Vidyavardhini's College of Engineering and Technology

Computer Engineering

Academic Year 2023-24

Machine Learning

Course Project Sem-VII

Email/SMS Spam Detection

Mentor
Dr. Megha Trivedi

Team Member

Dimple Khuman (Roll No. 04)

Divya Patil (Roll No. 08)

Aditi Sawant (Roll No. 12)

Introduction :

Email spam detection is a crucial aspect of modern communication, aiming to identify and filter out unsolicited, irrelevant, or potentially harmful emails from reaching users' inboxes. With the exponential growth of email traffic, efficient spam detection systems have become imperative to maintain productivity and security. These systems employ various techniques, including machine learning algorithms, natural language processing, and rule-based filters, to analyze email content.

Contents of PPT

1. Data cleaning
2. EDA
3. Text Preprocessing
4. Model building
5. Evaluation
6. Result
7. Conclusion

Dataset before cleaning

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
812	spam	Congratulations ur awarded either å£500 of CD ...	NaN	NaN	NaN
637	ham	When Ì_ login dat time... Dad fetching Ì_ home...	NaN	NaN	NaN
2341	ham	Tell dear what happen to you. Why you talking ...	NaN	NaN	NaN
824	ham	Have a good evening! Ttyl	NaN	NaN	NaN
1972	ham	Yes but can we meet in town cos will go to gep...	NaN	NaN	NaN

Data Cleaning

1. Removing unnecessary columns.
2. Renaming columns.
3. Handling missing values.
4. Removing duplicate rows.

Dataset after cleaning

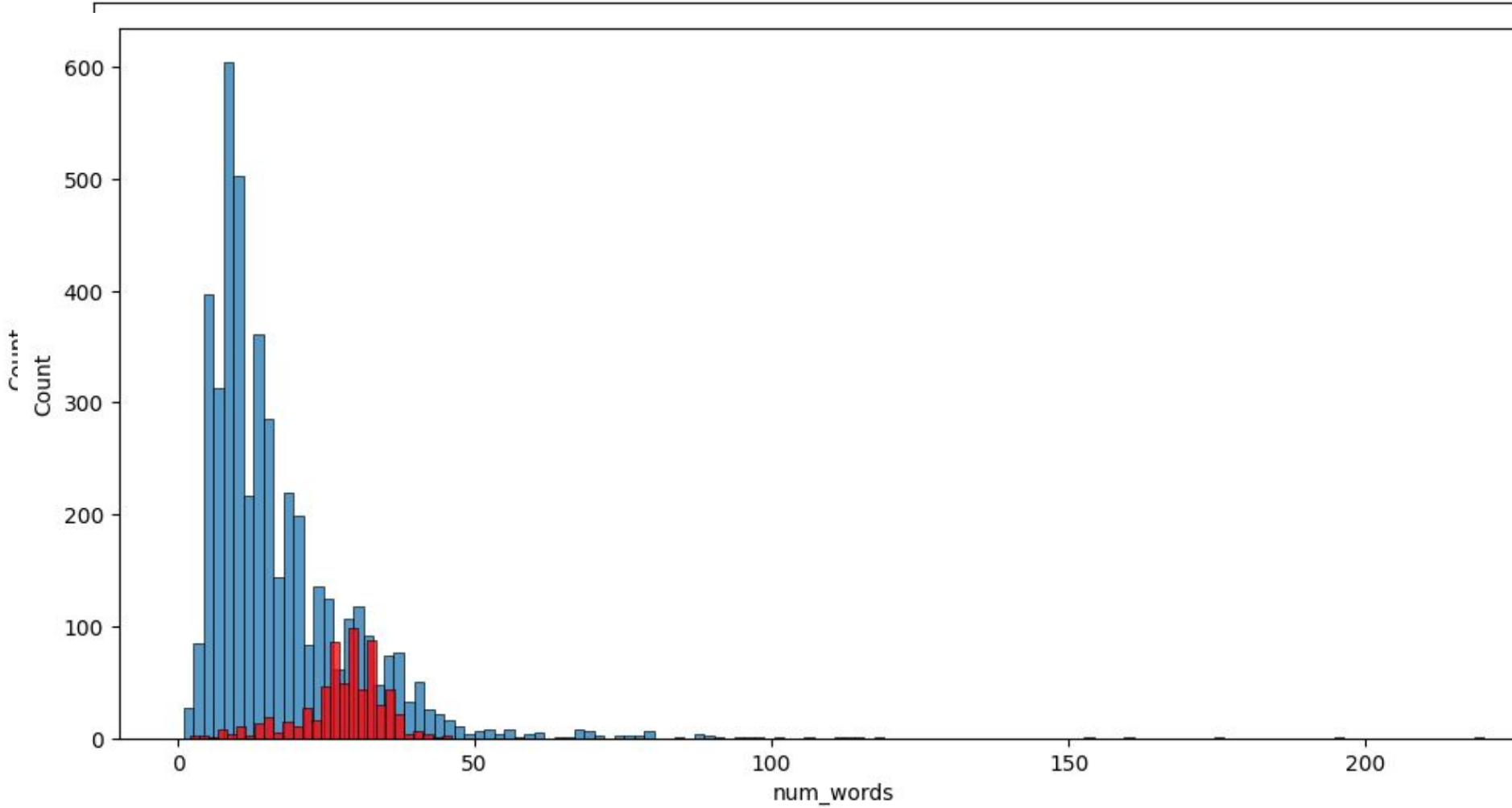
	target	text
4730	ham	K:)eng rocking in ashes:)
2748	ham	Send his number and give reply tomorrow mornin...
3946	ham	Sorry, went to bed early, nightnight
2379	ham	Good evening Sir, hope you are having a nice d...
5153	ham	Haven't left yet so probably gonna be here til...

