PYTHON PROJECT ON :

SPAM EMAIL DETECTION

Submitted By : Ashribad Das
College : Institute of Technical Education and Research (ITER)
Guided by : Mr.Sambit Subhasish Sahu

KEY CONCEPTS :

"Electronic mail" or email refers to a method of exchanging digital messages over a network, typically the internet.

Spam email :
These are the undesired email which are sent in bulk , especially for advertising , phishing , spreading malware , or other malicious activities.

Types of Spam Emails :

- **Advertising Spam : These are those emails which promotes products or services or schemes .**
- **Phishing Emails : These are designed to steal sensitive information such as usernames, passwords, or credit card details .**
- **Scam Emails : These emails attempt to deceive the recipients into sending certain informations (e.g., lottery scams).**
- **Malware Emails : These emails contain the attachments or links that , when opened , install the malicious software on the recipient's device.**

Spam email detection is an essential field in cybersecurity . By collecting prior information regarding spam , the techniques used in spam detection, and the threats as well as challenges faced in our day to day life , engineers can continue to improve spam filters , making email communication safer and more efficient for users all over the world .

OBJECTIVES

Key Components:

1. Data Collection: Obtain a dataset of labeled emails
2 .Data Preprocessing : Cleaning and preprocessing the email data , including tasks such as tokenization , removing stop words , and vectorizing text data.
3.Feature Engineering : Extract relevant features from the text data, such as word frequencies, to train the machine learning model.
4.Model Selection: Experiment with different machine learning algorithms
5. Model Training : Model on the preprocessed email dataset to learn patterns and characteristics of spam and non-spam emails.

Technologies and Tools:

· Python programming language for model development and data preprocessing .

· Jupyter Notebook for interactive development and documentation of the project .

Benefits:

· Improved Efficiency: Automates the process of filtering spam emails and saving time of the user .

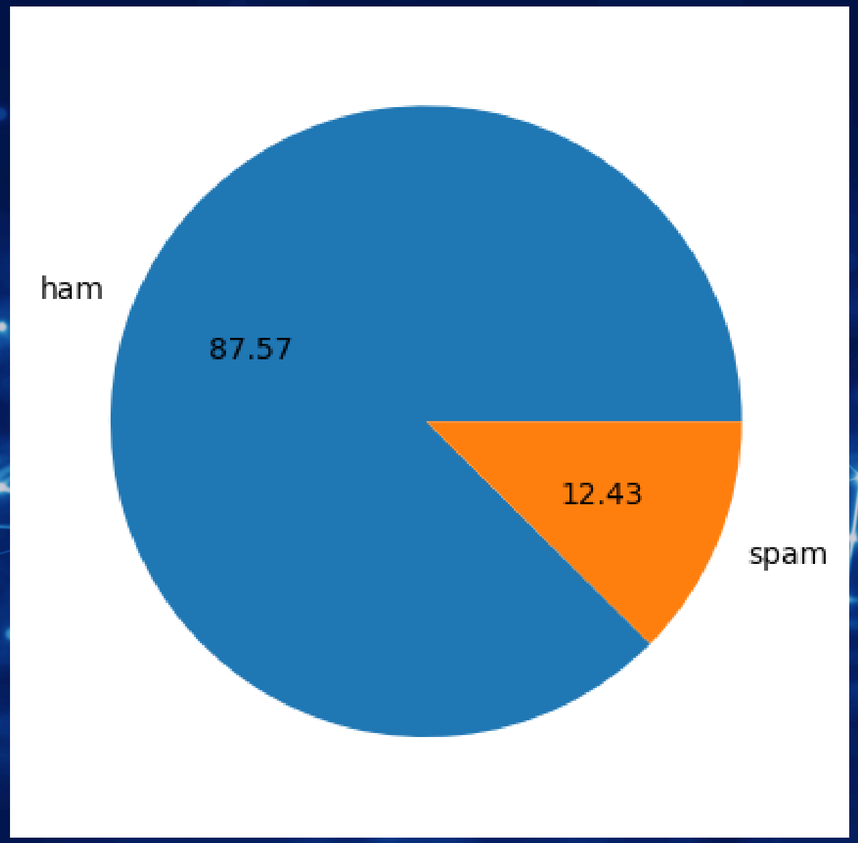
· Customization: The model can be fine-tuned to adapt to specific user preferences and evolving spam email patterns .

· Scalability: Can be scaled to handle large volumes of emails efficiently, making it suitable for both individual users as well as the organizations .

