NLP-517-CS

January 18, 2025

[15]:
```
import pandas as pd
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import nltk
```

[16]:
```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\aishwarya\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\aishwarya\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\aishwarya\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

[16]:
```
True
```

[17]:
```
file_path = "Tweets.csv"
df = pd.read_csv(file_path)
```

[18]:
```
print("Original Dataset:")
print(df.head())
```

```
Original Dataset:
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
```

```
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                                text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t…         NaN
2  @VirginAmerica I didn't today… Must mean I n…         NaN
3  @VirginAmerica it's really aggressive to blast…         NaN
4  @VirginAmerica and it's a really big bad thing…         NaN

              tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

[19]:
```
df = df[['text', 'airline_sentiment']]
```

[20]:
```
print("\nSentiment Distribution:")
print(df['airline_sentiment'].value_counts())
```

```
Sentiment Distribution:
negative    9178
neutral     3099
positive    2363
Name: airline_sentiment, dtype: int64
```

[21]:
```
stopwords_set = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()
```

[22]:
```
def preprocess_text(text):
    # Remove non-alphabetical characters
    text = re.sub(r'[^a-zA-Z]', ' ', text)
    # Convert text to lowercase
    text = text.lower()
    # Tokenize the text
    tokens = word_tokenize(text)
```

```
    # Remove stopwords and lemmatize tokens
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in␣
↪stopwords_set]
    return tokens
```

[23]:
```
df['Processed_Text'] = df['text'].apply(preprocess_text)
```

[24]:
```
print("\nProcessed Data:")
print(df[['text', 'Processed_Text', 'airline_sentiment']].head())
```

```
Processed Data:
                                                text  \
0                @VirginAmerica What @dhepburn said.
1  @VirginAmerica plus you've added commercials t…
2  @VirginAmerica I didn't today… Must mean I n…
3  @VirginAmerica it's really aggressive to blast…
4  @VirginAmerica and it's a really big bad thing…

                                      Processed_Text airline_sentiment
0                    [virginamerica, dhepburn, said]           neutral
1  [virginamerica, plus, added, commercial, exper…          positive
2  [virginamerica, today, must, mean, need, take,…           neutral
3  [virginamerica, really, aggressive, blast, obn…          negative
4           [virginamerica, really, big, bad, thing]           negative
```

[ ]: