Project Name	Building a Smarter AI-Powered Spam Classifier
Team Id	Proj_212177_Team_2
Date	October 11 2023
Maximum mark	

Project: SMS spam classifier

Introduction:

Data preprocessing is a crucial step in building an effective SMS spam classifier. In this phase, raw SMS data is transformed and prepared for machine learning analysis. It involves tasks such as text cleaning, tokenization, stop word removal, and more. By refining and structuring the data, we can improve the accuracy of our spam detection model and ensure it operates effectively in distinguishing between legitimate and unwanted messages. This introductory process sets the foundation for creating a robust SMS spam classifier.

Data preprocessing:

Data set for our project 'Building a Smarter AI-Powered Spam Classifier' is as follows

(https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset)

Code:

In[1]:

```
import numpy

import pandas as pd

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import re

import nltk

from nltk.corpus import stopwords

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer,TfidfTransformer

from sklearn.model_selection import train_test_split

from sklearn.naive_bayes import MultinomialNB

from sklearn.metrics import accuracy_score,classification_report,
confusion_matrix
```

In[2]:

```
df = pd.read_csv("C:/Users/ELCOT/Downloads/sms_spam.csv", encoding='ISO-8859-1')

df
```

Out[2]:

	type	text
0	ham	Hope you are having a good week. Just checking in
1	ham	K..give back my thanks.
2	ham	Am also doing in cbe only. But have to pay.
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...
4	spam	okmail: Dear Dave this is your final notice to...
...	...	...
5554	ham	You are a great role model. You are giving so ...
5555	ham	Awesome, I remember the last time we got someb...
5556	spam	If you don't, your prize will go to another cu...
5557	spam	SMS. ac JSco: Energy is high, but u may not kn...
5558	ham	Shall call now dear having food

5559 rows × 2 columns

In[3]:

```
df.head()
```

Out[3]:

	type	text
0	ham	Hope you are having a good week. Just checking in
1	ham	K..give back my thanks.
2	ham	Am also doing in cbe only. But have to pay.
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...
4	spam	okmail: Dear Dave this is your final notice to...

In[4]:

df.shape

Out[4]:

(5559, 2)

In[5]:

df.info()

Out[5]:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 5559 entries, 0 to 5558

Data columns (total 2 columns):

#	Column	Non-Null Count	Dtype
0	type	5559 non-null	object
1	text	5559 non-null	object

dtypes: object(2)

memory usage: 87.0+ KB

In[6]:

df.isnull().sum()

Out[6]:

type 0

text 0

dtype: int64

In[7]:

df.rename(columns={'v1':'target', 'v2':'text'}, inplace=True)

In[8]:

df.head()

Out[8]:

	type	text
0	ham	Hope you are having a good week. Just checking in
1	ham	K..give back my thanks.
2	ham	Am also doing in cbe only. But have to pay.
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...
4	spam	okmail: Dear Dave this is your final notice to...

In[9]:

df.duplicated().sum()

Out[9]:

408

In[10]:

```
df = df.drop_duplicates(keep='first')

df.duplicated().sum()
```

Out[10]:

0

In[11]:

```
df.drop(['Unnamed: 2','Unnamed : 3','Unnamed:
4'],axis=1,inplace=True)
```

Out[11]:

	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will Ì_ b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

5572 rows × 2 columns

In[12]:

```
df['type'].value_counts()
```

Out[12]:

