



**islington college**  
(इस्लिंग्टन कॉलेज)

**Module Code & Module Title**

**CC6057NI**

**Applied Machine Learning**

**60% Individual Coursework**

**Submission: Milestone 1**

**AY 2025 2026**

**Academic Semester: Autumn Semester 2025**

**Credit: 15 Credit Semester Long Module**

**Student Name: Sanskriti Agrahari**

**London Met ID: 23048503**

**College ID: NP01AI4A230002**

**Assignment Due Date: Monday, December 15, 2025**

**Assignment Submission Date: Sunday, December 14, 2025**

**Module Leader: Mr. Mahotsav Bhattarai**

*I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.*

# Table of Contents

<b>1.</b>	<b><i>Introduction</i></b>	<b>1</b>
1.1	Overview of the text classification project	1
1.2	Machine Learning and NLP Concepts Applied	1
<b>2.</b>	<b><i>Problem Domain</i></b>	<b>2</b>
2.1	Problem Description	2
2.2	Dataset Description and Background	2
2.3	Societal and Business Relevance	3
<b>3.</b>	<b><i>Solution</i></b>	<b>4</b>
3.1	Overview of the Proposed Solution	4
3.2	Text Preprocessing Techniques	4
3.3	Feature Extraction Methods	5
3.4	Machine Learning Models Implemented	5
3.5	Diagrammatic Representation	6
3.6	Development Process and Tools	8

## **TABLE OF FIGURES**

Figure 1 Diagrammatic Representation .....	7
--	---

## 1. Introduction

### 1.1 Overview of the text classification project

Text classification is one of the major tasks in Natural Language Processing, which means setting texts into predefined classes. With the rapid growth of e-commerce, thousands of volumes of customer-generated text are produced every single day in the form of reviews for certain products. Manually analysing this information becomes impracticable, hence developing an automated text classification system is needed.

This project focuses on the multi-class sentiment analysis of Amazon product reviews, whereby each review is tagged with a particular sentiment class: Negative, Neutral, or Positive. The goal is to bring forth a machine learning-based text classification model that can determine the polarity of sentiment from review text to scale customer opinion and feedback analysis.

### 1.2 Machine Learning and NLP Concepts Applied

NLP is a crucial area of study for dealing with unstructured data present in the form of text, which needs to be converted into a structure that can be easily processed by machine learning algorithms. In this research work, various NLP processing tasks would be performed for preprocessing the review text data. These would include tasks such as tokenization, removing stop words, and lemmatization.

The proposed work focuses on a supervised learning method for sentiment analysis using machine learning techniques, with the sentiment labels inferred from the ratings given by users. The text is then represented as numerical features using Term Frequency-Inverse Document Frequency (TF-IDF), measuring the importance of words in the given text data. Two models of machine learning algorithms would be developed for sentiment analysis: Multinomial Naive Bayes, a probabilistic model that is widely used for text-based classification tasks, and Logistic Regression, a discriminative model that is effective for data with high dimensional features.

## 2. Problem Domain

### 2.1 Problem Description

Multi-class sentiment classification of customer product reviews is the issue this project attempts to solve. The goal is to automatically classify a customer's textual review into one of three categories based on the sentiment it expresses: negative, neutral, or positive. Since neutral reviews frequently contain conflicting or ambiguous opinions, making sentiment boundaries less clear, multi-class sentiment classification adds more complexity than binary sentiment analysis. For the analysis of vast amounts of user-generated content on online platforms, accurate sentiment classification is crucial. It is not feasible to manually analyse customer opinions because e-commerce websites receive thousands of reviews every day. Scalable interpretation of customer feedback is made possible by automated sentiment analysis, which also promotes a deeper comprehension of customer behaviour and satisfaction.

### 2.2 Dataset Description and Background

The Amazon Fine Food Reviews dataset, which was acquired from Kaggle, will be used in this project. This dataset, which covers reviews from October 1999 to October 2012, is made up of customer reviews of high-end food items sold on Amazon. It is appropriate for sentiment analysis and text classification tasks because it includes actual customer feedback.

The dataset includes approximately 568,454 reviews, 256,059 unique users, 74,258 unique products. Each review contains product and user-related metadata, a numerical rating score, and a plain text review describing the customer's experience. For the purpose of this project, the following attributes will be primarily used: Text, which is the written customer review and Score, a numerical rating ranging from 1 to 5. To define the sentiment classes required for supervised learning, sentiment labels will be derived from the rating scores as follows:

- Ratings 1-2 labelled as Negative

- Rating 3 labelled as Neutral
- Ratings 4-5 labelled as Positive

The dataset is transformed into a three-class sentiment classification problem. Positive reviews are more common than neutral and negative reviews, indicating a natural class imbalance in the dataset. This imbalance poses a further difficulty for machine learning models, especially when it comes to correctly identifying minority classes, and reflects realistic customer reviewing behaviour. To ensure computational efficiency while maintaining the overall characteristics and distribution of the data, a representative subset of the reviews will be chosen for model development due to the dataset's size.

