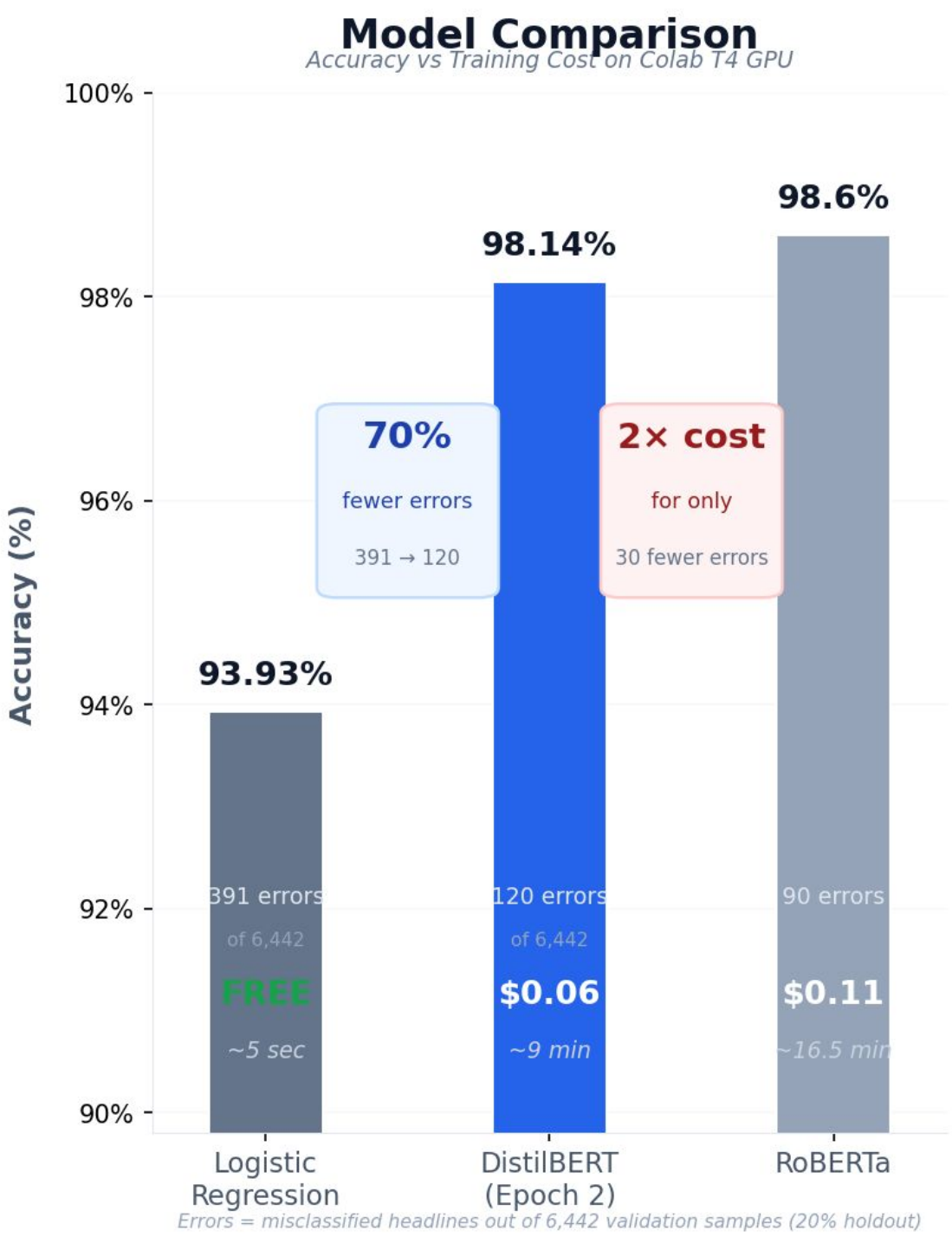
- Correctly classifies all but ~120 out of 6,442 validation headlines
- Selected Epoch 2 checkpoint (lowest validation loss, best generalization)

Expected on unseen data: ~97-98%

Alternatives Considered:

- Logistic Regression: 93.93% — excellent zero-cost baseline
- RoBERTa: 98.60% — double the cost for only +0.46% accuracy gain

Key decisions: Deduplication removed 1,950 headlines (5.7%). Minimal preprocessing outperformed complex pipelines across all models.



