9

```
# Used the UDF function to clean the tweets

clean_tweet_udf = f.udf(clean_tweet, t.StringType())
df = df.withColumn('cleaned_text', clean_tweet_udf(f.col('text')))
```

10

```
# remove rows where 'cleaned_text' is null and filter out rows where 'cleaned_text' is empty

df = df.dropna(subset=["cleaned_text"]).filter(f.col("cleaned_text") != "")
```

11

```
# Extracted the date from the 'timestamp' column, created a new 'date' column and converted it to the date format.

df = df.withColumn('date', f.to_date(f.substring('timestamp',1,10), 'yyyy-MM-dd'))
```

SENTIMENT ANALYSIS

13

```
# Initialize SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()
```

Calculate the compound score and classification of sentiment, if it is greater than 0 then it is positive, less than 0 it is negative, otherwise it is neutral

15

```
def sentiment_analysis(cleaned_text):
```