Text preprocessing using Spacy

Aim: To preprocess the dataset using Spacy

Description: spaCy is a free, open-source Python library that provides advanced capabilities to conduct natural language processing (NLP) on large volumes of text at high speed. It helps you build models and production applications that can underpin document analysis, chatbot capabilities, and all other forms of text analysis.

```
! pip install unidecode
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting unidecode
  Downloading Unidecode-1.3.6-py3-none-any.whl (235 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 235.9/235.9 kB 4.3 MB/s eta 0:00:00
Installing collected packages: unidecode
Successfully installed unidecode-1.3.6
```

```
!pip install spacy download en_core_web_sm
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: spacy in /usr/local/lib/python3.9/dist-packages (3.5.2)
Collecting download
  Downloading download-0.3.5-py3-none-any.whl (8.8 kB)
Requirement already satisfied: en_core_web_sm in /usr/local/lib/python3.9/dist-packages (3.5.0)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.9/dist-packages (from spacy) (2.0.7)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.9/dist-packages (from spacy) (2.4.6)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (1.0.4)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (1.22.4)
Requirement already satisfied: pathy>=0.10.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (0.10.1)
Requirement already satisfied: thinc<8.2.0,>=8.1.8 in /usr/local/lib/python3.9/dist-packages (from spacy) (8.1.9)
Requirement already satisfied: setuptools in /usr/local/lib/python3.9/dist-packages (from spacy) (67.6.1)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.9/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (3.3.0)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (2.27.1)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.9/dist-packages (from spacy) (3.0.8)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in /usr/local/lib/python3.9/dist-packages (from spacy) (1.10.7
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (23.1)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (0.7.0)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.9/dist-packages (from spacy) (2.0.8)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (4.65.0)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.9/dist-packages (from spacy) (1.1.1)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.9/dist-packages (from spacy) (6.3.0)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.9/dist-packages (from spacy) (3.1.2)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.9/dist-packages (from spacy) (1.0.9)
Requirement already satisfied: six in /usr/local/lib/python3.9/dist-packages (from download) (1.16.0)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/python3.9/dist-packages (from pydantic!=1.8,!=1.8.1,<1.11
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0,>=2.13.0->spacy
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0,>=2.13.0->s
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0,>=2.13.0->spacy) (3.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.9/dist-packages (from requests<3.0.0,>=2.13.0->spacy) (
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.9/dist-packages (from thinc<8.2.0,>=8.1.8->spacy) (0.7.
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.9/dist-packages (from thinc<8.2.0,>=8.1.8->spacy)
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.9/dist-packages (from typer<0.8.0,>=0.3.0->spacy) (8.1
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.9/dist-packages (from jinja2->spacy) (2.1.2)
Installing collected packages: download
Successfully installed download-0.3.5
```

```
import pandas as pd
import spacy
import matplotlib.pyplot as plt
from collections import Counter
import re
from sklearn.model_selection import train_test_split
from spacy.language import Language
from imblearn.over_sampling import SMOTE
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
data = pd.read_csv('/content/drive/MyDrive/NLP/Womens Clothing E-Commerce Reviews.csv')
```

```
data.info()
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               23486 non-null  int64 
 1   Clothing ID              23486 non-null  int64 
 2   Age                      23486 non-null  int64 
 3   Title                    19676 non-null  object
 4   Review Text              22641 non-null  object
 5   Rating                   23486 non-null  int64 
 6   Recommended IND          23486 non-null  int64 
 7   Positive Feedback Count  23486 non-null  int64 
 8   Division Name            23472 non-null  object
 9   Department Name          23472 non-null  object
 10  Class Name               23472 non-null  object
dtypes: int64(6), object(5)
memory usage: 2.0+ MB
```

```
data
```

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses
...	...	...	...	...	...	...	...	...	...	...	...
23481	23481	1104	34	Great dress for many occasions	I was very happy to snag this dress at such a ...	5	1	0	General Petite	Dresses	Dresses
23482	23482	862	48	Wish it was made of cotton	It reminds me of maternity clothes. soft, stre...	3	1	0	General Petite	Tops	Knits

```
data.isna().any()
```

```
Unnamed: 0                 False
Clothing ID                False
Age                        False
Title                       True
Review Text                 True
Rating                     False
Recommended IND            False
Positive Feedback Count    False
Division Name               True
Department Name             True
Class Name                  True
dtype: bool
```

```
import re
import string
import time
nlp = spacy.load("en_core_web_sm")
```

```
def spacy_preprocess(text):
  text=str(text)
  text = re.sub(r'http\S+', '', text)
```

```
    text = re.sub(r'@\w+', '', text)
    text = re.sub(r"\ [A-Za-z]*\.com", " ", text)
    text = re.sub(r"[^a-zA-Z0-9:$-,%.?!]+", ' ',text)
    text = re.sub(r"[|]", ' ',text)
    doc=nlp(text)
    tokens=[token.lemma_ for token in doc if token.is_punct == False and token.is_space == False and token.like_url == Fal
    text=" ".join(tokens)
    return text
```

```
data['Title'] = data['Title'].apply(spacy_preprocess)
```

```
data
```

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	767	33	nan	absolutely wonderful silky sexy comfortable	4	1	0	Initmates	Intimate	Intimates
1	1	1080	34	nan	love dress sooo pretty happen find store glad ...	5	1	4	General	Dresses	Dresses
2	2	1077	60	major design flaw	high hope dress want work initially order peti...	3	0	0	General	Dresses	Dresses
3	3	1049	50	favorite buy	love love love jumpsuit fun flirty fabulous ti...	5	1	0	General Petite	Bottoms	Pants
4	4	847	47	flattering shirt	shirt flattering adjustable tie perfect length...	5	1	6	General	Tops	Blouses
...	...	...	...	...	...	...	...	...	...	...	...
23481	23481	1104	34	great dress occasion	happy snag dress great price easy slip flatter...	5	1	0	General Petite	Dresses	Dresses
23482	23482	862	48	wish cotton	remind maternity clothe soft stretchy shiny ma...	3	1	0	General Petite	Tops	Knits
23483	23483	1104	31	cute	fit work glad able try store order online	3	0	1	General Petite	Dresses	Dresses

Colab paid products - Cancel contracts here

✓ 28s completed at 11:02 PM