✧ NLP CLASSIFICATION PROJECT

Fake or Real? Detecting News with NLP

Presented by:

Eya Cherif, Isis Hassan and Suzana Cracco

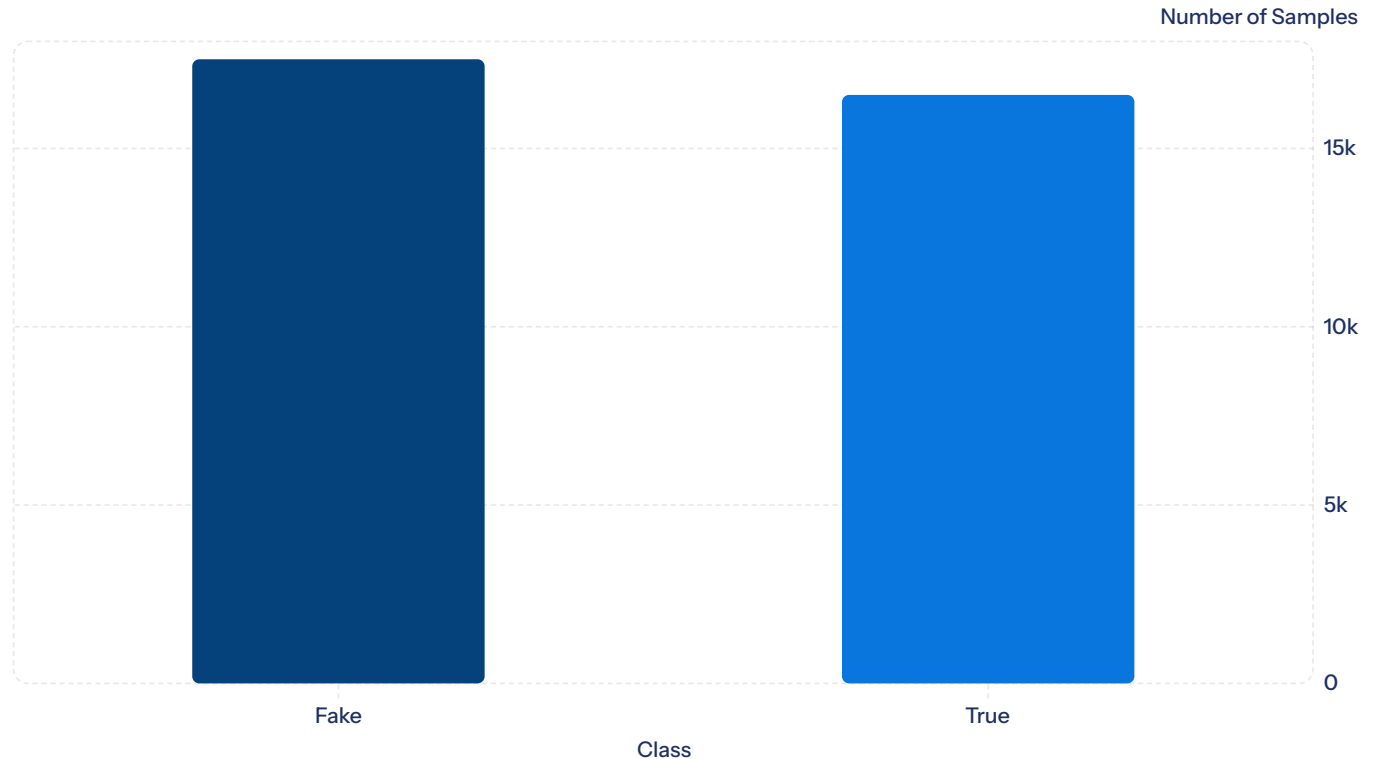Feb.2026

Problem Statement & Data

Can machines spot fake news?

Kaggle Fake News Detection dataset:
32,206 labeled articles combined

Custom test set of 100 headlines
50 real (2025 news from CNN, Wikipedia, Britannica) and 50 satirical (The Onion).

The distribution of samples per class shows a good balance in the training data.

