GENERATION GIRL
HOLIDAY CLUB [YEAR]

FINAL PROJECT

1 Maret 2024

Outlines

Introduction

Objectives

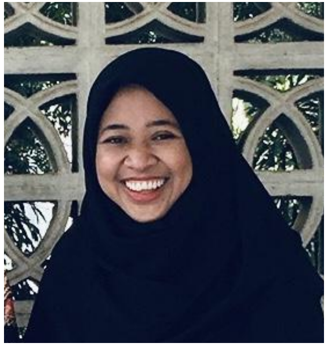
EDA

Model Choices

Best Model and Recommendation

Introduction

We Are



Ina

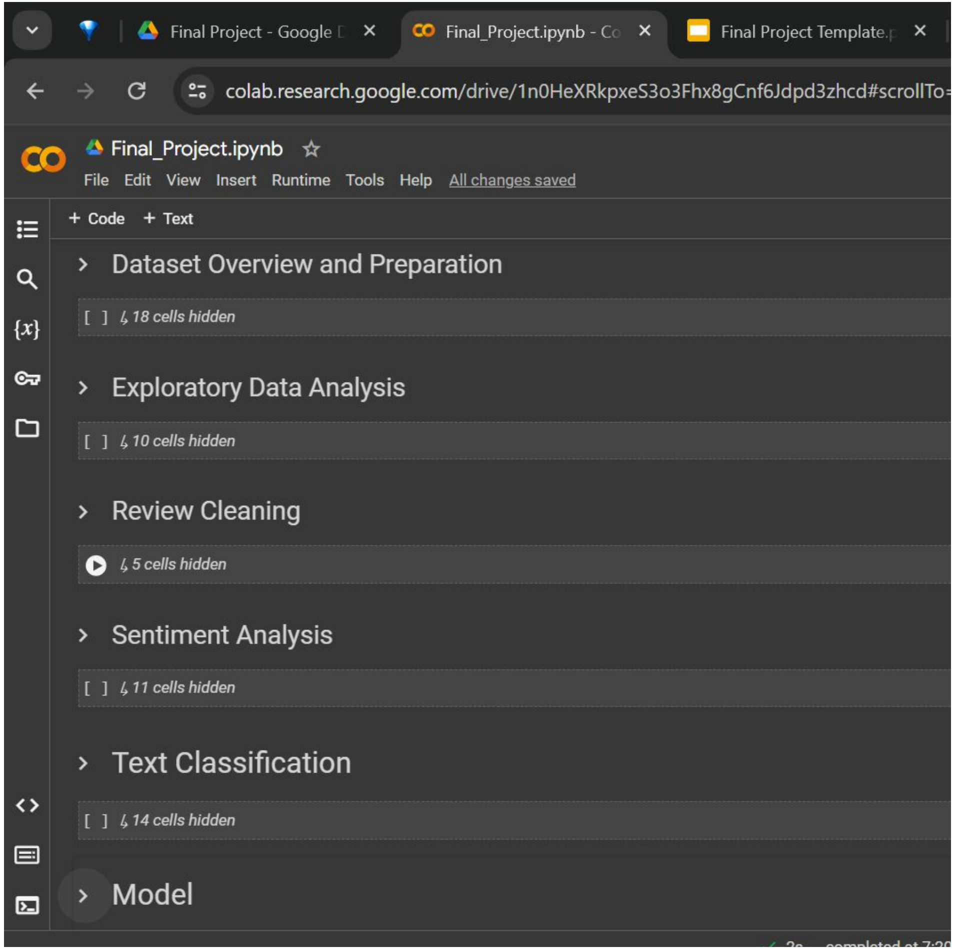
Ika

Objectives

Problem Statement and Objective

1. Untuk memahami pengalaman pengunjung Universal Studios di Florida, Jepang, dan Singapura menggunakan metode sentimen analisis
2. Untuk mengetahui apa saja hal yang perlu diperbaiki oleh Universal Studios dalam rangka meningkatkan kepuasan pengunjung yang nantinya diharapkan mampu meningkatkan keuntungan bisnis bagi Universal Studios

Steps


Final Project - Google
Final_Project.ipynb - Co
Final Project Template.
colab.research.google.com/drive/1n0HeXRkpxeS3o3Fhx8gCnf6Jdpd3zhcd#scrollTo=
Final_Project.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