Type	
Ham	4812
Spam	747

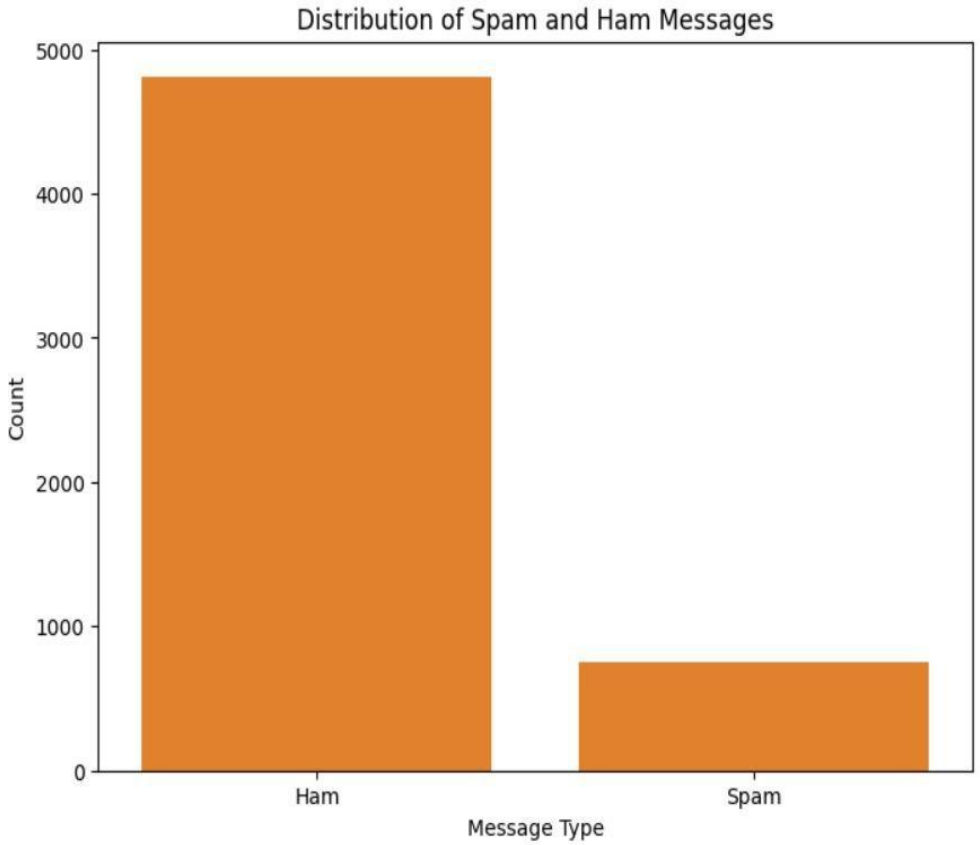
Name: count, dtype: int64

In[13]:

```
plt.figure(figsize=(8, 6))
sns.countplot(data=df, x='type')
df['type'].value_counts()
```

```
sns.countplot(data=df, x='type')
plt.xlabel('Message Type')
plt.ylabel('Count')
plt.title('Distribution of Spam and Ham Messages')
plt.xticks([0, 1], ['Ham', 'Spam'])
plt.show()
```

Out[13]:



In[14]:

```
import nltk

nltk.download('punkt')
```

Out[14]:

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\tawfe\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.

True
```

In[15]:df['num_characters'] = df['text'].apply(len)

df.head()

Out[15]:

	type	text	num_characters
0	ham	Hope you are having a good week. Just checking in	49
1	ham	K..give back my thanks.	23
2	ham	Am also doing in cbe only. But have to pay.	43
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...	150
4	spam	okmail: Dear Dave this is your final notice to...	161

In[16]:

```
# number of words
        df['num_words'] = df['text'].apply(lambda x:len(nltk.word_tokenize(x)))
        df.head()
```

Out[16]:

	type	text	num_characters	num_words
0	ham	Hope you are having a good week. Just checking in	49	11
1	ham	K..give back my thanks.	23	7
2	ham	Am also doing in cbe only. But have to pay.	43	12
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...	150	23
4	spam	okmail: Dear Dave this is your final notice to...	161	32

In[17]:

