

Fake News Detector 🕵️ (ML-Powered Fact-Checking System)

Key Stats


- 44,898 news articles analyzed (52% fake / 48% real)
- **99.8% accuracy** with Random Forest model
- Only **0.3% false positives** in testing

Tech Stack

- Python + NLTK (text cleaning)
- TF-IDF (5,000 features)
- Logistic Regression (98.7% acc) vs. Random Forest (99.8% acc)

Live Example

"Scientists found a human planet"

 **FAKE** (91% confidence)

Missing sources • Sensational claim • Matches fake patterns




Fake News Dataset Analysis

Dataset Composition




- **44,898 articles** total
- **52.3% Fake** (23,481) | **47.7% Real** (21,417)
-  *Nearly balanced – ideal for training*

Key Characteristics



Fake News (❌):

-  *Sensational headlines* ("BREAKING...")
-  *Opinionated language* ("chaos")
-  *Political focus* (75% of samples)



Real News (✅):

-  *Neutral tone* ("pledges to defend")
-  *Cited sources* (e.g., Reuters)
-  *Specific dates/locations*
-

Preprocessing Needs

-  *Clean*: CAPS, special chars ("Enoug-"), citations
-  *Standardize dates* (July 21 vs Apr 25)

Predictive Potential

-  *Headlines = Strong predictor*
-  *Subject tags add context* (e.g., "politics")



Data Split & Feature Analysis



Train-Test Split

- **35,918 samples** (80%) → Training 🏆
- **8,980 samples** (20%) → Testing 🔍
- ✅ *Ideal for reliable evaluation*



Feature Engineering

- **5,000 TF-IDF features** 📏
- *High-dimensional* but sparse → 🎯 **Logistic Regression recommended**
- ⚠️ Watch for overfitting $*(1:7 \text{ feature-to-sample ratio})^*$



Optimization Tips

- Tune `max_features` to reduce dimensionality
- Apply **L2 regularization** for stability
- Consider PCA if speed critical ⌚

Logistic Regression Performance





Key Metrics

- **98.7% Test Accuracy** (Near-perfect)
- **F1-Score: 0.99** (Both classes)
- **1.6% False Positives** ❌ → Fake news missed
- **1.0% False Negatives** ✅ → Real news flagged

Error Analysis

Actual \ Predicted	Fake ❌	Real ✅
Fake ❌	98.4%	1.6%
Real ✅	1.0%	99.0%

Recommendations

- **Boost Precision:** Add n-grams 
- **Tune Thresholds:** >95% confidence for "real" 
- **Human Review:** 1-2% borderline cases  



Random Forest Model - Supreme Performance

100 Record-Breaking Accuracy

- **99.78% Test Accuracy** (Beats LR's 98.69%)
- **0.3% False Positives** ❌ (Only 14 mistakes)
- **0.1% False Negatives** ✅ (Just 6 errors)

Error Comparison

Metric	Random Forest 🟢	Logistic Regression 🟡
Total Errors	20	118
FN Rate	0.1%	1.0%



Watchouts & Next Steps

- **Check for overfitting** 🤔 (100% train accuracy)
- **Production-ready** 🏢 for critical applications
- **Try on harder datasets** 💪





Fake Science News Detection

Suspicious Headline





"New planet could support human life"

 **Fake (90.94% confidence)**

Detection Clues

-  *Vague claims* ("new planet" unnamed)
-  *Sensational language*
-  *No credible sources* (NASA/ESA missing)
-  *Matches known fake patterns*

Verification Steps

1. Check NASA.gov 
2. Search peer-reviewed journals 
3. Consult astronomer forums  

9.06% Uncertainty

Could be premature real discovery



Fake News Detection - Ultimate Results



Model Showdown

	Accuracy	FP Rate	FN Rate
Random Forest 🌳	99.78%	0.3%	0.1%
Logistic Reg. 📊	98.69%	1.6%	1.0%



Case Study

"New planet supports life" →

❌ **Fake (91% confidence)**

- 🚩 No credible sources
- 📢 Sensational claim



Key Insights

- 🌳 **RF best for accuracy** (14 errors only)
- 📊 **LR better for debugging**
- 🔍 **Human review** for 85-95% confidence



THANK YOU!