### 2.3 Societal and Business Relevance

Sentiment analysis of customer reviews is important for e-commerce businesses today. Businesses can gather customer reviews to improve product quality and develop ways to enhance the end-user experience. By automating the classification of customer sentiment, businesses can efficiently evaluate large amounts of customer review data to make data-based business decisions regarding the improvement of their products, whether to continue to produce certain products due to quality assurance, or to improve their overall marketing strategy.

Socially, sentiment analysis systems aggregate customer opinion trends and help customers make informed purchasing decisions based upon their collective satisfaction levels with a product or service. Therefore, if customers can access accurate and reliable customer feedback using sentiment analysis technology, this will create a more open, trustworthy, and beneficial business environment for e-commerce websites.

### 3. Solution

#### 3.1 Overview of the Proposed Solution

The proposed solution follows a standard machine learning pipeline for text classification. The system takes raw customer review text as input and processes it through a series of Natural Language Processing steps to prepare the data for machine learning. The cleaned and processed text is then transformed into numerical features that can be understood by classification algorithms.

Using supervised learning, the transformed features will be used to train machine learning models that learn patterns associated with different sentiment categories. The trained models are capable of predicting whether an unseen Amazon product review expresses negative, neutral, or positive sentiment. Multiple models are implemented to allow comparison of performance and behaviour in a multi-class sentiment classification setting.

This pipeline-based approach ensures that the solution is modular, interpretable, and suitable for analysing large-scale real-world textual data such as customer reviews.

#### 3.2 Text Preprocessing Techniques

The text data are inherently unstructured as they contain lots of noise, which always hurts the performance of machine learning. Some steps of text preprocessing are applied before feature extraction to improve data quality for consistency.

First, conversion all review text to lowercase to maintain uniformity and avoid taking the same words as different tokens due to case differences. Then, removing punctuation, special characters, and unnecessary symbols to reduce the noise in the text.

Next comes tokenization, which means that the reviews are split into single words or tokens. This can enable the model to analyse text at a word level rather than just as raw sentences. Following tokenization, stop words are removed, which are commonly

occurring words such as "the", "and", and "is" that do not carry much semantic meaning to reduce dimensionality and focus on terms relevant to sentiment.

Finally, lemmatization is performed to return the words to their base or dictionary form, such as "eating" to "eat". This helps in clustering similar words together, reduce vocabulary size, and thereby improve generalization by making different grammatical forms of a word fall into the same feature.

These steps in preprocessing clean and standardize the text into a more workable feature to be extracted and enhance the effectiveness of machine learning models.

### 3.3 Feature Extraction Methods

Numerical input data is required by all machine learning algorithms, which corresponds directly to the need for the numerical representation of pre-processed text data. In this project, TF-IDF (term frequency-inverse document frequency), an important method of text feature extraction, was chosen for use.

TF-IDF provides a numerical representation of text data, where it determines how important a particular word is within a given document relative to the greater corpus of documents being analysed. This weighting provides more importance to those words that may appear frequently within an individual review yet appear relatively infrequently when considering all reviews of that same product, thus allowing the model to more accurately identify those words that will contribute more to accurately classifying the sentiment of the review. TF-IDF performs better when working with high-dimensional datasets, and implements well with more traditional machine learning models, such as Naive Bayes and logistic regression.

### 3.4 Machine Learning Models Implemented

Two different supervised machine learning approaches will be used to accomplish multi-class sentiment classification in this project. These models were chosen to compare and contrast performance with several algorithms, as well as provide useful insight into the

relative performances of a number of different types of algorithms on high-dimensional textual data.

The Multinomial Naive Bayes model is a probabilistic classifier that uses Bayes' Theorem to create the model and assumes that the features are conditionally independent. Multinomial Naive Bayes is known to perform well on text classification tasks, as it performs well with discrete input. As the simplest of machine learning algorithms, its ease of use makes it an excellent candidate for baseline sentient analysis activities.

The second model is Logistic Regression, a well-known supervised learning method. Logistic Regression has a discriminative approach, whereby a probability for each class is generated based on the linear combination of the various input features used to create the model. The use of the TF-IDF representation therefore allows Logistic Regression to work with a very high-dimensional dataset. Multi-class classification can be accomplished with Logistic Regression by using the multiple one-vs-all discriminator classifiers. As such, Logistic Regression typically provides improved results relative to other probabilistic approaches as long as the training dataset has sufficient sample size to determine the parameters of the decision boundary.

### 3.5 Diagrammatic Representation

The overall process of classification is illustrated in the flowchart.

Initially, all unprocessed Amazon product reviews will go through a text preprocessing step. This preprocessing step is where the unprocessed text will be cleaned, tokenized, with stop-words removed, and lemmatized to produce a standard format of text.

