মেট্রোপলিটন ইউনিভার্সিটি
Metropolitan University

Project Report

Course Title: Artificial Intelligence
Course code: SWE-315
Topic: Fake Review Detection Using Natural Language Processing and Machine Learning.

Submitted to:

Al Akram Chowdhury
Lecturer
Department of SWE
Metropolitan University, Sylhet

Submitted by:

Shakif Niaz
232-134-040

Date of Submission: 17/12/2025

Introduction:

The rapid growth of e-commerce has given rise to an overwhelming number of user-generated reviews. While reviews strongly influence consumer decisions, many platforms face the issue of fake or deceptive reviews crafted to manipulate product ratings. These fabricated reviews can mislead customers and damage platform credibility.

This project focuses on developing a **Fake Review Detector** using Natural Language Processing (NLP) and Machine Learning (ML). The model classifies reviews into two categories:

- CG (Computer-Generated / Fake)
- OR (Original / Real)

A complete frontend was developed using Streamlit, allowing users to paste any review and instantly receive predictions with confidence scores.

Objectives:

1. To preprocess textual review data using NLP techniques.
2. To extract meaningful linguistic and structural features from reviews.
3. To train a machine learning pipeline capable of detecting fake reviews.
4. To deploy an interactive web-based interface for real-time classification.

System Architecture:

The Fake Review Detector system consists of the following components:

- NLP Preprocessing Module
- Feature Extraction Module
- Machine Learning Pipeline
- Streamlit User Interface

The workflow used for this project:

1. Raw Review
2. Preprocessing
3. Feature Engineering
4. ML Model
5. Prediction Output

Dataset and Labeling:

The fake review detection model was trained using the Fake Reviews Dataset with 40,000 entries available on Kaggle. This dataset contains two labeled categories:

- CG (Computer-Generated / Fake) (20,000 Entries)
- OR (Original / Real) (20,000 Entries)

Each review is pre-labeled, enabling the model to learn linguistic and structural distinctions between genuine and fabricated reviews. The dataset is relatively balanced, making it appropriate for supervised machine learning tasks such as binary classification.

Dataset Details:

- Name**:** Fake Reviews Dataset
- Source**:** Kaggle
- Original Authors / Citation:
  Salminen, J., Kandpal, C., Kamel, A. M., Jung, S., & Jansen, B. J. (2022). *Creating and detecting fake reviews of online products*. Journal of Retailing and Consumer Services, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771
- Dataset Link**:** https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

Methodology:

The system preprocesses each review using classical NLP steps:

1. **NLP Preprocessing:**

Step	**Description**
Lowercasing	Converts text to lowercase
Tokenization	Splits text into words using nltk.word_tokenize()
Alphanumeric Filtering	Removes non-alphanumeric tokens
Stopword Removal	Eliminates meaningless words using NLTK stopword corpus
Stemming	Uses Porter Stemmer to reduce words to root form
Rejoining	Reconstructs cleaned tokens into transformed string

2. **Feature Engineering:**
   Each review is converted into a feature vector with:
   1. Transformed Text (after NLP cleaning)

2. Number of Characters
3. Number of Words
4. Number of Sentences

These features are generated using the featurize_single_review() function. The inclusion of structural features helps detect writing style differences between real and fabricated reviews.

3. **Machine Learning Pipeline:**
   A trained ML pipeline is loaded from: fake_reviews_pipe.pkl
   Although the internal architecture is not visible, typical components include:
   - TF-IDF vectorizer for text transformation
   - Oversampling/undersampling (if dataset imbalance exists)
   - Classifier algorithm, likely Logistic Regression / SVM / Random Forest

   The pipeline provides:
   - predict() : Final label (CG or OR)
   - predict_proba() : Confidence probabilities

Model Evaluation:

Performance metrics are displayed from model_metrics.json, including:

- Accuracy
- Precision for the CG class (Fake review precision)

High precision for the fake class reduces the chance of falsely accusing a real reviewer of fraud.

Machine Learning Model Comparison:

To identify the most suitable model for fake review detection, multiple machine learning algorithms were trained and evaluated using accuracy and precision (CG treated as the positive class). The results are shown below:

Model	**Accuracy**	**Precision**
K-Nearest Neighbors	67.73%	67.26%
Naïve Bayes	79.81%	77.44%
Decision Tree	69.09%	67.41%
Logistic Regression	88.50%	87.88%

Random Forest	84.96%	85.10%
AdaBoost	76.70%	75.75%
Bagging Classifier	82.28%	82.05%
Extra Trees Classifier	84.65%	85.17%
Gradient Boosting	80.31%	78.18%
XGBoost	87.15%	86.89%
Multinomial Naïve Bayes	79.81%	77.44%

Highlights:

- Logistic Regression performs the best overall, with Accuracy = 88.50% and Precision = 87.88%
- Ensemble methods (RF, ETC, XGB) also perform strongly.

User Interface:

The frontend is designed with a modern UI using custom CSS. Key features include:

Input:
A text area where users paste any review.

Output:

- Prediction Header (Fake Review / Real Review)
- Fake Probability Meter
- Real Probability Meter
- Numerical Confidence Scores

The interface combines aesthetics and usability with stylized progress bars and metrics.
Implementation: Streamlit UI code block from app.py.

Results and Discussion:

Strengths

- High accuracy in distinguishing fake vs. real reviews.
- Robust feature extraction combining linguistic and structural cues.
  Clean and interactive interface enabling real-time use.
- Simple preprocessing ensures fast predictions.

Limitations

- Model performance heavily depends on dataset quality.
  Stemming may distort word meaning.
- Does not use deep learning or contextual embeddings like BERT.

Future Improvements

- Upgrade to transformer-based models (BERT, RoBERTa).
- Add explainability (highlight suspicious parts of text).
- Expand the dataset with multilingual support.
- Add behavioural metadata (IP, timing, posting frequency).

Conclusion:

This project successfully demonstrates how Natural Language Processing and Machine Learning can be used to identify deceptive reviews. By implementing text cleaning, feature extraction, and a trained ML pipeline, the system provides reliable predictions along with user-friendly visualization.

The Fake Review Detector has practical applications in e-commerce, hospitality, and online marketplaces, contributing to maintaining trust and transparency.