Dataset Overview and Preparation

[ ] 18 cells hidden

Exploratory Data Analysis

[ ] 10 cells hidden

Review Cleaning

5 cells hidden

Sentiment Analysis

[ ] 11 cells hidden

Text Classification

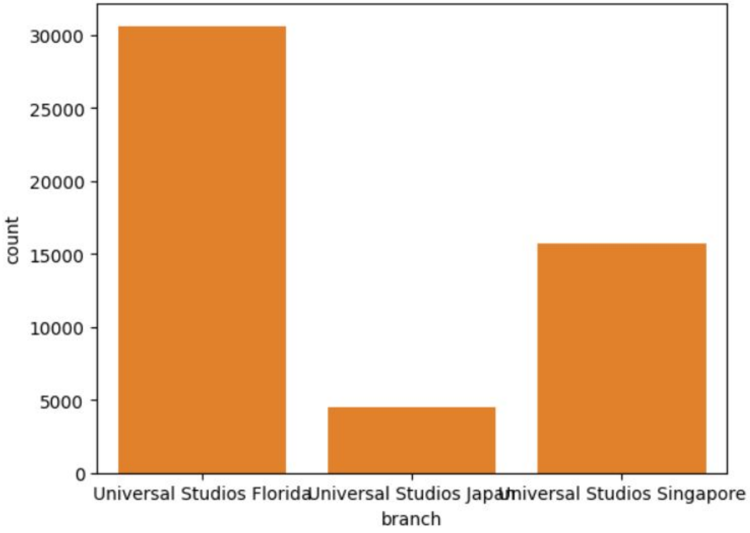
[ ] 14 cells hidden

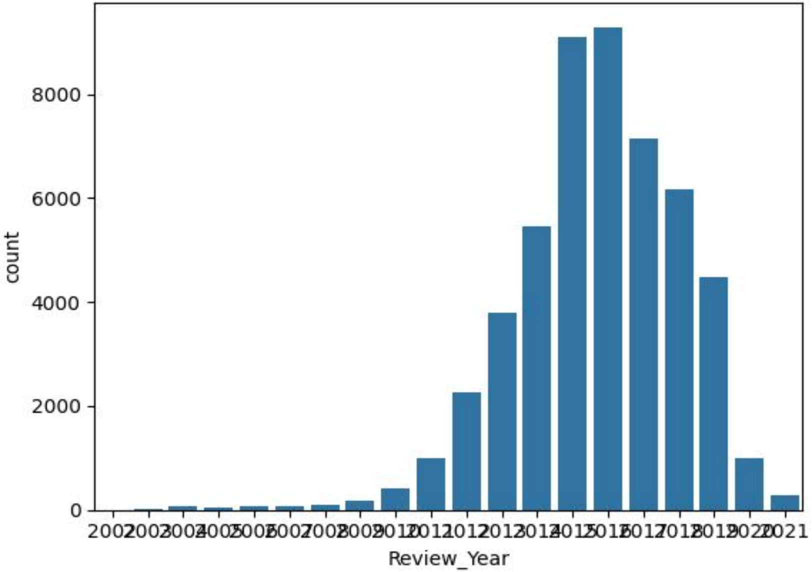
Model

Data Overview

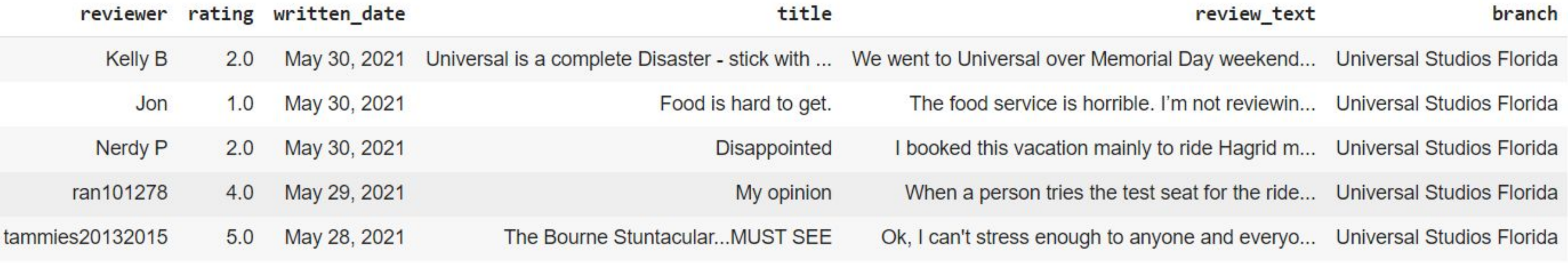
Dataset: https://www.kaggle.com/datasets/dwiknrd/reviewuniversalstudio

Terdapat sejumlah 50,904 TripAdvisor reviews Universal Studios dari 3 cabang berbeda, yaitu Florida, Jepang dan Singapura, dengan proporsi sebagai berikut:

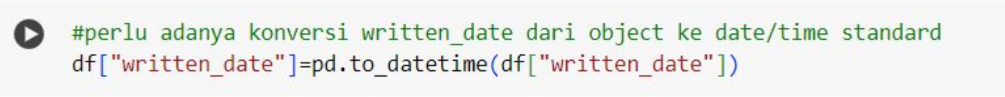




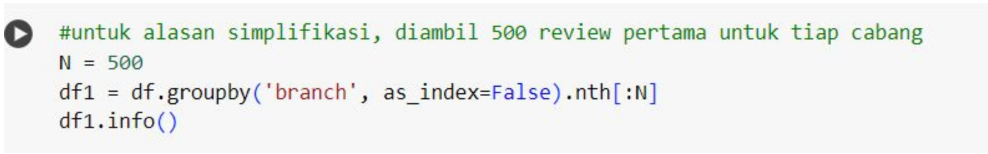
Review terdiri atas beberapa kolom informasi, yaitu



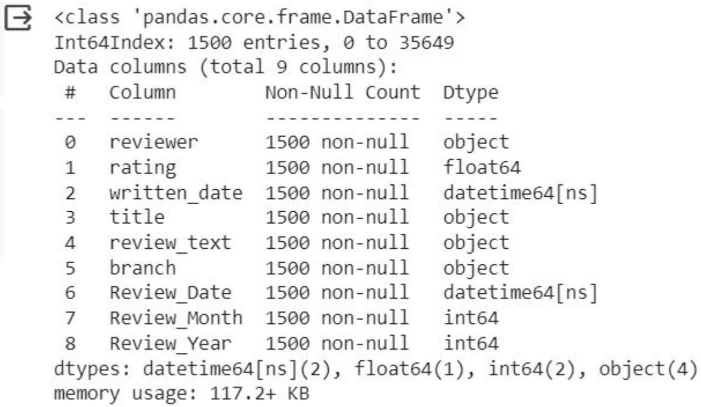
reviewer	rating	written_date	title	review_text	branch
Kelly B	2.0	May 30, 2021	Universal is a complete Disaster - stick with ...	We went to Universal over Memorial Day weekend...	Universal Studios Florida
Jon	1.0	May 30, 2021	Food is hard to get.	The food service is horrible. I'm not reviewin...	Universal Studios Florida
Nerdy P	2.0	May 30, 2021	Disappointed	I booked this vacation mainly to ride Hagrid m...	Universal Studios Florida
ran101278	4.0	May 29, 2021	My opinion	When a person tries the test seat for the ride...	Universal Studios Florida
tammies20132015	5.0	May 28, 2021	The Bourne Stuntacular...MUST SEE	Ok, I can't stress enough to anyone and everyo...	Universal Studios Florida



```
#perlu adanya konversi written_date dari object ke date/time standard
df["written_date"]=pd.to_datetime(df["written_date"])
```