Then, using TF-IDF vectorization, the clean text will be turned into a vector of numeric features. The two classifiers used to classify these features into a sentiment category will be Multinomial Naïve-Bayes and Logistic Regression. Finally, the trained classifiers will predict the sentiment of each review: positive, neutral, or negative.

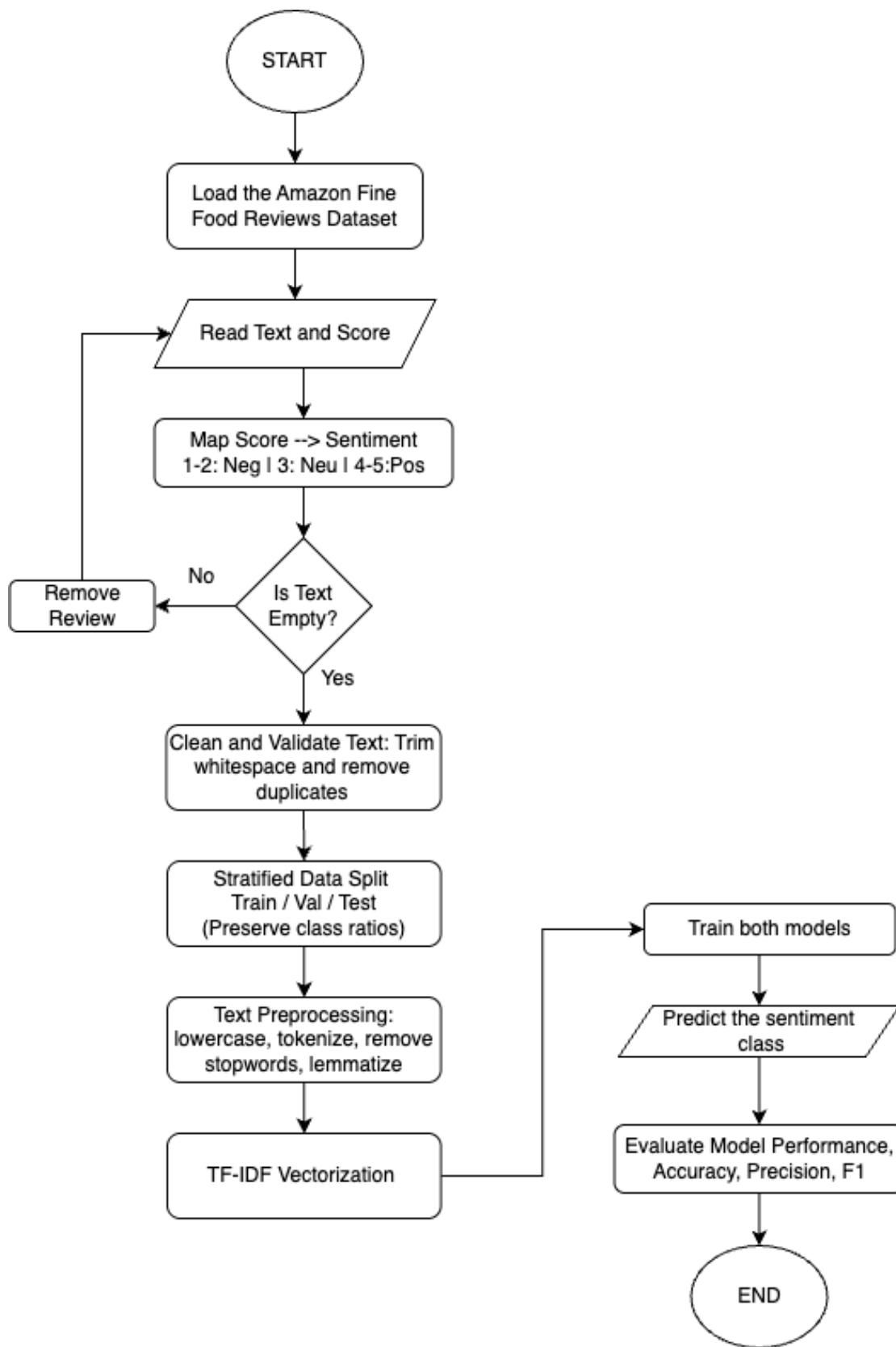


Figure 1 Diagrammatic Representation

### 3.6 Development Process and Tools

The development of the proposed system will take place in Python and Google Colab because it allows users to combine code with visuals, tests, and comments to create a unique document for conducting machine-learning experiments.

Libraries that are open source will be used throughout the system's development. In particular, libraries such as Pandas and NumPy facilitate data entry, cleaning, and manipulation, while NLTK or SpaCy will be used for tasks related to NLP like tokenisation, removing stop words, and lemmatising words. To assess and present models, graphs and visualisations will be created using plotting libraries such as Matplotlib and Seaborn. Scikit-Learn library will allow for the extraction of features and the developing of models, as it offers efficient implementations of TF-IDF vectorisation, along with many machine-learning algorithms including Multinomial Naïve Bayes and Logistic Regression.