Week 9: Neural Networks Basics

Assignment 9

Name: Minahil Irfan

Course: DataScience and AI

Tool Used: Python (Jupyter Notebook – Anaconda)

1. Introduction

The purpose of this assignment is to apply Artificial Neural Network (ANN) concepts to a real-world dataset and compare its performance with an earlier machine learning model. Sentiment analysis was performed on a large Twitter dataset to classify tweets into positive and negative sentiments. A Logistic Regression model was implemented as a baseline, which serves as a foundation for further ANN-based experimentation.

2. Dataset Description

The dataset used in this study consists of **1,599,982 Twitter messages** collected for sentiment analysis. The dataset contains the following six attributes:

- **target** – Sentiment label (0 = Negative, 4 = Positive)
- **ids** – Unique tweet identifier
- **date** – Timestamp of the tweet
- **flag** – Query flag
- **user** – Twitter username
- **text** – Tweet content

There were **no missing values** in any of the columns, making the dataset suitable for direct analysis without additional data cleaning.

3. Data Preprocessing

The text column was selected as the input feature, while the target column was used as the sentiment label. Since the dataset encodes positive sentiment as 4 and negative sentiment as 0, labels were kept in their original form for classification.

Text data was converted into numerical features using the **CountVectorizer** technique with a maximum of 5,000 features. This approach transforms textual data into a bag-of-words representation suitable for machine learning models.

4. Baseline Model: Logistic Regression

Logistic Regression was implemented as the baseline model to establish a performance benchmark before applying Artificial Neural Networks. The dataset was split into **80% training data** and **20% testing data**.

Evaluation Metrics Used

- Accuracy
- Precision
- Recall
- F1-score

5. Results and Performance Evaluation

The Logistic Regression model achieved the following results:

Accuracy

- **78.91%**

Classification Report

Sentiment	**Precision**	**Recall**	**F1-Score**
Negative (0)	0.80	0.77	0.78
Positive (4)	0.78	0.81	0.79
Overall Accuracy			**0.79**

The model demonstrated balanced performance across both sentiment classes, indicating effective learning from textual features.

6. Discussion

The results show that Logistic Regression performs well for large-scale text classification tasks and serves as a strong baseline model. However, Logistic Regression is limited in capturing complex, non-linear relationships in data. Artificial Neural Networks can potentially improve performance by learning deeper representations of text features.

This baseline model establishes a foundation for implementing ANN-based architectures in future stages of the project.

7. Conclusion

This assignment successfully applied a machine learning model to a large Twitter sentiment dataset. The Logistic Regression baseline achieved an accuracy of approximately **79%**, demonstrating reliable sentiment classification performance. The results provide a benchmark for future experimentation with Artificial Neural Networks, fulfilling the Week 9 project milestone of establishing a baseline model.

8. Tools and Technologies Used

- Python (Anaconda – Jupyter Notebook)
- Pandas, NumPy
- Scikit-learn
- Matplotlib, Seaborn

9. GitHub Repository

GitHub Link:
MinahilIrfan98/DataScience-AI