Configuration Manual

MSc Research Project
Data Analytics

Shreyas Akash Rao
Student ID: x23205342

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: … Shreyas Akash Rao ……..…………………………………………………………………

Student ID: … x23205342 ………………..………………………………………………………..……

Programme: … MSc in Data Analytics …………………..……….. **Year:** 2024-2025……….

Module: … MSc Research Project …………..…………………………………………………..………

Lecturer: … Christian Horn …………..…………………………………………………………………

Submission Due Date: … 29th January 2025 …………………………………………………………..……

Project Title: Improving fake review detection in e-commerce using combined analysis techniques

Word Count: … 581 ……………………………… **Page Count:** ……4…………………………..……..………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …… Shreyas Akash Rao ……………………………………..………………

Date: …… 27th January 2025 ……………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual for “Improving Fake Review Detection in E-Commerce Using Combined Analysis Techniques”

Shreyas Akash Rao

Student ID: x23205342

1. OVERVIEW

The objective of this project is to find out the fake reviews within the e-commerce dataset using an exploratory data analysis approach together with machine learning. In addition, the focus is on numeric features such as review counts, average ratings, helpful votes, and sentiment.

2. SYSTEM REQUIREMENTS

Ensure that the following requirements are in place to set up the project:
Hardware:

- Minimum 4 GB of RAM or more (Recommended 8 GB)
- CPU: Multi-core processor, Intel i5/i7 or higher
- GPU (Optional): CUDA-enabled GPU for acceleration (if applicable)

Software:

- Python 3.11.5
- Required Python libraries:
  1. **Pandas 2.2.3**: For advanced data manipulation and analysis using DataFrames.
  2. **NumPy 1.26.4**: For efficient numerical calculation and array operations.
  3. **Matplotlib 3.9.3**: For creating data visualizations.
  4. **Seaborn 0.13.2**: For generating statistical plots.
  5. **Wordcloud 1.9.4**: For generating word clouds from text data to visualize the most frequently occuring terms.
  6. **scikit-learn 1.6.0**: For implementing machine learning models and data preprocessing with updated algorithms.
  7. **imbalanced-learn 0.12.4**: For handling imbalanced datasets with techniques like oversampling and undersampling.
  8. **XGBoost 2.1.3**: For implementing high-performance gradient boosting models with advanced tuning options.

Environment:

- Preferred IDE: Google Colab or Jupyter Notebook
- Operating System: Windows/Linux/MacOS

Machine Learning Models Developed:

- Naïve Bayes: Optimized for probabilistic analysis using numeric features.
- Decision Tree: Configured for optimal depth to avoid overfitting while maintaining interpretability.
- Logistic Regression: Simplified linear model for quick and effective classification.
- Gradient Boosting: An ensemble learning method that builds models sequentially, where each model tries to correct the errors of the previous one.
- XGBoost: An optimized version of Gradient Boosting, XGBoost stands out due to its efficiency and scalability

3. INSTALLATION STEPS

1. **Python Installation**

   Download Python from python.org and ensure it is added to the system PATH.

2. **Install Required Libraries**

Use pip to install the libraries

pip install pandas NumPy matplotlib seaborn wordcloud sci-kit-learn imbalanced-learn xgboost

or

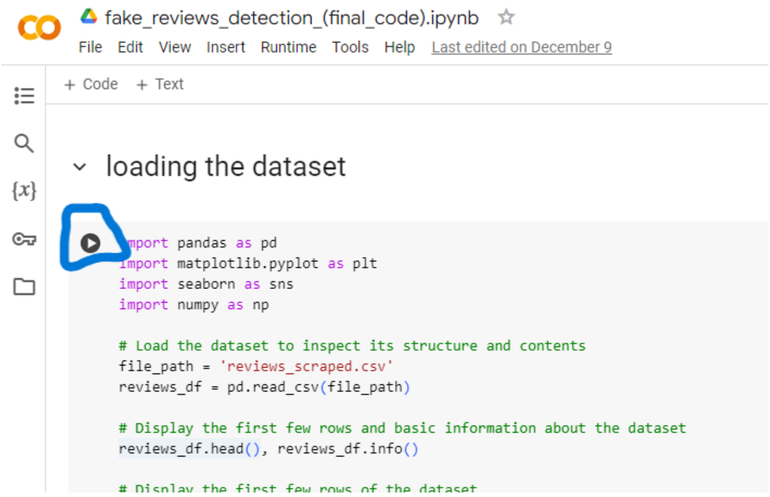
! pip install pandas NumPy matplotlib seaborn wordcloud sci-kit-learn imbalanced-learn xgboost

