Disney_EDA_Report

June 15, 2023

1 Disney Review Topic Modeling EDA Report

Author: Amelia Tang

1.1 Read Python Packages

[1]:
```
import numpy as np
import pandas as pd
import altair as alt
import string
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.corpus import wordnet
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from wordcloud import WordCloud, STOPWORDS
import tkinter
import matplotlib
import matplotlib.pyplot as plt
from PIL import Image
from textblob import TextBlob
from nltk.sentiment.vader import SentimentIntensityAnalyzer

alt.renderers.enable("html")
alt.data_transformers.disable_max_rows()
matplotlib.use('TkAgg')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
nltk.download('punkt')
nltk.download('brown')
nltk.download('vader_lexicon')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/amelia/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /Users/amelia/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
```

```
[nltk_data] Downloading package omw-1.4 to /Users/amelia/nltk_data…
[nltk_data]   Package omw-1.4 is already up-to-date!
[nltk_data] Downloading package punkt to /Users/amelia/nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package brown to /Users/amelia/nltk_data…
[nltk_data]   Package brown is already up-to-date!
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/amelia/nltk_data…
[nltk_data]   Package vader_lexicon is already up-to-date!
```

[1]: True

1.2 Read the data

1.2.1 There are 42,656 lines of non-null data. Among the three branches of Disney, the California branch has the most review data.

[2]:
```
disney_df = pd.read_csv("../data/raw/DisneylandReviews.csv")
disney_df
```

[2]:

	Review_ID	Rating	Year_Month	Reviewer_Location	\
0	670772142	4	2019-4	Australia	
1	670682799	4	2019-5	Philippines	
2	670623270	4	2019-4	United Arab Emirates	
3	670607911	4	2019-4	Australia	
4	670607296	4	2019-4	United Kingdom	
…	…	…	…	…	
42651	1765031	5	missing	United Kingdom	
42652	1659553	5	missing	Canada	
42653	1645894	5	missing	South Africa	
42654	1618637	4	missing	United States	
42655	1536786	4	missing	United Kingdom	

	Review_Text	Branch
0	If you've ever been to Disneyland anywhere you…	Disneyland_HongKong
1	Its been a while since d last time we visit HK…	Disneyland_HongKong
2	Thanks God it wasn t too hot or too humid wh…	Disneyland_HongKong
3	HK Disneyland is a great compact park. Unfortu…	Disneyland_HongKong
4	the location is not in the city, took around 1…	Disneyland_HongKong
…	…	…
42651	i went to disneyland paris in july 03 and thou…	Disneyland_Paris
42652	2 adults and 1 child of 11 visited Disneyland …	Disneyland_Paris
42653	My eleven year old daughter and myself went to…	Disneyland_Paris
42654	This hotel, part of the Disneyland Paris compl…	Disneyland_Paris
42655	I went to the Disneyparis resort, in 1996, wit…	Disneyland_Paris

[42656 rows x 6 columns]

[3]:
```
disney_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42656 entries, 0 to 42655
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Review_ID          42656 non-null  int64 
 1   Rating             42656 non-null  int64 
 2   Year_Month         42656 non-null  object
 3   Reviewer_Location  42656 non-null  object
 4   Review_Text        42656 non-null  object
 5   Branch             42656 non-null  object
dtypes: int64(2), object(4)
memory usage: 2.0+ MB
```

We collected the most data on the reviews for the California branch.

[4]:
```
alt.Chart(disney_df).mark_bar().encode(
    x='count()',
    y=alt.Y('Branch', sort='-x'))
```

[4]:
```
alt.Chart(…)
```

1.2.2 Here are some sample review texts:

[5]:
```
disney_df['Review_Text'][89]
```

[5]:
```
'Hong Kong Disneyland is definitely a must see. It is a smaller park but what it
loses in size it makes up for in quality. Hong Kong is currently home to the
only Marvel themed rides and they do not disappoint. Hong Kong has many of the
standard rides shows you expect from Disney but it also maintains a variety of
exclusives. Hong Kong also has a lot of exclusive merchandise especially pins.
Also most if not all of the rides are Bilingual. '
```

[6]:
```
disney_df['Review_Text'].head(10)
```

[6]:
```
0    If you've ever been to Disneyland anywhere you…
1    Its been a while since d last time we visit HK…
2    Thanks God it wasn   t too hot or too humid wh…
3    HK Disneyland is a great compact park. Unfortu…
4    the location is not in the city, took around 1…
5    Have been to Disney World, Disneyland Anaheim …
6    Great place! Your day will go by and you won't…
7    Think of it as an intro to Disney magic for th…
8    Feel so let down with this place,the Disneylan…
9    I can go on talking about Disneyland. Whatever…
Name: Review_Text, dtype: object
```

1.3 Preprocess corpus

1.3.1 To clean the corpus (Disney review texts) for analysis, I performed the following tasks:

- Changed all the letters to lowercase.
- Ensured consistent spelling of "hongkong".
- Replaced the abbreviation "n't" with "not".
- Removed all digits from the corpus.

[7]:
```
corpus=''
for i in disney_df['Review_Text']:
    corpus+=str(i)
```

[8]:
```
def preprocess(corpus):
    corpus = corpus.lower()
    corpus = corpus.replace('hk', 'hongkong')
    corpus = corpus.replace('hong kong', 'hongkong')
    corpus = corpus.replace("n't", " not")
    corpus = ''.join([i for i in corpus if not i.isdigit()])
    new_corpus =''
    for i in corpus:
        if i not in string.punctuation:
            new_corpus+=i
    return new_corpus
```

1.3.2 Lemmatization and Stemming

- Stemming is a crude method for cataloging related words, such as "boat," "boats," and "boating." One of the most popular stemming tools is Porter's Algorithm, developed by Martin Porter in 1980 in his paper titled "An algorithm for suffix stripping." The algorithm employs five phases of word reduction (Porter, 1980).

- Lemmatization looks beyond word reduction and considers the vocabulary, part of speech, and its use in a sentence.

[9]:
```
def lemmatize_stem(corpus):
    lemmatizer = WordNetLemmatizer()
    stemmer = PorterStemmer()
    new_corpus = lemmatizer.lemmatize(corpus)
    new_corpus = stemmer.stem(new_corpus)
    return new_corpus
```

1.3.3 Remove stopwords

Stopwords refer to commonly used words in a language that are often removed from text during natural language processing (NLP) tasks. These words are considered to have little or no significant meaning and are typically filtered out to focus on the more important and relevant words in a document or corpus. Examples of stopwords in English include "the," "is," "and," "in," "a," and

"an." The removal of stopwords helps reduce noise and improve the efficiency and effectiveness of NLP tasks such as text classification, information retrieval, and sentiment analysis. I also added customized stopwords related to Disney because they appear so frequently that don't have significant meanings.

[10]:
```
print(stopwords.words("english"))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what',
'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or',
'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above',
'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',
'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now',
'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn',
"hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't",
'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn',
"wouldn't"]
```

[11]:
```
stopwords_700 = pd.read_csv('stopwords.txt', header= None, delimiter = "\t")
stopwords_700 = set(stopwords_700[0])
```

[12]:
```
def remove_stopwords(corpus):
    words = nltk.word_tokenize(corpus)
    stop_words = set(stopwords.words("english"))
    stop_words_customize = {'disney', 'disneyland', 'land', 'park', 'parks',␣
 ↪'world', 'disneyworld', 'disney world'}
    stop_words_all = stop_words.union(stopwords_700, stop_words_customize)
    new_corpus = [word for word in words if not word in stop_words_all]
    return new_corpus
```

[13]:
```
new_corpus = preprocess(corpus)
new_corpus = lemmatize_stem(new_corpus)
new_corpus = remove_stopwords(new_corpus)
```