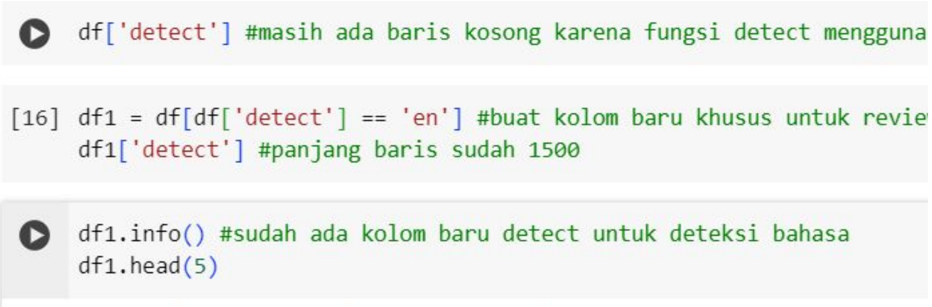
```
#untuk alasan simplifikasi, diambil 500 review pertama untuk tiap cabang
N = 500
df1 = df.groupby('branch', as_index=False).nth[:N]
df1.info()
```



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1500 entries, 0 to 35649
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   reviewer      1500 non-null   object
 1   rating        1500 non-null   float64
 2   written_date  1500 non-null   datetime64[ns]
 3   title         1500 non-null   object
 4   review_text   1500 non-null   object
 5   branch        1500 non-null   object
 6   Review_Date   1500 non-null   datetime64[ns]
 7   Review_Month  1500 non-null   int64
 8   Review_Year   1500 non-null   int64
dtypes: datetime64[ns](2), float64(1), int64(2), object(4)
memory usage: 117.2+ KB
```

Tidak ada null value

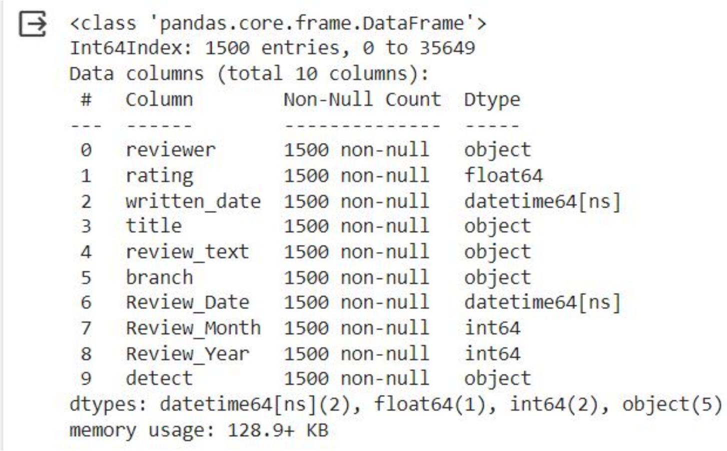
Menghilangkan Review Berbahasa Selain Inggris



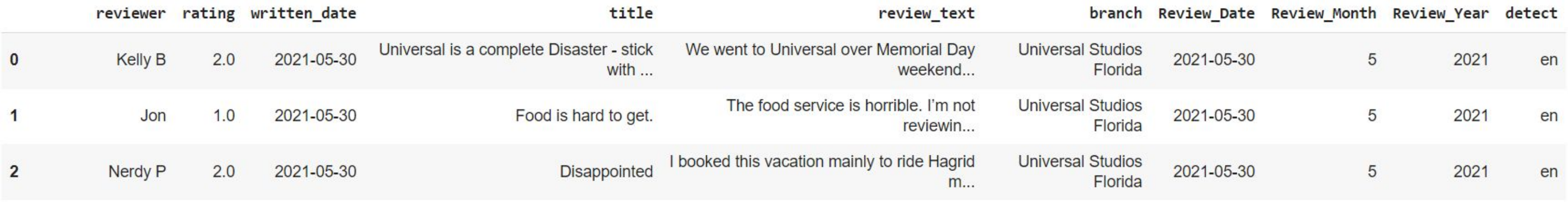
```
df['detect'] #masih ada baris kosong karena fungsi detect mengguna
```

```
[16] df1 = df[df['detect'] == 'en'] #buat kolom baru khusus untuk revie
     df1['detect'] #panjang baris sudah 1500
```

