### **SMS Spam Detection Analysis Report**

#### **1. Introduction**

This project builds and analyzes a model that can effectively classify SMS messages as either "spam" (unsolicited) or "ham" (legitimate). Using the SMS Spam Collection dataset, we applied various preprocessing techniques and machine learning algorithms to accomplish this goal.

#### **2. Dataset Description**

The dataset used for this project is the SMS Spam Collection Dataset, which consists of 5,574 messages labeled as either "spam" or "ham". The breakdown is as follows:

* Ham (legitimate): 87%
* Spam (unsolicited): 13%

The dataset has two columns:

* v1: Label (either 'ham' or 'spam')
* v2: The raw text of the SMS message.

#### **3. Data Preprocessing**

Several preprocessing steps were applied to clean and prepare the dataset for modeling:

* Lowercasing all text to ensure uniformity.
* Removal of special characters and punctuation.
* Tokenization of the text into words.
* Removal of stopwords.
* Stemming of words to their base form using the PorterStemmer.

These steps were applied to ensure the dataset was in an appropriate format for vectorization and classification.

#### **4. Model Building**

A **Logistic Regression** model was trained and evaluated on the dataset:

* **Logistic Regression**: This model was trained using a TF-IDF vectorizer to transform the text data into numerical features. The model achieved an accuracy of **96.59%** on the test data.

#### **5. Model Evaluation**

Model performance was evaluated using several metrics, including accuracy, precision, recall, and F1-score. A confusion matrix was also plotted to visually represent the model's performance. Below are the evaluation results:

* Accuracy: 96.59%
* Precision (spam): 0.99
* Recall (spam): 0.75
* F1-score (spam): 0.86

These metrics indicate that the model performs exceptionally well in distinguishing between legitimate and unsolicited SMS messages.

#### **6. Conclusion**

The Logistic Regression model achieved high accuracy on the SMS Spam Classification task. The model is a viable option for practical spam detection systems. Future work could involve exploring hyperparameter tuning and other text classification techniques to further improve the model's performance.

### **Submitted by:**

**Vishnu PV**, August Batch 2024