1.4 Most frequent words in the corpus

[14]:
```
from collections import Counter
new_disney_df = pd.DataFrame()
new_disney_df['word'] = list(Counter(new_corpus).keys())
new_disney_df['count']= list(Counter(new_corpus).values())
new_disney_df.sort_values(by=['count'], ascending=False).head(20)
```

[14]:

	word	count
10	rides	33898
24	time	28441
15	day	27421
420	ride	17498
38	great	15659
125	food	13985
161	kids	13824
83	place	13108
55	like	11868
201	see	11566
25	visit	11257
181	people	10878
214	went	10792
945	long	10086
184	fast	10001
409	pass	9690
402	wait	9565
365	back	9432
272	fun	9265
462	first	8777

1.5 WordCloud

1.5.1 Unigram WordCloud

I first generated a unigram WordCloud to visualize frequently appeared words.

[15]:
```
micky_mask = np.array(Image.open('mickey.png'))
```

[16]:
```
text = ''
for i in new_corpus:
    text+=' '
    text+=i
```

[17]:
```
wordcloud = WordCloud(stopwords = STOPWORDS,
                      mask = micky_mask,
                      random_state = 123,
                      collocations=False,
                      max_words = 100).generate(text)
```

[18]:
```
wordcloud.to_file('wordcloud.png')
```

[18]:
```
<wordcloud.wordcloud.WordCloud at 0x7fdd3900f490>
```

1.5.2 Bigram WordCloud

A bigram WordCloud is usually more informative than a unigram WordCloud because it provides same contexts to the words.

[19]:
```
bi_gram_cloud = WordCloud(stopwords = STOPWORDS,
                          background_color='white',
                          random_state = 123,
                          collocation_threshold = 3,
                          max_words = 100
                          ).generate(text)
```

[20]:
```
bi_gram_cloud.to_file('bi_gram_wordcloud.png')
```

[20]:
```
<wordcloud.wordcloud.WordCloud at 0x7fdd3900f850>
```

1.6 Sentiment Polarity Analysis

1.6.1 Using TextBlob

Polarity is defined between -1 and 1 with -1 defines a negative sentiment and 1 defines a positive sentiment.

[21]:
```
def clean(Review):
    Review = Review.str.lower()
    Review = Review.replace('hk', 'hongkong', regex = True)
    Review = Review.replace('hong kong', 'hongkong', regex = True)
    Review = Review.replace("n't", " not", regex = True)
    return Review
```

[22]:
```
disney_df['Review_Text'] = clean(disney_df['Review_Text'])
```

[23]:
```
disney_df['polarity'] = disney_df['Review_Text'].map(lambda text:␣
 ↪TextBlob(text).sentiment.polarity)
```

[24]:
```
disney_df.head(10)
```

[24]:

	Review_ID	Rating	Year_Month	Reviewer_Location	\
0	670772142	4	2019-4	Australia	
1	670682799	4	2019-5	Philippines	
2	670623270	4	2019-4	United Arab Emirates	
3	670607911	4	2019-4	Australia	
4	670607296	4	2019-4	United Kingdom	
5	670591897	3	2019-4	Singapore	

```
6  670585330       5     2019-4                India
7  670574142       3     2019-3             Malaysia
8  670571027       2     2019-4            Australia
9  670570869       5     2019-3                India

                                          Review_Text               Branch  \
0  if you've ever been to disneyland anywhere you…  Disneyland_HongKong
1  its been a while since d last time we visit ho…  Disneyland_HongKong
2  thanks god it wasn   t too hot or too humid wh…  Disneyland_HongKong
3  hongkong disneyland is a great compact park. u…  Disneyland_HongKong
4  the location is not in the city, took around 1…  Disneyland_HongKong
5  have been to disney world, disneyland anaheim …  Disneyland_HongKong
6  great place! your day will go by and you wo no…  Disneyland_HongKong
7  think of it as an intro to disney magic for th…  Disneyland_HongKong
8  feel so let down with this place,the disneylan…  Disneyland_HongKong
9  i can go on talking about disneyland. whatever…  Disneyland_HongKong

   polarity
0  0.243981
1  0.236131
2  0.160498
3  0.189286
4  0.266667
5 -0.065476
6  0.185000
7  0.054722
8 -0.067284
9  0.204497
```

