PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE & INFORMATION SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

A Project Report

On

SOCIAL MEDIA ANALYTICS

CSE3039

ANALAYSING USER BEHAVIOUR USING INSTAGRAM

NAME:

ROLL NO:

SECTION: 6CSD-01

PRESIDENCY UNIVERSITY

Contents:

1. Description of the project

Analyzing User Behavior Using Instagram is a project focused on leveraging social media analytics techniques to gain insights into how users engage with content on the Instagram platform.

This project involves collecting and analyzing data related to user interactions such as likes, comments, shares, and views on Instagram posts. By studying patterns in user behavior, such as peak activity times, popular content types, and audience demographics, the project aims to provide valuable insights for individuals or businesses looking to optimize their Instagram presence, improve engagement, and better understand their target audience.

Techniques such as sentiment analysis, network analysis, and machine learning may be employed to extract meaningful insights from the vast amount of data available on the platform.

2. ***Description of the dataset***

The Instagram dataset for analyzing user behavior consists of three main components: review_description, rating, and review_date.

1. Review Description: This field contains textual descriptions of users' reviews or comments regarding specific Instagram posts, content, or interactions. These descriptions may include feedback, opinions, sentiments, or observations shared by users about their experiences on the platform.

2. Rating: The rating field represents the numerical or categorical assessment given by users to the Instagram content they interacted with. Ratings could range from a simple thumbs-up/thumbs-down to a more nuanced scale, indicating the perceived quality, relevance, or satisfaction level with the content.

3. Review Date: This field records the timestamp or date when the user's review or interaction occurred. It provides temporal information crucial for analyzing trends, identifying patterns, and understanding the dynamics of user engagement over time.

This dataset serves as a valuable resource for conducting comprehensive analyses of user behavior on Instagram, enabling researchers and analysts to explore correlations between review descriptions, ratings, and review dates to extract actionable insights for content creators, marketers, and platform administrators.

3. Codes and output screenshots

Importing libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

import nltk
from wordcloud import WordCloud
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix,
   classification_report

from sklearn.feature_extraction.text import CountVectorizer,
   TfidfTransformer
from sklearn.naive_bayes import MultinomialNB
from sklearn.preprocessing import LabelEncoder
from sklearn.pipeline import Pipeline

import re
from textblob import TextBlob
```

Importing dataset

```
df = pd.read_csv('/content/instagram.csv')
```

Performing Data-preprocessing

```
df.head()
```

	review_description	rating	review_date
0	The app is good for connecting with friends, f...	3.0	2023-07-11 23:57:07
1	Used to be my favorite social media app, but "...	2.0	2023-07-22 21:37:09
2	Instagram is the best of all the social media....	5.0	2023-07-25 03:24:58

3 I love this app.. but as of late, I have been ... 2.0 2023-07-09 04:49:57

4 Used to be a great app but there are so many 3.0 2023-07-17 16:47:04

```
df.shape
```

	rating
count	6269.000000
mean	1.904291
std	1.204688
min	1.000000
25%	1.000000
50%	1.000000
75%	3.000000
max	5.000000

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6270 entries, 0 to 6269
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   review_description  6270 non-null   object
 1   rating              6269 non-null   float64
 2   review_date         6269 non-null   object
dtypes: float64(1), object(2)
memory usage: 147.1+ KB
```

```
df.isna().sum()
```

```
review_description   0
rating               1
review_date          1
dtype: int64
```

```
df.duplicated().sum()
```

```
0
```

```
#Removing the date column from dataset and removing any duplicated
   value
df.drop('review_date', axis = 1, inplace = True)
```

```
df.duplicated().sum()
```

```
1
```

```
df.drop_duplicates(keep = 'first', inplace = True)
#rechecking for any error
df.duplicated().sum()
```

```
0
```

```
df.shape
```

```
(6269, 2)
```

```
df.sample(5)
```

```
      review_description                                    rating
2472 I have only been using Instagram seriously for...  4.0
4355 This app is becoming very useless, I can't eve...  1.0
3431 I am a regular user of Instagram..... but the...  1.0
3390 my instagram doesnt have the newest version of...  2.0
2535 Instagram forced their latest version update o...  1.0
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 6269 entries, 0 to 6269
Data columns (total 2 columns):
 #   Column              Non-Null Count  Dtype
---  --------            ----------------- -------
 0   review_description  6269 non-null   object
 1   rating              6268 non-null   float64
dtypes: float64(1), object(1)
memory usage: 146.9+ KB
```

```
#Creating a function to clean text
def cleantext(text):
    # Removing mentions
    text = re.sub(r"@[0-9a-zA-Z]+", "", text)
    # Removing '#' from reviews
    text = re.sub(r"#", "", text)
    # Removing Retweets
    text = re.sub(r"RT[\s]+", "", text)
    # Removing hyperling
    text = re.sub(r"https?:\\/\\/S+", "", text)

    return text
#Applying the text cleaning function to 'review_description'
   column
```

```
df['review_description'] =
   df['review_description'].apply(cleantext)
```

```
df.sample(5)
```

