



Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering

Name : Dream Patel
Roll No. : 33
Experiment No. 3
To perform data cleaning on social media data using python. Hardware/Software Requirement: Windows Operating System
Date of Performance: 14/02/2024
Date of Submission: 24/02/2024



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Aim : To perform data cleaning on social media data using python.

Objective : To analyze data cleaning on social media data using Python is to prepare the data for meaningful analysis, ensure data integrity and quality, and facilitate efficient and ethical use of the data for generating actionable insights.

Theory :

Data cleaning is a crucial step in the data preprocessing pipeline. It involves identifying and correcting errors, inconsistencies, and inaccuracies in the data to improve its quality and reliability. In the context of social media data, which is often unstructured and noisy, data cleaning becomes even more essential.

- **Ensure Data Quality:** The primary objective of data cleaning is to ensure that the data is accurate, consistent, and reliable. Social media data can contain various types of errors such as misspellings, grammatical mistakes, and inconsistencies that need to be addressed.
- **Handle Missing Values:** Social media data often contains missing values due to incomplete user inputs or data collection processes. Data cleaning involves identifying and handling these missing values appropriately, either by imputation or removal.
- **Remove Duplicates:** Social media data may contain duplicate entries, such as duplicate posts or comments. Removing duplicates ensures that each piece of information is unique and prevents redundancy in the dataset.
- **Standardize Formats:** Social media data can have diverse formats for representing dates, times, and other structured information. Data cleaning involves standardizing these formats to facilitate analysis and comparison across different data points.
- **Text Cleaning and Preprocessing:** Since social media data often consists of text data, cleaning and preprocessing text is essential. This may include removing special characters, URLs, hashtags, mentions, and other noise, as well as tokenization, lemmatization, and removing stopwords to prepare the text for analysis.
- **Ensure Consistency and Uniformity:** Data cleaning ensures that the data is consistent and uniform across different attributes and records. This consistency is crucial for accurate analysis and modeling.
- **Enhance Analytical Results:** Clean data leads to more accurate and reliable analytical results. By removing errors and inconsistencies, data cleaning improves the quality of insights derived from social media data analysis.



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

- **Compliance and Ethical Considerations:** Data cleaning may also involve ensuring compliance with regulations such as GDPR (General Data Protection Regulation) and addressing ethical considerations such as privacy concerns when dealing with sensitive user data in social media datasets.

Handle Missing Values: Check for missing values and decide how to handle them. Options include dropping rows with missing values, filling them with a default value, or using more sophisticated methods like interpolation.

```
# Drop rows with missing values
data.dropna(inplace=True)
# Fill missing values with a default value
data.fillna(0, inplace=True)
```

Remove Duplicates: Remove any duplicate rows in the dataset

```
data.drop_duplicates(inplace=True)
```

Text Cleaning: Preprocess text data by removing special characters, URLs, hashtags, mentions, and performing other text cleaning tasks.

```
def clean_text(text):
    # Remove URLs
    text = re.sub(r'http\S+', '', text)
    # Remove special characters and punctuation
    text = re.sub(r'^\w\s]', '', text)
    # Remove numbers
    text = re.sub(r'\d+', '', text)
    # Convert to lowercase
    text = text.lower()
    return text
data['clean_text'] = data['text'].apply(clean_text)
```

Tokenization and Lemmatization/Stemming: Tokenize the text and perform lemmatization or stemming to standardize words.

```
#from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
def tokenize_and_lemmatize(text):
    tokens = word_tokenize(text)
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
    return lemmatized_tokens
data['tokenized_text'] = data['clean_text'].apply(tokenize_and_lemmatize)
```

Program:

```
import pandas as pd
import numpy as np
```

Reading Dataset

```
data=pd.read_csv('/content/Iphone 15 Pro Max- Buy Products Online at Best Price in India - All Categories _ Flipkart.com.csv')
data.head(10)
```



	Name	Price	Ratings	Reviews
0	NaN	NaN	NaN	NaN
1	Apple iPhone 15 Pro Max (Natural Titanium, 256...	₹1,48,900	492 Ratings	45 Reviews
2	Apple iPhone 15 Pro Max (Blue Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
3	Apple iPhone 15 Pro Max (White Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
4	Apple iPhone 15 Pro Max (Black Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
5	Apple iPhone 13 Pro Max (Silver, 1 TB)	₹1,79,900	2,060 Ratings	198 Reviews
6	Apple iPhone 12 Pro Max (Gold, 256 GB)	₹1,29,900	1,270 Ratings	102 Reviews
7	Apple iPhone 12 Pro Max (Pacific Blue, 128 GB)	₹1,19,900	1,270 Ratings	102 Reviews
8	Apple iPhone 15 Pro Max (Blue Titanium, 1 TB)	₹1,88,900	492 Ratings	45 Reviews
9	Apple iPhone 11 Pro Max (Midnight Green, 64 GB)	₹1,09,900	1,101 Ratings	99 Reviews



Next steps:

[Generate code with data](#)

 [View recommended plots](#)


Duplicate Value Removal

```
duplicate_rows=data[data.duplicated()]
num_duplicates=len(duplicate_rows)
print("Number of Duplicate Rows: -" + str(num_duplicates))
```



Number of Duplicate Rows: -5

```
df_without_duplicate=data.drop_duplicates()
df_without_duplicate.to_csv('NewFile.csv',index=False)
```

```
data1=pd.read_csv("/content/NewFile.csv")
data1.head(10)
```