[25]:
```
(alt.Chart(disney_df, title ="Polarity Distribution Using TextBlob").mark_bar().
 ↪encode(
    alt.X('polarity', bin=alt.Bin(maxbins=30)),
    y='count()'))
```

[25]:
```
alt.Chart(…)
```

[26]:
```
print('3 random reviews with the positive sentiment polarity: \n')
pos_review = disney_df.loc[disney_df.polarity > 0, ['Review_Text']].sample(3).
 ↪values
for review in pos_review:
    print(review)
```

```
3 random reviews with the positive sentiment polarity:

['disneyland is the best place in world! i loved every moment. a well oiled
machine, fantastic staff, everyone is truly happy, the list goes on. everyone
needs to visit disneyland at least once in their lifetime. i am now officially
```

```
disney obsessed. amazing!']
['this is a no brainer : if you are in paris with children, go an visit
disneyland paris park. your children will love it and the same experience as in
orlando is available.  the access through metro (rer) is quite convenient.']
["just spent two days at disneyland and disney studios, paris, with my nieces
and sisters. the children (aged 6 and 8) absolutely loved it and said it was
the best day of their lives .once you've accepted that you have to queue for at
least 25 minutes for everything (rides, meeting characters, food) and that it's
all geared around pester power, it is really enjoyable.cons: the massive queues
for everything, pushing and shoving around the characters (mainly the adults,
sadly), pester power everywhere you turn, dreadful overpriced junk food served
unbeliebably slowly.pros: clean, relentlessly cheerful, very professional actors
playing the characters (they were so lovely, patient and kind with the
children), music everywhere, zones with different themes and  feel  to them,
rides are great fun when you finally get on (especially the buzz lightyear laser
blast), plenty of toilets, staff generally very pleasant and friendly.we went in
school term time to avoid the worst of the crowds and it was still heaving.
suggest going at the most quiet time you can and getting free  fast pass
tickets in advance where possible.the parade (every day at 7.00pm) was
brilliant; you get to see most disney characters so you wo not come away
disappointed. i honestly take my hat off to the cast members who play the
characters, they stay totally in character at all times and make it really quite
magical."]
```

[27]:
```
print('3 random reviews with the negative sentiment polarity: \n')
neg_review = disney_df.loc[disney_df.polarity < 0, ['Review_Text']].sample(3).
  ↪values
for review in neg_review:
    print(review)
```

```
3 random reviews with the negative sentiment polarity:

['this was my first time visiting an actual disneyland and i felt a bit let
down, as other people have posted there was a lack of characters roaming the
park which i feel loses its uniqueness, alot of effort and details have gone
into most…more']
['had a great time, but it was packed with teens grads. if you have small kids
avoid going on a  grad night  day. the park is overflowing with gangs of un
chaperoned, hormone driven grads in like colored t shirts running around,
screaming for no reason,and cutting lines endlessly. ']
['eurodisney is a great experience if you dont go to florida, california or rust
(europepark) in germany, where the level of entertainment, technique and
greatness are in a higher class. at our visit the eurodisneypark unfortunately
suffered a number of maitenance closures and break downs.especially in
wonderland, waitingtimes are tough. in the studios it   s only af few of the
rides though, that operates with up to 1 hour waitingtime. the disneyland
express should run until near the closing og the park, and not stop already at
17. the park needs shady benches for babysitting patents and grandparents, spacy
```

```
restaurants and non  deep fried dishes. where is the good american
hotdog???boots with cold beverages opens to late and closes too early. ']
```

1.6.2 Using NLTK VADER

- Valence Aware Dictionary for Sentiment Reasoning (VADER) is a model used to analyze the sentiment of text that are sensitive to polarity and emotional intensity (Hutto, 2015)
- NLTK has a direct implementation of the VADER model.

