Email-spam-classifier

End to end code for the email spam classifier project

The main goal of this project is to detect the class of text message that is Spam or Ham based on the content of the message. Spam is a junk mail/message, or an unsolicited mail/message. Spam sms are also those unwanted, unsolicited sms that are not intended for a specific receiver. The good, perfect, and official mails are known as ham. Our objective is to alert users from spam or junk mail/messages so that they can save themselves from being cheated from spammers.

Data Collection

Spam Ham Message Collection Set from UCI Machine Learning Repository For Data Set: https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

Data Pre-Processing

•Lower Case
•Tokenization
•Removing Special Characters
•Removing Stop Words and Punctuation
•Stemming

DPR: Spam Ham Detection

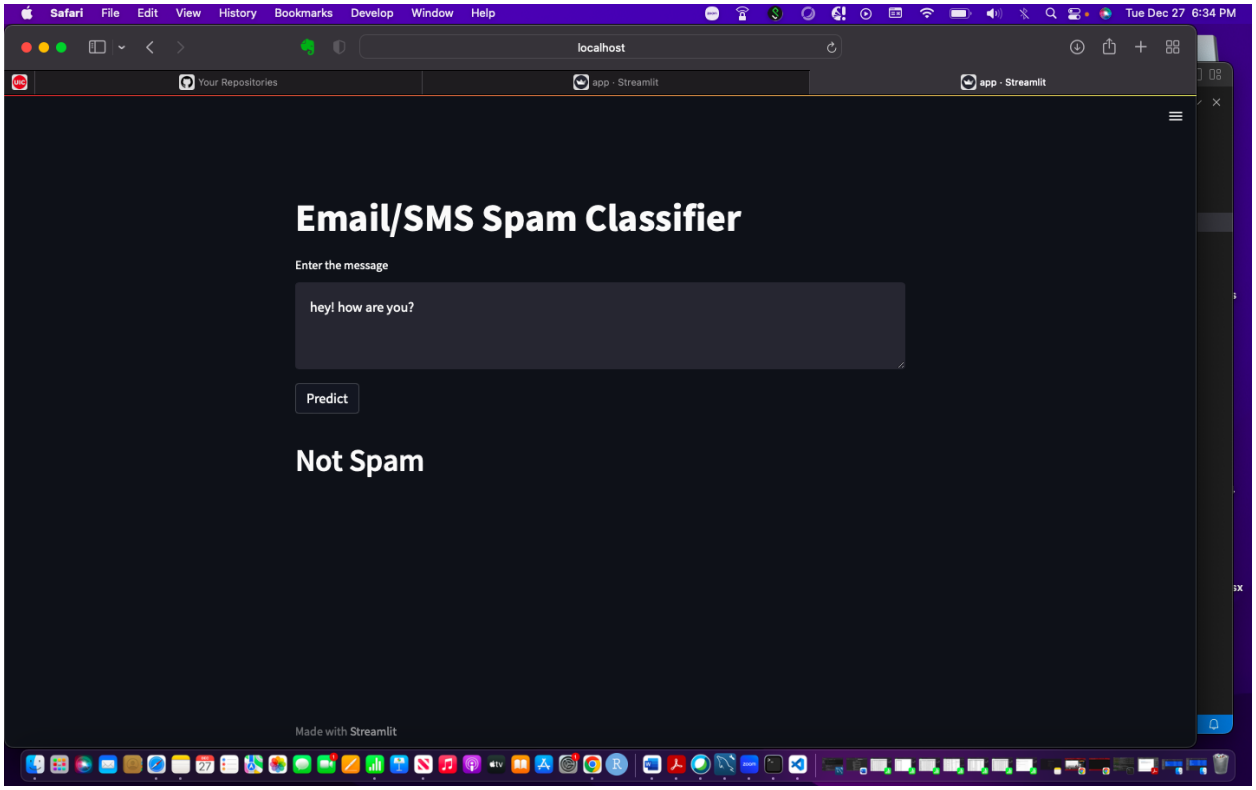
Data Cleansing

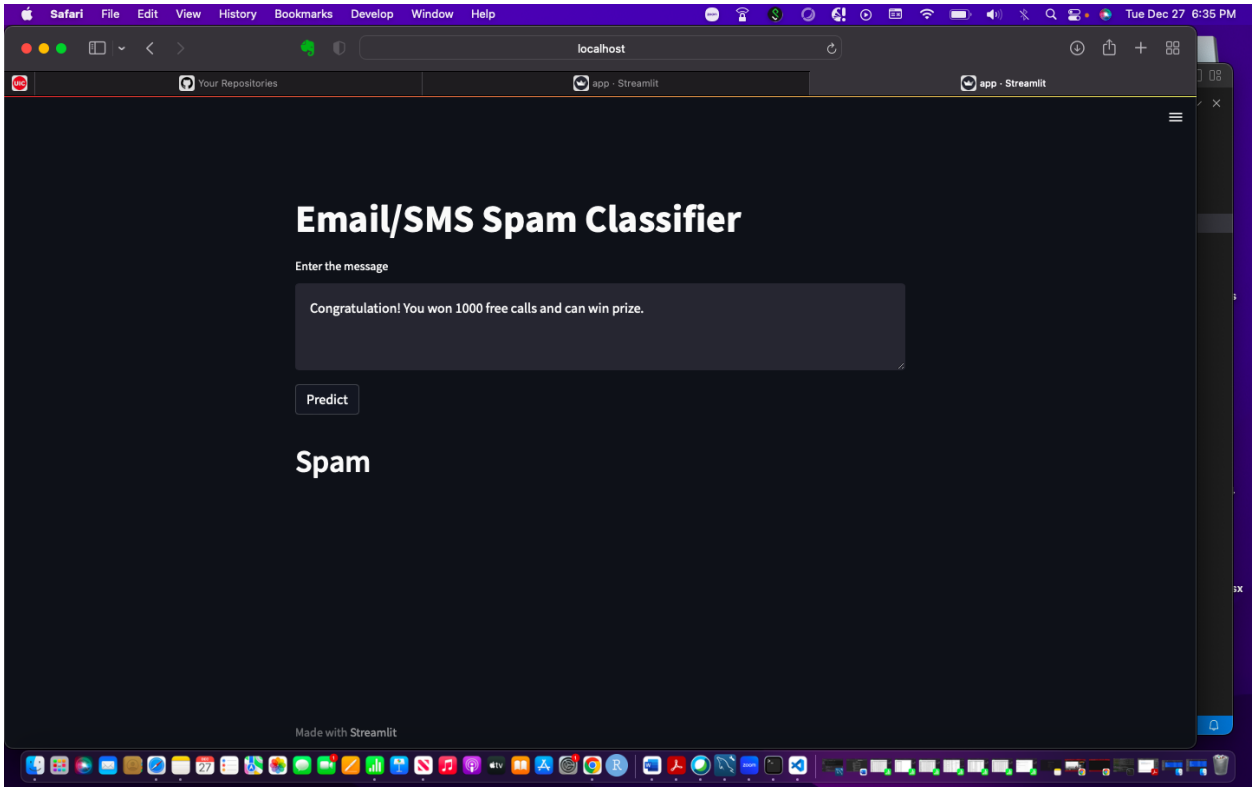
•Drop unnecessary columns.
•Drop duplicate rows from the dataset.
•Rename required columns.
•Encode target column using Label Encoder.

Text Preprocessing

•Lower Case
•Tokenization
•Removing Special Characters
•Removing Stop Words and Punctuation
•Stemming

Overview-


Safari File Edit View History Bookmarks Develop Window Help
Tue Dec 27 6:34 PM
localhost
Your Repositories
app · Streamlit
app · Streamlit
Email/SMS Spam Classifier
Enter the message
hey! how are you?
Predict
Not Spam
Made with Streamlit


Safari File Edit View History Bookmarks Develop Window Help
Tue Dec 27 6:35 PM
localhost
Your Repositories
app · Streamlit
app · Streamlit
Email/SMS Spam Classifier
Enter the message
Congratulation! You won 1000 free calls and can win prize.
Predict
Spam
Made with Streamlit

Model Creation and Evaluation

•Various classification models were created.

•Algorithms used are Naive Bayes, Logistic Regression, Decision Tree Classifier, Random Forest
Classifier, Ada-Boost Classifier, Bagging Classifier etc.

•Multinomial Naïve Bayes classifier has given good accuracy and precision.

•Model performance evaluated based on accuracy, precision.

•In TFIDF Vectorizer and Countvectorizer, We have selected TFIDF.

Multinomial Naive Bayes Algorithm

Multinomial Naive Bayes algorithm is a probabilistic learning method that
is mostly used in Natural Language Processing (NLP).
The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.
Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.
Advantages:

•It is easy to implement as you only have to calculate probability.

•You can use this algorithm on both continuous and discrete data.

•It is simple and can be used for predicting real-time applications.

•It is highly scalable and can easily handle large datasets.