Youtube Spam Classifier.

22nd august,,2022

project presentation

Advisor

Ameena Najeeb

Student

Shaik Irshad Ahmed

Group -8

Table of Contents

I Libraries included.

II Data preprocessing/ cleaning

III Data visualization

IV Model building

V Model Depoyment & Conclusion

PROBLEM STATEMENT

Youtube Spam and Genuine Comment Classification.

Task: Binary classification problem to classify comments to spam or ham (not spam)

Input: Youtube comments

Output: Spam (or) Ham (Not spam)

- This approach would aid in constructing an algorithm for blocking/removing inadmissible and undesirable comments from the view. In this project, NLP techniques are used to preprocess the data, before training the classification algorithm.

DATA DESCRIPTION

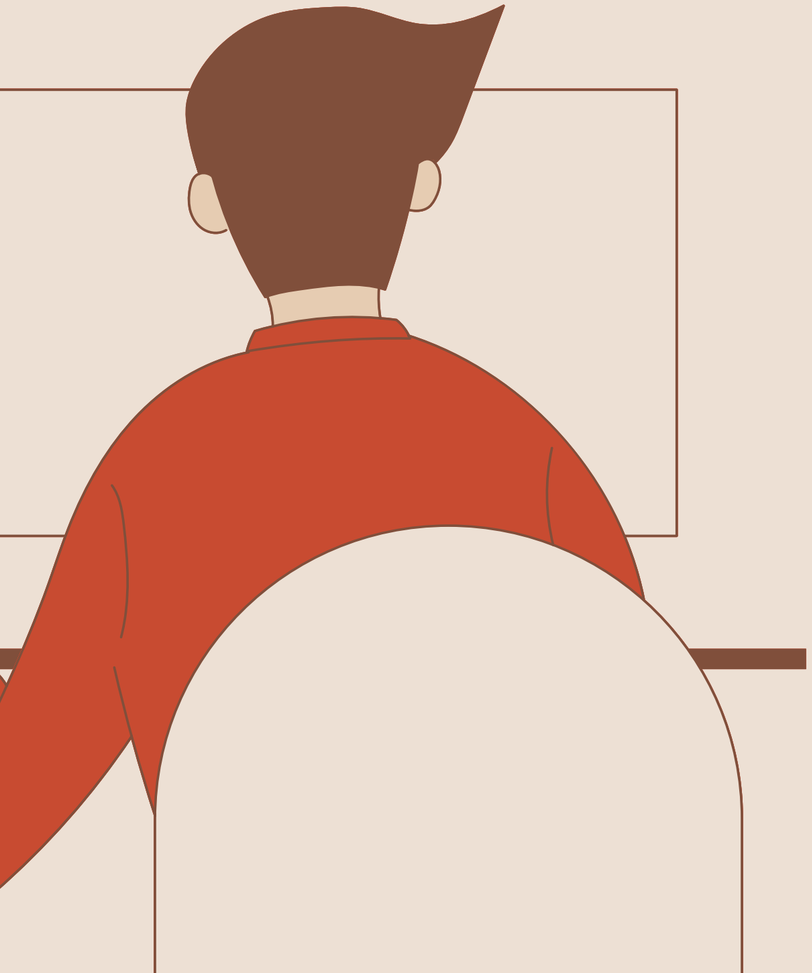
The dataset is obtained from UCI Repository. It has five csv files composed of 1,956 tweet data extracted from five videos that were among the 10 most viewed on the collection period from 2015.The downloaded data folder has 5 data sets (csv files) :

1. Youtube01-Psy
2. Youtube02-KatyPerry
3. Youtube03-LMFAO
4. Youtube04-Eminem
5. Youtube05-Shakira

The collection is composed by one CSV file per dataset, where each line has the following attributes:

COMMENT_ID, AUTHOR, DATE, CONTENT, TAG And Target class which has 0/1 values.

