PROJECT ABSTRACT

Fake Product Review Detection using NLP

A natural-language classifier that separates genuine product reviews from deceptive (fake) ones using linguistic signals learned from the Deceptive Opinion Spam corpus, with an interactive analyzer demo.

01. Title

Fake Product Review Detection using NLP

02. Introduction / Background

Fake reviews distort buying decisions and marketplace trust, and they are written to sound genuine — which makes them a natural text-classification problem. The Deceptive Opinion Spam corpus (Ott, Choi, Cardie & Hancock, ACL 2011) provides 1,600 labeled hotel reviews: 800 fake ones commissioned from crowd workers and 800 truthful ones mined from TripAdvisor, all for Chicago hotels. Published research on this corpus shows that fake and truthful reviews differ systematically in linguistic style: deceptive reviews overuse superlatives and first-person singular pronouns, underuse concrete specifics such as numbers and spatial detail, and show different sentiment patterns. This project turns those findings into a student-buildable, fully explainable classifier: TF-IDF unigram/bigram features with logistic regression, where every weight maps to an n-gram the student can inspect and discuss.

03. Problem Statement

Students studying NLP rarely work with a real labeled deception corpus; most spam-detection demos train on toy data or call a black-box API, leaving the student unable to explain what the model learned. The project addresses this by building the complete pipeline on the published Deceptive Opinion Spam corpus: preprocessing, TF-IDF feature extraction, regularized logistic regression with cross-validated tuning, and honest held-out evaluation — plus an interactive demo that analyzes any pasted review with a per-signal linguistic breakdown.

04. Objectives

- Build a reproducible training pipeline on the real 1,600-review Deceptive Opinion Spam corpus with documented splits.
- Train an explainable TF-IDF + logistic-regression classifier and tune regularization with 5-fold cross-validation.
- Evaluate once on a held-out test split (accuracy, precision, recall, F1, confusion matrix).
- Extract the top fake-indicative and genuine-indicative n-grams for viva-ready interpretation.
- Ship an interactive web demo that classifies pasted reviews with a per-signal linguistic breakdown.

- Deliver the complete student kit: notebook, trained model, demo, report, PPT and viva Q&A.;

05. Proposed Methodology

The system is developed in three stages. (1) Data: the corpus is loaded with its labels, cleaned (lowercasing, punctuation stripping, stopword removal) and split 60/20/20 into stratified train, validation and held-out test sets. (2) Modeling: a TF-IDF vectorizer over unigrams and bigrams (design target ~10,000 features) feeds an L2-regularized logistic regression; the regularization strength C is tuned with 5-fold cross-validation on the validation data. (3) Evaluation and demo: the final model is scored exactly once on the held-out test split, and the same preprocessing + vectorization + model chain powers the browser demo, which adds six computed linguistic signals (superlative density, exclamation intensity, first-person overuse, lack of specifics, repetition, ALL-CAPS emphasis) for the UI breakdown.

06. System Architecture / Working

The training notebook is the core artifact: corpus loading, preprocessing, TF-IDF vectorization, model training with cross-validation, and evaluation with plots. The trained vectorizer and classifier are serialized for inference. The web demo mirrors this chain in JavaScript for instant analysis: pasted text is tokenized, the six linguistic signals are computed transparently, and a documented scoring rule combining them produces the fake/genuine verdict with confidence. A session history table records every analysis. The demo's rule-based scoring is explicitly documented as an illustration of the trained model's top signals, not the model itself.

07. Hardware / Software Requirements

- Python 3 with scikit-learn, pandas, NumPy, Matplotlib (training & evaluation)
- Jupyter Notebook for the training pipeline
- Deceptive Opinion Spam corpus (Ott et al., 2011) — publicly available for research
- Any modern web browser for the interactive demo (single HTML file, runs offline)
- Laptop CPU sufficient — training takes minutes, inference is effectively instant

08. Expected Outcomes

The project delivers a trained fake-review classifier evaluated on the held-out test split with accuracy, precision, recall, F1 and a confusion matrix; the design target is ~89% accuracy, matching published bigram-model benchmarks on this corpus — the built-to-order run's measured figure is reported after training, never claimed in advance. Deliverables: the complete training/evaluation notebook, serialized TF-IDF vectorizer and classifier, the interactive analyzer web demo with sample reviews, top-weighted n-gram lists with interpretation notes, the project report PDF, PPT and viva Q&A; document.

09. Applications

Marketplace review-moderation assistance; e-commerce trust-and-safety tooling; teaching text classification, feature engineering and model evaluation; a base for extending to

restaurant or product-review domains with domain-specific retraining.

10. Future Scope

- Trained on hotel reviews — performance on other domains is expected to drop without retraining.
- The model reads linguistic style, not facts; careful liars can evade it and flowery truthful reviews can score suspicious.
- ~89% is a design target matching published benchmarks, not a measured claim.
- Very short reviews (under ~20 words) carry too little signal; the demo warns on these.
- Future scope: transformer fine-tuning (BERT), multilingual/code-mixed support, fact-consistency features.
- The project gives the student a complete, honest NLP build: a real labeled corpus, an explainable model whose every weight can be discussed, rigorous evaluation, and a demo that makes linguistic deception detection tangible. It is scoped as a teaching project in fake-review linguistics — not a production moderation system — and that honest framing is part of what makes it strong viva material.