PROJECT ABSTRACT

Fake Review Detection using LSTM

Two-layer LSTM classifier that detects deceptive hotel reviews, trained on Ott et al.'s Deceptive Opinion Spam Corpus (1,600 reviews) and delivered as a built-to-order student project with code, training notebook, demo app, report, PPT and viva kit.

01. Title

Fake Review Detection using LSTM

02. Introduction / Background

Online reviews drive purchasing decisions, which makes them a target: sellers post glowing fake reviews of their own products while competitors post damning ones. Humans are notoriously poor at spotting the fakes — in the Cornell studies behind this project, human judges barely beat chance. The field's classic benchmark is the Deceptive Opinion Spam Corpus (Ott, Choi, Cardie and Hancock, ACL 2011): 1,600 Chicago hotel reviews, half truthful (from TripAdvisor) and half deceptive (written by paid Mechanical Turk workers). This project trains a two-layer LSTM classifier on that corpus. Unlike bag-of-words models, the LSTM reads the review in order, so negation and word order genuinely affect the verdict.

03. Problem Statement

Review platforms need automated ways to flag opinion spam, but deception hides in subtle linguistic patterns — first-person pronoun overuse, superlatives, missing concrete detail — that no human-written rule set captures reliably. The project addresses this by learning those patterns with a sequence model trained on the gold-standard corpus, and by making each verdict explainable through word-level highlighting so the model's reasoning can be inspected and questioned.

04. Objectives

- Train a two-layer LSTM (64 units each, 128-d embeddings, dropout 0.4) on the Deceptive Opinion Spam Corpus with stratified 80/10/10 splits.
- Reach a design target of approximately 88-90% accuracy on the held-out split, with the notebook logging accuracy, precision, recall, F1 and the confusion matrix every epoch.
- Train a TF-IDF + logistic regression baseline in the same notebook to quantify what sequence modeling adds.

- Deliver the complete student kit: source code, training notebook, demo app with verdict highlighting, report, PPT and viva Q&A document.

05. Proposed Methodology

Reviews are lowercased, tokenized and padded to 400 tokens with a ~20,000-word vocabulary built from the training split. Learned 128-d embeddings feed two stacked 64-unit LSTM layers with 0.4 dropout; the final hidden state passes through a dense sigmoid layer. Binary cross-entropy is minimized with Adam (lr 1e-3) and early stopping on validation F1 guards against overfitting the small corpus. The baseline pipeline vectorizes the same splits with TF-IDF and fits logistic regression for the comparison the report presents.

06. System Architecture / Working

The system has three stages. (1) Data pipeline: the 1,600 labeled reviews are loaded with truthful/deceptive labels and split 80/10/10 with stratification over sentiment and label. (2) Model: the embedding + stacked-LSTM + sigmoid network is trained with early stopping; the best checkpoint is kept. (3) Demo application: a Flask app tokenizes a pasted review, runs inference, and renders the verdict with confidence, word-level highlighting (deceptive cues, truthful anchors, stylistic flags) and a linguistic-cue dashboard of pronoun, superlative, exclamation and concrete-detail counts.

07. Hardware / Software Requirements

- Python 3.10, TensorFlow/Keras, NLTK, scikit-learn
- NumPy, pandas, Matplotlib, Seaborn
- Jupyter notebook for the buyer-run training and evaluation
- Flask demo web app
- Deceptive Opinion Spam Corpus (Ott et al., public research download)
- CPU training is practical (corpus is small); no GPU required

08. Expected Outcomes

The deliverable is a design-target-driven build. The training notebook computes accuracy, precision, recall, F1 and the confusion matrix on the held-out split every epoch; the report presents these as the build's measured results against the design target of approximately 88-90% accuracy (the corpus's classic SVM baseline is ~89%, Ott et al. ACL 2011). Expected outcomes: a working review verdict tool with explainable highlighting, an LSTM-vs-baseline comparison, and an honest error analysis of which reviews still fool the model.

09. Applications

- Review-platform moderation assistance (prototype scope)

- Teaching material for NLP preprocessing, sequence modeling and the opinion-spam literature
- Reference build for other deception-detection tasks (fake news, phishing text)
- Explainability case study: word-level highlighting of classifier decisions

10. Future Scope

- Retrain on product-review data to test domain transfer beyond hotels
- Add transformer-based classifiers (BERT) for comparison with the LSTM
- Incorporate reviewer metadata (review history, timing) alongside text
- Extend to multilingual reviews with multilingual embeddings