[28]:
```
sid = SentimentIntensityAnalyzer()
```

[29]:
```
disney_df['NLTK_polarity'] = disney_df['Review_Text'].map(lambda text: sid.
↪polarity_scores(text)['compound'])
```

[30]:
```
disney_df.head(10)
```

[30]:

	Review_ID	Rating	Year_Month	Reviewer_Location	\
0	670772142	4	2019-4	Australia	
1	670682799	4	2019-5	Philippines	
2	670623270	4	2019-4	United Arab Emirates	
3	670607911	4	2019-4	Australia	
4	670607296	4	2019-4	United Kingdom	
5	670591897	3	2019-4	Singapore	
6	670585330	5	2019-4	India	
7	670574142	3	2019-3	Malaysia	
8	670571027	2	2019-4	Australia	
9	670570869	5	2019-3	India	

	Review_Text	Branch	\
0	if you've ever been to disneyland anywhere you…	Disneyland_HongKong	
1	its been a while since d last time we visit ho…	Disneyland_HongKong	
2	thanks god it wasn t too hot or too humid wh…	Disneyland_HongKong	
3	hongkong disneyland is a great compact park. u…	Disneyland_HongKong	
4	the location is not in the city, took around 1…	Disneyland_HongKong	
5	have been to disney world, disneyland anaheim …	Disneyland_HongKong	
6	great place! your day will go by and you wo no…	Disneyland_HongKong	
7	think of it as an intro to disney magic for th…	Disneyland_HongKong	
8	feel so let down with this place,the disneylan…	Disneyland_HongKong	
9	i can go on talking about disneyland. whatever…	Disneyland_HongKong	

	polarity	NLTK_polarity
0	0.243981	0.7069
1	0.236131	0.9892
2	0.160498	0.9920
3	0.189286	0.8489
4	0.266667	0.2846
5	-0.065476	0.9653
6	0.185000	0.7489

```
7  0.054722          0.8345
8 -0.067284          0.5195
9  0.204497          0.9645
```

[31]:
```
(alt.Chart(disney_df, title ="NLTK Polarity Score Distribution").mark_bar().
 ↪encode(
    alt.X('NLTK_polarity', bin=alt.Bin(maxbins=30)),
    y='count()'))
```

[31]:
```
alt.Chart(…)
```

[32]:
```
print('3 random reviews with the positive sentiment polarity using NLTK: \n')
pos_review_NLTK = disney_df.loc[disney_df.NLTK_polarity > 0, ['Review_Text']].
 ↪sample(3).values
for review in pos_review_NLTK:
    print(review)
```

```
3 random reviews with the positive sentiment polarity using NLTK:

["i was traveling to hongkong on business and went to the park as a solo
traveler. this is my first experience to a disney park but i've been to
universal, ferrari world, dream world etc. in comparison i thought this is more
suitable for kids. very few selective rides for adults. when you check the park
map you see they have elaborated even little things like walk through snow
whites castle, wishing well, fairy garden etc but actually there's nothing much
to experience. i went there around 10 am and since i had e ticket my entry was
easy. you have machines to get the ticket. by 3pm i had finished all different
sections of the park and finished rides for adults. hyperspace ride in tomorrow
land and big grizzly mountain ride will give you adrenaline rush. then there are
some magical musical show with mickey and few other like toy story parachutes
fall. there's a ferry ride and visit to tarzan's tree house which was so boring.
i took adventure park boat ride and it is also recommended for kids. there's a
train ride around the park which is good for you to get an idea of all
attractions and how crowded they are. i could not do lion king concert since i
had a flight to catch. at 3pm the parade started. it was short and only about 6
carriages and was over within 30 mins. i left after this but heard there are
light shows in the night but i could not wait since i had a flight. but overall
i thought you do not need full day for this park unless you have kids. you can
either come at 10 and leave by 4 or come around 12 and stay back for night
shows. it never extremely crowded. queues were bearable. i went on a weekday so
probably weekend could be different. i think this is good as a disney theme park
but it's very small so plan your day accordingly. recommended for kids but not
much for adults "]
['been here june 2015 with my 2 kids for 2 days. we have last visited 2013
without mystic point. we really had a lot of fun with the rides, characters,
parade,fireworks, and the night parade.we also had a great time exploring the
food and taking picture with the frozen characters ( take note you have to get
```

```
ticket in advance to have picture with them.']
['ok …rides shut no notice before you go ..no vegetarian only
restaurants…toilets clean but could do with attendants to stay on top…could
do with designated smoking areas..long queues average about 90minutes and as you
can only get one fast pass at a time you have not much choice…no characters
just wandering around like what you see on adverts …they have designated times
for a few of them but in 10 hours never saw any…very expensive …did go to
the buzz lightyear pizza very good food and price but had to queue for 40
minutes to get a table this was at one pm as we has a fast pass to star tours
for 3pm..liked the pirates ride and the star tour also the back lot in warner
lot …staff walking around guides etc are very friendly polite and really the
best …mixed feelings really but have done once and never again ..if you go by
train it stops at the entrance car parks are a bit of a walk away and not
free.']
```

