Data Preprocessing

In [1]:
```
# Package imports
import pandas as pd
import numpy as np
from time import time
from bs4 import BeautifulSoup
import spacy
!pip install unidecode
!pip install word2number
import unidecode
from word2number import w2n
!pip install contractions
import contractions

# Loading the data into pandas dataframe
tweetsdf = pd.read_csv(r"C:\Users\srini\OneDrive\Desktop\Advanced Machine Lea
tweetsdf.info()
```

```
Requirement already satisfied: unidecode in c:\users\srini\anaconda3\lib\si
te-packages (1.3.4)
Requirement already satisfied: word2number in c:\users\srini\anaconda3\lib
\site-packages (1.1)
Requirement already satisfied: contractions in c:\users\srini\anaconda3\lib
\site-packages (0.1.68)
Requirement already satisfied: textsearch>=0.0.21 in c:\users\srini\anacond
a3\lib\site-packages (from contractions) (0.0.21)
Requirement already satisfied: anyascii in c:\users\srini\anaconda3\lib\sit
e-packages (from textsearch>=0.0.21->contractions) (0.3.1)
Requirement already satisfied: pyahocorasick in c:\users\srini\anaconda3\li
b\site-packages (from textsearch>=0.0.21->contractions) (1.4.4)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   tweet_id                      14640 non-null  float64
 1   airline_sentiment             14640 non-null  object
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object
 6   airline_sentiment_gold        40 non-null     object
 7   name                          14640 non-null  object
 8   negativereason_gold           32 non-null     object
 9   retweet_count                 14640 non-null  int64
 10  text                          14640 non-null  object
 11  tweet_coord                   1019 non-null   object
 12  tweet_created                 14640 non-null  object
 13  tweet_location                9907 non-null   object
 14  user_timezone                 9820 non-null   object
dtypes: float64(3), int64(1), object(11)
memory usage: 1.7+ MB
```

In [2]:
```
tweetsdf.head()
```

Out[2]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereaso
0	5.700000e+17	neutral	1.0000	NaN	
1	5.700000e+17	positive	0.3486	NaN	
2	5.700000e+17	neutral	0.6837	NaN	
3	5.700000e+17	negative	1.0000	Bad Flight	
4	5.700000e+17	negative	1.0000	Can't Tell	

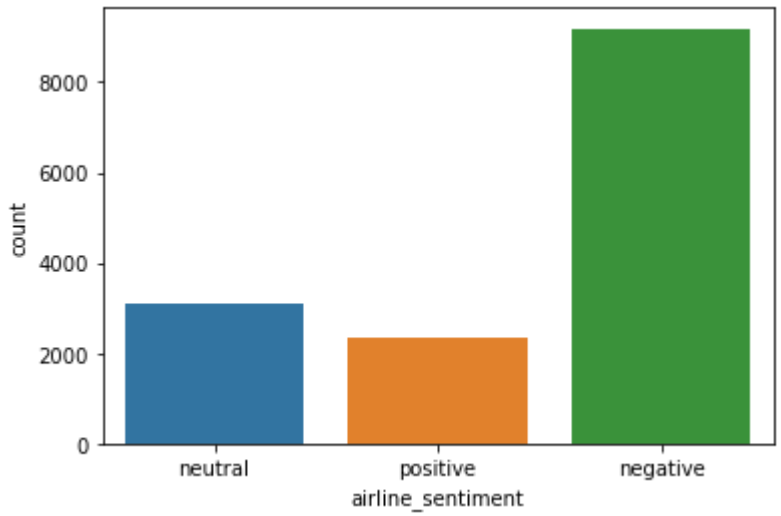
In [3]:
```
# Drop columns
tweetsdf.drop(columns = ['tweet_id', 'airline_sentiment_gold', 'negativereasc
```

In [4]:
```
# NUmber of each sentiment reviews
print('Number of negative tweets:', tweetsdf[tweetsdf['airline_sentiment'] ==
print('Number of positive tweets:', tweetsdf[tweetsdf['airline_sentiment'] ==
print('Number of neutral tweets:', tweetsdf[tweetsdf['airline_sentiment'] ==
```

```
Number of negative tweets: 9178
Number of positive tweets: 2363
Number of neutral tweets: 3099
```

In [5]:
```
import seaborn as sns
sns.countplot(x = "airline_sentiment", data = tweetsdf)
```

Out[5]: <AxesSubplot:xlabel='airline_sentiment', ylabel='count'>



In [6]:
```
# Replacing 'neutral' & 'positive' with 'non-negative' respectively
tweetsdf['airline_sentiment'].replace('positive', 'non-negative', inplace=Tru
tweetsdf['airline_sentiment'].replace('neutral', 'non-negative', inplace=True
tweetsdf.head()
```

