Explainable Fake News Detection

Comparing Intrinsically Interpretable vs Post-Hoc Explainability

Decision Trees vs BERT + SHAP

Soham Sharma

Purdue University | sharm710@purdue.edu

Presentation Outline

1. Background & Motivation
- The misinformation crisis and why explainability matters
- Current research landscape

2. Research Question & Hypotheses
- Intrinsic vs post-hoc explainability
- Expected trade-offs

3. Two Approaches
- Approach A: Decision Trees with SHAP (intrinsically interpretable)
- Approach B: BERT with SHAP (post-hoc explainability)

4. Experimental Design & Evaluation

5. Expected Outcomes

The Misinformation Crisis

Why This Matters: Fake news poses serious societal risks

- COVID-19 pandemic: Fake medical treatments endangered lives
- Political impact: Top 20 fake news stories during 2016 US election reached more people than top 20 true stories
- Economic harm: Fake White House explosion story wiped out $130 billion in stock value
- Trust erosion: Public confidence in mass media has dramatically degraded

The LLM Threat Multiplier:

ChatGPT has reached 800 million weekly users. LLMs now enable mass generation of synthetic fake news across 95+ languages at unprecedented scale.

The Explainability Challenge: High-accuracy systems (85-99%) are black boxes. Users, moderators, and regulators need to understand WHY a decision was made.

Current Research Landscape

Evolution of Fake News Detection:

Traditional Machine Learning

- Decision Trees, Random Forests: 70-85% accuracy
- Manual feature engineering
- Naturally interpretable
- Limited capacity for complex patterns

Deep Learning (Black Box)

- BERT, LSTM, Transformers: 85-99% accuracy
- Automatic feature learning
- No interpretability by default
- FakeBERT achieved 98.9% accuracy

Post-Hoc Explainability

- LIME, SHAP applied to black boxes
- Provides explanations after training
- dEFEND: Attention-based explanations
- Question: How reliable are these explanations?

My Project

Systematic comparison of intrinsically interpretable models vs post-hoc explainability. Which provides better balance of accuracy and trustworthy explanations?

Research Question & Hypotheses

Central Research Question:

What is the optimal balance between accuracy and explainability in fake news detection: intrinsically interpretable models (Decision Trees) or high-accuracy models with post-hoc explanations (BERT + SHAP)?

Key Questions to Answer:

- How much accuracy do we sacrifice for intrinsic interpretability?
- Are SHAP explanations from BERT as trustworthy as decision tree rules?
- Do both approaches identify similar features as important?
- Which approach is more suitable for real-world deployment (content moderation, journalism)?

Hypotheses:

H1: BERT will achieve 10-15% higher accuracy than Decision Trees
H2: Decision Trees will provide more stable and consistent explanations across similar inputs

Two Approaches: Overview

Comparing Intrinsic vs Post-Hoc Explainability

Approach A: Decision Tree + SHAP

(Intrinsically Interpretable)

- Features: TF-IDF, sentiment, readability, Author information
- Decision tree classifier (depth 8-10)
- SHAP adds quantitative feature attribution
- Every prediction has clear IF-THEN path

Why This Matters:

Built-in transparency; explanations can't contradict model logic

Approach B: BERT + SHAP

(Post-Hoc Explainability)

- BERT pre-trained language model
- Fine-tuned on fake news dataset
- SHAP applied post-hoc to explain predictions
- Learns complex contextual patterns

Why This Matters:

High accuracy; but explanations are estimations of black-box logic

Approach A: Decision Tree + SHAP (Intrinsic)

Architecture Pipeline:

Article Text → Feature Extraction → Decision Tree → SHAP Explainer → Prediction + Explanation

Feature Engineering:

- TF-IDF: Top 1000-5000 terms to capture vocabulary
- Sentiment: Polarity & subjectivity (TextBlob)
- Readability: Flesch-Kincaid grade level
- Metadata: Length, has author (binary)

Why Simple Features:

Humans can understand what each feature measures and why it matters

Decision Tree Configuration:

- scikit-learn DecisionTreeClassifier
- Max depth: 8-10 (human-readable constraint)
- 5-fold cross-validation grid search

SHAP TreeExplainer:

- Exact Shapley values (no approximation)
- Local & global explanations
- Quantifies each feature's contribution

The Accuracy-Interpretability Trade-off

Core Question: What is the 'cost' of interpretability?

Expected Trade-off Curve:

As tree depth increases from 4 to 12, we gain accuracy but lose interpretability. At depth 10-12+, trees become too complex for humans. At depth 4-6, accuracy suffers significantly.