```
df['num_sentences'] = df['text'].apply(lambda x:len(nltk.sent_tokenize(x)))
    df[['num_characters', 'num_words', 'num_sentences']].describe()
```

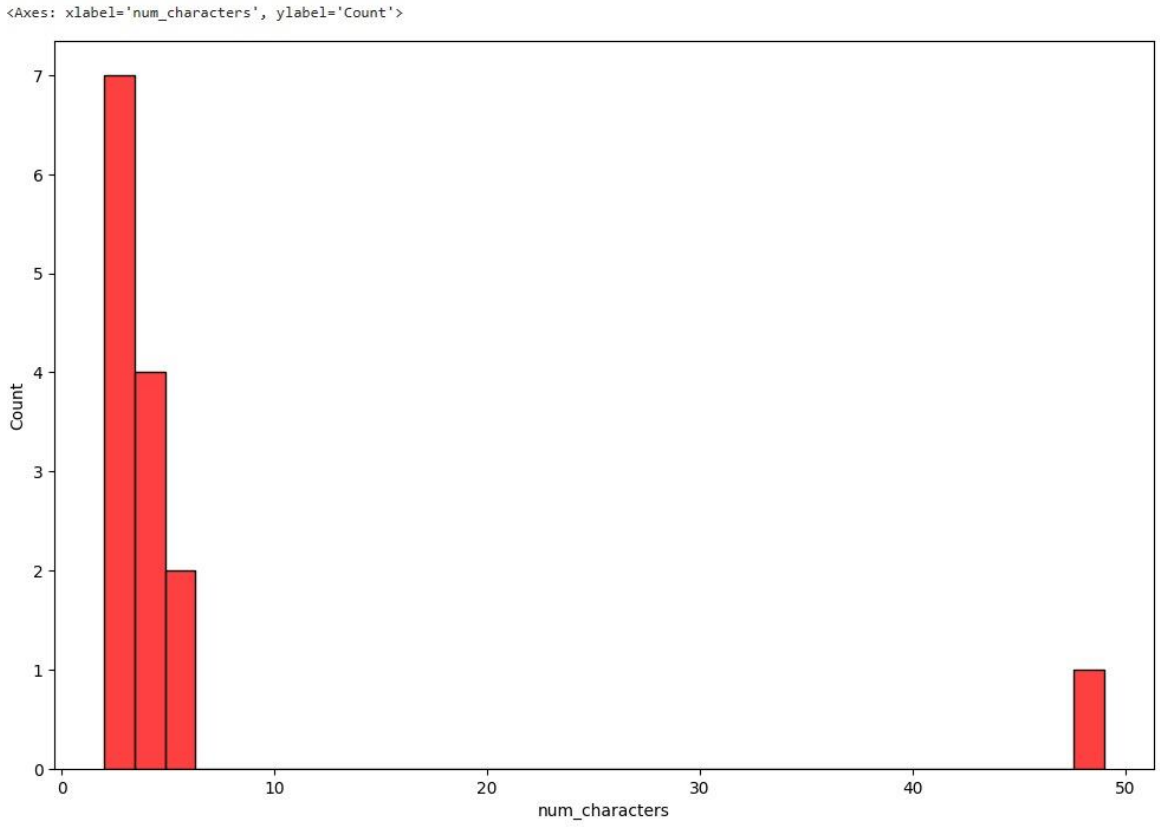
Out[17]:

	type	text	num_characters	num_words
0	ham	Hope you are having a good week. Just checking in	49	11
1	ham	K..give back my thanks.	23	7
2	ham	Am also doing in cbe only. But have to pay.	43	12
3	spam	complimentary 4 STAR Ibiza Holiday or Â£10,000...	150	23
4	spam	okmail: Dear Dave this is your final notice to...	161	32

In[18]:

```
import seaborn as sns
plt.figure(figsize=(12, 8))
sns.histplot(df[df['num_words']==0]['num_characters'])
sns.histplot(df[df['num_words']==1]['num_characters'], color='red')
```

Out[18]:



In[19]:

df.describe()

Out[19]:

	num_characters	num_words	num_sentences
count	5559.000000	5559.000000	5559.000000
mean	79.893326	18.382443	2.006296
std	59.200791	13.167199	1.540083
min	2.000000	1.000000	1.000000
25%	35.000000	9.000000	1.000000
50%	61.000000	15.000000	2.000000
75%	121.000000	27.000000	3.000000
max	910.000000	196.000000	38.000000

Explanation:

Import necessary libraries for preprocessing.Drop unnecessary columns from the dataframe.define the shape of the dataset.get the information of the dataframe.rename the columns Drop duplicated values.Create a bar plot to visualize the distribution of spam and ham messages.display the length of the each text rows.Count the number of words in each rows.Create a Bar plot to visualize the num words and num characters.finally Describe the DataFrame.

Conclusion:

Data preprocessing is a critical step in developing a spam classifier,ensuring that the model can effectively differentiate spam and legitimate messages.