|  |
| --- |
|  |
| SMS Spam Detection Documentation |
| Text Classification Model |

|  |
| --- |
| Lee Jaryl  222441U |

**About the Data:**

The ‘spam.csv’ dataset used in training the model contains one set of SMS messages in English of 5,574 messages, classified as ham (legitimate) or spam. The 5574 messages consists of a collection of 425 SMS spam messages that was manually extracted from the Grumbletext Web site, a subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), a list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis and finally 1,002 SMS ham messages and 322 spam messages from the SMS Spam Corpus v.0.1 Big. The Dataframe has two relevant columns, ‘v1’ which indicates the class of message as ham or spam and the second column ‘v2’ is for the string of text. There is significantly more ham than spam messages in the dataset.

**Relevance:**

This data is relevant in training the model as it only consists of spam and ham SMS text messages, which compared to for example, emails may have different contexts and structure that may affect the accuracy of the model. The specificity of the dataset will greatly increase the model’s accuracy and hence its relevance to training it for spam detection.

**Irregularities:**

* **Irrelevant Columns:** Columns Unnamed: 2, Unnamed: 3, and Unnamed: 4 are null and should be dropped.
* **Unicode Decoding Error:** The utf-8 codec could not decode certain bytes correctly and caused some data to be corrupted.

**Data Preparation:**

Data preparation is crucial to ensure the quality and integrity of the dataset before training the machine learning model. It includes tasks such as cleaning the data, handling missing values, and transforming the data into a suitable format for model training. This step can affect the accuracy of the final model.

**Steps:**

1. Data Cleaning

* Remove irrelevant columns (Unnamed: 2, Unnamed: 3, Unnamed: 4).
* Ensure there are no missing values.

1. Data Exploration

* Explore the data to gain insights into the characteristics of spam versus non-spam messages.
* Find out if the most frequent words are stop words.

1. Data Preprocessing

* Remove the stop words to improve the performance of the model as it will allow it to focus on more meaningful words.
* Encode labels as the model requires numerical input instead of categorical data.

**Model Training:**

The process of selecting and optimizing the model to perform its best at a given task. This performance can be measured later with different metrics.

**Steps:**1. Data Splitting:

* Split the dataset into training and testing sets to evaluate model performance later.

2. Model Selection:

* Select the most appropriate machine learning algorithm for the task.

3. Hyperparameter Tuning:

* Optimize the model parameters to improve performance.

**Model Evaluation:**

Testing the model on the test set created earlier to determine which algorithm and parameters performed best.

This performance can be determined by the following metrics:

* Accuracy: The ratio of correctly predicted instances to the total instances.
* Precision: The ratio of true positive predictions to the total predicted positives.
* Recall: The ratio of true positive predictions to the actual positives.
* F1 Score: The harmonic mean of precision and recall, providing a balance between them.

In our case, we want to use the model with the highest possible test precision and from those models, the one with the highest accuracy. This is because spam messages will be blocked by the SMS spam detection system while the ham will be let through, the consequence of accidentally blocking a ham message is much greater than accidentally letting through a spam message. Hence, having no false positives to prevent ham messages from being flagged as spam is the utmost priority, only then followed by its accuracy.