Three Regions of Operation:

1. **High Interpretability Zone (Depth 4-6):** Easy to understand, but low accuracy
2. **Balanced Zone (Depth 8-10):** Reasonable accuracy, still human-readable
3. **Complex Zone (Depth 12+):** Higher accuracy, but hard to interpret

Approach B: BERT + SHAP (Post-Hoc)

Architecture Pipeline:

Article Text → BERT Tokenizer → BERT Fine-tuned Model → Classification Head→ SHAP Explainer → Prediction + Explanation

BERT Model Configuration:

- Base model: bert-base-uncased or DistilRoBERTa
- Input: Max 512 tokens per article
- Fine-tuning: 2-4 epochs on FakeNewsNet
- Learning rate: 2e-5
- Output: Binary classification (fake/real)

SHAP for BERT:

- SHAP Partition Explainer
- Approximates Shapley values via sampling
- Token-level importance scores
- Highlights influential words/phrases
- Limitations: Computationally expensive

Key Difference from Approach A:

BERT learns its own features (hidden representations). SHAP explains these learned features, not hand-crafted ones. This creates a gap between model logic and explanation.

Experimental Design

Dataset: FakeNewsNet (PolitiFact + GossipCop)

80/20 stratified train-test split, ensuring class balance

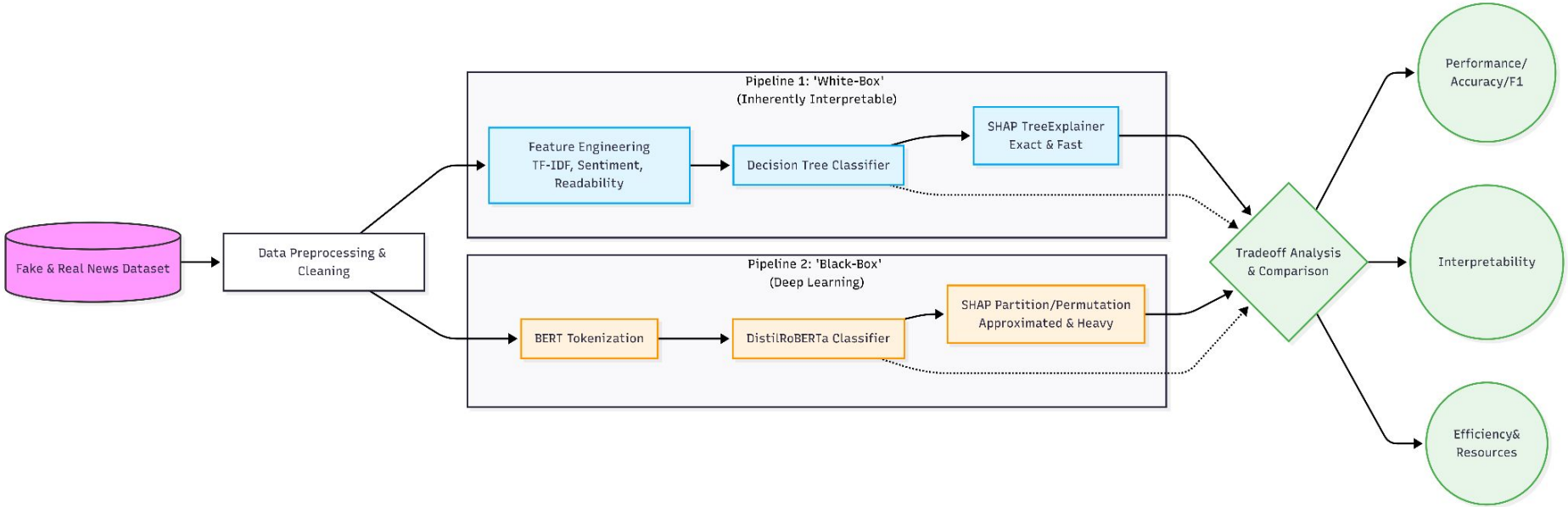
Evaluation Metrics:

- Accuracy Metrics: Accuracy, Precision, Recall, F1-score
- Explainability Metrics: Feature importance consistency, SHAP value stability across similar inputs
- Efficiency Metrics: Training time, inference time, explanation generation time

Five Key Experiments:

1. Baseline Comparison: Train both approaches, measure accuracy gap
2. Explanation Consistency: Test SHAP stability on similar articles
3. Feature Importance Overlap: Do both identify same key features?
4. Trade-off Analysis: Vary tree depth (4-12); plot accuracy vs interpretability

System Diagram



Expected Insights

What this project will demonstrate:

- Comparison of intrinsic vs post-hoc explainability in fake news detection with quantified accuracy-interpretability trade-offs
- Concrete measurements of SHAP reliability: Does BERT + SHAP provide trustworthy explanations or just plausible-sounding ones?
- Identification of optimal decision tree depth for accuracy and interpretability
- Evidence-based deployment guidelines: which approach for which stakeholder and use case
- Discovery of whether both approaches identify the same linguistic features (sentiment, sensational language) or if BERT learns fundamentally different patterns

Thank You