LIBRARIES INCLUDED

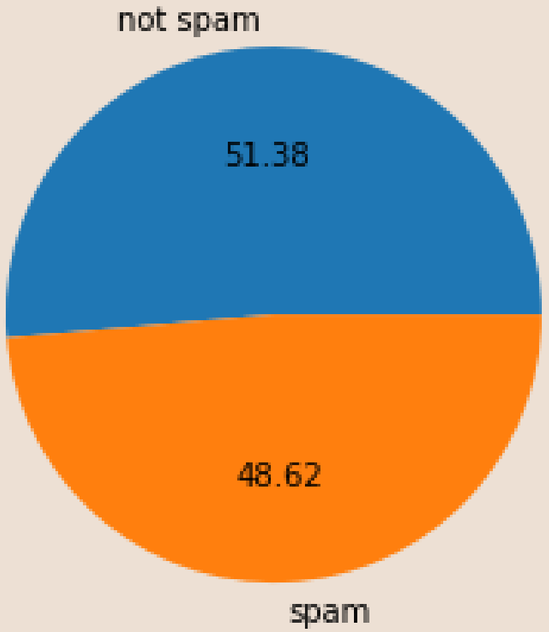


- OS
- Pandas
- Numpy
- Matplotlib
- Seaborn
- NLTK
- Sklearn
- Pickle
- Flask

22nd august,2022

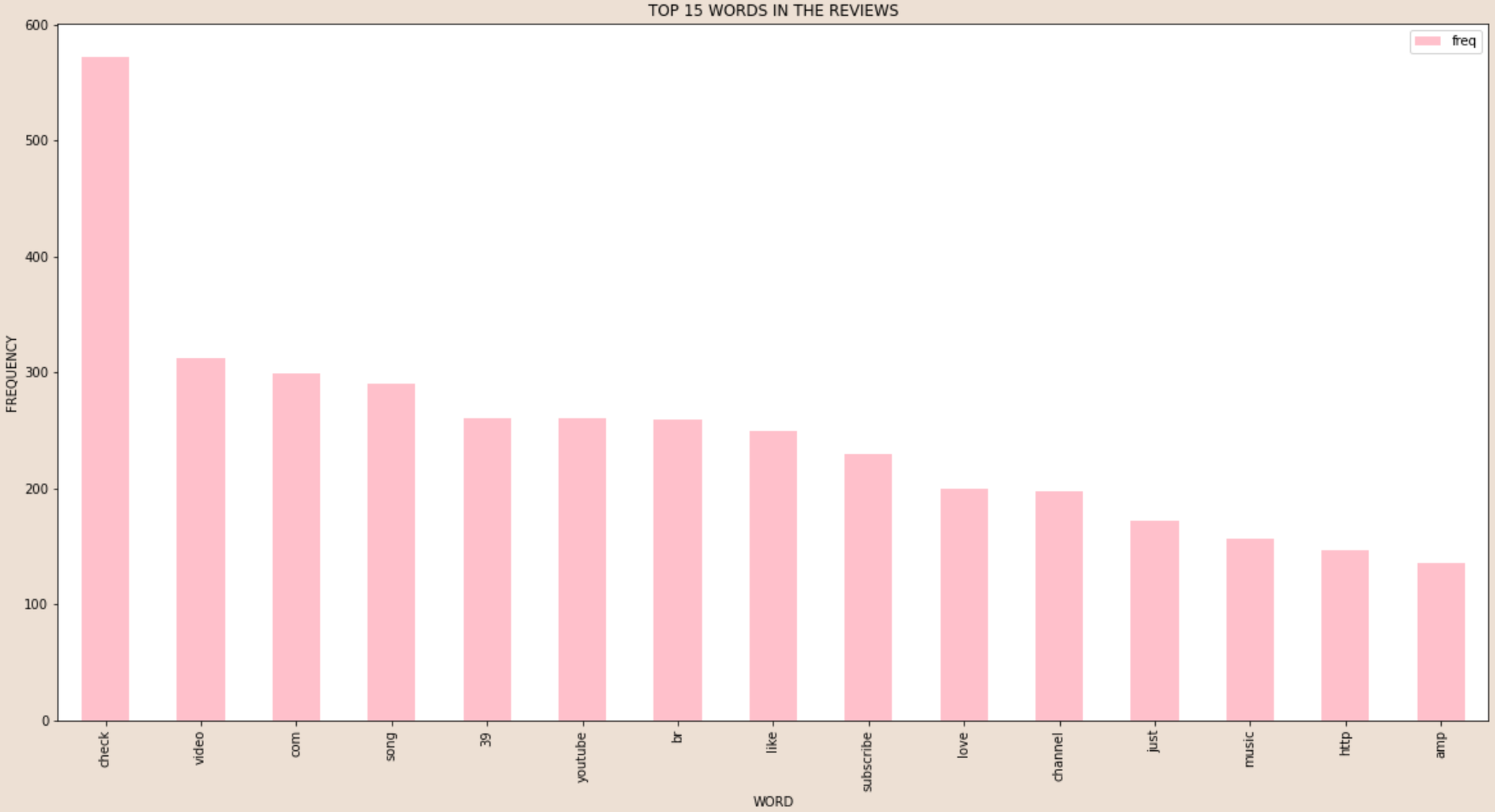
DATA PREPROCEESING AND CEANING

About 78% of the data that has length more than 95 are spam. Average length of spam comments is more than non-spam comments.



