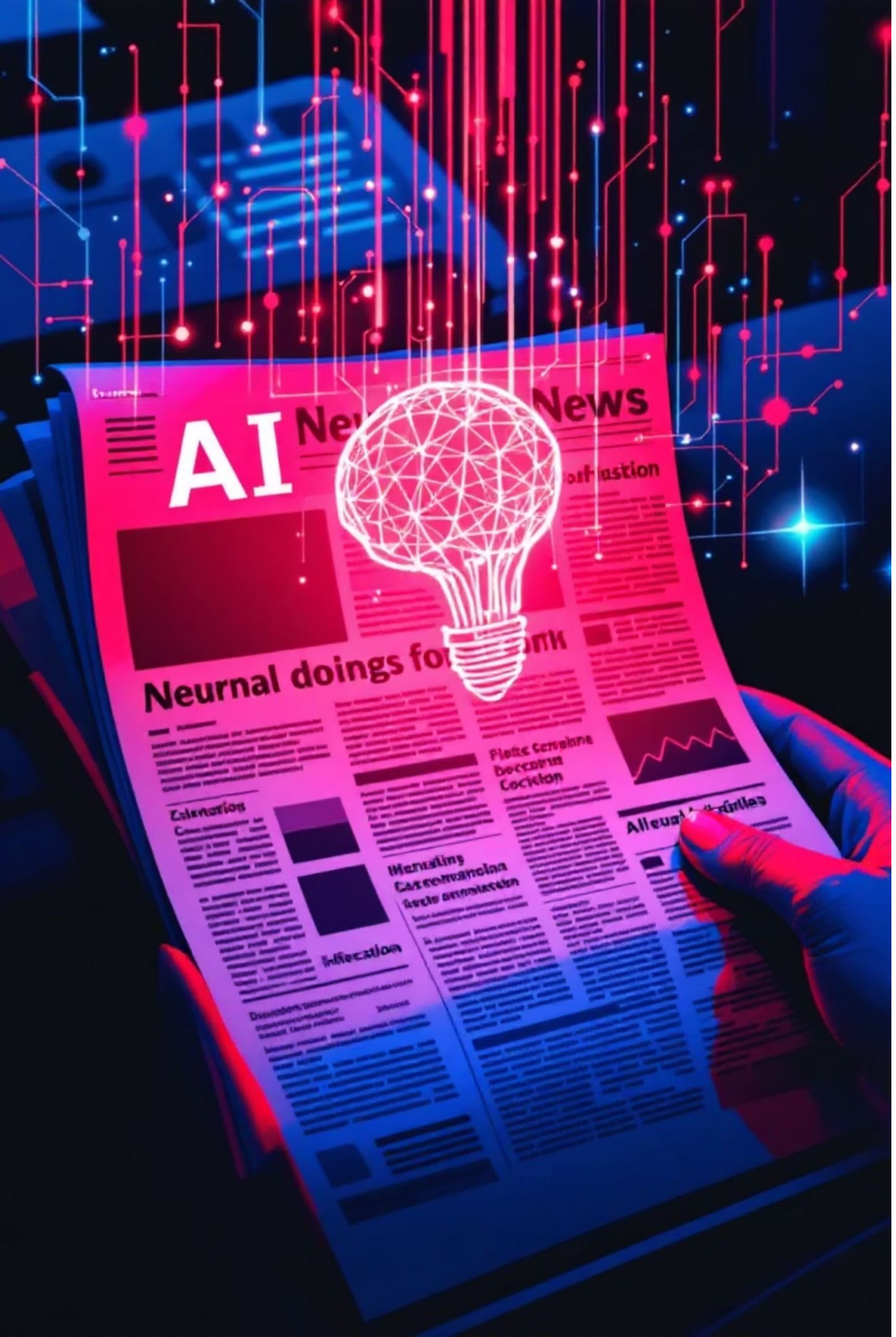


Fake vs Real News Detection

Using Natural Language Processing

Team:
Chantal Silva
Girish Kumar
Mitesh Parab

20 November 2025

Can Machines Detect Clickbait & Fake News from Titles Alone?