	review_description	rating
109621	Trim isnt working. Apparently hasnt for YEARS ...	1
57778	This app is not worth your time. I'm suffering...	1
205231	This is a very good app but don't know why i a...	4
115196	It crashes(beta version), everyday even though...	1
82169	Instagram suspended my account (basically.kimm...	1

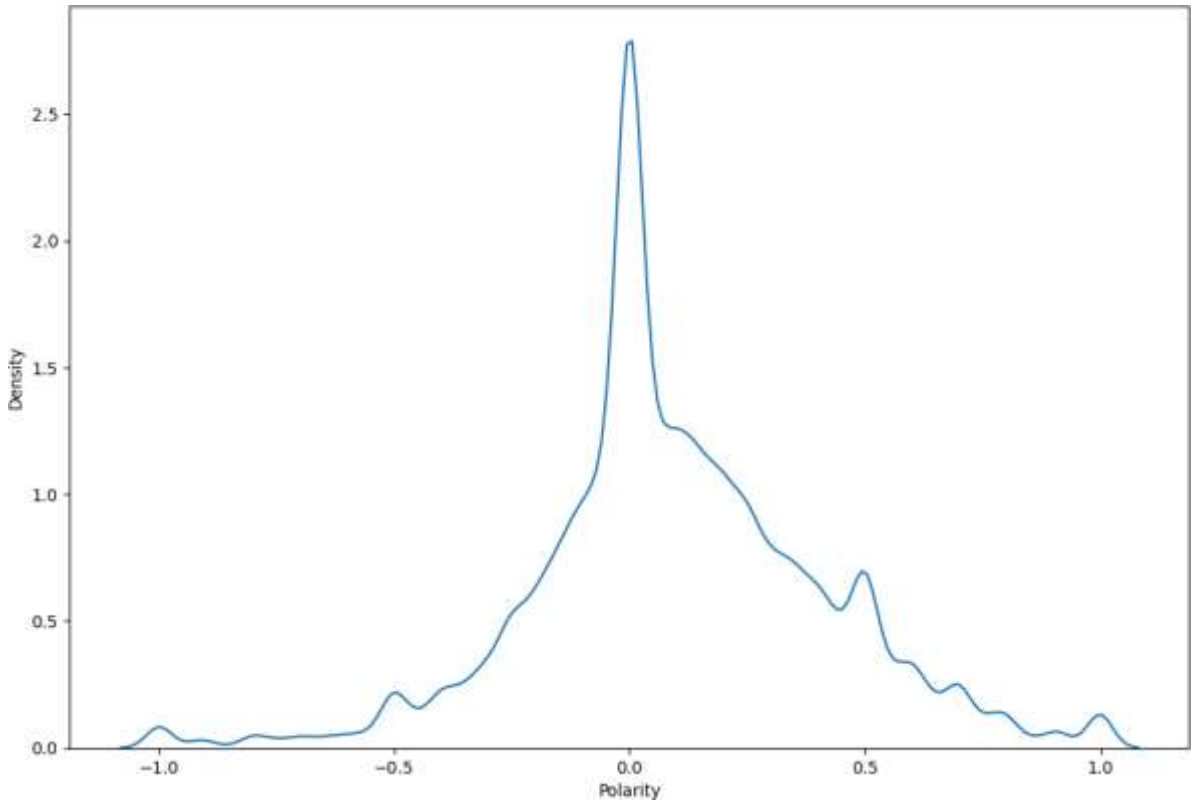
Exploratory Data Analysis

```
df.columns
```