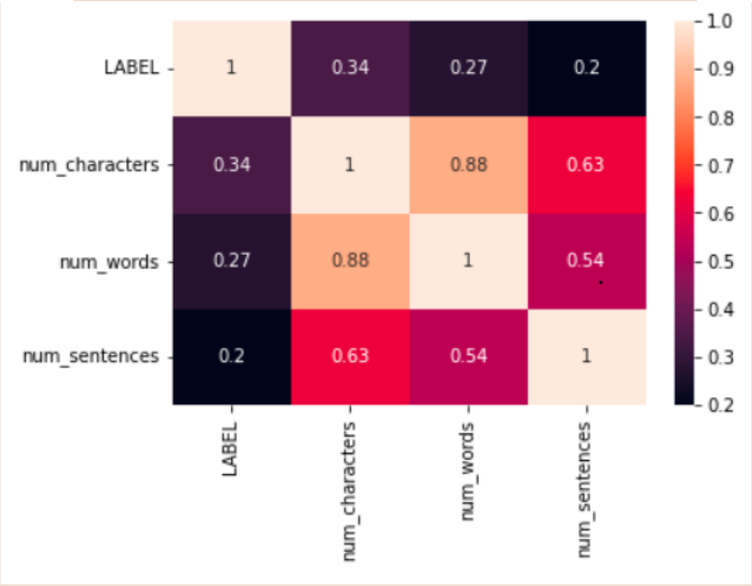
	TEXT	LABEL	num_characters	num_words	num_sentences
0	We pray for you Little Psy ♡	0	29	7	1
1	This awesome song needed 4 years to reach to 8...	0	118	23	2
2	The rap: cool Rihanna: STTUUPID	0	36	7	1
3	Hey dickwad - we're all africans. The colo...	1	233	58	3
4	i am from Brazil please subscribe my channel I...	1	58	11	1