3. **Prepare the Dataset:**

- Ensure the dataset (reviews_scraped.csv) is available in the same directory as the code.
- Data source: https://www.kaggle.com/datasets/sofiazowormazabal/amazon-fake-reviews-scrapped

4. RUNNING THE CODE

To run the code, run every cell in Google Collab/Jupyter Notebook. Select the cell you want to run then click run.


fake_reviews_detection_(final_code).ipynb
File Edit View Insert Runtime Tools Help Last edited on December 9
+ Code + Text
loading the dataset

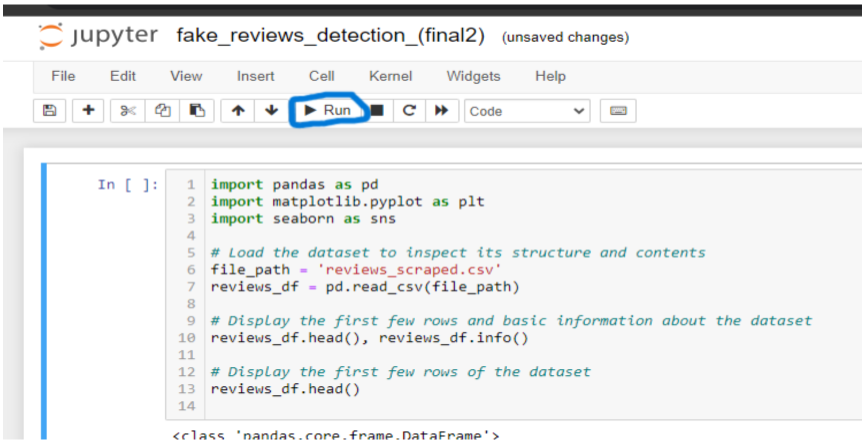
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Load the dataset to inspect its structure and contents
file_path = 'reviews_scraped.csv'
reviews_df = pd.read_csv(file_path)

# Display the first few rows and basic information about the dataset
reviews_df.head(), reviews_df.info()

# Display the first few rows of the dataset
```

Figure 1: running code in Google Collab


jupyter fake_reviews_detection_(final2) (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset to inspect its structure and contents
file_path = 'reviews_scraped.csv'
reviews_df = pd.read_csv(file_path)

# Display the first few rows and basic information about the dataset
reviews_df.head(), reviews_df.info()

# Display the first few rows of the dataset
reviews_df.head()

```

```
<class 'pandas.core.frame.DataFrame'>
```

Figure 2: running code in Jupiter notebook

Main Functionalities

1. Data Preprocessing

- Handles missing values by filling or removing them.
- Converts text data into numeric features for model training.

2. Exploratory Data Analysis (EDA)

- Generates visualizations such as histograms, count plots, and correlation heatmaps.

3. Model Training and Evaluation

- Trains models including Random Forest, Logistic Regression, Decision Tree, Naive Bayes, Gradient Boosting, and XGBoost.
- Evaluate models using metrics like accuracy, precision, recall, and F1-score.

4. Balancing Data

- Utilizes SMOTEENN to handle class imbalances for better model performance.

5. Visualization

- Generates ROC curves, confusion matrices, and word clouds for analysis.

5. EXPECTED OUTPUTS

- **Data Preprocessing:**
  - Summary statistics and visualizations to understand the dataset.
- **EDA Results:**
  - Visualizations: Heatmaps, histograms, word clouds.
- **Model Performance:**
  - Metrics like accuracy, precision, recall, F1 score.
  - ROC curves to compare models.

6. TROUBLESHOOTING

Issue	**Solution**
Missing required libraries	Run pip install <library_name> or !pip install <library_name> .
Dataset not found	Ensure reviews_scraped.csv is in the project directory.
Memory errors	Reduce dataset size or use sampling. Ensure sufficient RAM is available.
Slow model training	Consider using a GPU or limiting the dataset for testing.