34,152 Labeled Titles

Training dataset with fake and real news classifications

9,984 Test Titles

Unlabeled data to predict with highest accuracy

The Challenge

Titles are short, sensational, and full of deceptive tricks

Data Snapshot

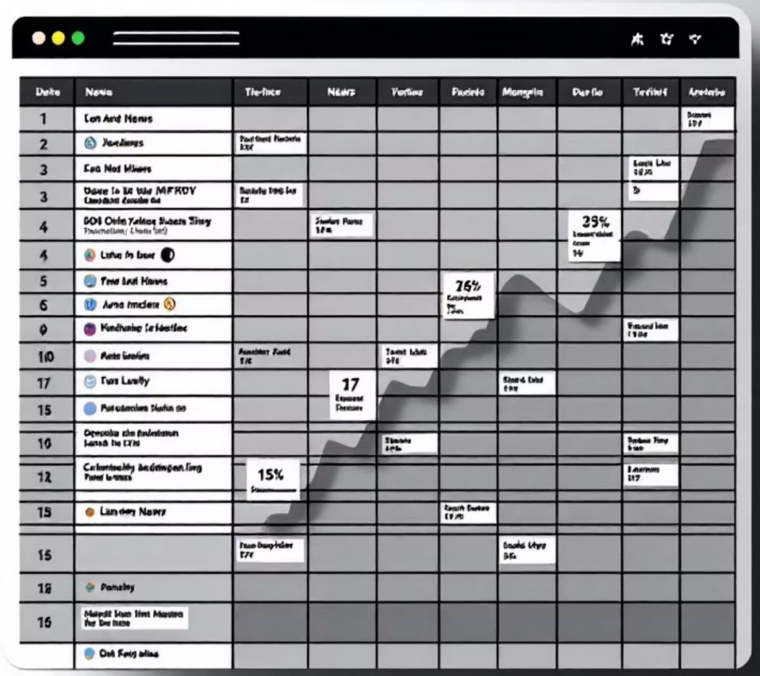
Training Data

- **34,152 labeled news titles**

- Fake titles use "wow!", "shocking", "boom!", all caps

- Real titles are neutral and factual

  "germany's fdp look to fill schaeuble's big shoes"



Test Data

- 9,984 unlabeled titles
- Unknown labels (marked as 2) need prediction

Data Preprocessing

- Lowercasing
- Punctuation
- Whitespace
- Stopword removal
- TF-IDF / BOW / Word2Vec vectorization

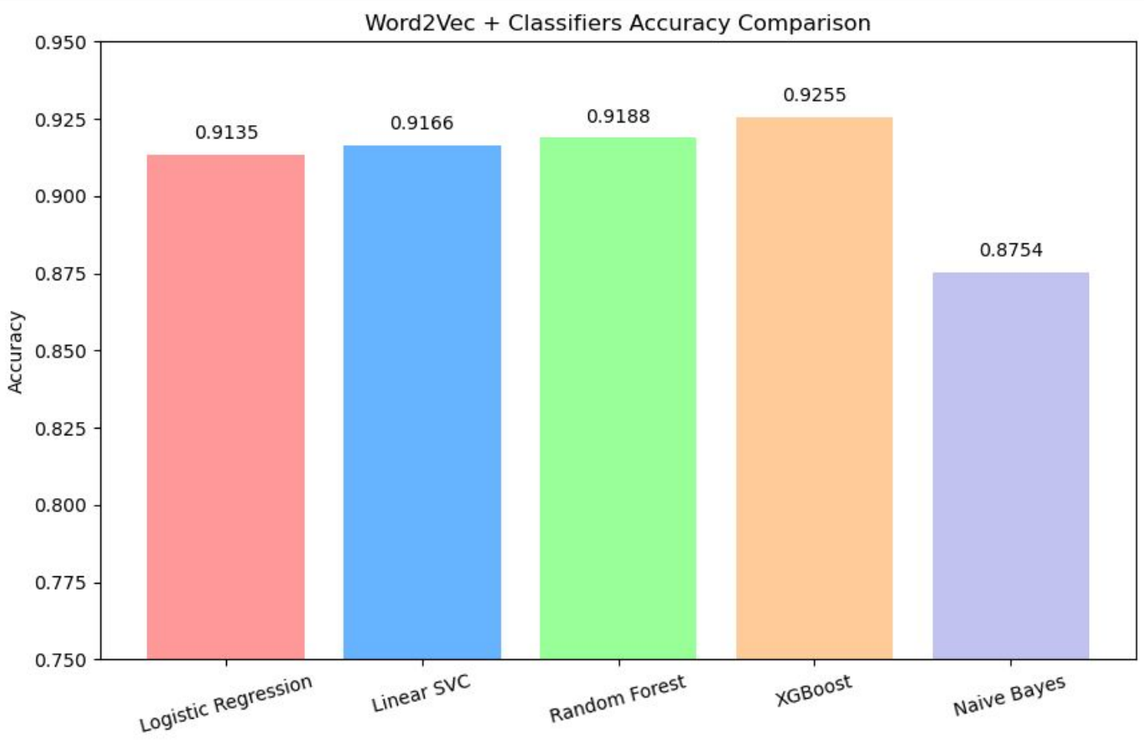
Model Selection

- LogisticRegression
- MultinomialNB
- LinearSVC
- RandomForest
- XGBoost

Word2Vec – Learning Word Meaning from Context

- Learns 300-dim, similar words = vector
- Trained on 34k titles to capture word relationship
- Tested across multiple models

Word2Vec – Learning Word Meaning from Context



Classification Report - XGBoost (Word2Vec)

	precision	recall	f1-score
Fake News	0.9236	0.9323	0.9279
Real News	0.9275	0.9183	0.9229
accuracy	0.9255	0.9255	0.9255
macro avg	0.9256	0.9253	0.9254
weighted avg	0.9255	0.9255	0.9255

Word2Vec - Validation on Unknown data

Predicted 4959 fake and 5025 real titles

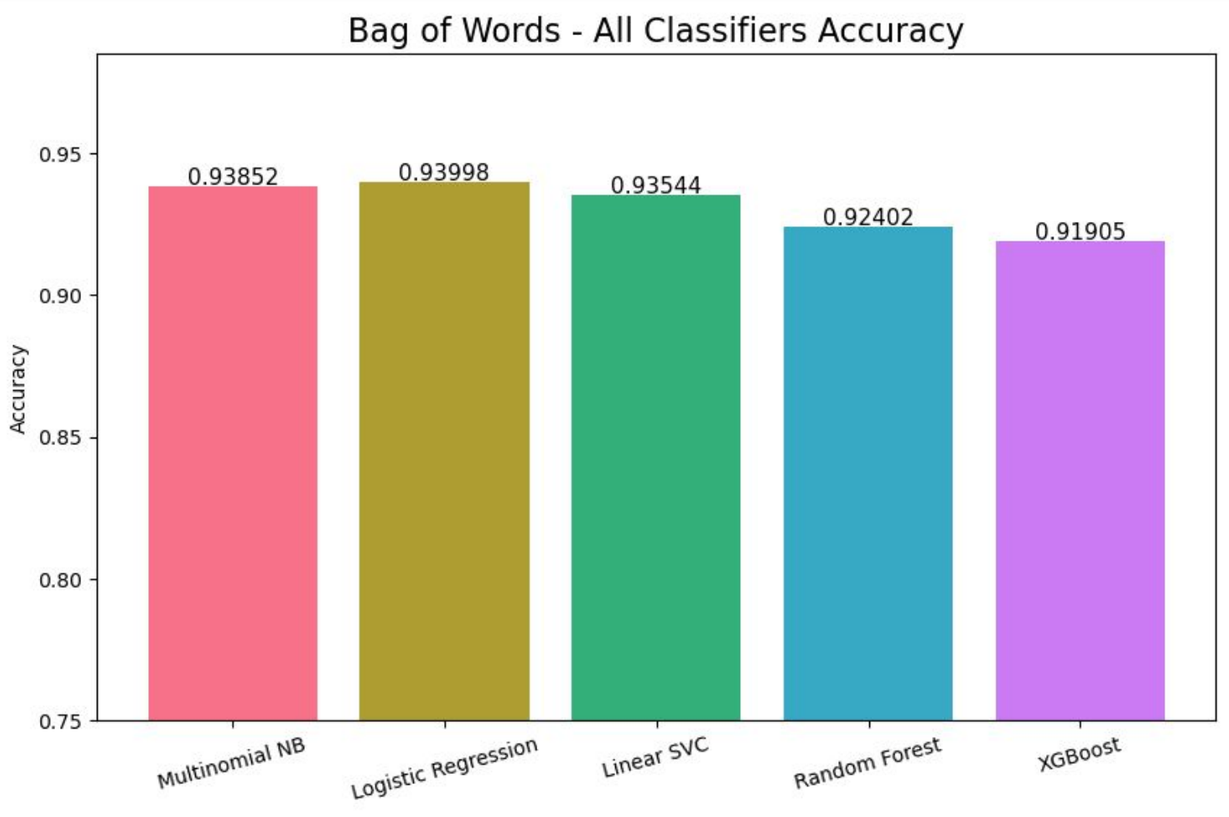
Bag of Words (BOW)