[33]:
```
print('3 random reviews with the negative sentiment polarity using NLTK: \n')
neg_review_NLTK = disney_df.loc[disney_df.NLTK_polarity < 0, ['Review_Text']].
 ↪sample(3).values
for review in neg_review_NLTK:
    print(review)
```

```
3 random reviews with the negative sentiment polarity using NLTK:

['of course, expect to pay at least 10 per person per meal. either eat early
before 11:00 or 7:00 and than the lines are not so bad. unless you like sweet
potatoe fries do not eat at the bear resturant as they did not offer any other
fries so i was disappointed with my lunch. but we always eat at the mexican
resturant which is near thunder mountain and the mark twain. we have never been
disappointed with the meals.']
["i was worried about visiting with nearly 2yr old and was right to worry. the
queues were stupid and there was no shade, so we're standing in a queue for
nearly 1 hour with a 5yr old and a 1yr old in the sun with no buggy (had to
leave it before we joined the queue)   absolute nightmare in my opinion. i gave
up and told my hubby to take our older child on stuff and i hung around with the
baby waiting on them.most of the staff were rubbish. no customer service skills
what so ever. i was waiting for some food with my baby and when it came to my
turn the girl got up and started talking to one of her colleagues about her
mobile phone. this went on for a good 5 mins and i walked away then she shouted
me to come back. eh no thanks!!the actual food in the place is below average and
you pay a premium price as you can imagine. quite a lot of the attractions were
shut!i do not regret going, but not sure i will visit this particular one
again."]
['the fireworks   fantasmic   was difficult to move around in the evening before
and after the fireworks']
```

1.6.3 TextBlob vs NLTK VADER

Similarities

- In total, 38,800 (about 90% of the data set) polarity scores obtained by using TextBlob and NLTK VADER indicate the same polarity (positive / negative).
- Both methods are lexicon-based. In the lexicon approach, words are mapped to sentiment, and the overall sentiment of a sentence is determined by aggregating the sentiment of each individual term.
- Lexicon sentiment analysis produces a polarity score ranging from -1 to 1, where -1 indicates highly negative sentiment and 1 indicates highly positive sentiment. A value close to 0 indicates a neutral sentiment.

Differences

- In total, 3,856 (about 10% of the data set) polarity scores obtained by using TextBlob and NLTK VADER indicate the same polarity (positive / negative).
- VADER is specifically designed for analyzing social media content. As a result, VADER dedicates significant effort to identify sentiments in content commonly found on social media platforms, including emojis, repetitive words, and punctuation marks (such as exclamation marks).

Finding I investigated the reviews on which TextBlob and NLTK VADER disagreed by drawing random samples. These reviews contain both negative and positive sentiments, which could be the reason why the two methods cannot agree on the overall sentiment.