	Name	Price	Ratings	Reviews
0	NaN	NaN	NaN	NaN
1	Apple iPhone 15 Pro Max (Natural Titanium, 256...	₹1,48,900	492 Ratings	45 Reviews
2	Apple iPhone 15 Pro Max (Blue Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
3	Apple iPhone 15 Pro Max (White Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
4	Apple iPhone 15 Pro Max (Black Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews
5	Apple iPhone 13 Pro Max (Silver, 1 TB)	₹1,79,900	2,060 Ratings	198 Reviews
6	Apple iPhone 12 Pro Max (Gold, 256 GB)	₹1,29,900	1,270 Ratings	102 Reviews



Next steps:

[Generate code with data1](#)

 [View recommended plots](#)

```
duplicate_rows=data1[data1.duplicated()]
num_duplicates=len(duplicate_rows)
print("Number of Duplicate Rows: -" + str(num_duplicates))
```

Number of Duplicate Rows: -0

Removal of NaN Values

```
data1.dropna(inplace=True)
```

```
data1.head(10)
```

	Name	Price	Ratings	Reviews	
1	Apple iPhone 15 Pro Max (Natural Titanium, 256...	₹1,48,900	492 Ratings	45 Reviews	
2	Apple iPhone 15 Pro Max (Blue Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews	
3	Apple iPhone 15 Pro Max (White Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews	
4	Apple iPhone 15 Pro Max (Black Titanium, 256 GB)	₹1,48,900	492 Ratings	45 Reviews	
5	Apple iPhone 13 Pro Max (Silver, 1 TB)	₹1,79,900	2,060 Ratings	198 Reviews	
6	Apple iPhone 12 Pro Max (Gold, 256 GB)	₹1,29,900	1,270 Ratings	102 Reviews	
7	Apple iPhone 12 Pro Max (Pacific Blue, 128 GB)	₹1,19,900	1,270 Ratings	102 Reviews	

Next steps:

Generate code with data1

☒ View recommended plots

Text Cleaning

```
data1['Ratings']=data1['Ratings'].str.replace(' Ratings', '')
data1.head(10)
```

	Name	Price	Ratings	Reviews	
1	Apple iPhone 15 Pro Max (Natural Titanium, 256...	₹1,48,900	492	45 Reviews	
2	Apple iPhone 15 Pro Max (Blue Titanium, 256 GB)	₹1,48,900	492	45 Reviews	
3	Apple iPhone 15 Pro Max (White Titanium, 256 GB)	₹1,48,900	492	45 Reviews	
4	Apple iPhone 15 Pro Max (Black Titanium, 256 GB)	₹1,48,900	492	45 Reviews	
5	Apple iPhone 13 Pro Max (Silver, 1 TB)	₹1,79,900	2,060	198 Reviews	
6	Apple iPhone 12 Pro Max (Gold, 256 GB)	₹1,29,900	1,270	102 Reviews	
7	Apple iPhone 12 Pro Max (Pacific Blue, 128 GB)	₹1,19,900	1,270	102 Reviews	
8	Apple iPhone 15 Pro Max (Blue Titanium, 1 TB)	₹1,88,900	492	45 Reviews	
9	Apple iPhone 11 Pro Max (Midnight Green, 64 GB)	₹1,09,900	1,101	99 Reviews	
10	Apple iPhone 15 Pro Max (White Titanium, 512 GB)	₹1,68,900	492	45 Reviews	

Next steps:

Generate code with data1

☒ View recommended plots

```
data1['Reviews']=data1['Reviews'].str.replace('Reviews', '')
data1.head(10)
```

	Name	Price	Ratings	Reviews	
1	Apple iPhone 15 Pro Max (Natural Titanium, 256...	₹1,48,900	492	45	
2	Apple iPhone 15 Pro Max (Blue Titanium, 256 GB)	₹1,48,900	492	45	
3	Apple iPhone 15 Pro Max (White Titanium, 256 GB)	₹1,48,900	492	45	
4	Apple iPhone 15 Pro Max (Black Titanium, 256 GB)	₹1,48,900	492	45	
5	Apple iPhone 13 Pro Max (Silver, 1 TB)	₹1,79,900	2,060	198	
6	Apple iPhone 12 Pro Max (Gold, 256 GB)	₹1,29,900	1,270	102	
7	Apple iPhone 12 Pro Max (Pacific Blue, 128 GB)	₹1,19,900	1,270	102	
8	Apple iPhone 15 Pro Max (Blue Titanium, 1 TB)	₹1,88,900	492	45	
9	Apple iPhone 11 Pro Max (Midnight Green, 64 GB)	₹1,09,900	1,101	99	
10	Apple iPhone 15 Pro Max (White Titanium, 512 GB)	₹1,68,900	492	45	

Next steps:

Generate code with data1

☒ View recommended plots

Start coding or [generate](#) with AI.



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Conclusion:

Python was employed to preprocess Flipkart's social media data on iPhone models. Duplicate entries were eliminated to maintain data integrity, and extraneous text such as "Ratings" and "Reviews" was filtered out. This efficient data cleaning standardized the dataset for subsequent analysis, underscoring Python's vital role in preprocessing social media data for reliability and analytical readiness. Future exploration might involve advanced cleaning techniques and addressing missing data to elevate the quality of social media data analysis.