Faculty of Computer and Information Science
Ain Shams University

2022170175	زيد هاني محمد صلاح الدين
2022170029	أحمد عمرو الحسيني أمين
2022170038	أحمد محمد مجاهد
2022170050	ارساني عادل قسطندي تاوضروس
2022170036	احمد محمد عبد الرؤوف ابراهيم
2022170256	على أشرف ابراهيم سيد

Overview

Classify emails as "spam" or "valid" based on text.

Features like word frequencies, presence of spam keywords, and sender information.

SVM excels at handling high-dimensional data often present in text classification tasks.

Preprocessing

- First, we explore through the dataset to see its features, call function head () and tail () to see data at each row, seeing basic information about each column, for numerical columns we need description for them, so we call function describe () , to know (count , mean std, min , max.....etc.).

- Then we need to check if nulls exist and manipulate them by dropping cells which contain nulls or replace their value with median value if numeric or mod if categorial.

- After that we need to drop duplicated rows if they exist because it makes redundancy in the data.

- There are two columns which have same target, to show is the text is spam or not spam so column named “label” is the same column named “label num” only difference is one is categorial and another is numerical, so keep numerical columns so a better thing is to drop column “label.”

- Last step we need to select the feature which the models will train on them, we inserted new column has name “size text "to see if length of a text will show us if the text is spam or not spam,

- But there was no difference, so we needed to try another way, we checked every text that was sent to how many people we found that the range between “send emails” in case of spam is not intersected with the range of “sent emails “in case of ham so we didn’t use “sent emails”.

- best way to define any text whether if it is spam or not is by text itself, most of text that was defined spam contains strange words and unrelated to English words, so here was the key to define the text spam or not.

- Our feature selection decided to be text and applying all models based on column “Text.”

Training data

- We need two things before starting modeling on data, first convert column "Text" to something numerical so ai models can work on that column, second, technique to find the text which contains many words unrelated to English.

- TfidfVectorizer does exactly what we need, we applied this technique TfidfVectorizer does exactly what we need, we applied this technique to convert column "Text" to numerical by mathematical equation that calculates frequency of strange words divided by total words.

- At the same time, it does train the data to be ready to apply different models.

Models

- We have done main three requirements' models which are:

1. Logistic regression
2. Decision tree classifier
3. SVM

- Additional to them, we added another two models to try different accuracies which are:
    1. KNN
    2. Random Forest classifier

- So, in total there are 5 AI (Artificial Intelligence) models.

- We take each model and apply to it the features and the target to fit in data, then display classification report that contains confusion matrix which tells us precision and recall

- After that we display overall accuracy for both the train and test data for each model.

- We need to use a test called K-fold cross-validation in which the dataset is divided into k subsets or folds, the model is trained and evaluated k times, using a different fold as the validation set each time.

- Finally, we display overall accuracy both for the train and test data for each model.

Deployment

- Gui is made by web using HTML, CSS and JavaScript as the frontend and python flask as the backend

- Ai is ready to classify each new entered text to the system is spam or not spam.

- Users can select any of existing AI models and enter text and see whether text is spam or not spam and.

- Users can also see the selected models' test accuracy and train accuracy.