**Spam Ham Classifier**

***Architecture Documentation***

# **Architecture**

Data Insertion into Database

Data Ingestion

Start

Data for Recommendation

Transforming Text Data

Data Preprocessing

Data Validation

Splitting Data

Model Pusher

Model Evaluation

Model Building

NLKT Techniques

Data From User

Application Start

Pushing app to Cloud

Cloud Setup

Data Transformation

Model Call for Prediction

End

Spam Classification

# **Architecture Description**

## **Data Description**

This corpus has been collected from free or free for research sources at the Internet:

* A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site.
* A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore.
* A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis.

This dataset contains a total of 5574 texts classified as ham or spam.

## **Data Insertion into Database**

1. Database Creation and connection – Create a database with name passed.
2. Inserting data in JSON format in MongoDB.

## **Data Ingestion**

Imported data directly from database and created a DataFrame (tabular form) out of the JSON formatted data.

## **Splitting Data**

Divided data into three parts – Train set, Test set and Validation set.

## **Data Validation**

Creating perimeters to ensure any upcoming data is of same format as the current one. If conditions aren’t met then program must stop here, else proceed to next steps.

## **Data Pre-processing**

Data Pre-processing or Data Transformation steps includes Null value handling, removing duplicated values etc.

## **Transforming Text Data**

Transforming the text so as to make it ready for model building. Lower casing the text, stop words removal, punctuation removal, Tokenization, stemming.

## **Model Building**

We will use TFIDF first to vectorize the text data. Then we will find the best model. We will calculate the Accuracy Score, Precision Score & False Positive section of Confusion Matrix. Whichever algorithm gives us max accuracy & precision score as well as least False Positive value in confusion matrix table (preferably 0), we will pick that model.

## **Model Evaluation**

We will evaluate the model using Test Data we segregated in Data Ingestion. If model new model outperforms ongoing model, then we will replace it.

## **Model Pusher**

After picking best model, we push the best performing model, transformer & encoder files into other directory to be easily picked during operation.

## **Cloud Setup**

We will be deploying the model on Streamlit Cloud service.

## **Pushing App to Cloud**

After cloud setup, we push the app to cloud and get it configured as per the use case to get text data from the user easily.

## **Data From User**

A text field provided on the app allows user to input text.

## **Data Transformation**

Received data will get transformed using NLTK techniques.

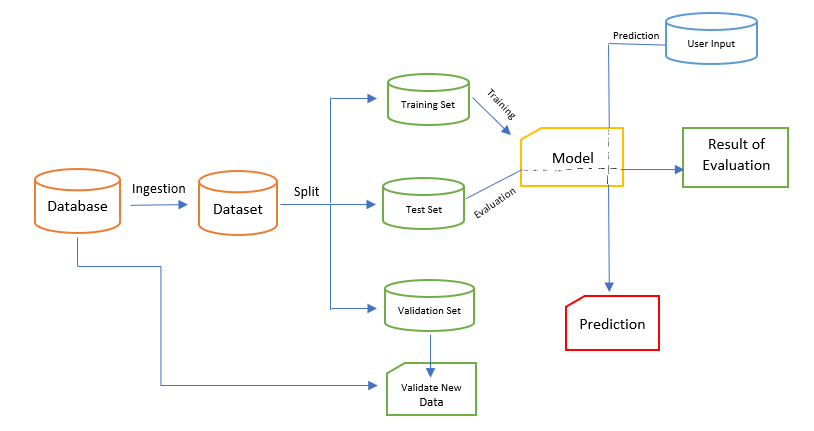
## **Model Call for Prediction**

Data gets through the model which instantly classifies whether the text is a ham or spam.

## **Spam Classification**

As soon the user inputs text, within a second user gets the display of result on the app front page that whether the text is spam or ham.

# **Model Training/Validation Workflow**



# **User I/O workflow**

**Check the info**

Failure

Exception

Handling

Start

User inputs

Text data

End

Display Result

On Web App

Prediction using Model

NLTK Text

Transformation

Data

Pre-Processing

Success