```
df1.info() #sudah ada kolom baru detect untuk deteksi bahasa
df1.head(5)
```

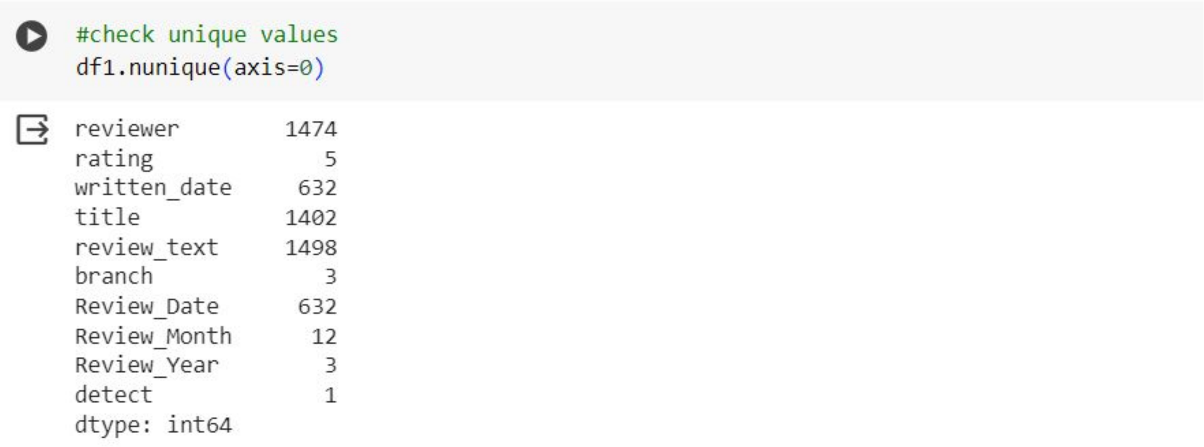


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1500 entries, 0 to 35649
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   reviewer      1500 non-null   object
 1   rating        1500 non-null   float64
 2   written_date  1500 non-null   datetime64[ns]
 3   title         1500 non-null   object
 4   review_text   1500 non-null   object
 5   branch        1500 non-null   object
 6   Review_Date   1500 non-null   datetime64[ns]
 7   Review_Month  1500 non-null   int64
 8   Review_Year   1500 non-null   int64
 9   detect        1500 non-null   object
dtypes: datetime64[ns](2), float64(1), int64(2), object(5)
memory usage: 128.9+ KB
```



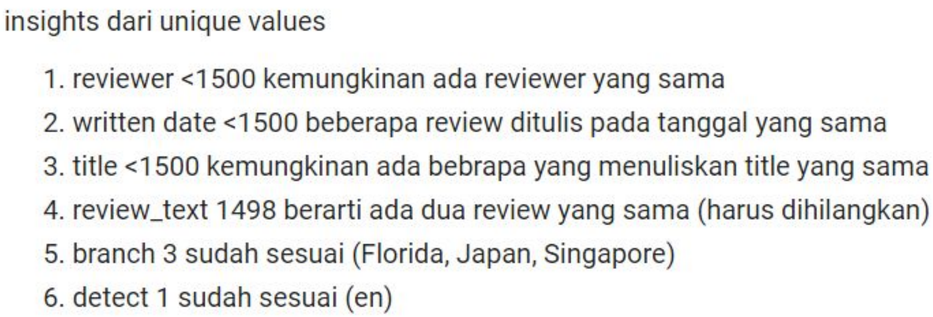
	reviewer	rating	written_date	title	review_text	branch	Review_Date	Review_Month	Review_Year	detect
0	Kelly B	2.0	2021-05-30	Universal is a complete Disaster - stick with ...	We went to Universal over Memorial Day weekend...	Universal Studios Florida	2021-05-30	5	2021	en
1	Jon	1.0	2021-05-30	Food is hard to get.	The food service is horrible. I'm not reviewin...	Universal Studios Florida	2021-05-30	5	2021	en
2	Nerdy P	2.0	2021-05-30	Disappointed	I booked this vacation mainly to ride Hagrid m...	Universal Studios Florida	2021-05-30	5	2021	en

Pengecekan Unique Values



```
#check unique values
df1.nunique(axis=0)
```

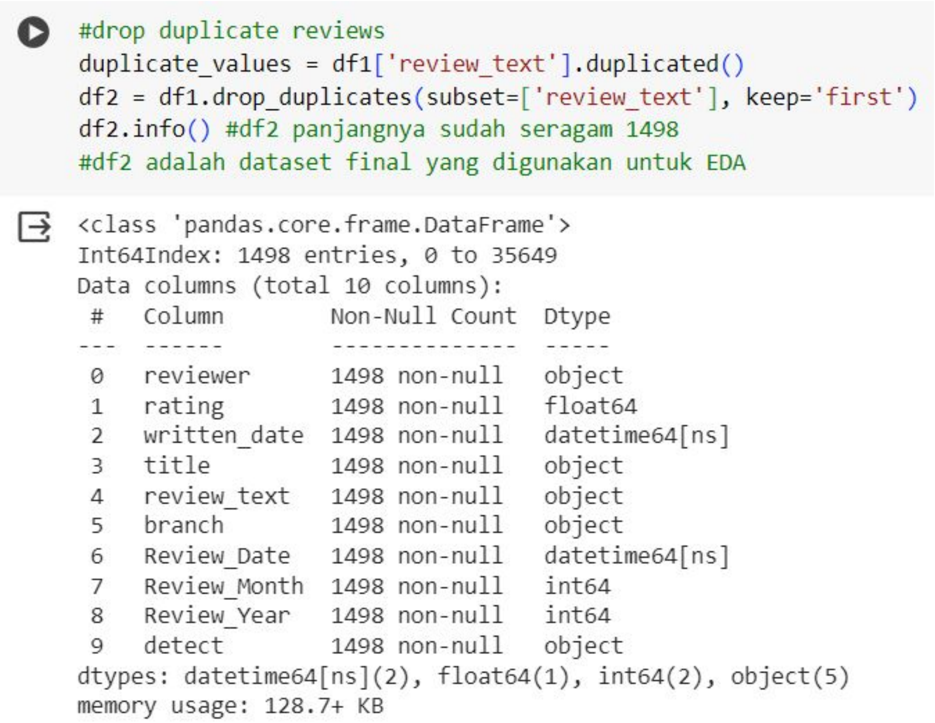
```
reviewer        1474
rating             5
written_date     632
title           1402
review_text     1498
branch             3
Review_Date      632
Review_Month      12
Review_Year        3
detect             1
dtype: int64
```