Reference: https://pub.towardsai.net/textblob-vs-vader-for-sentiment-analysis-using-python-76883d40f9ae#:~:text=A%20critical%20difference%20between%20TextBlob,exclamation%20marks%2C%20for%2

[34]:
```
disney_df['compare'] = disney_df['polarity'].mul(disney_df['NLTK_polarity']).
↪ge(0)
```

[35]:
```
disney_df.head(10)
```

[35]:

	Review_ID	Rating	Year_Month	Reviewer_Location	\
0	670772142	4	2019-4	Australia	
1	670682799	4	2019-5	Philippines	
2	670623270	4	2019-4	United Arab Emirates	
3	670607911	4	2019-4	Australia	
4	670607296	4	2019-4	United Kingdom	
5	670591897	3	2019-4	Singapore	
6	670585330	5	2019-4	India	
7	670574142	3	2019-3	Malaysia	
8	670571027	2	2019-4	Australia	
9	670570869	5	2019-3	India	

	Review_Text	Branch	\
0	if you've ever been to disneyland anywhere you…	Disneyland_HongKong	
1	its been a while since d last time we visit ho…	Disneyland_HongKong	
2	thanks god it wasn t too hot or too humid wh…	Disneyland_HongKong	
3	hongkong disneyland is a great compact park. u…	Disneyland_HongKong	
4	the location is not in the city, took around 1…	Disneyland_HongKong	

```
5  have been to disney world, disneyland anaheim …  Disneyland_HongKong
6  great place! your day will go by and you wo no…  Disneyland_HongKong
7  think of it as an intro to disney magic for th…  Disneyland_HongKong
8  feel so let down with this place,the disneylan…  Disneyland_HongKong
9  i can go on talking about disneyland. whatever…  Disneyland_HongKong
```

```
   polarity  NLTK_polarity  compare
0  0.243981         0.7069     True
1  0.236131         0.9892     True
2  0.160498         0.9920     True
3  0.189286         0.8489     True
4  0.266667         0.2846     True
5 -0.065476         0.9653    False
6  0.185000         0.7489     True
7  0.054722         0.8345     True
8 -0.067284         0.5195    False
9  0.204497         0.9645     True
```

[36]:
```
disney_df['compare'].value_counts()
```

[36]:
```
True     38800
False     3856
Name: compare, dtype: int64
```

[37]:
```
print('3 random reviews with different sentiment by TextBlob and NLTK Vader:␣
 ↪\n')
diff_sent_review = disney_df.loc[disney_df['compare'] == 0, ['Review_Text']].
 ↪sample(3).values
for review in diff_sent_review:
    print(review)
```

```
3 random reviews with different sentiment by TextBlob and NLTK Vader:

['as always the great joy for the kids makes up for the ridiculously high costs
of food and drink! still a winner for the kids   our 3 year old grandson and our
25 year old daughter and her husband. fast pass is a must if you do not want to
queue for hours. but rides deliver and staff were i have no idea how! the food
and drinks are atrociously expensive and the food inside is dreadful. dinner at
pirates of the caribbean was 2.5 hours of misery, warm food and dreadful
service. how do they get it so wrong. eat in the village just outside the park
so much better! ']
["visited when it rained in the morning only lace to shelter is the shops.. none
of the character s we're available. parades shows were cancelled or delayed so
long that ckids got bored. biggest waste of money. go to a non disney theme park
that has decent rides. kids will have a better day"]
['lots of things to do and see at disneyland. cast members are very helpful.
they are always updating rides and events. park is immaculate. it truly is the
```

```
happiest place on earth. we try to go at least once a year. only downside is the
park hoppers are so expensive.']
```

[38]:
```
disney_df.to_csv('../data/sensitivity.csv') # save the polarity / sensitivity␣
↪results for further analysis using ChatGPT
```

References Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014. Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137. https://doi.org/10.1108/eb046814