National College of Ireland

Configuration Manual

MSc Research Project
Programme Name

Alan Thomas
Student ID: x23288944

School of Computing
National College of Ireland

Supervisor: Eamon Nolan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Alan Thomas
Student ID:	X23188944
Programme:	Msc in Data Analytics **Year:** Jan 2024
Module:	Msc Research Project
Lecturer:	Eamon Nolan
Submission Due Date:	13th December 2024
Project Title:	Online Reviews and Product Sales: A Sentiment Analysis Approach
Word Count:	6927 **Page Count:** 23

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: Alan Thomas

Date: 13/12/2024

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

Alan Thomas
Student ID: x23188944

1 Introduction

This Configuration Manual documents the setup, configurations, tools, and processes used in the research project titled "Online Reviews and Product Sales: A Sentiment Analysis Approach.' The project looks at how review can predict product price and sales utilising Machine learning and Data visualisations. Some of the information within this paper aims at explaining the methodologies used, the software requirements, and the difficulties found during the implementation.

2 Tools and Technologies

Hardware Requirements:

- **Processor:** Intel Core i5 or higher
- **RAM:** 16 GB (recommended for handling large datasets)
- **Storage:** 256 GB SSD or higher
- **Operating System:** Windows 10 or equivalent

Software Requirements:

1. **Programming Language:** Python 3.9
2. **Development Environment:** Jupyter Notebook
3. **Libraries and Frameworks:**
    - **Machine Learning:** scikit-learn (Logistic Regression, Random Forest, Support Vector Machine, XGBoost)
    - **Sentiment Analysis:** VADER
    - **Data Handling:** Pandas, NumPy
    - **Data Visualization:** Matplotlib, Plotly, Dash
    - **Text Preprocessing:** NLTK, TfidfVectorizer
4. **Visualization Tool:** Interactive dashboard developed using Dash and Plotly.

Dataset:

- **Source:** Publicly available datasets on Kaggle. (click here to view dataset)

- **Content:** Product reviews and sales data, with attributes such as review text, scores, and metadata.
- **Preprocessing Steps:**
    - Removed duplicates and irrelevant data.
    - Handled missing values by replacing scores with median values and timestamps with mean values.
    - Applied tokenization, stopword removal, and text vectorization

3 Implementation Workflow

1. Data Collection:

- Combined datasets of product reviews and sales data.
- Preprocessed data to ensure quality and relevance.

2. Machine Learning Models:

- Implemented and evaluated four models:
    - Logistic Regression: Achieved the highest accuracy of 86.46%.
    - Random Forest: Accuracy of 82.50%, suitable for handling noisy data.
    - Support Vector Machine: Accuracy of 86.29%, effective for high-dimensional spaces.
    - XGBoost: Accuracy of 84.64%, capable of capturing complex relationships.
- Metrics used for evaluation included accuracy, precision, recall, and F1-score.

3. Sentiment Analysis:

- Classified reviews as positive, neutral, or negative using VADER.
- Enhanced insights through correlation analysis of sentiment scores and sales data.

4. Dashboard Development:

- Developed an interactive dashboard to visualize:
    - Sentiment trends over time.
    - Correlation between sales and sentiment.
    - Product-specific insights

4 Challenges and Resolutions

1. Large Dataset Handling:

- Challenge: Processing millions of reviews and sales records was computationally intensive.

- Resolution: Optimized data pipelines and reduced dimensionality using TfidfVectorizer.

2. Fake Review Detection:

- Challenge: Identifying sarcastic or manipulated reviews was difficult.
- Resolution: Incorporated heuristic methods and sentiment analysis tools like VADER.

3. Dashboard Scalability:

- Challenge: Real-time analysis for large datasets.
- Resolution: Optimized backend processing and caching mechanisms.

References

[1] Medill Spiegel Research Center, *How Online Reviews Influence Sales*.

[2] J. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, Aug. 2006.

[3] N. Hu, L. Liu, and J. Zhang, "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Information Technology and Management*, vol. 9, no. 3, pp. 201–214, Sep. 2008.

[4] X. Zhang, F. Keegan, and S. Keller, "Trustworthiness of online reviews and sales performance: A dynamic view," *Journal of Retailing and Consumer Services*, vol. 17, no. 1, pp. 7–16, Jan. 2010.

[5] H. Cui, Y. Jiang, and M. Liu, "Text mining for online reviews: Discovering factors affecting trust and sales," *Journal of Business Research*, vol. 85, pp. 135–145, 2018.

[6] X. Chen, Q. Zhang, and J. Wang, "Sentiment analysis and machine learning for predicting product sales," *Expert Systems with Applications*, vol. 92, pp. 24–35, 2019.

[7] Y. He, L. Zhou, and Z. Li, "Temporal pattern recognition in online reviews using RNNs," *Neural Computing and Applications*, vol. 31, no. 7, pp. 2155–2165, 2020.

[8] K. Zhou, S. Yang, and W. Xu, "BERT-based transformer models for sentiment analysis and sales prediction," *Proceedings of the International Conference on Data Mining*, pp. 45–52, 2021.