insights dari unique values

1. reviewer <1500 kemungkinan ada reviewer yang sama
2. written date <1500 beberapa review ditulis pada tanggal yang sama
3. title <1500 kemungkinan ada bebrapa yang menuliskan title yang sama
4. review_text 1498 berarti ada dua review yang sama (harus dihilangkan)
5. branch 3 sudah sesuai (Florida, Japan, Singapore)
6. detect 1 sudah sesuai (en)

Menghilangkan Duplicate Reviews



```python
#drop duplicate reviews
duplicate_values = df1['review_text'].duplicated()
df2 = df1.drop_duplicates(subset=['review_text'], keep='first')
df2.info() #df2 panjangnya sudah seragam 1498
#df2 adalah dataset final yang digunakan untuk EDA
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1498 entries, 0 to 35649
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   reviewer      1498 non-null   object
 1   rating        1498 non-null   float64
 2   written_date  1498 non-null   datetime64[ns]
 3   title         1498 non-null   object
 4   review_text   1498 non-null   object
 5   branch        1498 non-null   object
 6   Review_Date   1498 non-null   datetime64[ns]
 7   Review_Month  1498 non-null   int64
 8   Review_Year   1498 non-null   int64
 9   detect        1498 non-null   object
dtypes: datetime64[ns](2), float64(1), int64(2), object(5)
memory usage: 128.7+ KB
```

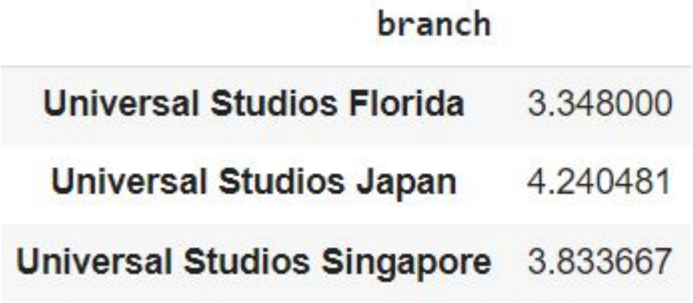
Jadi total baris data yang digunakan adalah 1498