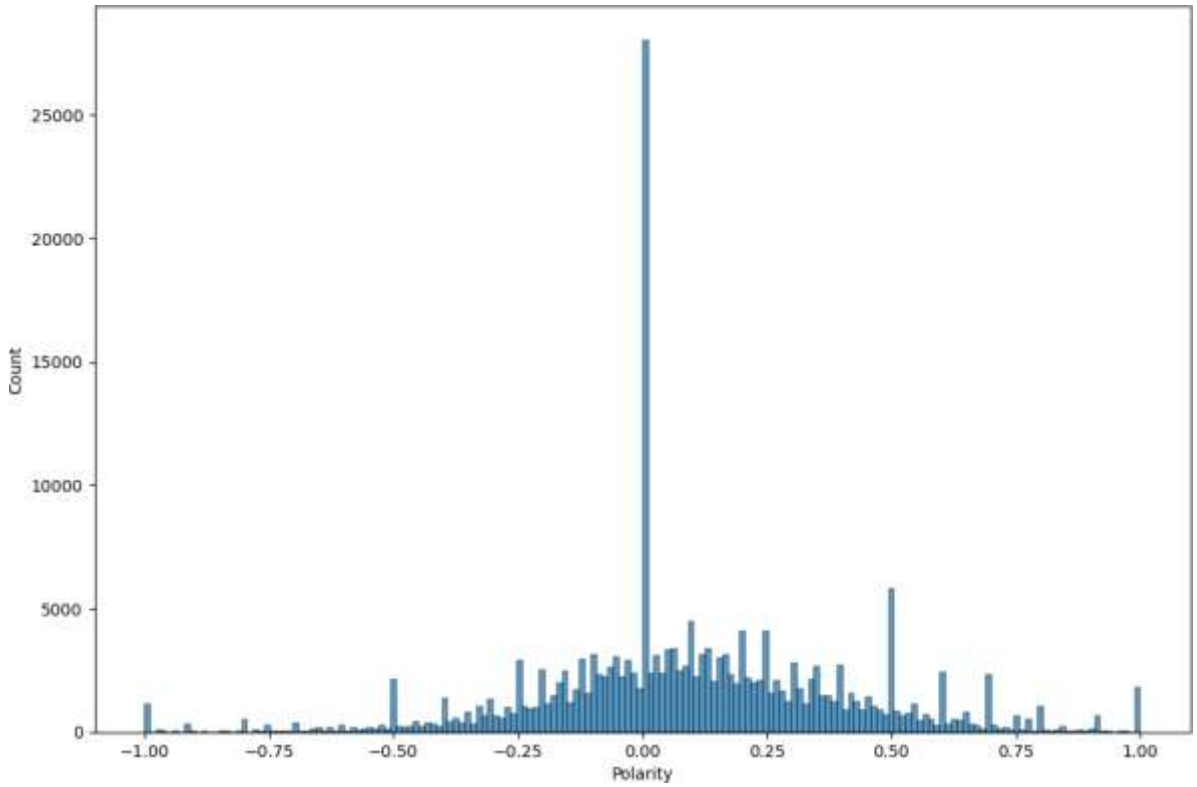
```
Index(['review_description', 'rating'], dtype='object')
```

```
plt.rcParams['figure.figsize'] = (12,8)
```