Out[6]:

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

In [7]:
```
# Finding the duplicate values
tweetsdf.duplicated().sum()
```

Out[7]: 205

In [8]:
```
# Dropping duplicates
tweetsdf = tweetsdf.drop_duplicates(keep='first')
tweetsdf.duplicated().sum()
```

Out[8]:
```
0
```

In [9]:
```
# Checking for any null values
tweetsdf.isnull().all()
```

Out[9]:
```
airline_sentiment    False
text                 False
dtype: bool
```

In [10]:
```
tweetsdf.head()
```

Out[10]:

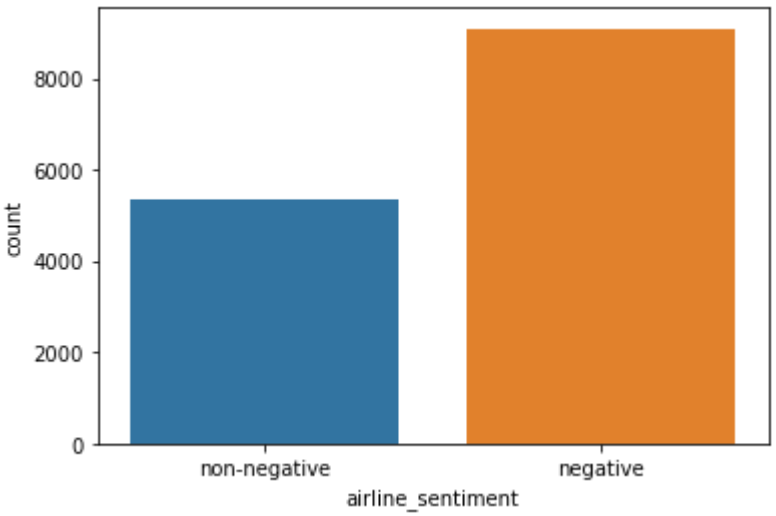
	airline_sentiment	text
0	non-negative	@VirginAmerica What @dhepburn said.
1	non-negative	@VirginAmerica plus you've added commercials t...
2	non-negative	@VirginAmerica I didn't today... Must mean I n...
3	negative	@VirginAmerica it's really aggressive to blast...
4	negative	@VirginAmerica and it's a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

In [11]:
```
tweetsdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14435 entries, 0 to 14639
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   airline_sentiment  14435 non-null  object
 1   text               14435 non-null  object
dtypes: object(2)
memory usage: 338.3+ KB
```

In [12]:
```
import seaborn as sns
sns.countplot(x = "airline_sentiment", data = tweetsdf)
```

Out[12]: <AxesSubplot:xlabel='airline_sentiment', ylabel='count'>



In [13]:
```
tweetsdf.text[73]
```

Out[13]: '@VirginAmerica your airline is awesome but your lax loft needs to step up its game. $40 for dirty tables and floors? http://t.co/hy0VrfhjHt' (http://t.co/hy0VrfhjHt')

In [14]:
```
tweetsdf.text[233]
```

Out[14]: "@VirginAmerica, the only airline based in Silicon Valley! #disruption #FCmostinnovative #incubator @FastCompany's http://t.co/wU3LbCNcr9" (http://t.co/wU3LbCNcr9")

In [15]:
```
tweetsdf.text[24]
```

Out[15]: '@VirginAmerica you guys messed up my seating.. I reserved seating with my friends and you guys gave my seat away ... 😡 I want free internet'

In [16]:
```
# Remove any URL's from the data
import re
def remove_URL(sample):
    """Remove URLs from a sample string"""
    return re.sub(r"http\S+", "", sample)

tweetsdf['no_URL'] = tweetsdf['text'].apply(lambda x: [remove_URL(word) for w
tweetsdf.head()
```

Out[16]:

	airline_sentiment	text	no_URL
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...

In [17]:
```
tweetsdf.no_URL[73]
```

Out[17]:
```
['@VirginAmerica',
 'your',
 'airline',
 'is',
 'awesome',
 'but',
 'your',
 'lax',
 'loft',
 'needs',
 'to',
 'step',
 'up',
 'its',
 'game.',
 '$40',
 'for',
 'dirty',
 'tables',
 'and',
 'floors?',
 '']
```