Methods (preprocessing)

Dataset: 34,152 headlines (51.5% fake / 48.5% real)
→ 32,206 after deduplication

Used (helped):

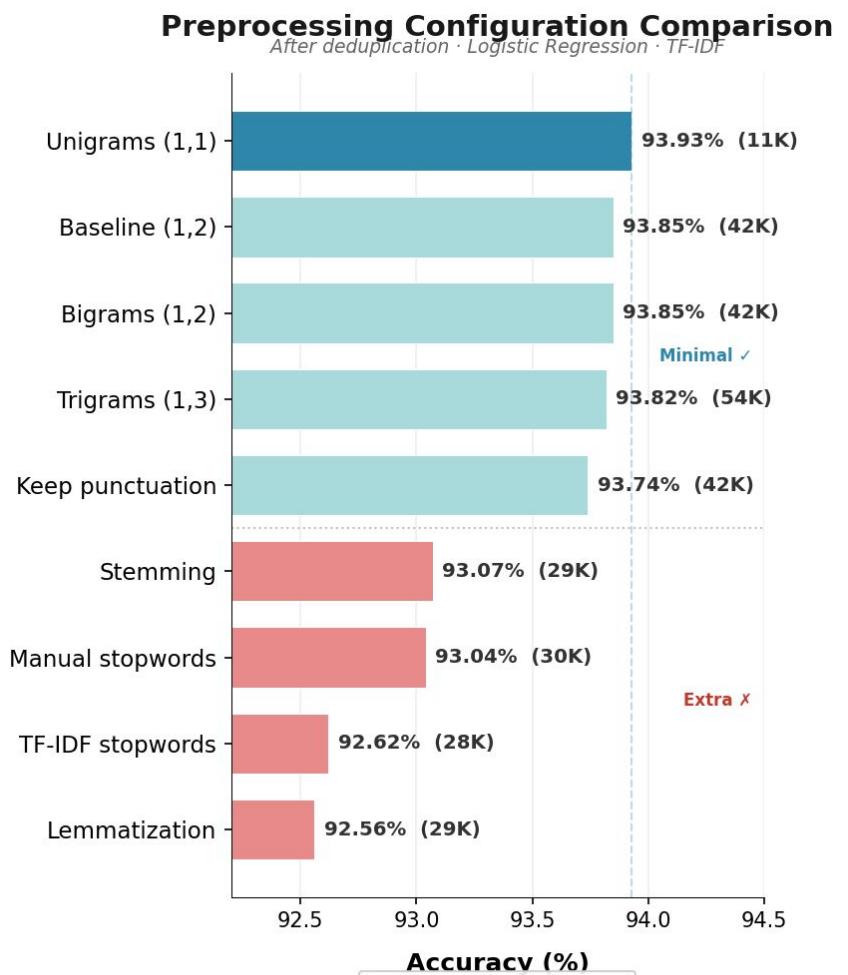
- Tab character in the title (appeared only in fake news) - good predictor
- Apostrophe encoding fixes

