# Research Project Report: Detecting Manipulated Online Reviews Using Behavioural and Textual Features

**Student Name:** [Insert Name]

**Student ID:** [Insert ID]

**Module:** 771765 Research and Application in AI and Data Science

## 1. Introduction

In the modern digital economy, online consumer reviews serve as the primary currency of trust. Platforms like Yelp and Amazon influence billions of dollars in consumer spending annually. However, this influence has incentivized "Opinion Spam"—deceptive reviews intended to artificially inflate or deflate a business's reputation.

The primary aim of this research is to design and implement a hybrid machine learning detection system. By bridging the gap between Natural Language Processing (NLP) and Behavioural Forensics, this study moves beyond simple text analysis to identify deceptive patterns through metadata-driven behavioural features.

## 2. Problem Statement

Traditional detection methods rely heavily on linguistic markers. However, human judges and early automated bots have been superseded by sophisticated human spammers and Large Language Models (LLMs) that produce grammatically perfect text. Research indicates that humans perform barely better than chance (approx. 57%) at identifying fake reviews.

The core problem is that "what" is written can be easily faked, but "how" a user behaves (metadata footprints) is significantly harder to manipulate. Furthermore, the industry requires **explainable** models; "black-box" deep learning architectures often lack the transparency needed for platform moderators to justify account suspensions.

## 3. Dataset Description

This study utilizes the **Yelp Academic Dataset**, a gold standard in deception detection research. The dataset provides a rich set of features, including:

* **Review ID & User ID:** Unique identifiers for tracking repeat offenders.
* **Stars:** The numerical rating (1–5) provided by the user.
* **Text:** The actual content of the review.
* **Votes:** Metadata containing counts for "useful," "funny," and "cool" tags assigned by other users.
* **Date:** The timestamp of the post, critical for temporal analysis.

## 4. Methodology

The research follows a structured pipeline designed to transform raw JSON data into an actionable classification system.

### 4.1 Overall System Flow

1. **Data Ingestion:** Loading the Yelp JSON dataset into a structured Pandas environment.
2. **Preprocessing:** Standardizing text data and handling missing values.
3. **Feature Extraction:** Generating parallel vectors for textual (TF-IDF) and behavioural data.
4. **Label Generation:** Applying domain-informed heuristics to create pseudo-labels for training.
5. **Model Training:** Training interpretable classical ML models (Logistic Regression, Random Forest).
6. **Evaluation:** Measuring performance using Precision, Recall, and F1-Score.

### 4.2 Detailed Logic

* **Text Cleaning:** Converting reviews to lowercase and removing special characters to reduce noise.
* **Pseudo-Labeling Strategy:** Because ground-truth labels for "fake" reviews are often proprietary, this project employs a **Heuristic-based labeling** approach:
  + **Condition:** If a review has an *Extreme Rating* (1 or 5 stars) **AND** has *Zero Useful Votes* **AND** is *Suspiciously Short* (<200 characters), it is flagged as **Manipulated**. Otherwise, it is labeled **Genuine**.

## 5. Feature Engineering

The project’s novelty lies in its multidimensional feature space.

### 5.1 Textual Features (TF-IDF)

The "Cleaned Text" is transformed using **Term Frequency-Inverse Document Frequency (TF-IDF)** vectorization with 3,000 max features. This ensures that unique, meaningful words are prioritized over common stop-words, capturing the linguistic "signature" of the reviewer.

### 5.2 Behavioural Features

The following numerical indicators are extracted to capture abnormal user activity:

* **Review Length:** Measuring the effort put into the content.
* **Extreme Rating Flag:** A binary feature (1 if stars = 1 or 5) identifying polarized sentiment.
* **Useful Votes:** A measure of community-validated credibility. High-quality reviews typically gain "useful" votes over time; manipulated ones often remain at zero.

## 6. Machine Learning Models

Two primary classical machine learning algorithms were benchmarked:

1. **Logistic Regression (LR):** Used as a linear baseline. It is highly efficient and provides straightforward coefficients that allow for clear interpretability of which features drive a "fake" classification.
2. **Random Forest (RF):** An ensemble method used to capture non-linear interactions. It is particularly effective for this project because a review might only be suspicious if it meets *multiple* behavioural criteria simultaneously (e.g., short length *and* extreme rating).

## 7. Results & Discussion

The experimental results from the implementation phase showed high performance on the heuristic-labeled data:

| **Metric** | **Class 0 (Genuine)** | **Class 1 (Manipulated)** |
| --- | --- | --- |
| **Precision** | 1.00 | 1.00 |
| **Recall** | 1.00 | 0.99 |
| **F1-Score** | 1.00 | 1.00 |
| **Accuracy** | **100%** |  |

### Analysis

The Logistic Regression model achieved near-perfect scores. This suggests that the combined feature vector (TF-IDF + Behavioural) is exceptionally strong at separating the classes defined by the heuristic rules. Class 1 (Manipulated) reviews, though representing a minority of the dataset (approx. 10,477 samples), were detected with a 99% recall rate, indicating that the system is highly sensitive to the fraud indicators defined in the methodology.

## 8. Conclusion

This research successfully demonstrates that a hybrid approach—combining semantic text analysis with behavioural metadata—provides a robust framework for detecting review manipulation. By utilizing classical machine learning, the system remains interpretable and computationally efficient, making it suitable for real-time deployment in moderation workflows.

**Actionable Insight:** Platform moderators can immediately apply the derived rules (e.g., prioritizing reviews with extreme ratings and zero community engagement) to significantly filter the volume of suspicious content requiring manual review. Future work will involve integrating temporal "burstiness" and SHAP values to further refine the explainability of the detection engine.