Data Preprocessing Pipeline

Lemmatization
Optional: reduce words to their root form.

Tokenization
Split cleaned text into individual tokens.

Text Cleaning
Remove punctuation, stop words and lowercase text.

Vectorization
Convert tokens to BoW, TF-IDF, or BERT vectors.

Tokenization Techniques

Bag of Words

Counts word occurrences. Simple but loses word order and context.
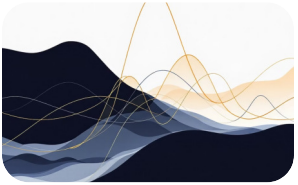
TF-IDF

Weighs terms by frequency and rarity. Better at surfacing distinctive words.
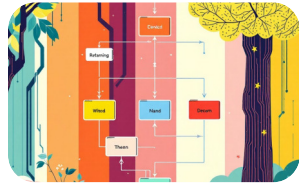
BERT Transformer

Contextual embeddings from a pretrained language model. Captures meaning and nuance.
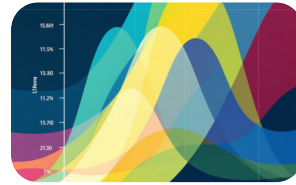
Model Architectures

Logistic Regression

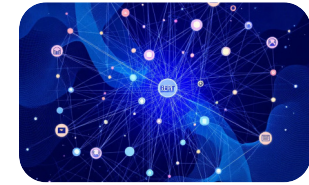Linear baseline. Fast training, interpretable coefficients.

XGBoost

Gradient-boosted trees. Handles sparse text features well.

MultinomialNB

Naïve Bayes for text. Strong with BoW/TF-IDF features.

BERT

Fine-tuned transformer. Learns deep contextual representations.

Made with GAMMA
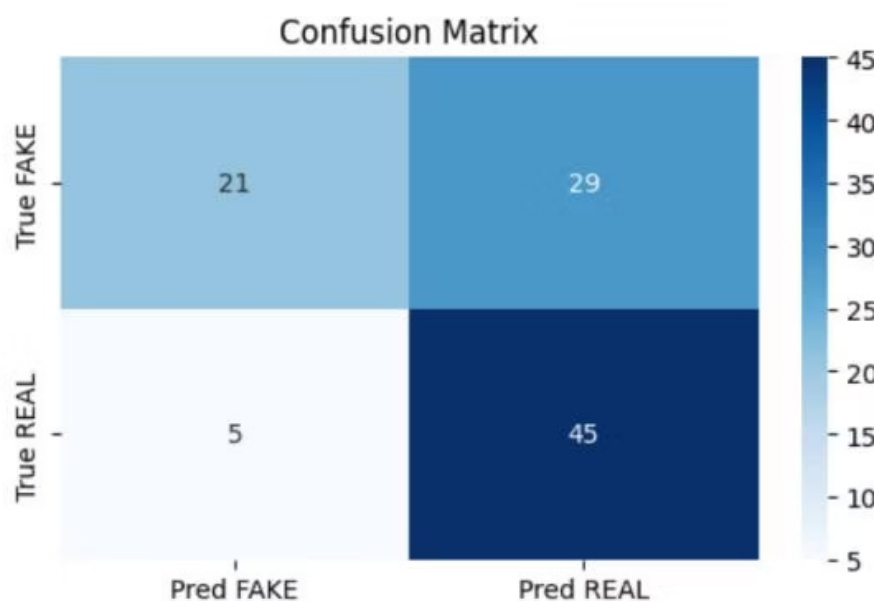
Validation vs Test Performance

All models achieved **high validation accuracy (90–98%)**, but performance dropped on our custom 2025 test set: **Generalization challenges**

| Model	Val Acc (80 Train/20 Test Split)	Test Acc (generated dataset)
TF–IDF + XGB (NoLem)	0.88	0.30
TF–IDF + XGB (Lem)	0.90	0.64
BoW + XGB (NoLem)	0.93	0.42
BoW + XGB (Lem)	0.94	0.39
BERT (bert-base-uncased)	**0.98**	0.64
BERT + Swear Flag	**0.98**	0.66
TF–IDF + LogReg	0.93	0.69
BoW + LogReg (tuned)	0.94	**0.73**
TF–IDF + LogReg	0.93	0.69
Naive Bayes (BoW)	0.96	0.31
Naive Bayes (TF–IDF)	**0.98**	0.3
DistilBERT (128)	0.78	—
DistilBERT (256)	**0.98**	—

