Week 10 – Sequence-Aware Model (Fast)

Student Name: Minahil Irfan
Roll Number: 2225165126
Date: January 4, 2026

1. Introduction

This report presents the development and evaluation of a sequence-aware machine learning model for sentiment analysis using a subset of a large text dataset. The task aimed to implement a fast and efficient model capable of classifying text messages into positive and negative sentiment classes.

Sequence-aware models take into account the order of words, capturing contextual relationships that improve classification performance compared to models that treat words independently. In this assessment, a **Logistic Regression model** was applied using **n-gram features**, which allows the model to recognise sequences of words (unigrams and bigrams).

2. Dataset

The dataset used was a CSV file containing labelled text data. For this assessment:

- Only the column text and target were considered.
- A random sample of 20,000 instances was extracted to reduce computation time while maintaining sufficient data for training and evaluation.
- The target column was preprocessed to convert any labels with value 4 to 1, ensuring a binary classification problem (0 = negative, 1 = positive).

3. Methodology

3.1 Data Preparation

- The dataset was split into training and testing sets with an 80-20 split.
- Text vectorisation was performed using **CountVectorizer**, capturing both **unigrams and bigrams** (n-gram range of 1–2).
- The vocabulary was limited to the **3,000 most frequent features** to improve computational efficiency.

3.2 Model Training

- A **Logistic Regression model** was trained using the vectorised text features.
- The max_iter parameter was set to 1000 to ensure model convergence.

3.3 Evaluation Metrics

The model was evaluated using the following metrics:

- **Accuracy**: Overall correctness of the model on the test set.
- **Precision, Recall, F1-Score**: For both classes (0 and 1) to assess model performance on individual sentiment categories.

4. Results

The model achieved the following performance on the test set:

Accuracy: 0.7555

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.76	0.72	0.74	1943
1	0.75	0.79	0.77	2057
Accuracy	-	-	0.76	4000
Macro Avg	0.76	0.75	0.75	4000
Weighted Avg	0.76	0.76	0.76	4000

The results indicate that the model performs well on both sentiment classes, achieving balanced precision and recall values. The use of n-grams helped capture context in the text, which contributed to the improved classification accuracy.

5. Conclusion

This assessment successfully implemented a sequence-aware model for sentiment classification. Key takeaways include:

- Logistic Regression, combined with n-gram features, is an effective baseline for text classification tasks.
- Limiting the feature set to 3,000 n-grams allowed for faster computation without significantly impacting accuracy.
- The model achieved a solid performance of **75.5% accuracy** on the test set, demonstrating its ability to distinguish between positive and negative sentiments.

Future improvements could include the use of **deep learning models** such as LSTM or Transformers to better capture long-range dependencies in text sequences.

6. GitHub Repository

MinahilIrfan98/DataScience-AI