- Counts occurrences of words & n-grams parameters hypertuning
- No semantics — just exact matches
- Captures full phrases perfectly (“you won t believe”, “breaking news”)
- Preserves sensational wording exactly
- Trained and ran on

Bag of Words (BOW)

Classification Report - Logistic Regression (BOW)

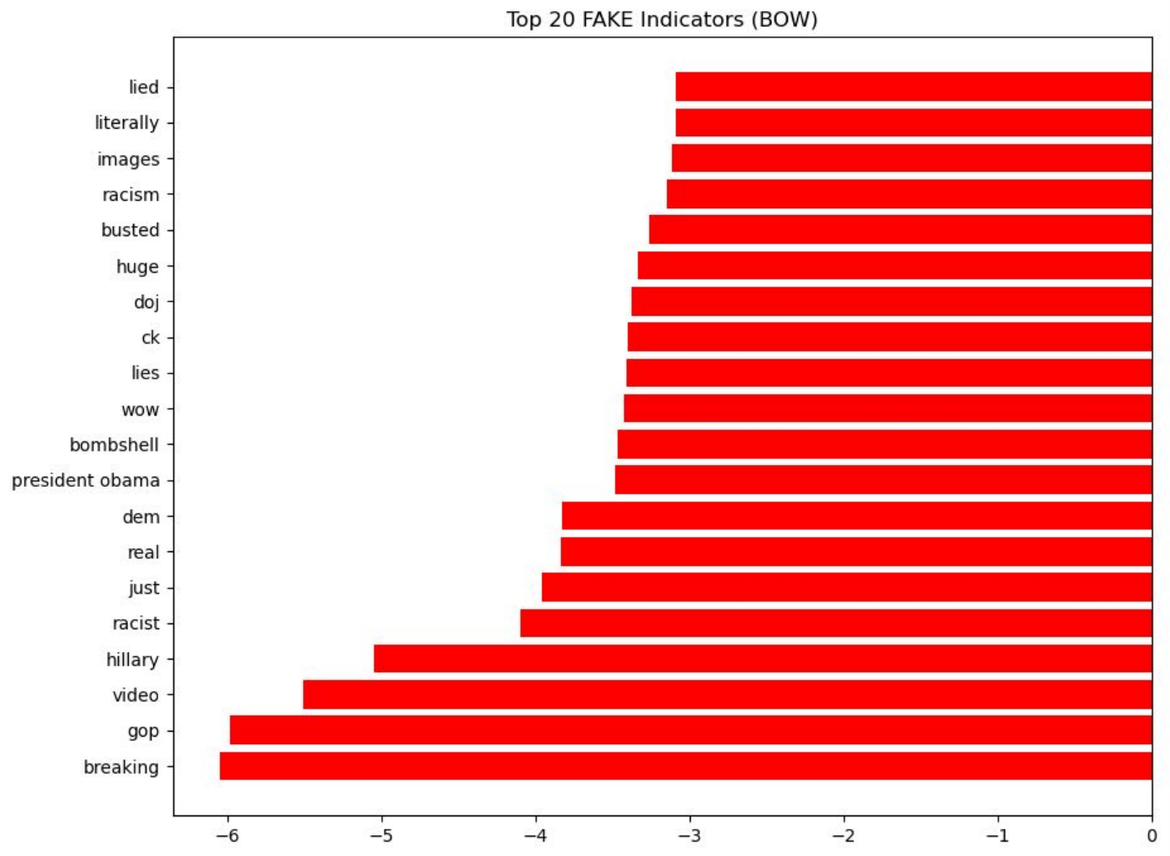
	precision	recall	f1-score
Fake News	0.9525	0.9297	0.941
Real News	0.9274	0.9508	0.939
accuracy	0.94	0.94	0.94
macro avg	0.9399	0.9403	0.94
weighted avg	0.9403	0.94	0.94