In [18]:
```
# Remove any html tags from the data
def strip_html_tags(text):
    soup = BeautifulSoup(text, "html.parser")
    stripped_text = soup.get_text(separator=" ")
    return stripped_text

tweetsdf['no_html'] = tweetsdf['no_URL'].apply(lambda x: [strip_html_tags(wor
tweetsdf.head()
```

```
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "..." looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "." looks like a filename, not markup. You should proba
bly open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "/" looks like a filename, not markup. You should proba
bly open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: ".." looks like a filename, not markup. You should prob
ably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "...." looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "con" looks like a filename, not markup. You should pro
bably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "images" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "data" looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "Data" looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "....." looks like a filename, not markup. You should p
robably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "test" looks like a filename, not markup. You should pr
obably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "TEST" looks like a filename, not markup. You should pr
```

```
obably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "............." looks like a filename, not markup. You
should probably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "data." looks like a filename, not markup. You should p
robably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "......." looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
C:\Users\srini\anaconda3\lib\site-packages\bs4\__init__.py:332: MarkupResem
blesLocatorWarning: "......" looks like a filename, not markup. You should
probably open this file and pass the filehandle into Beautiful Soup.
  warnings.warn(
```

Out[18]:

	airline_sentiment	text	no_URL	no_html
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...

In [19]:
```
# Remove accented characters
def remove_accented_chars(text):
    """remove accented characters from text, e.g. café"""
    text = unidecode.unidecode(text)
    return text

tweetsdf['no_accentchar'] = tweetsdf['no_html'].apply(lambda x: [unidecode.un
tweetsdf.head()
```

Out[19]:

	airline_sentiment	text	no_URL	no_html	no_accentchar
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...

In [21]:
```
tweetsdf.no_accentchar[24]
```

Out[21]:
```
['@VirginAmerica',
 'you',
 'guys',
 'messed',
 'up',
 'my',
 'seating..',
 'I',
 'reserved',
 'seating',
 'with',
 'my',
 'friends',
 'and',
 'you',
 'guys',
 'gave',
 'my',
 'seat',
 'away',
 '...',
 '',
 'I',
 'want',
 'free',
 'internet']
```

In [22]:
```
tweetsdf.no_accentchar[18]
```

Out[22]:
```
['I', '', 'flying', '@VirginAmerica.', '']
```

In [23]:
```
# Expanding contractions 'you've to you have'
!pip install contractions
import contractions
def expand_contractions(text):
    text = contractions.fix(text)
    return text
```

```
Requirement already satisfied: contractions in c:\users\srini\anaconda3\lib
\site-packages (0.1.68)
Requirement already satisfied: textsearch>=0.0.21 in c:\users\srini\anacond
a3\lib\site-packages (from contractions) (0.0.21)
Requirement already satisfied: pyahocorasick in c:\users\srini\anaconda3\li
b\site-packages (from textsearch>=0.0.21->contractions) (1.4.4)
Requirement already satisfied: anyascii in c:\users\srini\anaconda3\lib\sit
e-packages (from textsearch>=0.0.21->contractions) (0.3.1)
```

In [25]:
```
tweetsdf['no_contract'] = tweetsdf['no_accentchar'].apply(lambda x: [contract
tweetsdf.head()
```

Out[25]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	nc
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@Virg @
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@Virg plus, added,
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@Virg tod
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@Virg aggre
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@Virg a rea

In [26]:
```
tweetsdf['no_contract_str'] = [' '.join(map(str, l)) for l in tweetsdf['no_co
tweetsdf.head()
```

Out[26]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	nc
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@Virg @
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@Virg plus, added,
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@Virg tod
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@Virg aggre
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@Virg a rea

In [27]:
```
# Remove punctuations
import string
def remove_punct(text):
    text_nonpunct = "".join([char for char in text if char not in string.punc
    return text_nonpunct

tweetsdf['no_punc_text'] = tweetsdf['no_contract_str'].apply(lambda x: remove
tweetsdf.head()
```

Out[27]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@
3	negative	@VirginAmerica it's really aggressive to	[@VirginAmerica, it's, really, aggressive, to	[@VirginAmerica, it's, really, aggressive, to	[@VirginAmerica, it's, really, aggressive, to	[@

In [29]:
```
tweetsdf.no_punc_text[18]
```

Out[29]:
```
'I  flying VirginAmerica '
```

In [30]:
```
# Tokenization of the data
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords, wordnet
from nltk.stem import WordNetLemmatizer
tweetsdf['tokenized'] = tweetsdf['no_punc_text'].apply(word_tokenize)
tweetsdf.head()
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\srini\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[30]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@

In [31]:
```
tweetsdf.tokenized[18]
```

Out[31]:
```
['I', 'flying', 'VirginAmerica']
```

In [32]:
```
# convert text to lower case
tweetsdf['lower'] = tweetsdf['tokenized'].apply(lambda x: [word.lower() for w
tweetsdf.head()
```

Out[32]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@
3	negative	@VirginAmerica it's really aggressive to	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@

In [33]:
```
tweetsdf.lower[18]
```

Out[33]:
```
['i', 'flying', 'virginamerica']
```

In [36]:
```
# remove stop words
nltk_stopwords = nltk.corpus.stopwords.words('english')
nltk_stopwords.remove('but')
nltk_stopwords.remove('no')
nltk_stopwords.remove('not')