	target	text
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

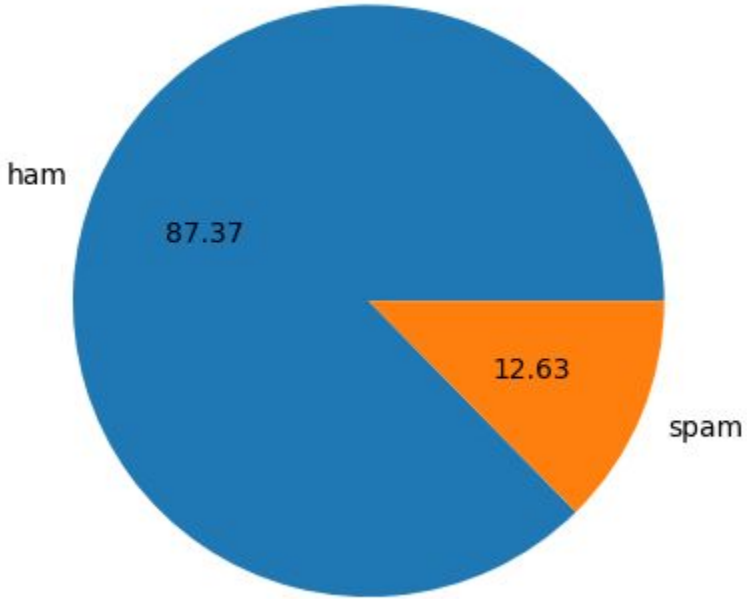
Exploratory Data Analysis (EDA):