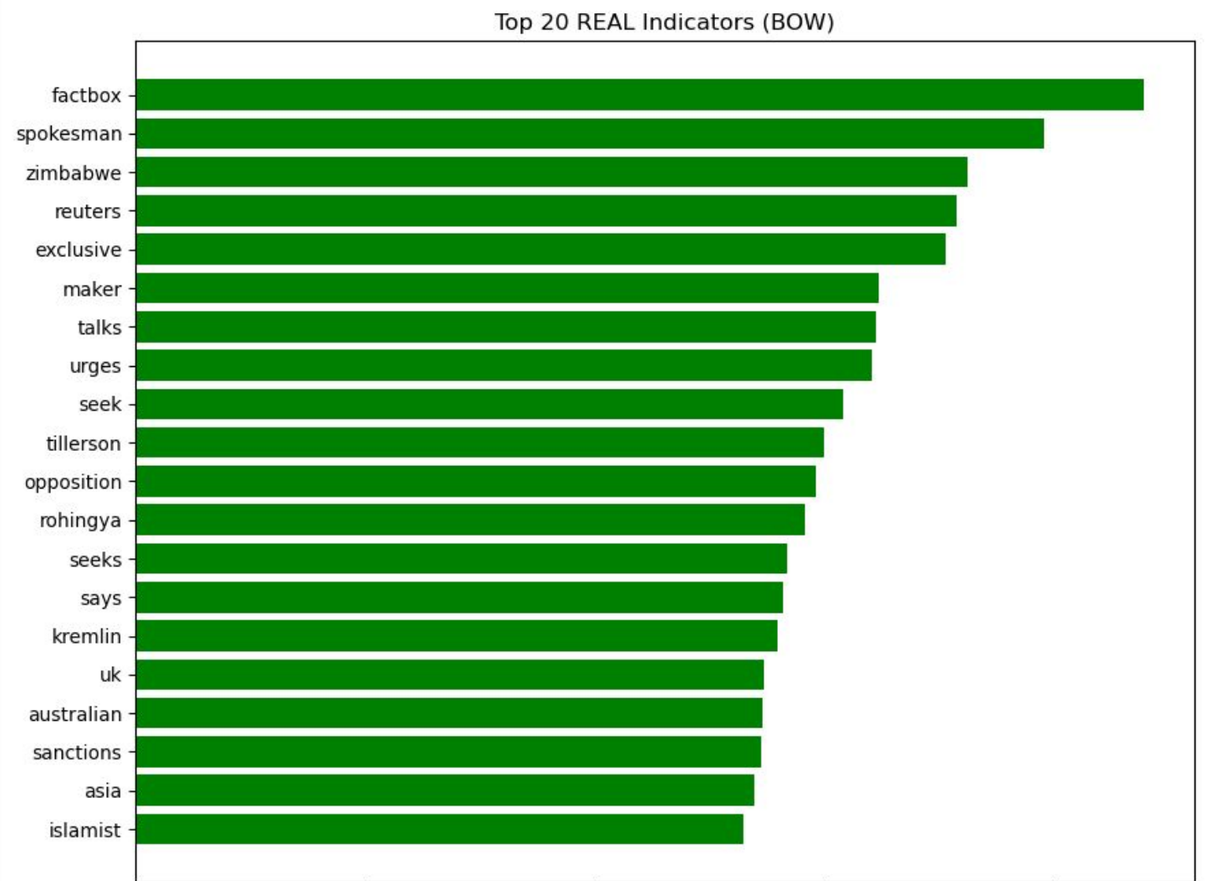


Bag of Words (BOW) - Validation on Unknown data

Predicted 4733 fake and 5251 real titles

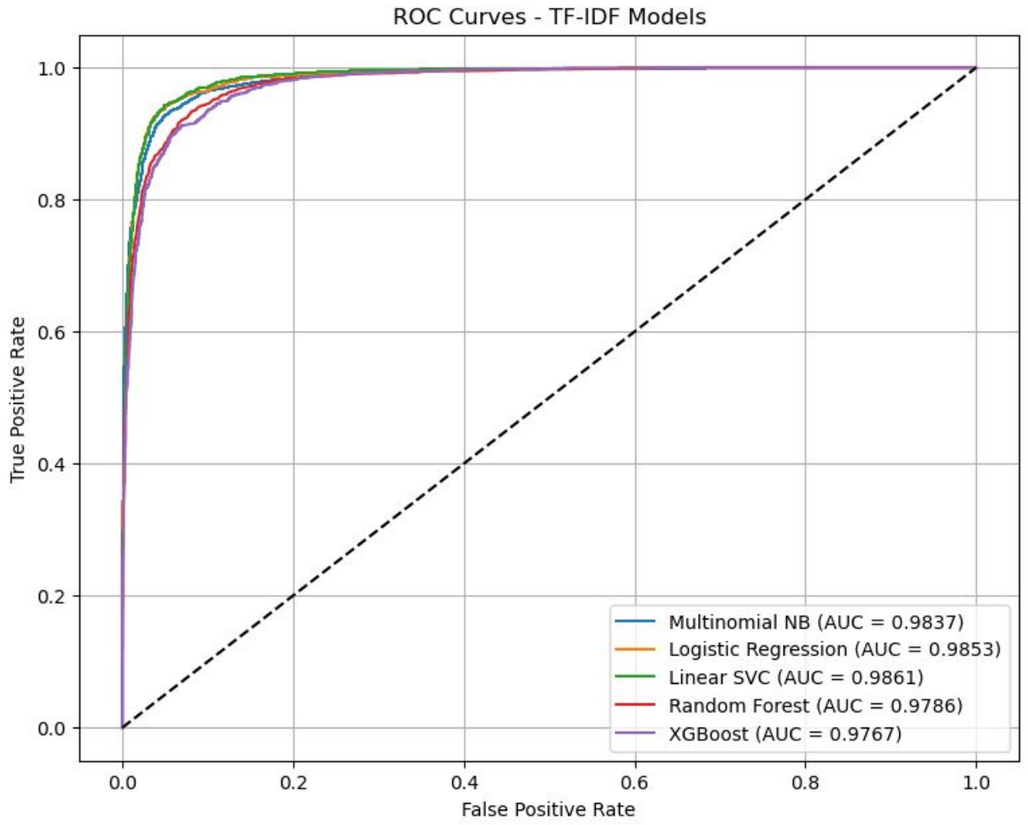
Bag of Words (BOW)