Confusion Matrices — Key Models

Confusion Matrices (Test)

BERT + Swear Flag

LogisticRegression (BoW)
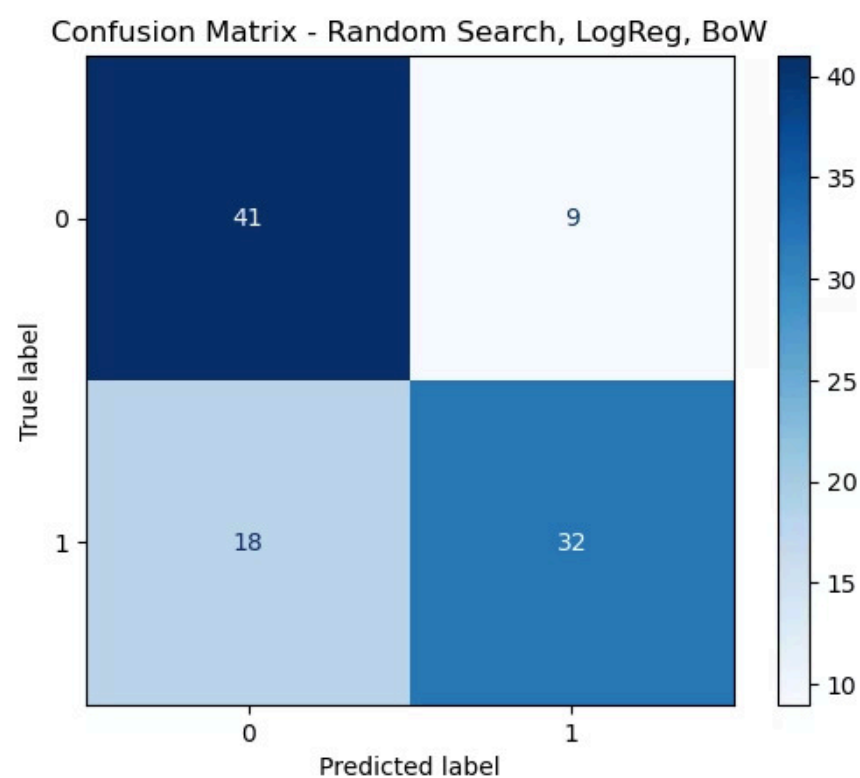
Key Takeaways

1 **Validation ≠ Real-World Performance**

All models scored 90–98% on validation but dropped significantly on unseen 2025 headlines, revealing the generalization gap.

2 **Preprocessing Matters (Selectively)**

Lemmatization was critical for XGBoost (30% → 64% accuracy), but other models achieved good results with minimal preprocessing.

3 **Simplicity Often Wins**

Logistic Regression achieved the best test accuracy at 69%, outperforming more complex models.

4 **BERT Shows Promise**

Transformers outperformed traditional methods, and adding a swear word flag improved generalization.

