

Innovation

Phase 2 Project

Modules and Explanation

The Application consists of three modules.

- i. UI
 - ii. Machine Learning
 - iii. Data Processing
- 15

I. UI Module

- a. This Module contains all the functions related to UI(user interface).
- b. The user interface of this application is designed using Streamlit library from python based packages.
- c. The user inputs are acquired using the functions of this library and forwarded to data processing module for processing and conversion.
- d. Finally the output from ML module is sent to this module and from this module to user in visual form.

II. Machine Learning Module

- a. This module is the main module of all three modules.
- b. This modules performs everything related to machine learning and results analysis.
- c. Some main functions of this module are
 - i. Training machine learning models.
 - ii. Testing the model
 - iii. Determining the respective parameter values for each model.
 - iv. Key-word extraction.
 - v. Final output calculation
- d. The output from this module is forwarded to UI for providing visual response to user

III. Data Processing Module

- a. The raw data undergoes several modifications in this module for further process.
- b. Some of the main functions of this module includes
 - i. Data cleaning
 - ii. Data merging of datasets
 - iii. Text Processing using NLP
 - iv. Conversion of text data into numerical data(feature vectors).
 - v. Splitting of data.
- c. All the data processing is done using Pandas and NumPy libraries.
- d. Text processing and text conversion is done using NLP, K and scikit-learn libraries

Requirements

Hardware Requirements

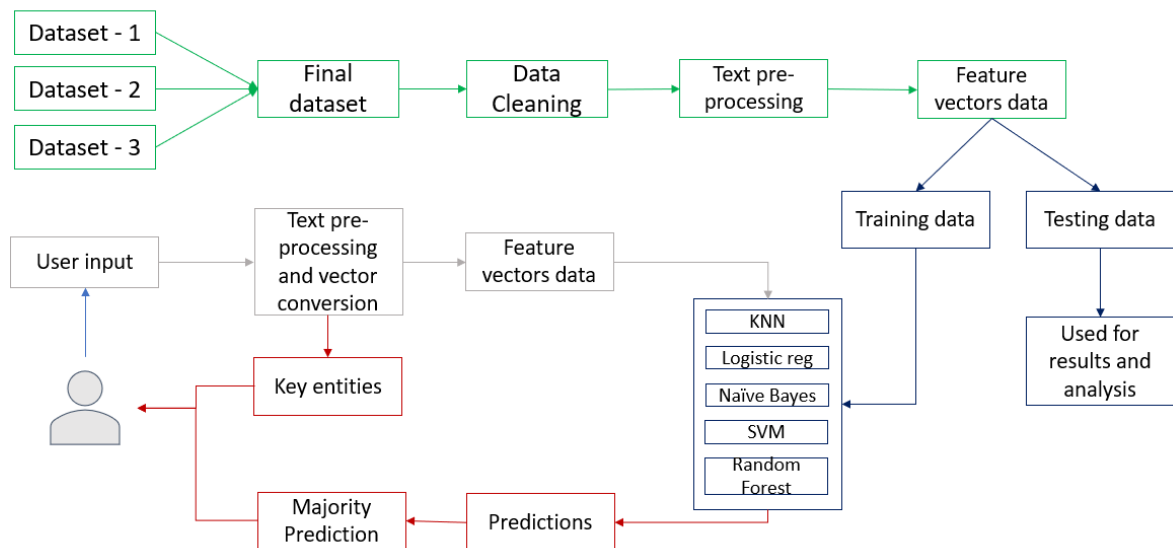
- PC/Laptop
- Ram – 8 Gig
- Storage – 100-200 Mb

Software Requirements

OS – Windows 7 and above

Code Editor – Pycharm, VS Code, Built in IDE
Anaconda environment with packages nltk, numpy, pandas, sklearn, tkinter, nltk data.
Supported browser such as chrome, firefox, opera etc..

WorkFlow



Data Collection and Description

- Data plays an important role when it comes to prediction and classification, the more the data the more the accuracy will be.
- The data used in this project is completely open-source and has been taken from various resources like Kaggle and UCI
- For the purpose of accuracy and diversity in data multiple datasets are taken. 2 datasets containing approximately over 12000 mails and their labels are used for training and testing the application.
- 6000 spam mails are taken for generalisation of data and to increase the accuracy.

Data Description

Dataset : enronSpamSubset.

Source : Kaggle

Description : this dataset is part of a larger dataset called enron. This dataset contains a set of spam and non-spam emails with 0 for non spam and 1 for spam in label attribute.

Composition :

- Unique values : 9687
- Spam values : 5000

- Non-spam values : 4687

Dataset : lingspam.

Source : Kaggle

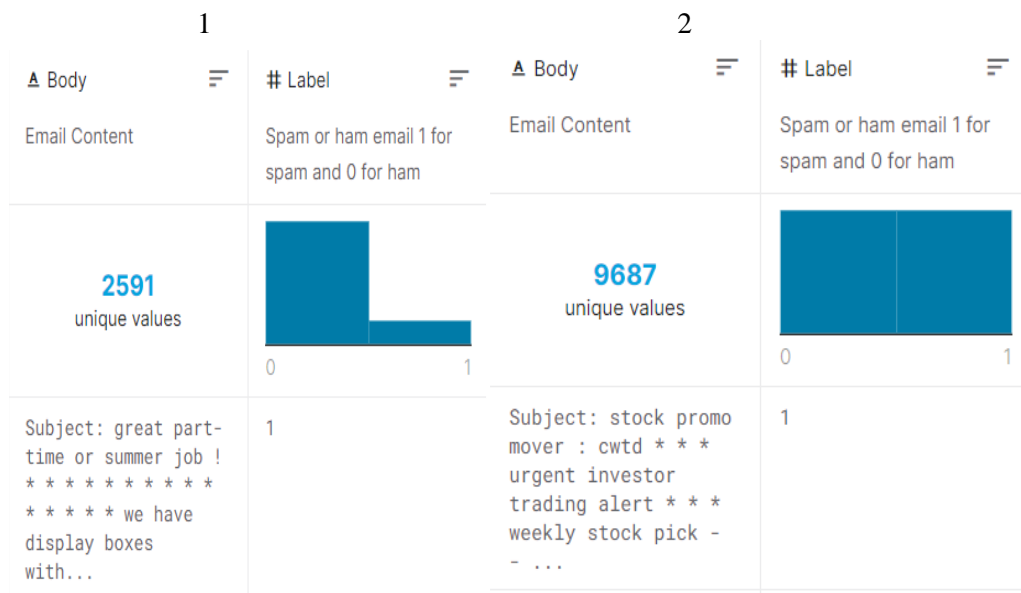
Description : This dataset is part of a larger dataset called Enron1 which contains emails classified as spam or ham(not-spam).

Composition :

Unique values : 2591

Spam values : 419

Non-spam values : 2172



Fig; lingspam