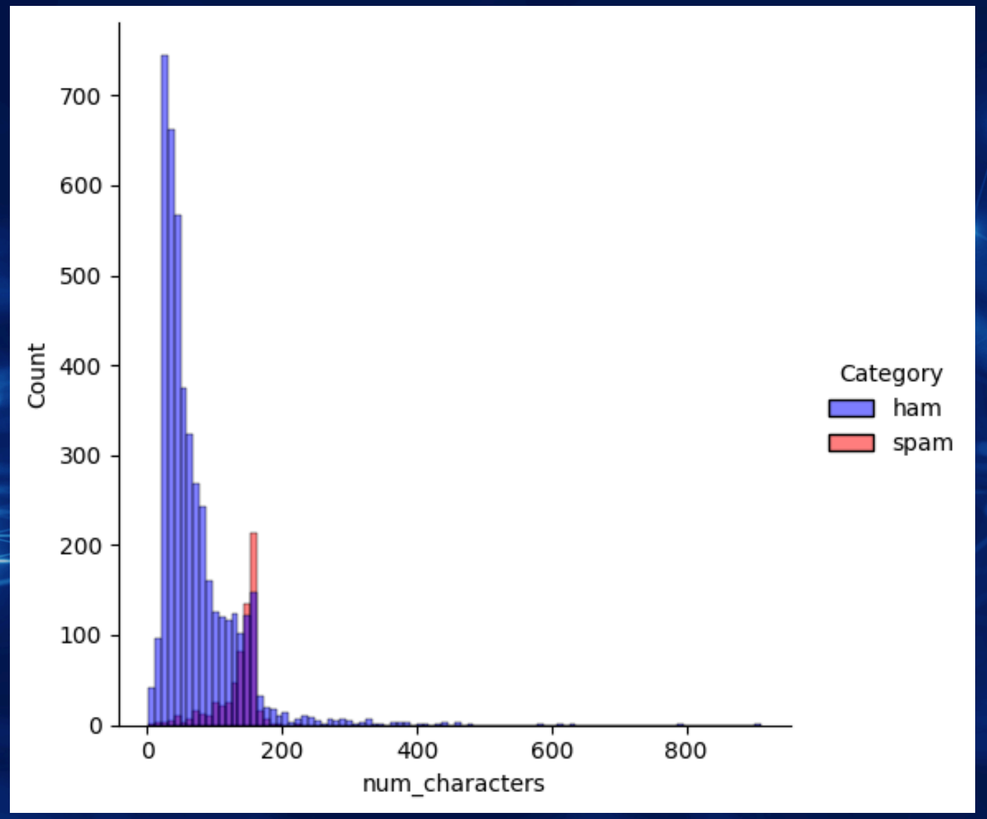
Importing Libraries

1.Numpy : Fundamental library for numerical computation
2.Pandas : Powerful library for data manipulation and analysis
3.Seaborn : Data visualisation library which is based on Matplotlib
4.Matplotlib: Python library used for creating static, interactive, and animated visualizations. It is especially useful for generating plots, histograms, bar charts, scatter plots , etc . .
5.NLTK: The Natural Language Toolkit is a popular python library for working with human language data .
6.Xgboost:This library deals with high performance and efficiency in various machine learning tasks .
7.Scikit-learn :It is often abbreviated as sklearn and is a powerful and widely-used Python library for machine learning. It provides simple and efficient tools for data mining, data analysis, and building machine learning models.
8.Joblib: Library for serializing and deserealizing Python objects .
9.Streamlit: A library that allows us to create the interactive web applications for the machine learning projects and models .

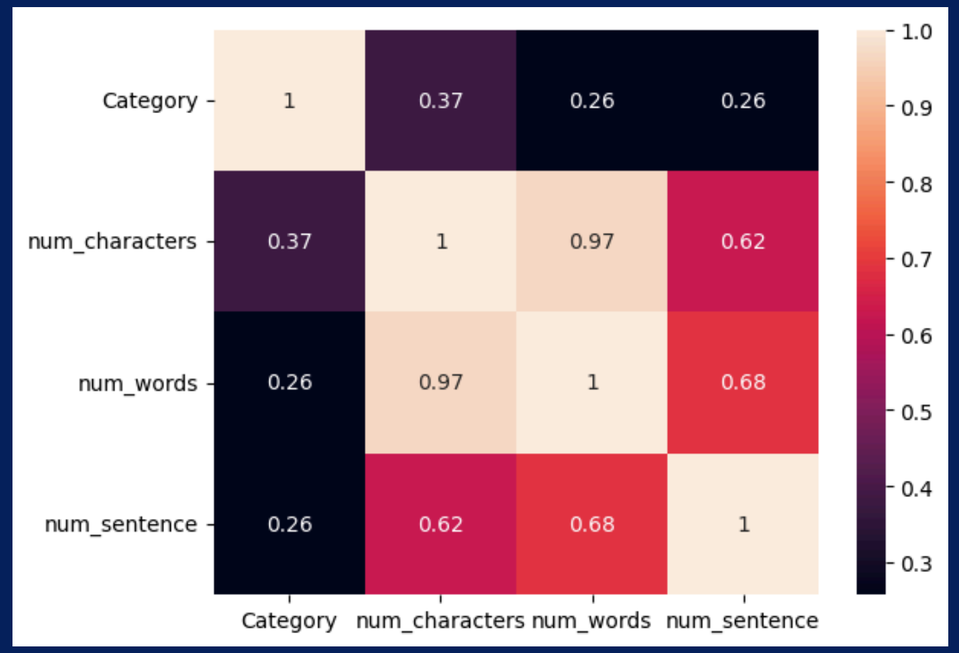
plt.pie(df['Category'].value_counts(),labels = ['ham','spam'],autopct = '%0.2f')