5 **Next Steps**
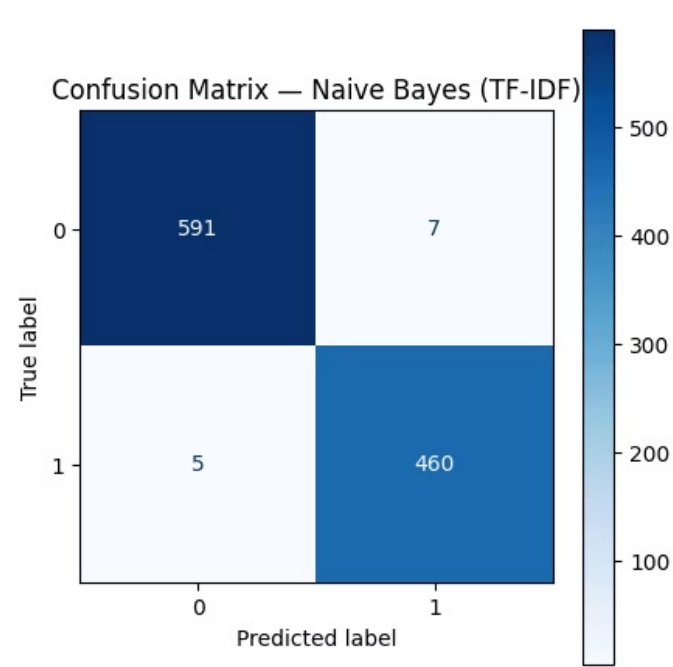
Larger, more diverse training data and domain adaptation techniques could close the generalization gap.
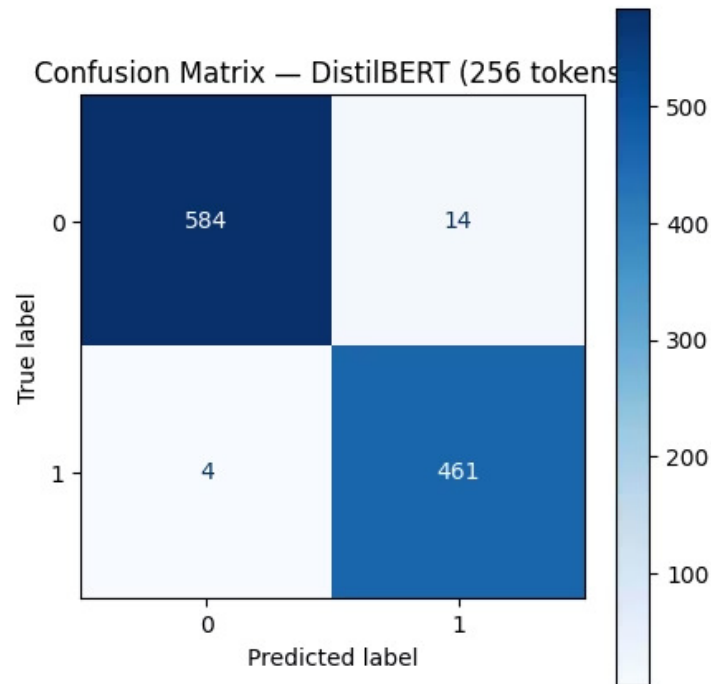
Appendix