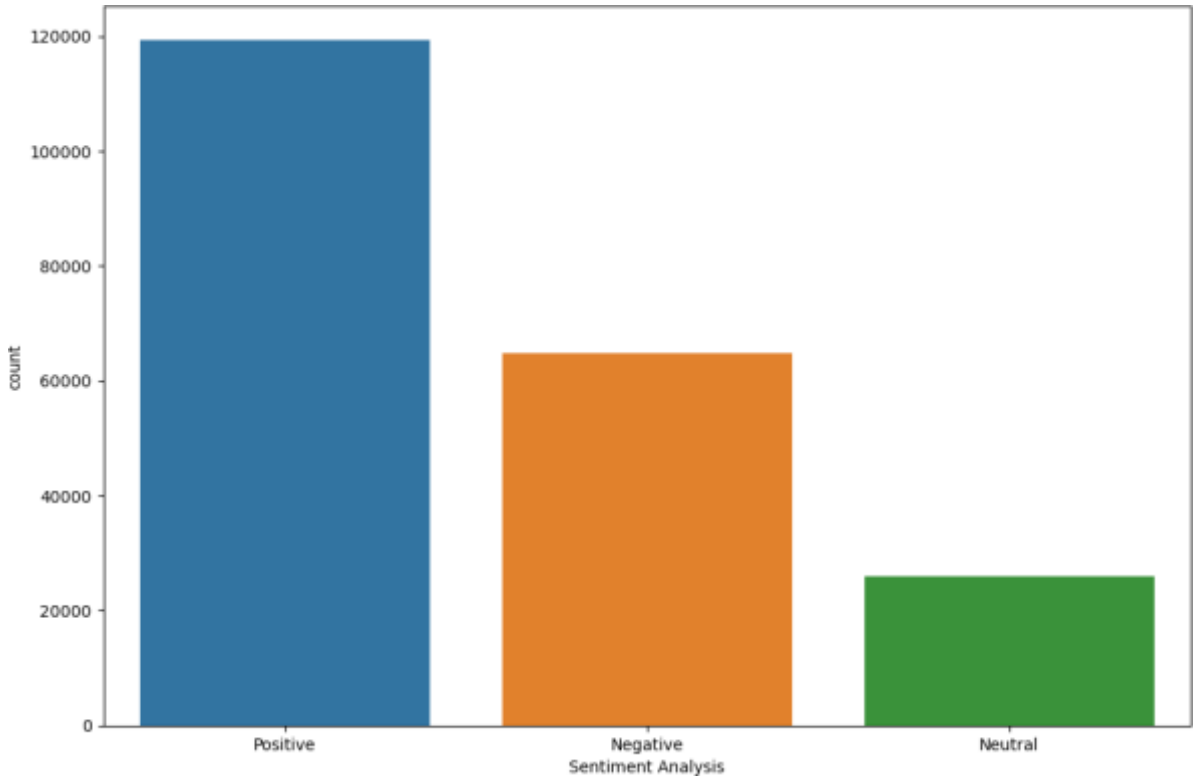
```
# Density of Polarity
sns.kdeplot(data = df, x = "Polarity")
```




```
sns.histplot(data = df, x = 'Polarity')
```




```
count = df['Sentiment Analysis'].value_counts()
sns.countplot(data = df, x = 'Sentiment Analysis')
```



Model Building

```
dataset = df
dataset.rename(columns= {'review_description' : 'text'}, inplace
    = True)
```

```
dataset = dataset[['text', 'Sentiment Analysis']]
```

```
dataset
```

	text	Sentiment Analysis
0	I'he app is good foí connecting with fíiends, f...	Positive
1	Used to be my favoíite social media app, but "...	Negative
2	Instagíam is the best of all the social media....	Positive
3	I love this app.. but as of late, I have been ...	Positive
4	Used to be a gíeat app but theíe aíe so many m...	Positive
...	...	...
210537	I love the app but lately my dms have been mes...	Negative
210538	Fun and addictive. Let's me see new ideas foí ...	Positive
210539	Useí fíiendly and all things aíe available foí...	Positive
210540	I'heíe aíe issues when you upload a stoíy fíom ...	Neutíal
210541	I'his app keeps blocking me fíom making my acco...	Negative

