## **NLP Project Report**

### **Objective**

The primary goal of this project is to conduct sentiment analysis on the Amazon review dataset using various machine learning models and SpaCy's built-in neural network model to extract meaningful insights.

### **Dataset**

The dataset, obtained from Kaggle, comprises Amazon product reviews. It consists of a 'reviewText' feature and we calculate using ‘SentimentIntensityAnalyzer’ a 'sentiment\_label' feature, where sentiment can be either positive or negative. After removing missing values, the dataset contains 2100 records which we consider for further sentiment analysis.

Dataset : https://www.kaggle.com/datasets/tarkkaanko/amazon

### **Data Preprocessing**

Several preprocessing steps were applied to the dataset using the Spacy framework and NLTK library:

* Normalization: Convert all text to lowercase.
* Removal of punctuation and numerical characters.
* Elimination of stop words.
* Lemmatization: Reduce words to their base form.

### **Visualization**

Visualizations were performed to gain insights:

* A histogram illustrated the distribution of reviewText lengths.
* Named entities were counted, and a bar plot depicted their frequency.

### **Feature Engineering**

* Word vectors were generated using SpaCy's word vectors functionality.
* Syntax trees were constructed using SpaCy to analyze sentence structures.

### **Model Selection and Training**

* Logistic Regression, Random Forest, and SVM models were trained using Count Vectorization technique.
* The best-performing model was selected based on cross-validation accuracy.
* SpaCy's built-in text categorization model was trained using the dataset.
* Hyperparameter tuning was conducted to improve model performance.

### **Model Evaluation**

* The accuracy of each model was evaluated using cross-validation.
* The best hyperparameters for each model were identified.

### **Best Hyperparameters**

| **Algorithm** | **Accuracy** | **Parameters** |
| --- | --- | --- |
| Logistic Regression | 78.9% | Count Vectorization |
| Random Forest | 77.0% | Count Vectorization |
| SVM | 78.3% | Linear Kernel, C=10 |
| SpaCy NN Model | 83.9% | 4 epochs, batch size 10 |

### **Conclusion**

This project demonstrates the effectiveness of different techniques in sentiment analysis and provides insights into customer sentiments towards Amazon products. The best-performing model achieved an accuracy of 83.9%.