EDA

Data Insights - Through Exploration

Rata-rata rating secara umum adalah 3.8

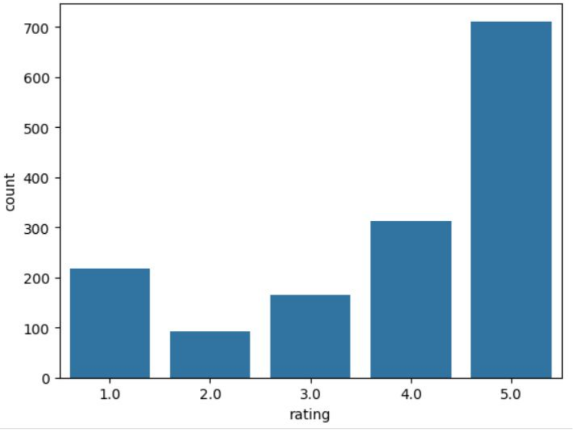
Rata-rata rating tiap cabang studio

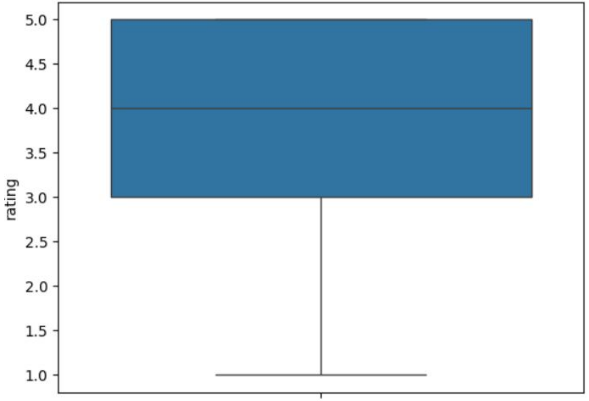


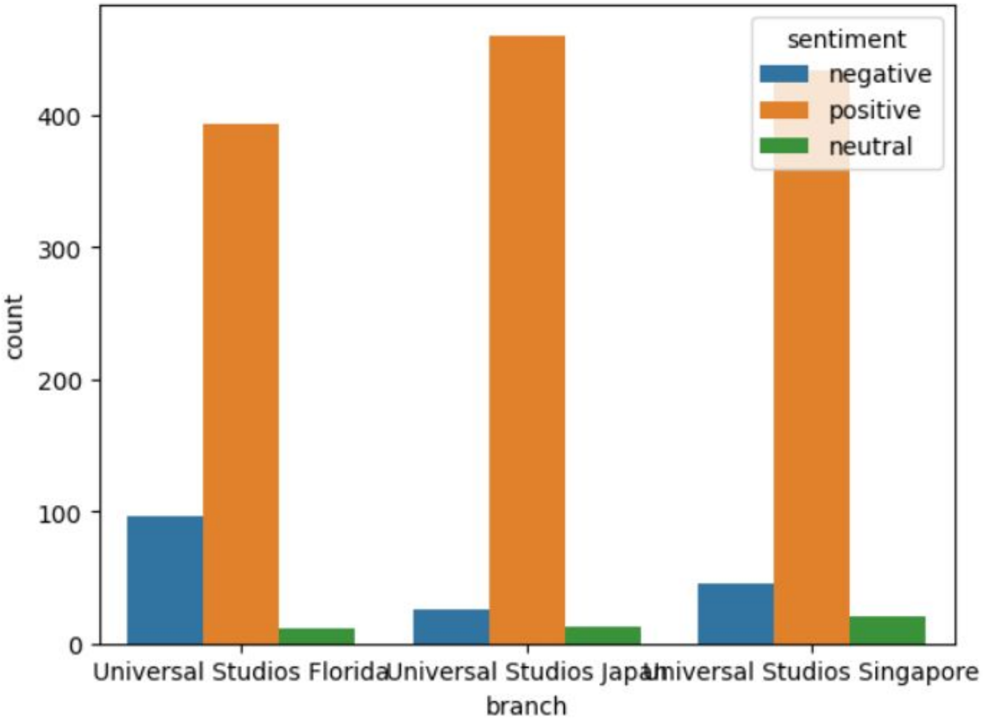
branch	
Universal Studios Florida	3.348000
Universal Studios Japan	4.240481
Universal Studios Singapore	3.833667

Meski demikian, rating 5.0 merupakan rating yang paling sering muncul

Visualisasi boxplot juga menunjukkan data yang cenderung besar di nilai Q3 (rating yang besar)



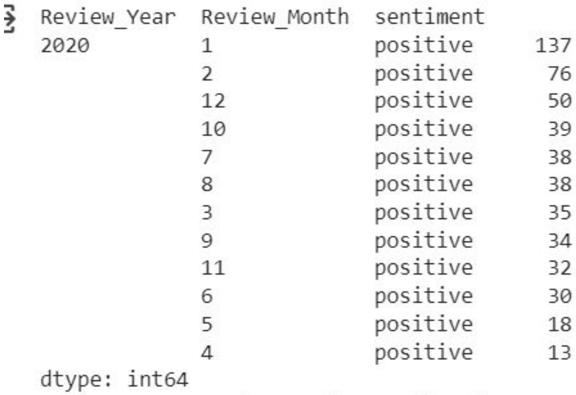




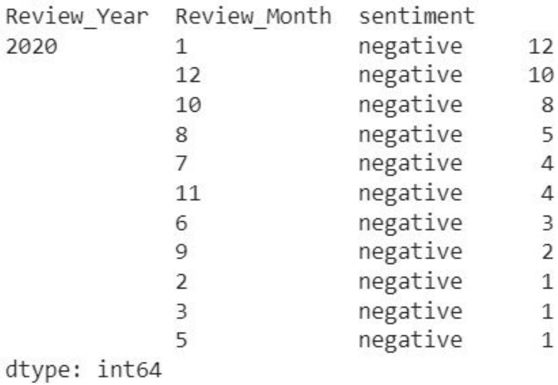
Hasil analisis sentimen juga menunjukkan >50% pengunjung memiliki pengalaman yang positif

Jika dilihat berdasarkan review per bulan pada tahun 2020, maka bulan Januari menjadi bulan dengan review positif dan negatif terbanyak secara bersamaan

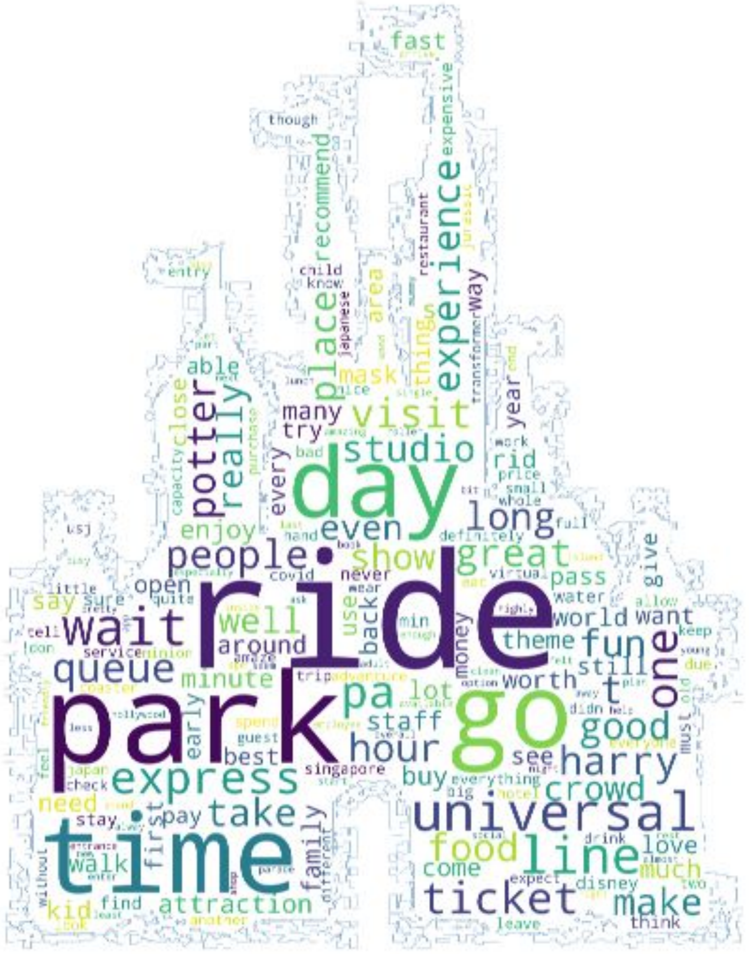
Hal ini kemungkinan dikarenakan pada bulan Januari merupakan musim puncak bersamaan dengan momen perayaan tahun baru.