DATA VISUALIZATION

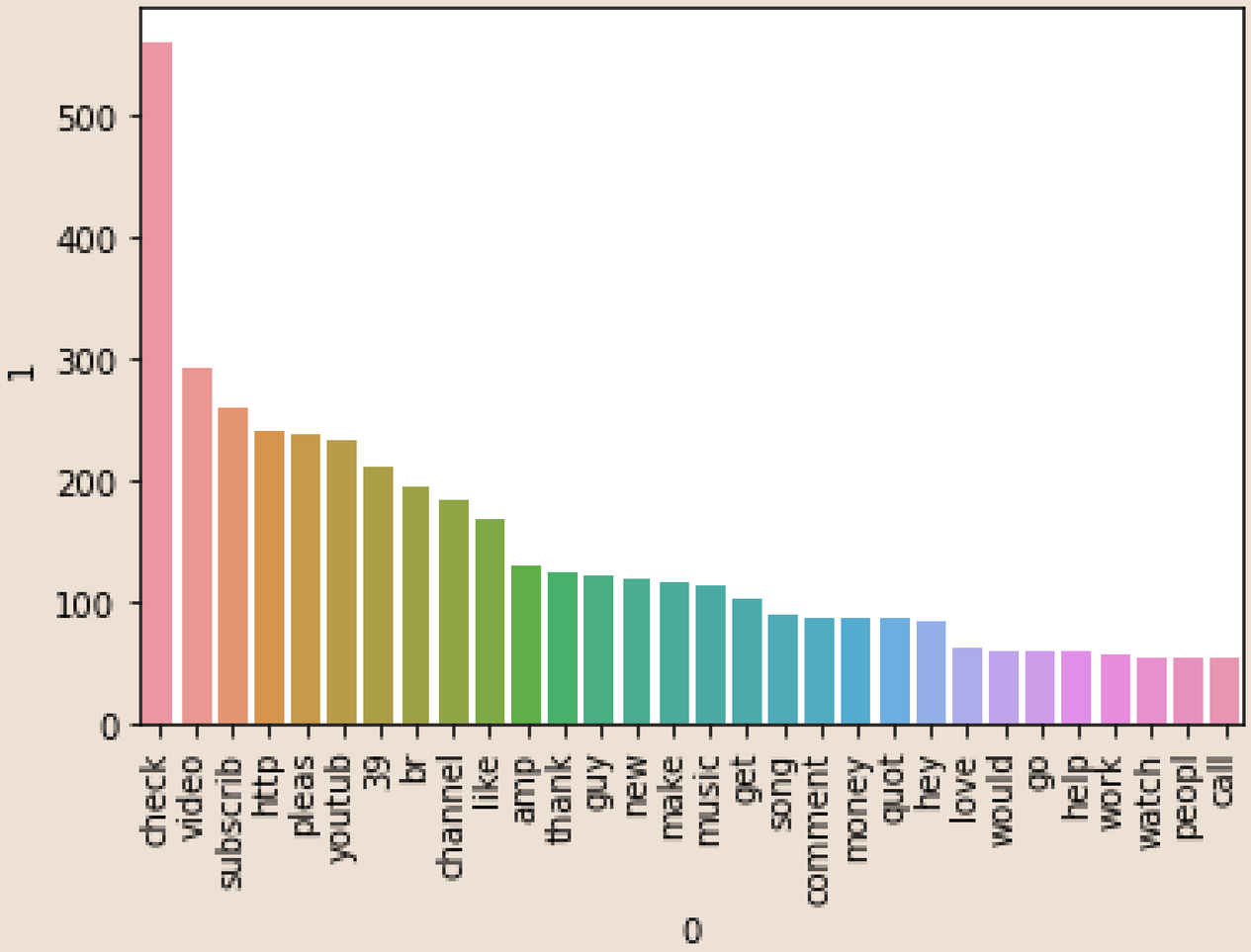


spam words

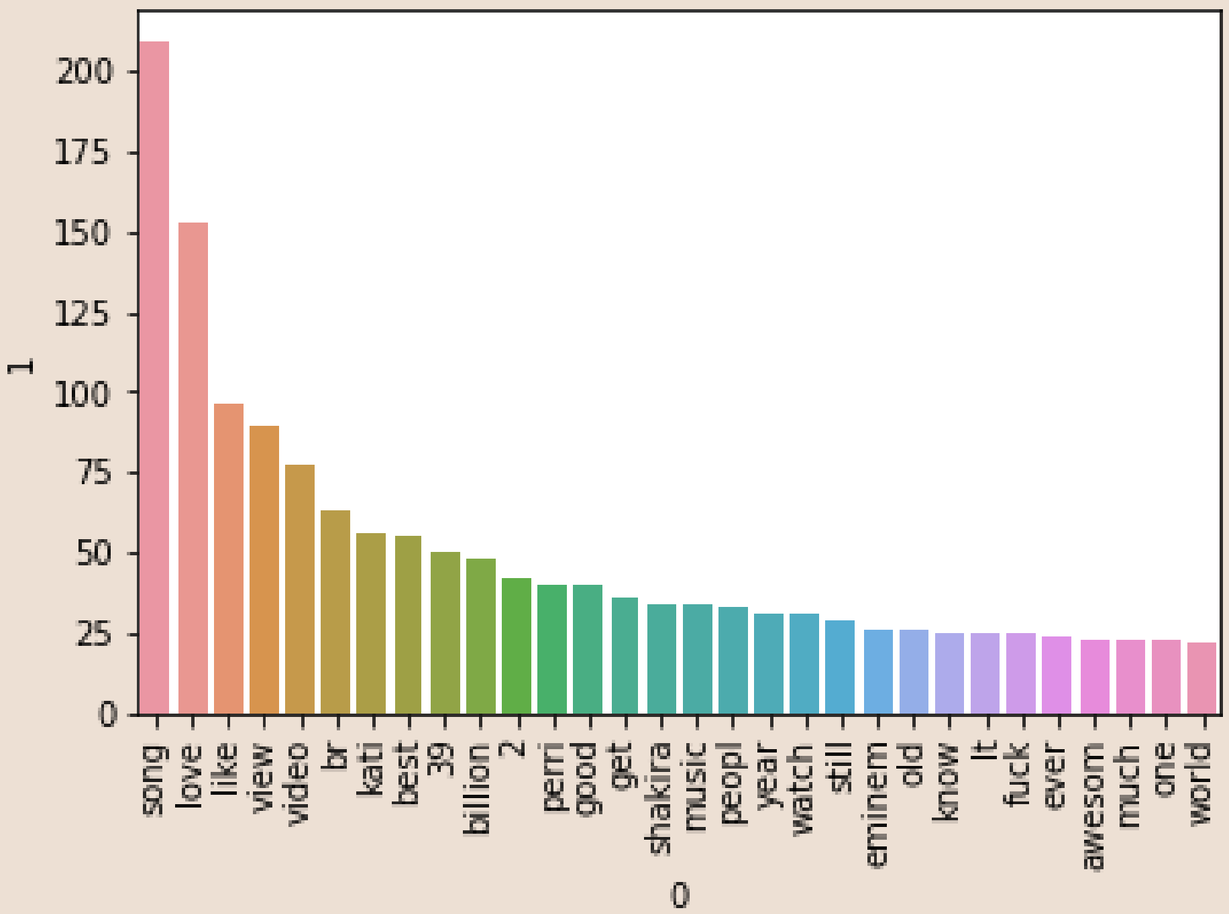
Not spam words



Spam_corpus



Not_Spam_corpus



MODEL BUILDING

NOT_SPAM comments have an average of 49.64 characters.
SPAM comments have an average of 137.34 characters.

Machine learning models used:

- KNeighborsClassifier
- DesicionTreeClassifier
- RandomForestClassifier
- SVC
- Naive bayes &MultinomiaNB

Machine learning Acurracy Models:

```
from sklearn.neighbors import KNeighborsClassifier,KNeighborsRegressor
KNNC = KNeighborsClassifier()
KNNC.fit(X_train,y_train)
print(f"Train Accuracy of model {KNNC.score(X_train,y_train)*100} %")
print(f"Test Accuracy of model {KNNC.score(X_test,y_test)*100} %")
```

```
Train Accuracy of model 89.46564885496183 %
Test Accuracy of model 85.44891640866872 %
```

```
from sklearn.tree import DecisionTreeClassifier, DecisionTreeRegressor
dtc = DecisionTreeClassifier()
dtc.fit(X_train,y_train)
print(f"Train Accuracy of model {dtc.score(X_train,y_train)*100} %")
print(f"Test Accuracy of model {dtc.score(X_test,y_test)*100} %")
```

```
Train Accuracy of model 100.0 %
Test Accuracy of model 93.96284829721363 %
```

```
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB()
clf.fit(X_train,y_train)
print(f"Train Accuracy of model {clf.score(X_train,y_train)*100} %")
print(f"Test Accuracy of model {clf.score(X_test,y_test)*100} %")
```

```
Train Accuracy of model 96.6412213740458 %
Test Accuracy of model 93.03405572755418 %
```

```
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
Rfc= RandomForestClassifier()
Rfc.fit(X_train,y_train)
print(f"Train Accuracy of model {Rfc.score(X_train,y_train)*100} %")
print(f"Test Accuracy of model {Rfc.score(X_test,y_test)*100} %")
```