TF-IDF

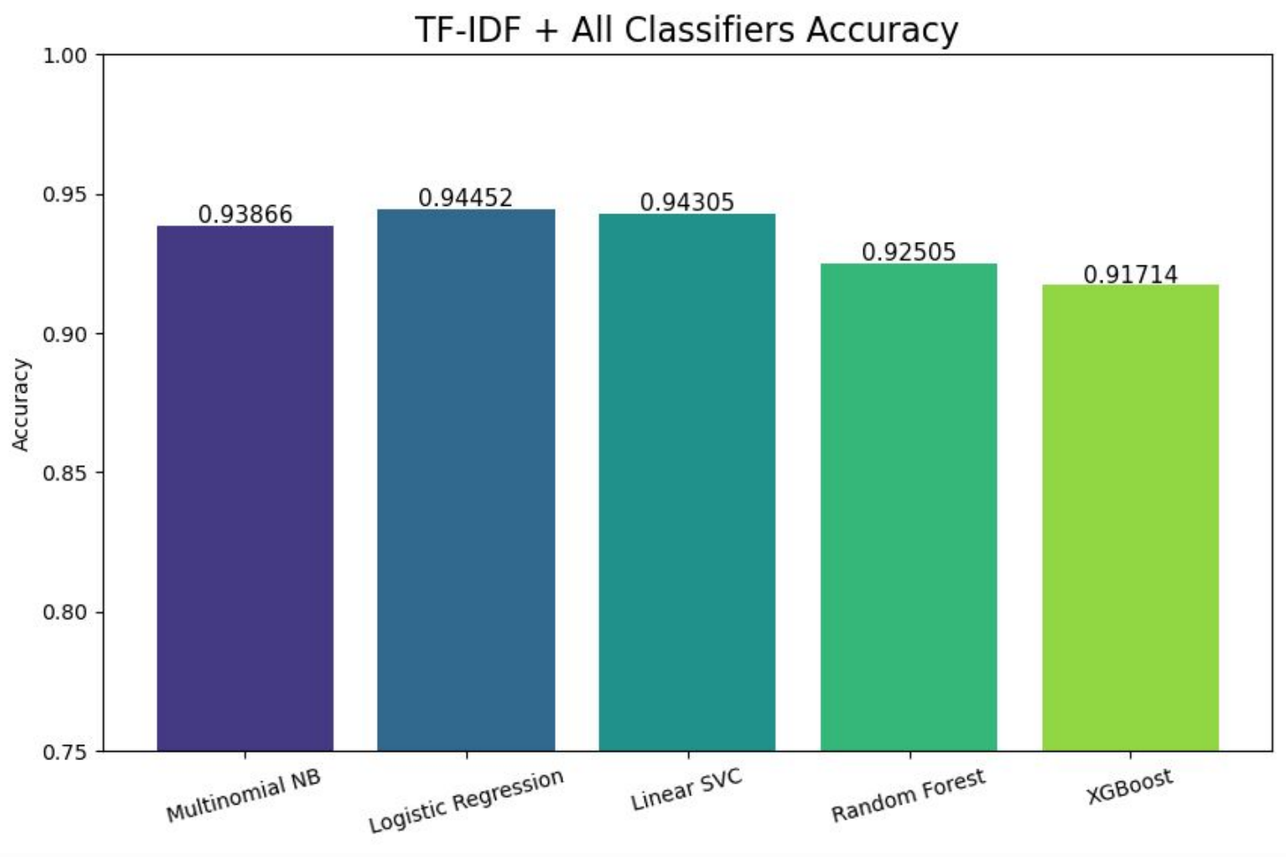
- Counts n-grams (1-3) like BOW
- Weights rare/discriminative terms higher (“boom!”, “shocking” → high score in fake) (“says”, “minister” → high score in real)
- sublinear_tf + stop_words = clean signal