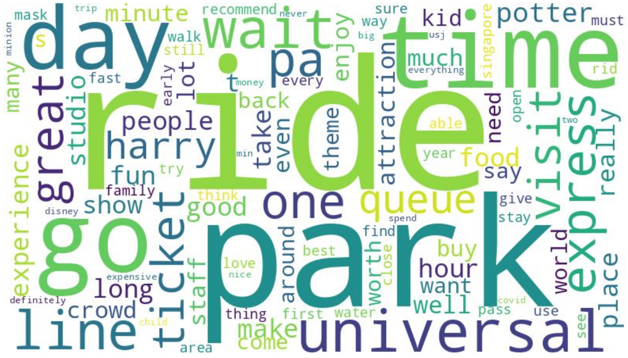
```
Review_Year  Review_Month  sentiment
2020         1             positive     137
             2             positive      76
             12            positive      50
             10            positive      39
             7             positive      38
             8             positive      38
             3             positive      35
             9             positive      34
             11            positive      32
             6             positive      30
             5             positive      18
             4             positive      13
dtype: int64
```



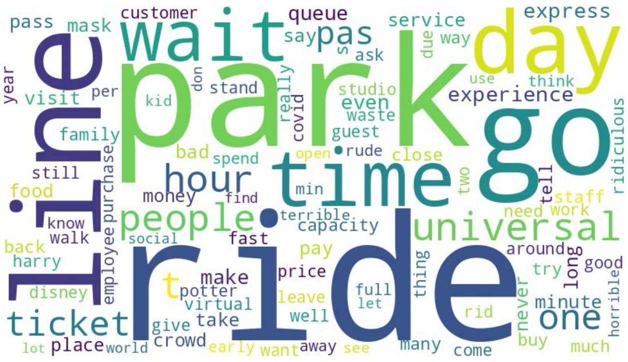
```
Review_Year  Review_Month  sentiment
2020         1             negative     12
             12            negative     10
             10            negative      8
             8             negative      5
             7             negative      4
             11            negative      4
             6             negative      3
             9             negative      2
             2             negative      1
             3             negative      1
             5             negative      1
dtype: int64
```



Positif: RIDE, PARK



Negatif: RIDE, PARK, LINE, WAIT, TIME

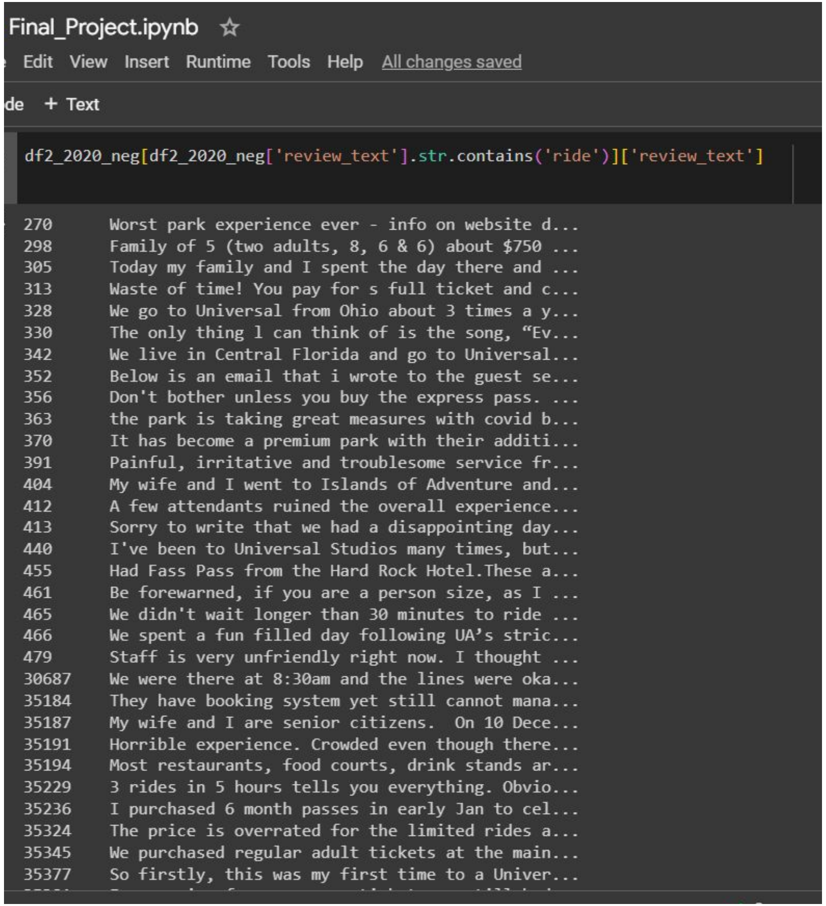


More Insights

Adanya sentimen positif dan negatif yang berimbang antara kata PARK dan RIDE

Sementara untuk sentimen negatif tampak ada kata LINE, WAIT, dan TIME yang kemungkinan menggambarkan adanya pengalaman kurang menyenangkan terkait antrian dan waktu tunggu

Perlu ada evaluasi lebih lanjut terhadap wahana permainan (RIDE) di Universal Studio untuk mengetahui apa yang dapat di improve dan sudah cukup baik menurut pengunjung terkait wahana permainan