Confusion Matrices — Key Models
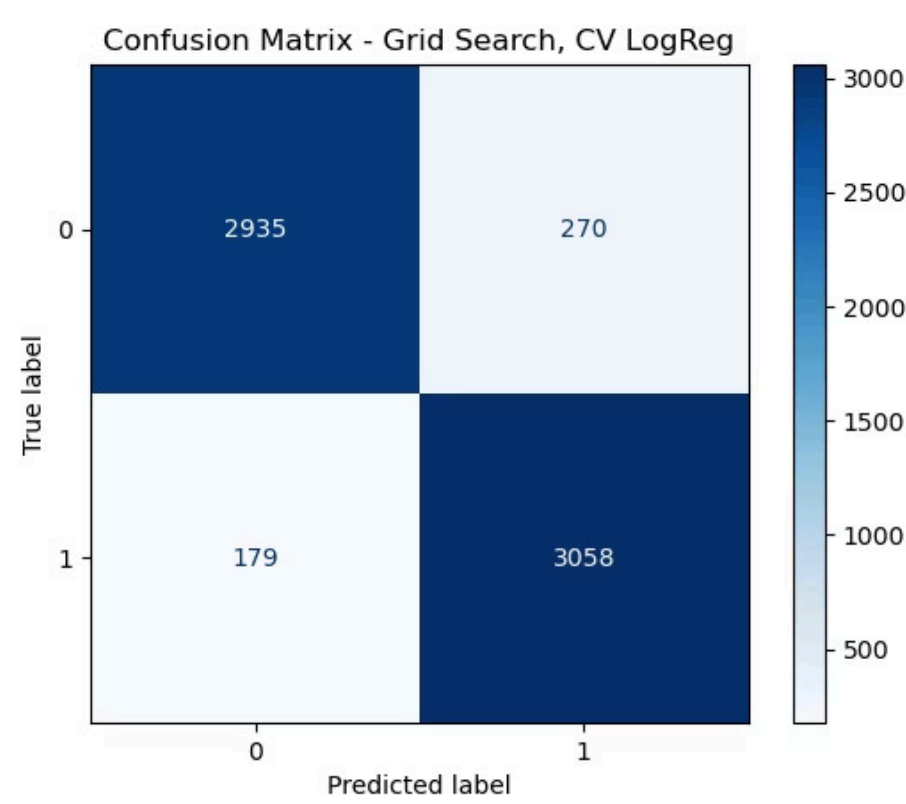
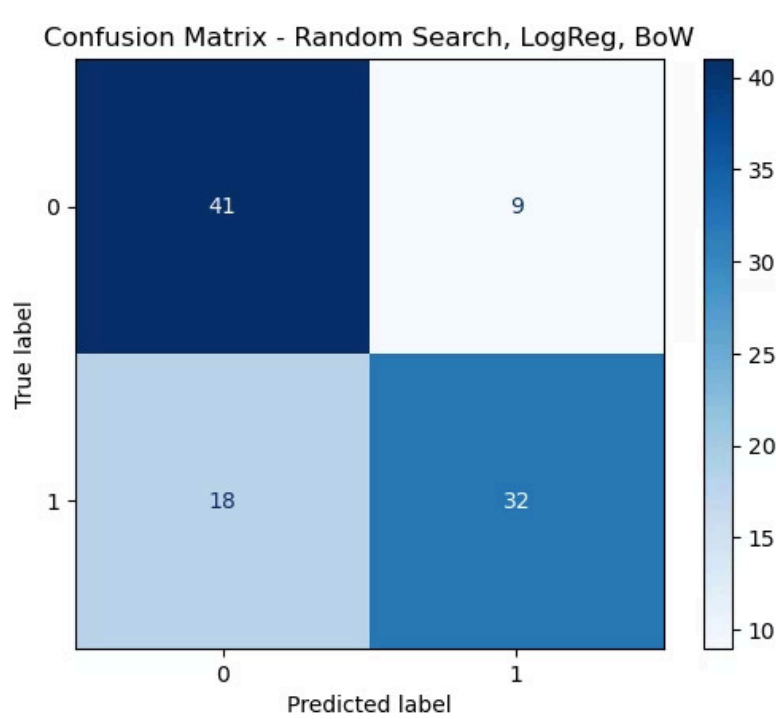## Confusion Matrices (Validation)

Naive Bayes (TF-IDF)

DistilBERT (256 tokens)