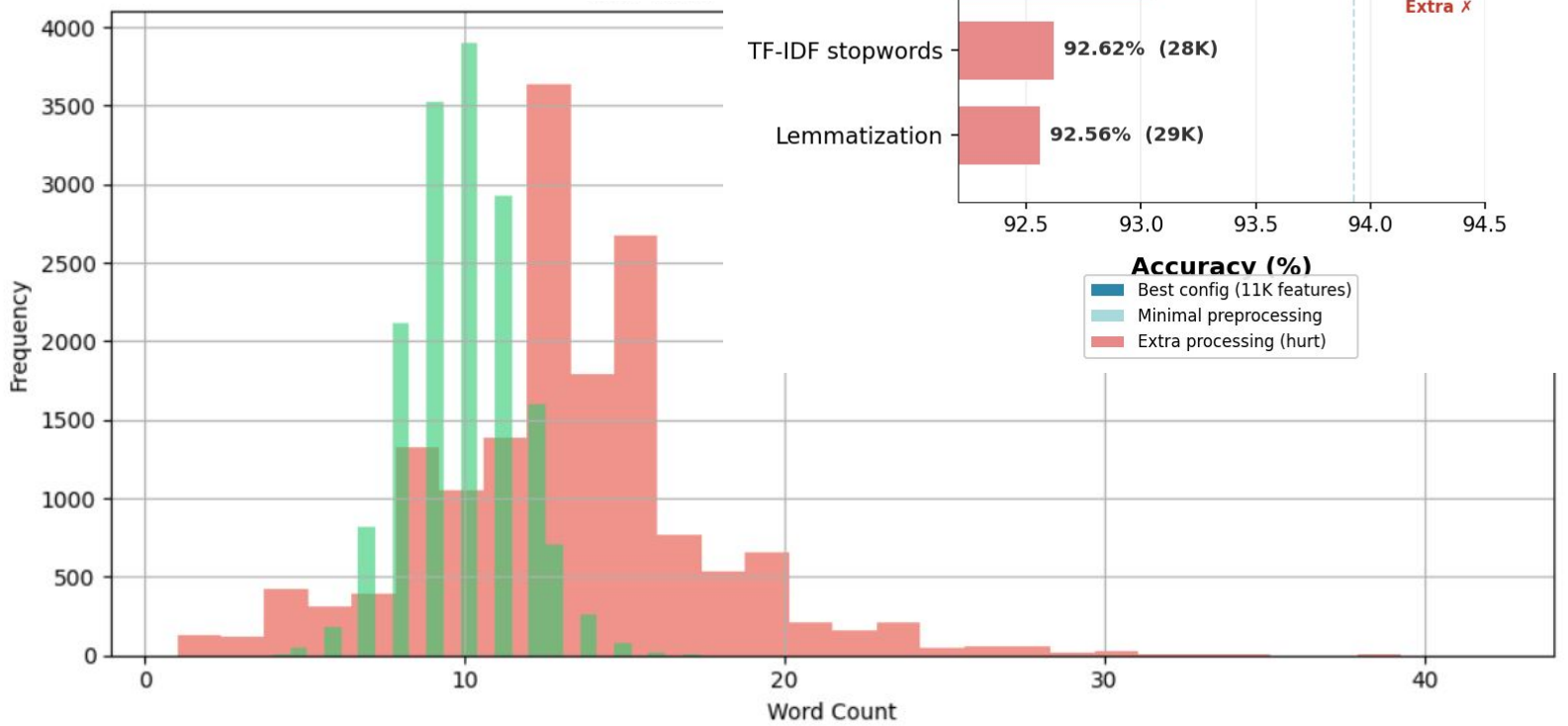
9 preprocessing configs tested via Logistic Regression:

Tested and discarded:

- Stopword removal — actually **hurt** performance (words like "the", "is" are predictive)
- Stemming / Lemmatization — no improvement
- Bigrams — worse than unigrams alone

Key finding: Simplest pipeline won every time. This informed our decision to keep transformer preprocessing minimal too.





Methods (models)