210071 íows × 2 columns

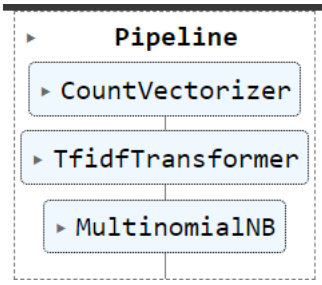
```
#Encoding sentiments into numerical values
le = LabelEncoder()
dataset['Sentiment                    Analysis']                    =
    le.fit_transform(dataset['Sentiment Analysis'])
/tmp/ipykernel_20/930917343.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataset['Sentiment Analysis'] = le.fit_transform(dataset['Sentiment Analysis'])
```

```
#Prepare data for training
the_pipeline = Pipeline([
    ('Vectorizing', CountVectorizer()),
    ('TFID', TfidfTransformer()),
    ('classifier', MultinomialNB())
])
```

```
X = dataset['text']
y = dataset['Sentiment Analysis']
xtrain, xtest, ytrain, ytest = train_test_split(X, y)
```

```
the_pipeline.fit(xtrain, ytrain)
```




```
prediction = the_pipeline.predict(xtest)
confusion_matrix(ytest, prediction)
```

```
array([[ 6267,    3, 10067],
       [  581,   70,  5820],
       [  685,   10, 29015]])
```

```
print(classification_report(ytest, prediction))
```

```
precision    recall  f1-score   support

           0       0.83      0.38      0.53     16337
           1       0.84      0.01      0.02      6471
           2       0.65      0.98      0.78     29710

    accuracy                           0.67     52518
   macro avg       0.77      0.46      0.44     52518
weighted avg       0.73      0.67      0.61     52518
```

```
from sklearn.metrics import accuracy_score
accuracy_score(ytest, prediction)
```

```
0.6731406374957157
```

```
#Creating a function to calculate sentiments
def calculate_sentiments(ds):
```

```
    sentiments = TextBlob(ds['review_description']).sentiment
                return       pd.Series([sentiments.subjectivity,
   sentiments.polarity])
```

```
#Generate a word cloud from cleaned data
#Applying calculate_sentiment function to df

df[['Subjectivity', 'Polarity']] = df.apply(calculate_sentiments,
   axis = 1)
```

```
df.sample(5)
```

review_description	rating	Subjectivity	Polarity
4618Some bugs in the recent update, hope the insta...	5.0	0.450000	-0.200000
5809I talk as an artist trying to grow on Instagra.. .	1.0	0.620000	-0.361667
5774Instagram have restricted my two accounts with.	1.0	0.900000	0.600000
1031Picture formatting is broken and has been for ...	1.0	0.390455	-0.143030
5018Worked good and dandy untill it just stopped s...	3.0	0.550000	0.350000

```
all_words = " ".join(words for words in df['review_description'])
```

```
wordcloud = WordCloud(width = 1000,
                      height = 1000,
                      random_state = 21,
                      max_font_size= 119).generate(all_words)
plt.figure(figsize = (20,20), dpi= 80)
plt.imshow(wordcloud, interpolation= 'bilinear')
plt.axis('off')
plt.show()
```

4. Conclusion

Here is a concise breakdown of what I accomplished in the provided code:

Imported Essential Libraries: I started by importing crucial Python libraries such as NumPy, Pandas, Matplotlib, Seaborn, NLTK, WordCloud, and scikit-learn.

Loaded Dataset: I loaded a dataset containing Instagram reviews from a CSV file.

Explored the Data: I conducted an initial exploration of the dataset, understanding its dimensions, summary statistics, basic information, as well as the presence of missing values and duplicates.

Data Cleaning: I proceeded to clean the data by removing the 'review_date' column, eliminating duplicated records, and sanitizing the 'review_description' text by eliminating mentions, hashtags, retweets, and hyperlinks.

Performed Sentiment Analysis: I employed the TextBlob library to compute sentiment subjectivity and polarity for each review description.

Generated Word Cloud: I created a word cloud visualization based on the processed review descriptions, offering a visual representation of common terms.

Categorized Sentiments: I defined a function to categorize sentiments (negative, neutral, positive) relying on polarity values. This categorization was then applied to produce a new 'Sentiment Analysis' column.

Visualized Data: I employed various visualizations, including KDE plots and histograms, to showcase the distribution of polarity and the count of different sentiment categories.

Implemented Machine Learning: I transformed sentiment labels into numerical values, prepped the data for model training, and constructed a pipeline incorporating vectorization, TF-IDF transformation, and a Multinomial Naive Bayes classifier.

Trained and Evaluated Model: I trained the pipeline on a designated training set, utilized it to make predictions on a separate test set, and gauged the model's performance via metrics like a confusion matrix, classification report, and accuracy score.

In summary, I conducted a comprehensive sentiment analysis of Instagram reviews, undertook data cleaning and preparation, visualized sentiment distributions, and crafted a machine learning model for sentiment classification.