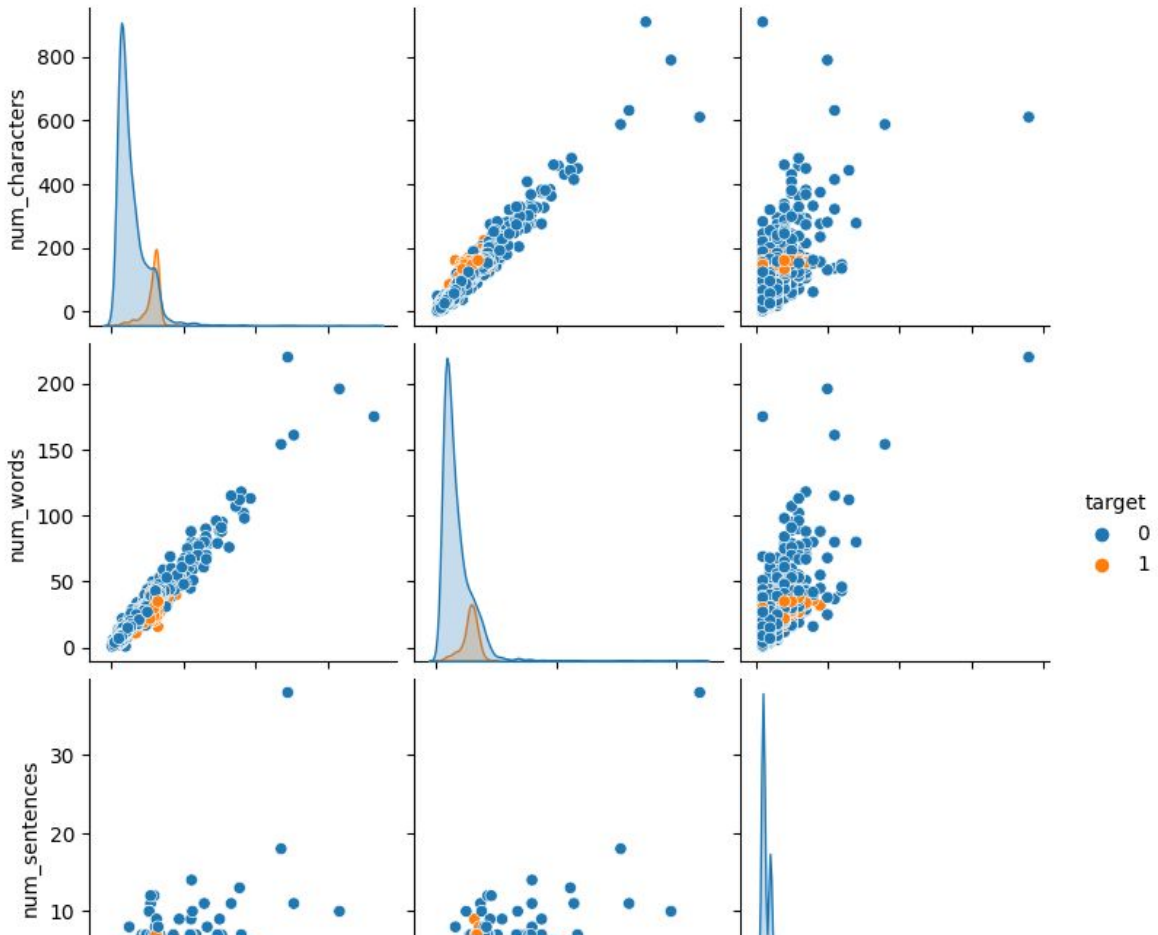
EDA involves understanding your data through statistical and visual techniques.

1. **Pie Chart**: A pie chart was used to visualize the distribution of the "ham" and "spam" labels in the dataset, showing the data's class balance.
2. **Histograms**: Histograms were used to visualize the distribution of the number of characters, words, and sentences in the text data. Separate histograms were created for "ham" and "spam" classes.
3. **Pair Plot**: A pair plot was used to visualize pairwise relationships between numerical features like the number of characters, words, and sentences.
4. **Heatmap**: A heatmap was used to visualize the correlation matrix of numerical features in the dataset.


Count
600
500
400
300
200
100
0
0
50
100
150
200
num_words


target
num_characters
num_words
num_sentences
1
0.38
0.26
0.26
0.38
1
0.97
0.62
0.26
0.97
1
0.68
0.26
0.62
0.68
1
target
num_characters
num_words
num_sentences
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3


ham
87.37
12.63
spam


num_characters
num_words
num_sentences
0
200
400
600
800
0
50
100
150
200
10
20
30
target
0
1

Text Preprocessing:

Text preprocessing is a crucial step in preparing text data for machine learning models, ensuring that they can effectively learn from the data and make informed classifications.

- Lowercasing.
- Tokenization.
- Removing special characters, punctuation, and stopwords.
- Stemming.

Data Preprocessing Results

Original Text (Before Preprocessing):

"I'm gonna be home soon and I don't want to talk about this stuff anymore tonight, k? I've cried enough today."

Text After Preprocessing:

"gon na home soon want talk stuff anymor tonight k cri enough today"

Word Clouds

Word clouds are generated for both spam and ham messages using the WordCloud library, providing a visual representation of the most common words in each category.

Model Building:

Several machine learning models are trained and evaluated:

- Naive Bayes (GaussianNB, MultinomialNB, BernoulliNB).
- Other classifiers (Support Vector Classifier, Decision Tree, Logistic Regression, Random Forest, AdaBoost, Bagging, Extra Trees, Gradient Boosting, XGBoost).
- Each model's accuracy and precision are calculated.