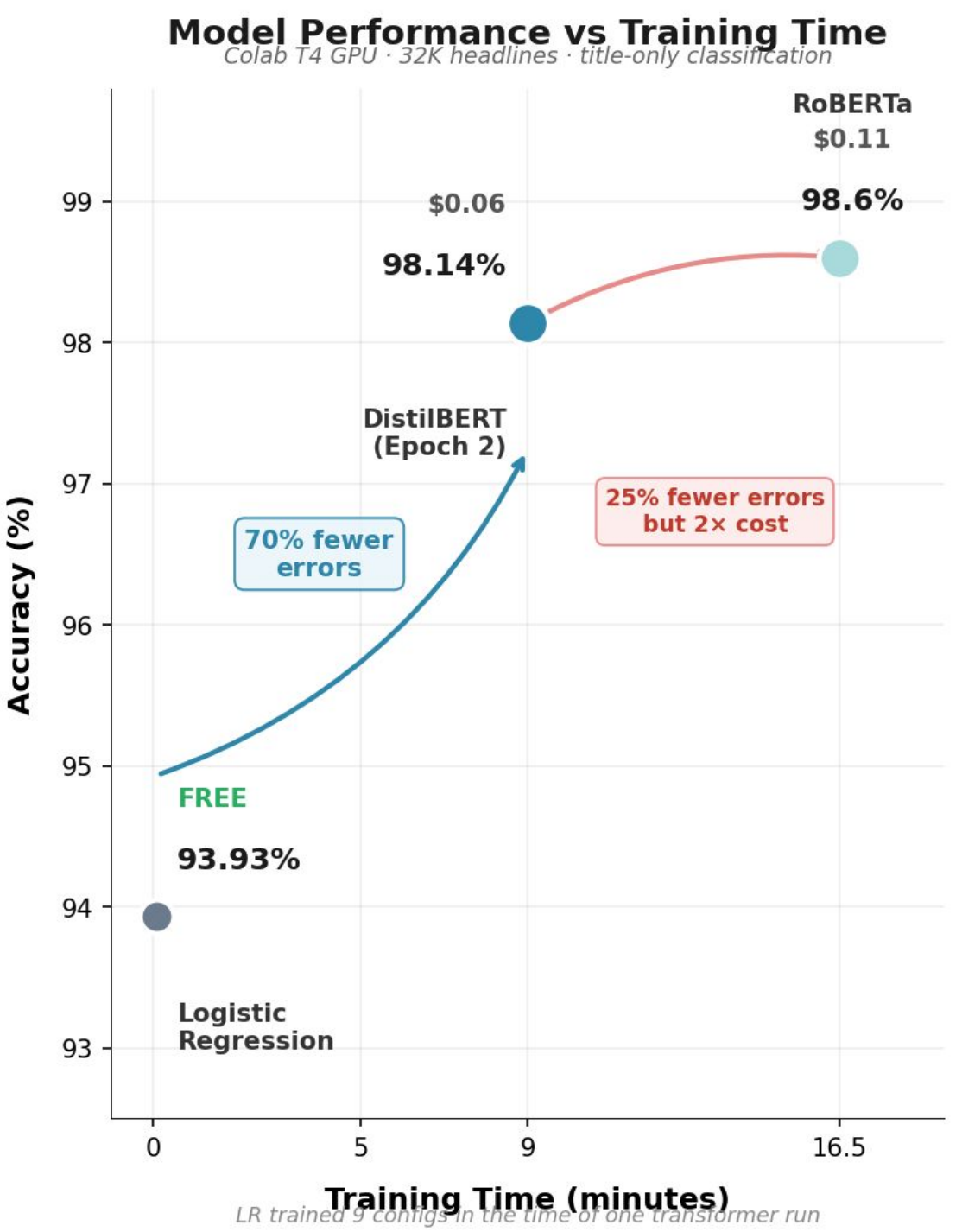
Three model families tested:

- **Logistic Regression:** 93.93% — free, CPU, ~5 sec, tested 9 configurations
- **DistilBERT (66M params):** 98.14% — ~$0.06, T4 GPU, ~9 min
- **RoBERTa (125M params):** 98.60% — ~$0.11, T4 GPU, ~16.5 min

The story is in the gaps:

- LR → DistilBERT: **70% fewer errors** — the transformative jump
- DistilBERT → RoBERTa: **25% fewer errors** at 2x cost — diminishing returns

LR deserves credit: At zero cost, we ran 9 experiments in the time one transformer takes. It taught us the data before we committed GPU resources.



Selected model

Why DistilBERT over RoBERTa:

- Double cost ($0.06 → $0.11) for only 24 fewer errors out of 6,442
- 40% smaller and faster at inference — matters at scale

Why Epoch 2 over Epoch 3:

- Epoch 2 validation loss: 0.0795 (lowest — best calibrated)
- Epoch 3 validation loss: 0.0946 (rising — overfitting signal)
- Accuracy gain Epoch 2→3: only +0.09%

Results: 3,142 TN / 63 FP / 51 FN / 3,186 TP

Honest limitation: Test set was used as validation set — reported accuracy is slightly optimistic. A proper 3-way split would give more conservative estimates.

Takeaways

Key Learnings:

1. **Less preprocessing = better results** — simplest pipeline won across all models
2. **Biggest gains come from first upgrade** — LR to DistilBERT cuts 70% of errors; DistilBERT to RoBERTa adds little at double the cost
3. **Watch validation loss, not just accuracy** — it caught overfitting that accuracy missed
4. **Start with fast baselines** — LR in 5 minutes shaped every decision we made after

Challenges: Test-as-validation setup in transformers, local GPU environment issues

Next steps: Proper 3-way data split, ensemble of LR + transformer, longer token sequences (currently 64)