TF-IDF + All classifiers

Classification Report - Logistic Regression

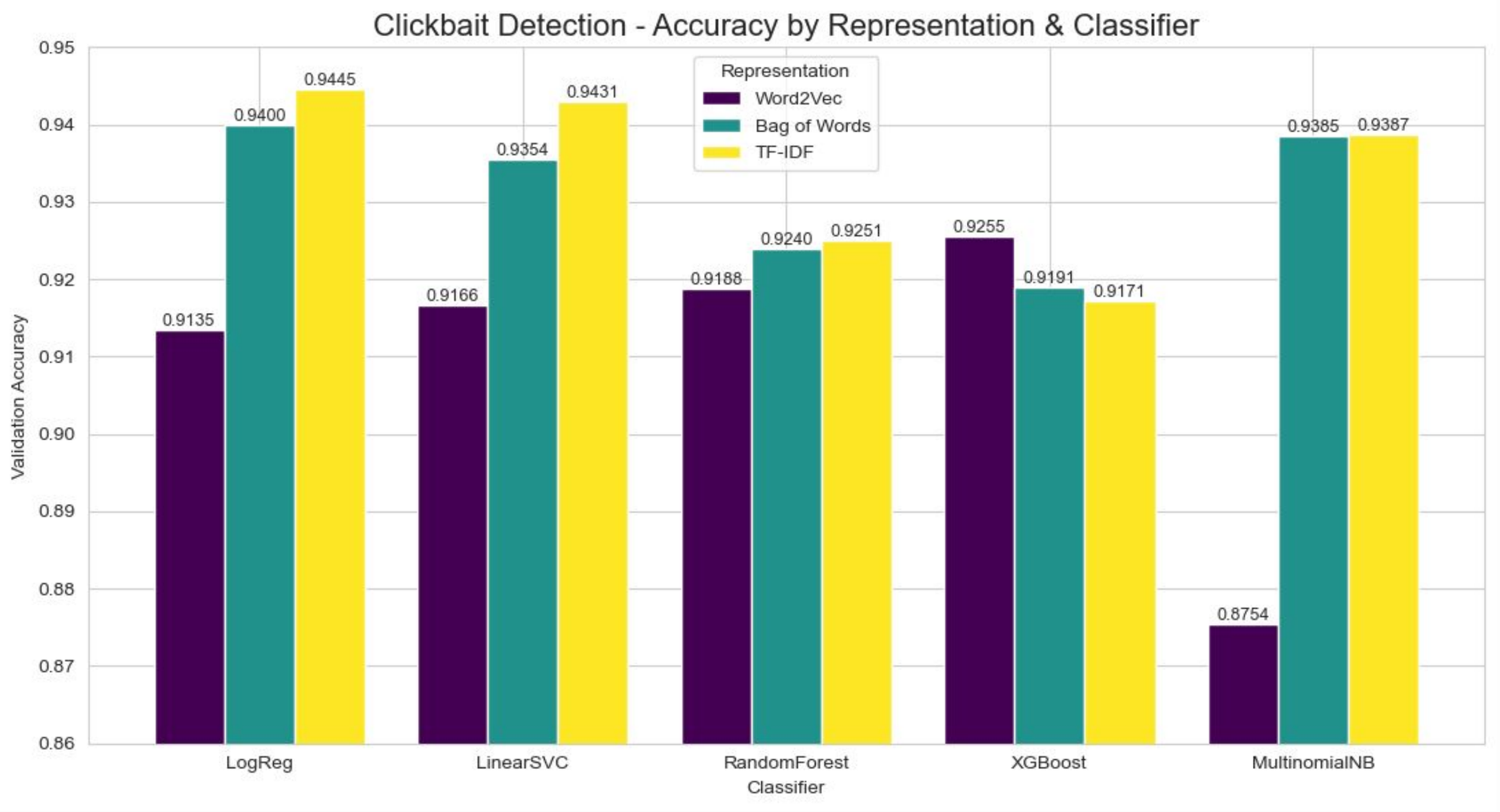
	precision	recall	f1-score
Fake News	0.9511	0.9405	0.9458
Real News	0.9377	0.9487	0.9432
accuracy	0.9445	0.9445	0.9445
macro avg	0.9444	0.9446	0.9445
weighted avg	0.9446	0.9445	0.9445



TF-IDF - Validation on Unknown data

Predicted 4751 fake and 5233 real titles

Accuracy Comparison



Top Fake vs Real Indicators

Top FAKE Signals

- boom!
- says
- wow!
- shocking
- you won't believe
- holy moly
- hilarious
- breaking:

Top REAL Signals

- minister
- government
- according to
- president
- reuters
- official
- statement
- announced

Thank you!