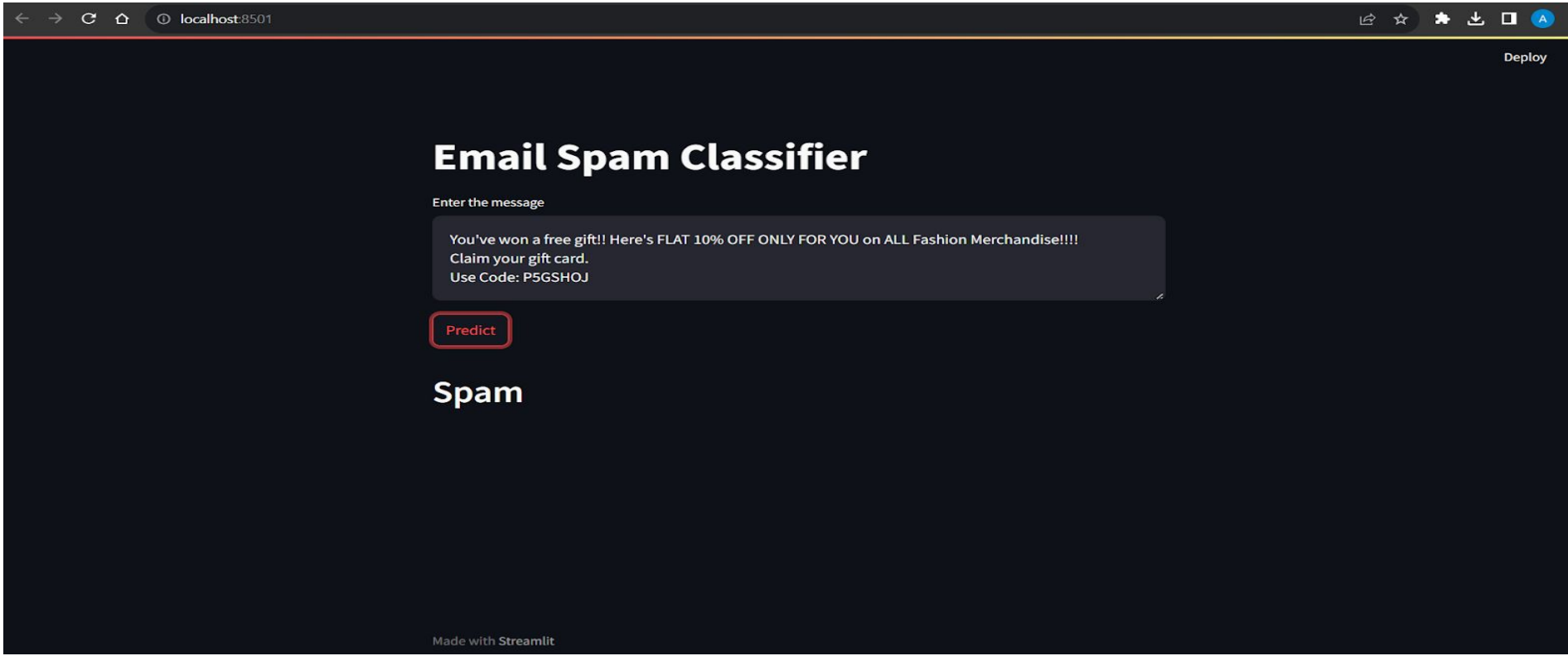
Model Comparison:

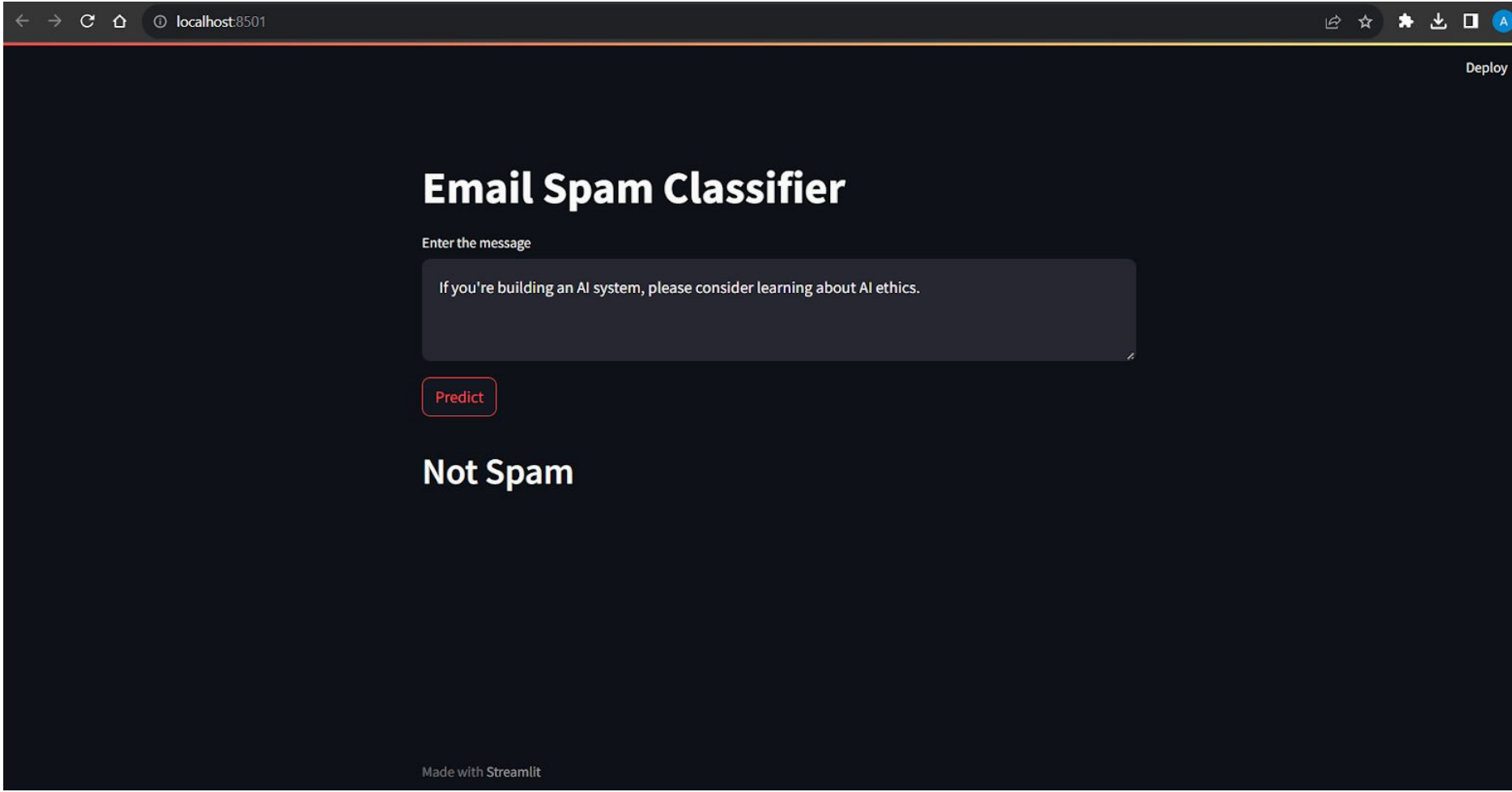
The script compares the performance of different models based on accuracy and precision. It creates a DataFrame to summarize these metrics.

Model Improvement:

The code explores potential model improvements, such as changing TF-IDF's max_features parameter and scaling features (although the scaling part is commented out).

Result


localhost:8501
Deploy
Email Spam Classifier
Enter the message
You've won a free gift!! Here's FLAT 10% OFF ONLY FOR YOU on ALL Fashion Merchandise!!!!
Claim your gift card.
Use Code: P5GSHOJ
Predict
Spam
Made with Streamlit


localhost:8501
Deploy
Email Spam Classifier
Enter the message
If you're building an AI system, please consider learning about AI ethics.
Predict
Not Spam
Made with Streamlit

Conclusion

The proposed "Email/SMS Spam Detection" system represents a significant leap forward in the realm of digital communication. This system aims to address the ever-growing challenge of spam, which disrupts our inboxes, poses security risks, and compromises the efficiency of digital interactions.By implementing advanced machine learning algorithms, user-friendly interfaces, and customization options, the system offers a comprehensive solution to these issues. With the ability to identify and classify messages as "Spam" or "Not Spam," the system empowers users to regain control over their digital communication environment.

Thank You!