sns.displot(df,x = 'num_characters',hue = 'Category',palette = {'ham':'blue','spam':'red'})



```
sns.heatmap(df.drop(['Message'],axis = 1).corr(),annot = True)
```



seaborn.heatmap is a function in the Seaborn library used for creating heatmaps in Python. It's often used to visualize the values in a matrix-like format, such as confusion matrices or correlation matrices.

ACCURACY AND PRECISION OF MODELS :

Accuracy of GaussianNB after max_features is 0.8713278495887191
Precison of GaussianNB after max_features is 0.5

Accuracy of BernoulliNB after max_features is 0.981786133960047
Precison of BernoulliNB after max_features is 0.9795918367346939

Accuracy of MultinomialNB after max_features is 0.9764982373678026
Precison of MultinomialNB after max_features is 0.8874458874458875

Accuracy of KNN after max_features is 0.9171562867215041
Precison of KNN after max_features is 0.9875

Accuracy of ExtraTreesClassifier after max_features is 0.9788484136310224
Precison of ExtraTreesClassifier after max_features is 0.9692307692307692

Accuracy of RandomForestClassifier after max_features is 0.9741480611045829
Precison of RandomForestClassifier after max_features is 0.9679144385026738

Accuracy of XGBClassifier after max_features is 0.9770857814336075
Precison of XGBClassifier after max_features is 0.9455445544554455

MODEL SELECTED :
RANDOM FOREST CLASSIFIER

Bernoulli NB Model : Best in terms of accuracy of 0.9818
KNN Model : Best in terms of the Second metric

Accuracy of RandomForestClassifier after max_features is 0.9741480611045829
Precison of RandomForestClassifier after max_features is 0.9679144385026738
So RandomForestClassifier is the best model selected .

