

PROJECT ABSTRACT

Fake News Detection using Transformers

A RoBERTa classifier fine-tuned on the ISOT fake-news dataset scores articles for misinformation risk with six explainable linguistic signals.

01. Title

Fake News Detection using Transformers

02. Introduction / Background

Misinformation spreads faster than fact-checks can follow, and readers cannot reliably judge an unfamiliar article in the moment. Automated detection is genuinely hard — style alone does not prove falsehood — which is why this project is built with unusual honesty about its limits.

RoBERTa fine-tuned on ISOT (44,898 labeled real and fake articles) learns stylistic fingerprints of fabricated news; a transparent signal layer scores sensationalism, emotional manipulation, attribution quality, source credibility, headline–body consistency and shouting patterns. The demo presents a risk assessment with evidence, like a spell-checker for credibility.

03. Problem Statement

Readers share misinformation because nothing flags it at reading time, while existing detectors give unexplained verdicts and over-claim. The project addresses this with an explainable risk assessment — every score decomposed into inspectable signals with the driving phrases highlighted — and an explicit uncertain band with fact-checking guidance instead of forced classifications.

04. Objectives

- Fine-tune RoBERTa-base on the ISOT Fake News dataset for real/fake classification.
- Implement six documented linguistic signals with per-signal scoring.
- Render a verdict card: likely-real / uncertain / likely-fake with confidence.
- Highlight signal-driving phrases inside the article text.
- Check publishing domains against a documented credibility reference list.
- Evaluate accuracy/precision/recall/F1 on ISOT's held-out split with error analysis on satire and well-written fakes.

05. Proposed Methodology

Article text splits into headline and body. RoBERTa-base fine-tuned on ISOT outputs a fake-vs-real probability. In parallel, six rule-based signals compute: sensational vocabulary density, emotion-lexicon hits, named-source attribution, source-list match, headline–body semantic overlap, capitalization patterns. Classifier probability and signals combine with documented weights into risk bands. Evaluation reports standard metrics plus a candid error analysis.

06. System Architecture / Working

Article → cleaning/splitting → RoBERTa classifier + signal extractors → weighted risk score → verdict card with per-signal bars and in-text highlighting. The demo analyzes pasted articles or preloaded real/fake samples; borderline articles get the uncertain verdict with verification guidance. The report states the tool assists readers and must not be used for automated content removal.

07. Hardware / Software Requirements

- Python 3.11, PyTorch, Hugging Face Transformers (RoBERTa), spaCy, scikit-learn
- ISOT Fake News Dataset (44,898 articles: 21,417 real / 23,481 fake)
- Flask + HTML/CSS/JS demo
- Single GPU for fine-tuning (~2–4 h); CPU inference under 3 s/article

08. Expected Outcomes

A detector with accuracy/F1 on ISOT's held-out split reported as achieved figures after training (design targets until then); an explainable demo; a documented error analysis showing where style-based detection fails.

09. Applications

- Reader-side credibility assistance while browsing news
- Media-literacy teaching with visible linguistic signals
- First-pass triage for fact-checking workflows (human in loop)
- Viva demonstration of transformer classification with honest limits

10. Future Scope

- Claim-level verification against knowledge bases
- Cross-lingual misinformation detection
- Temporal adaptation as news style drifts
- Browser extension for in-page risk badges