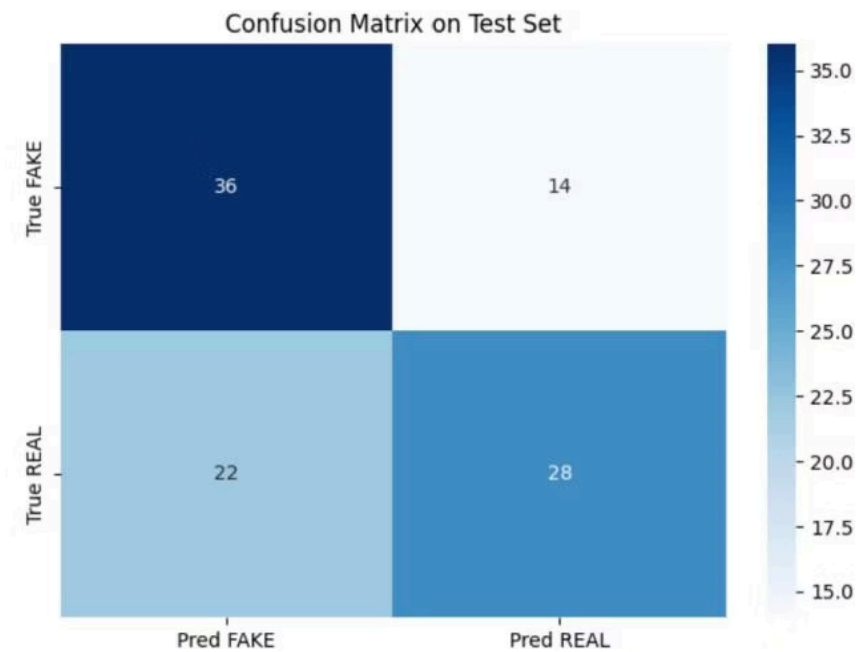### LogisticRegression (Bow) - Val

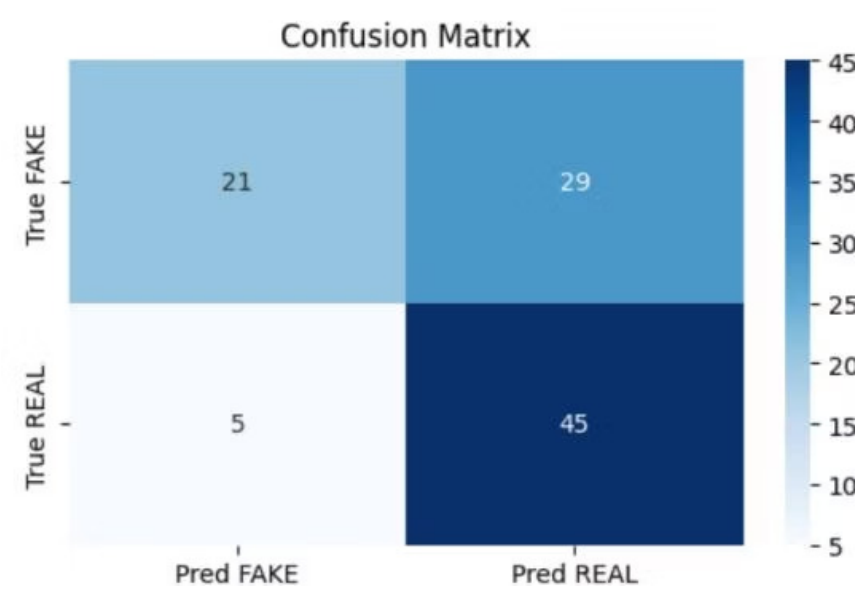### LogisticRegression (Bow) - Test - keep

Confusion Matrices (Test)

TF-IDF + XGB (Lem)


BERT + Swear Flag — Test

Additional Model Experiments

To explore model performance further, we evaluated additional Naive Bayes and DistilBERT configurations. These results highlight their validation accuracy and class-specific recall before deployment to our custom test set.

Model Performance Metrics (Validation)

Model	Val Acc	F1-Score	Recall (Class 0)	Recall (Class 1)
Naive Bayes (BoW)	0.960	0.960	0.980	0.980
Naive Bayes (TF-IDF)	0.980	0.990	0.980	0.990
DistilBERT (128)	0.780	0.780	0.640	0.980
DistilBERT (256)	0.980	0.980	0.980	0.990