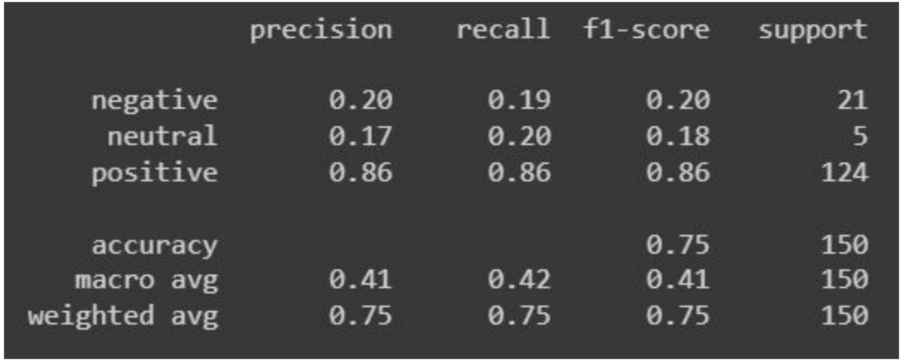
```
df2_2020_neg[df2_2020_neg['review_text'].str.contains('ride')]['review_text']
```

```
270      Worst park experience ever - info on website d...
298      Family of 5 (two adults, 8, 6 & 6) about $750 ...
305      Today my family and I spent the day there and ...
313      Waste of time! You pay for s full ticket and c...
328      We go to Universal from Ohio about 3 times a y...
330      The only thing l can think of is the song, “Ev...
342      We live in Central Florida and go to Universal...
352      Below is an email that i wrote to the guest se...
356      Don't bother unless you buy the express pass. ...
363      the park is taking great measures with covid b...
370      It has become a premium park with their additi...
391      Painful, irritative and troublesome service fr...
404      My wife and I went to Islands of Adventure and...
412      A few attendants ruined the overall experience...
413      Sorry to write that we had a disappointing day...
440      I've been to Universal Studios many times, but...
455      Had Fass Pass from the Hard Rock Hotel.These a...
461      Be forewarned, if you are a person size, as I ...
465      We didn't wait longer than 30 minutes to ride ...
466      We spent a fun filled day following UA’s stric...
479      Staff is very unfriendly right now. I thought ...
30687    We were there at 8:30am and the lines were oka...
35184    They have booking system yet still cannot mana...
35187    My wife and I are senior citizens.  On 10 Dece...
35191    Horrible experience. Crowded even though there...
35194    Most restaurants, food courts, drink stands ar...
35229    3 rides in 5 hours tells you everything. Obvio...
35236    I purchased 6 month passes in early Jan to cel...
35324    The price is overrated for the limited rides a...
35345    We purchased regular adult tickets at the main...
35377    So firstly, this was my first time to a Univer...
```

Model Choices

Model Description + Results

TF-IDF Vectorizer

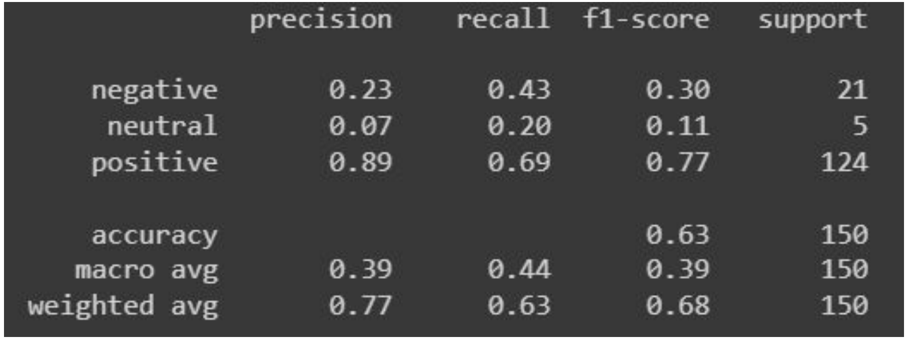


```
              precision    recall  f1-score   support

    negative       0.20      0.19      0.20        21
     neutral       0.17      0.20      0.18         5
    positive       0.86      0.86      0.86       124

    accuracy                           0.75       150
   macro avg       0.41      0.42      0.41       150
weighted avg       0.75      0.75      0.75       150
```

CountVectorizer



```
              precision    recall  f1-score   support

    negative       0.23      0.43      0.30        21
     neutral       0.07      0.20      0.11         5
    positive       0.89      0.69      0.77       124

    accuracy                           0.63       150
   macro avg       0.39      0.44      0.39       150
weighted avg       0.77      0.63      0.68       150
```

Skor presisi, recall dan F1-score lebih tinggi untuk sentimen positif dibandingkan sentimen negatif menunjukkan bahwa model lebih akurat dalam mengidentifikasi teks positif dibandingkan teks negatif.

Akurasi tergolong tinggi dan menunjukkan bahwa model cukup handal dalam memprediksi kategori teks

Best Model and Recommendation

Best Model + Insight

1. Tingkat akurasi pada TF-IDF sebesar 75% berarti bahwa model TF-IDF mampu memprediksi kategori teks dengan benar dalam 75% dari seluruh sampel data yang digunakan untuk pengujian dibanding dengan CountVectorizer.
2. TF-IDF lebih sensitif dalam memprediksi sentiment positive, menunjukkan tingkat error yang lebih rendah bila ada data baru.
3. Dari perbandingan dua model yang menggunakan metode ekstraksi fitur yang berbeda dapat diketahui bahwa presisi dan recall dari TF-IDF lebih baik dibanding count vectorizer. Ini bisa dilihat dari F1 Score pada Classification Report.

Q & A

Thank You!