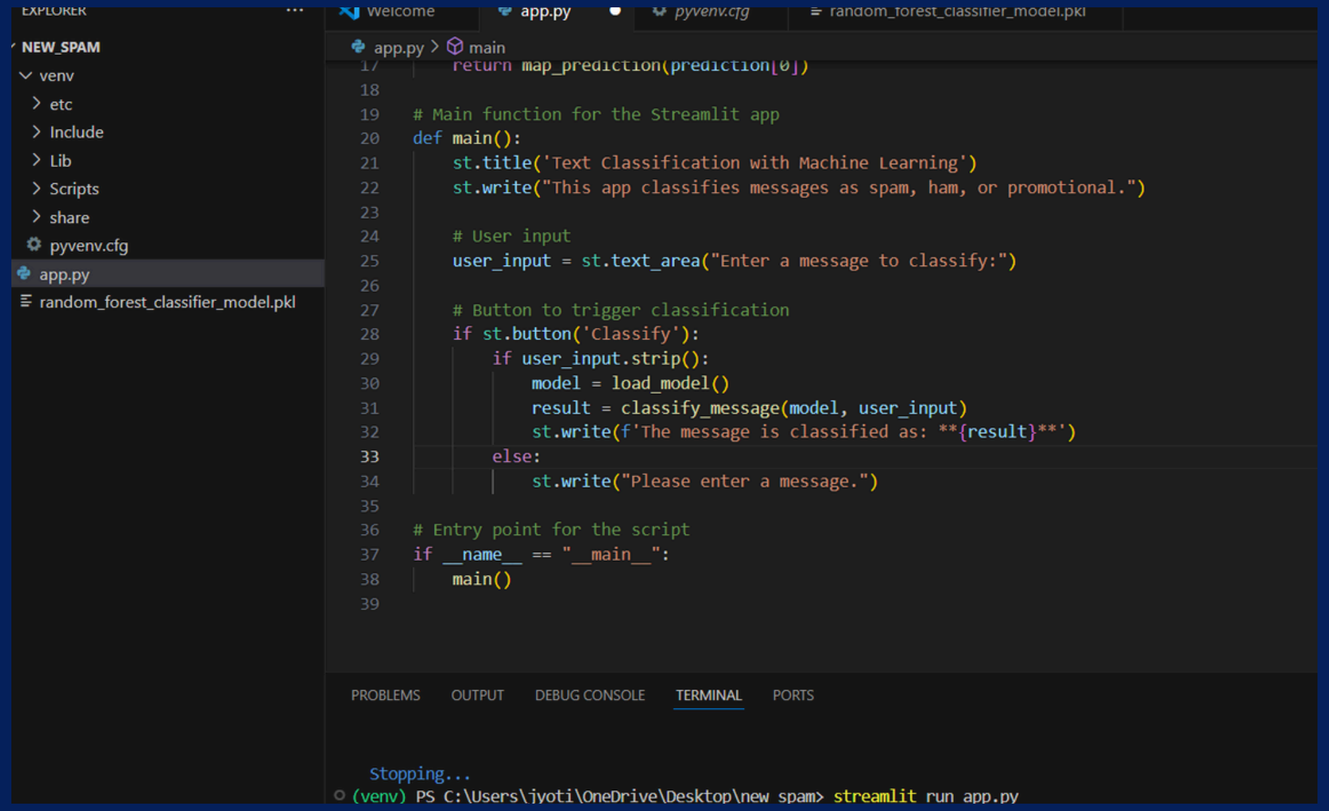
Random Forest Classifier

The Random Forest Classifier is an ensemble learning method which is used for classification tasks .

It operates by constructing a multitude of decision trees during training and outputting the mode of the classes for classification of the individual trees.

It combines multiple decision trees which improve accuracy and robustness.

STREAMLIT: A LIBRARY THAT ALLOWS US TO CREATE THE INTERACTIVE WEB APPLICATIONS FOR THE MACHINE LEARNING PROJECTS AND MODELS .



Applications of spam email detection

1. Email Filtering: Automatically filtering out spam messages from users' inboxes .
2. Reducing Phishing Attacks: Identifying and blocking phishing emails that attempt to steal personal information or credentials.
3. Improving Email Efficiency: Reducing the time and effort users spend on managing spam .
4. Protecting Networks: Preventing spam from overwhelming email servers .
5. Enhancing Security: Detecting and blocking emails that might contain malware or harmful links.
6. Compliance and Legal Requirements: Helping organizations comply with regulations regarding email communication and data protection.
7. Marketing Optimization: Allowing businesses to filter out spam and unwanted promotional emails, ensuring that marketing messages reach their intended audience.
8. User Experience: Providing a better overall experience for users by minimizing interruptions from unwanted emails.

LIMITATIONS OF SPAM EMAIL DETECTION

1. Legitimate emails might be mistakenly classified as spam, causing important messages to be missed.
2. Some spam emails might bypass filters and still appear in users' inboxes.
3. Spammers constantly evolve their tactics to bypass filters, requiring continual updates and retraining of detection models.
4. Language and Context: Filters might struggle with spam in less common languages or with context-specific content, leading to reduced accuracy.
5. Privacy Concerns: Analyzing email content for spam can raise privacy issues, especially if sensitive information is involved.

FUTURE DEVELOPMENTS AND APPLICATIONS OF MACHINE LEARNING IN SPAM EMAIL DETECTION

1. **Enhanced Accuracy:** Advances in ML algorithms, such as deep learning, could improve the accuracy of spam detection by better understanding context, semantics, and user behavior.
2. **Adaptive Learning:** Implementing models that continuously learn and adapt to new spam tactics and emerging threats in real-time, ensuring that spam filters remain effective.
3. **Personalized Filtering**: Tailoring spam filters to individual user preferences and behaviors, enhancing the relevance and accuracy of spam detection for each user.
4. **Integration with Security Systems:** Combining spam detection with other security systems, such as antivirus software and intrusion detection systems, for a more robust security solution.
5. **Explainable AI:** Developing models that offer explanations for their decisions, helping users and administrators understand why certain emails were classified as spam.

CONCLUSION :

This project has established a robust framework for spam email detection, combining machine learning techniques and utilizing a Random Forest model in conjunction with user-friendly tools by setting a strong precedent for future developments in email security.

THANK YOU FOR YOUR ATTENTION