Validation F1-Scores
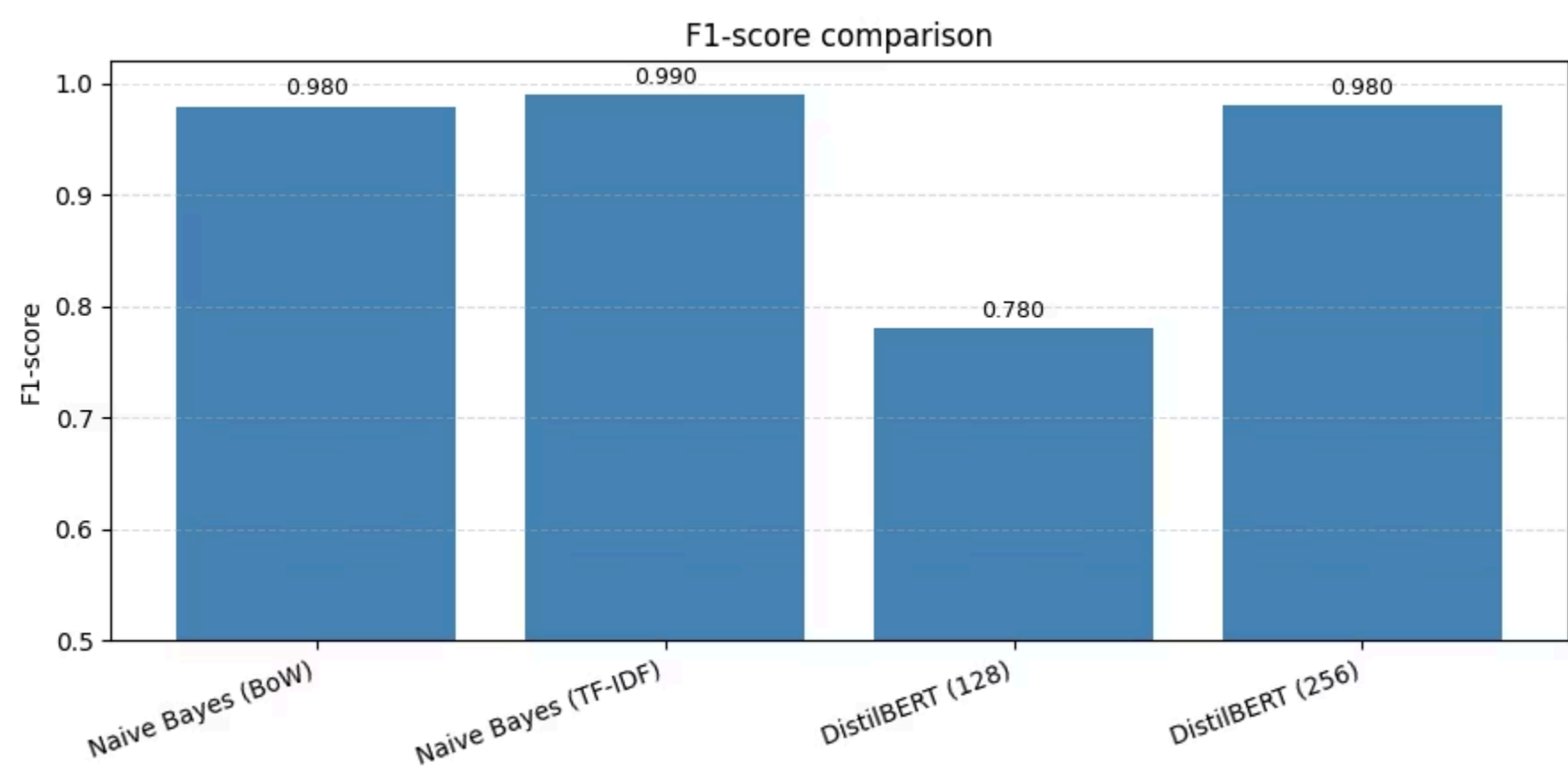
DistilBERT models, especially with longer sequence lengths (256 tokens), demonstrated superior F1-scores during validation.

Made with GAMMA

Our Custom Test Headlines

Real Headlines (2025)

- Trump inaugurated as 47th President
- LA wildfires burn 50,000+ acres
- Nvidia loses ~$600B after DeepSeek
- Israel–Hamas ceasefire agreement

Satirical Headlines (The Onion)

- "Man Who Bumped Tesla... Sentenced To Death"
- "Archaeologists Uncover Ancient Race Of Skeleton People"
- "No Way To Prevent This, Says Only Nation..."
- "Zoo Gorilla Looks Bored Reading Wuthering Heights"

50 real headlines from CNN, Wikipedia, Britannica/CBS · 50 fake from The Onion across politics, science, culture.