stop_words = set(stopwords.words('english'))
tweetsdf['no_stopwords'] = tweetsdf['lower'].apply(lambda x: [word for word i
tweetsdf.head()
```

Out[36]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@
3	negative	@VirginAmerica it's really aggressive to	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@

In [38]:
```
tweetsdf.no_stopwords[18]
```

Out[38]:
```
['flying', 'virginamerica']
```

In [39]:
```
tweetsdf.no_stopwords[328]
```

Out[39]:
```
['virginamerica',
 'shrinerack',
 'seattle',
 'bound',
 'wifey',
 'got',
 'duffle',
 'vday',
 'keeper',
 'holla']
```

In [40]:
```
# parts of speech of each word for lemmatization purpose
nltk.download('averaged_perceptron_tagger')
tweetsdf['pos_tags'] = tweetsdf['no_stopwords'].apply(nltk.tag.pos_tag)
tweetsdf.head()
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\srini\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[40]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today	[@VirginAmerica, I, didn't, today	[@VirginAmerica, I, didn't, today	[@VirginAmerica, I, didn't, today	[@

In [41]:
```
tweetsdf.pos_tags[18]
```

Out[41]:
```
[('flying', 'VBG'), ('virginamerica', 'NN')]
```

In [42]:
```
nltk.download('wordnet')
def get_wordnet_pos(tag):
    if tag.startswith('J'):
        return wordnet.ADJ
    elif tag.startswith('V'):
        return wordnet.VERB
    elif tag.startswith('N'):
        return wordnet.NOUN
    elif tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN
tweetsdf['wordnet_pos'] = tweetsdf['pos_tags'].apply(lambda x: [(word, get_wo
tweetsdf.head()
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\srini\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[42]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	nc
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@Virg @
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@Virg plus, added,
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@Virg tod
3	negative	@VirginAmerica it's really aggressive to blast...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@VirginAmerica, it's, really, aggressive, to,...	[@Virg aggre
4	negative	@VirginAmerica and it's a really big bad thing...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@VirginAmerica, and, it's, a, really, big, ba...	[@Virg a rea

In [43]:
```
tweetsdf.wordnet_pos[18]
```

Out[43]:
```
[('flying', 'v'), ('virginamerica', 'n')]
```

In [44]:
```
# Lemmatization of data
wnl = WordNetLemmatizer()
tweetsdf['lemmatized'] = tweetsdf['wordnet_pos'].apply(lambda x: [wnl.lemmati
tweetsdf.head()
```

Out[44]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@
3	negative	@VirginAmerica it's really aggressive to	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@

In [45]:
```
tweetsdf.lemmatized[18]
```

Out[45]:
```
['fly', 'virginamerica']
```

In [46]:
```
tweetsdf['lemmatized_str'] = [' '.join(map(str, l)) for l in tweetsdf['lemmat
tweetsdf.head()
```

Out[46]:

	airline_sentiment	text	no_URL	no_html	no_accentchar	
0	non-negative	@VirginAmerica What @dhepburn said.	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@VirginAmerica, What, @dhepburn, said.]	[@
1	non-negative	@VirginAmerica plus you've added commercials t...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@VirginAmerica, plus, you've, added, commerci...	[@ ac
2	non-negative	@VirginAmerica I didn't today... Must mean I n...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@VirginAmerica, I, didn't, today..., Must, me...	[@
3	negative	@VirginAmerica it's really aggressive to	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@VirginAmerica, it's, really,	[@

In [47]:
```
tweetsdf.lemmatized_str[18]
```

Out[47]:
```
'fly virginamerica'
```

In [49]:
```
# Drop columns
tweetsdf.drop(columns = ['text', 'no_URL', 'no_html', 'no_accentchar', 'no_cc
```

In [50]:
```
tweetsdf.to_csv('Finaldf1.csv')
```

In [32]:
```
tweetsdf.head()
```

Out[32]:

	airline_sentiment	lemmatized_text_str
0	non-negative	virginamerica dhepburn say
1	non-negative	virginamerica plus added commercial experience...
2	non-negative	virginamerica today must mean need take anothe...
3	negative	virginamerica really aggressive blast obnoxiou...
4	negative	virginamerica really big bad thing