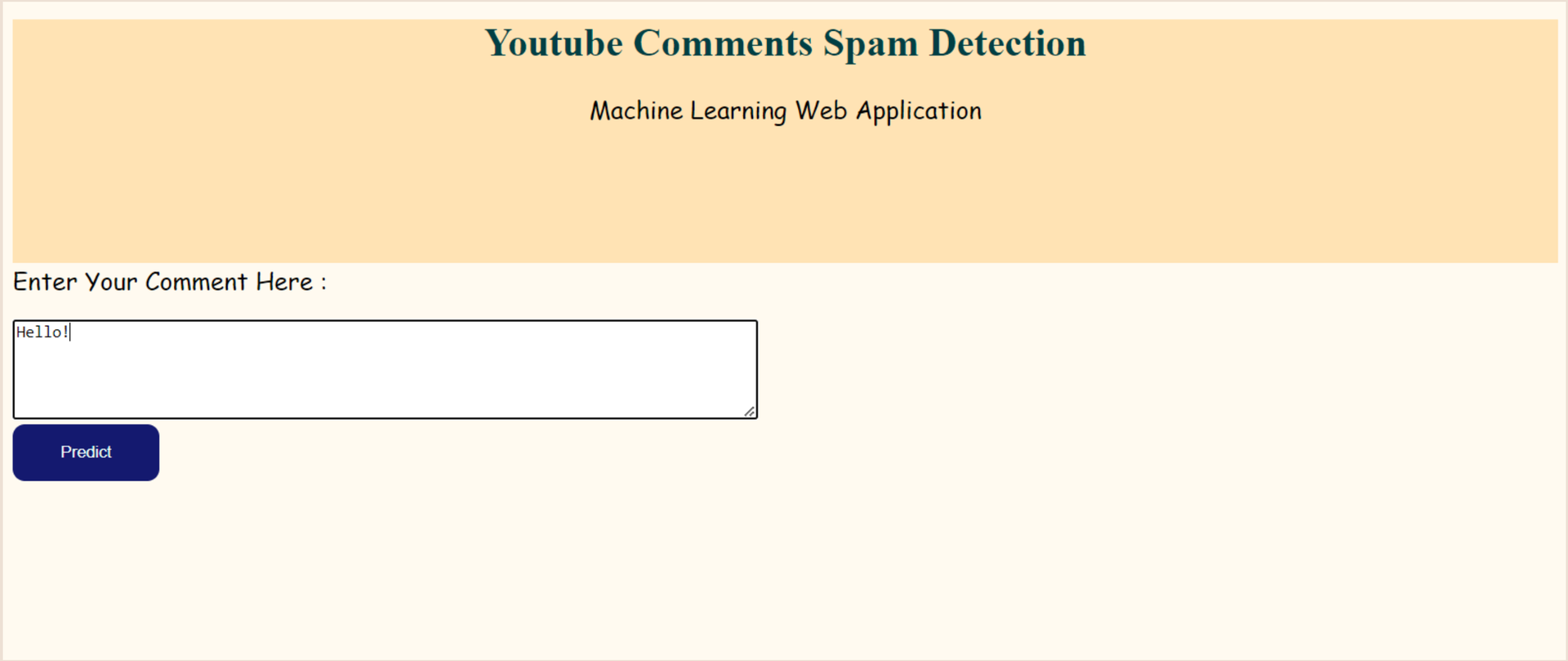
```
Train Accuracy of model 100.0 %
Test Accuracy of model 95.5108359133127 %
```

```
from sklearn.svm import SVC
from pandas.core.common import random_state
svc = SVC(random_state=101)
svc.fit(X_train,y_train)
print(f"Train Accuracy of model {svc.score(X_train,y_train)*100} %")
print(f"Test Accuracy of model {svc.score(X_test,y_test)*100} %")
```

```
Train Accuracy of model 97.63358778625954 %
Test Accuracy of model 93.49845201238391 %
```

MODEL DEPLOYMENT:

Flask is used to deploy the model into web.



output:


Youtube Comments Spam Detection
Machine Learning Web Application
Your Comment is Classified as :
Spam Comment

Thank you!