SMS_detector

October 17, 2024

[1]:
```
import re
import nltk
import warnings
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from sklearn.pipeline import Pipeline
from nltk.stem import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import precision_score, recall_score,␣
 ↪classification_report, accuracy_score, f1_score

warnings.filterwarnings('ignore')
```

[2]:
```
df = pd.read_csv(r"C:\Users\91969\Downloads\spam.csv", encoding='latin1')
df.head()
```

[2]:
```
     v1                                                 v2 Unnamed: 2  \
0   ham  Go until jurong point, crazy.. Available only …         NaN
1   ham                      Ok lar… Joking wif u oni…         NaN
2  spam  Free entry in 2 a wkly comp to win FA Cup fina…         NaN
3   ham  U dun say so early hor… U c already then say…         NaN
4   ham  Nah I don't think he goes to usf, he lives aro…         NaN

  Unnamed: 3 Unnamed: 4
0        NaN        NaN
1        NaN        NaN
2        NaN        NaN
3        NaN        NaN
```

```
4        NaN        NaN
```

[3]:
```
df = df.drop(columns = ["Unnamed: 2","Unnamed: 3","Unnamed: 4"], axis=1)
df.rename(columns = {"v1":"Target", "v2":"Text"}, inplace = True)
df.head()
```

[3]:

	Target	Text
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…

[4]:
```
df.shape
```

[4]:
```
(5572, 2)
```

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Target  5572 non-null   object
 1   Text    5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

[6]:
```
df.isnull().sum()
```

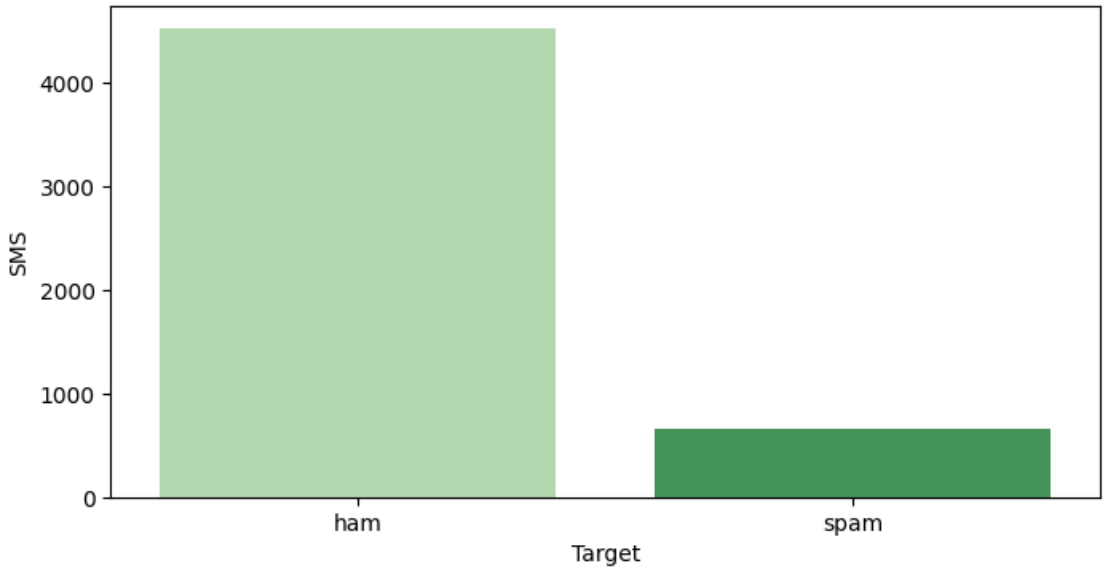
[6]:
```
Target    0
Text      0
dtype: int64
```

[7]:
```
df.duplicated().sum()
```

[7]:
```
np.int64(403)
```

[8]:
```
df.drop_duplicates(inplace=True)
```

[9]:
```
plt.figure(figsize=(8,4))
sns.countplot(data=df, x='Target', palette='Greens')
plt.xlabel("Target")
plt.ylabel("SMS")
plt.show()
```


4000
3000
2000
1000
0
SMS
ham
spam
Target

[ ]: