8144-SUDHARSAN ENGINEERING COLLEGE

SUDHARSHAN EDUCATIONAL TRUST

Creating Pathways to Wisdom

SUDHARSAN
ENGINEERING COLLEGE

REGISTER NUMBER: 814421243024

NAME: ROHINI M

DEGREE: BTECH

BRANCH: ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

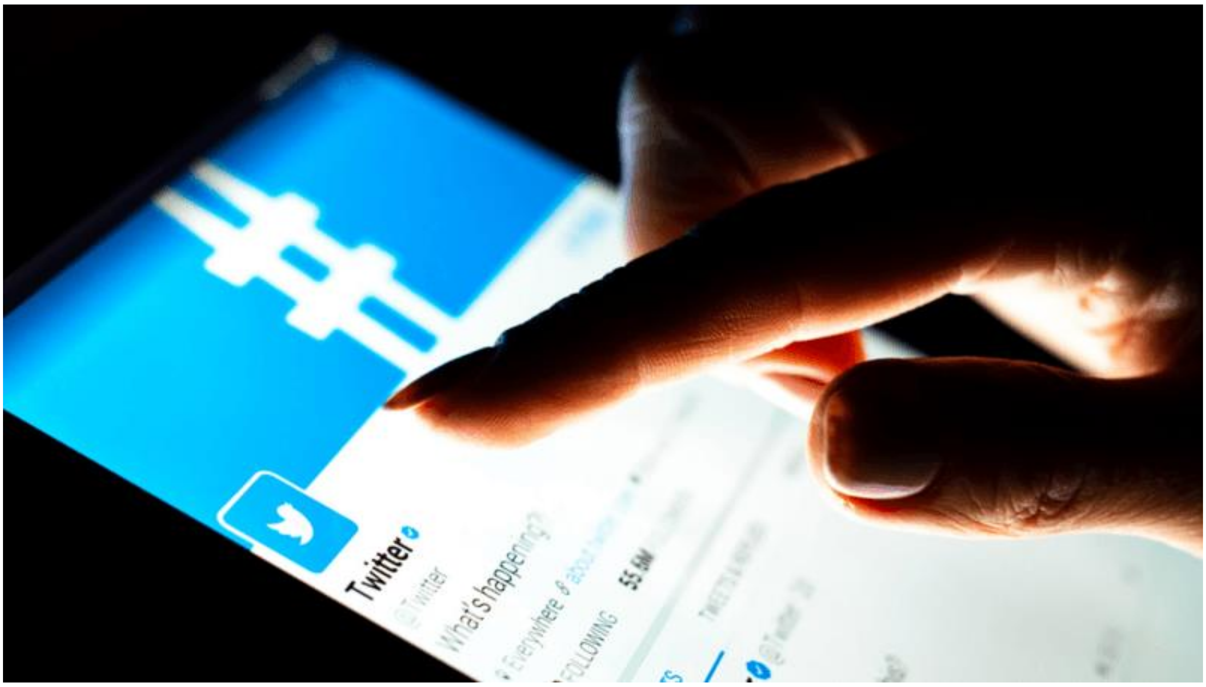
PROJECT TITLE: SENTIMENT ANALYSIS FOR MARKETING

SENTIMENT ANALYSIS FOR MARKETING USING PYTHON

PHASE 5 SUBMISSION DOCUMENT

Phase 5: **Project Documentation & Submission**

Topic: **In this section we will document the complete project and prepare it for submission.**



INTRODUCTION:

- In today's fast-paced and hyper-competitive business landscape, understanding customer sentiment is paramount for the success of marketing strategies.

- Sentiment analysis, a subfield of natural language processing (NLP), has emerged as a powerful tool for businesses to gain valuable insights into customer opinions, emotions, and attitudes.

- This project, "Sentiment Analysis for Marketing," aims to harness the potential of sentiment analysis to revolutionize the way companies approach marketing campaigns and customer engagement.

- In the digital age, customers share their opinions and experiences on a multitude of platforms, including social media, review websites, and online forums.

- These user-generated content pieces, which are often rich in sentiment, provide a goldmine of information for marketers.

- Extracting and understanding this information can help companies tailor their marketing efforts, improve customer satisfaction, and drive business growth.
- The primary goal of this project is to develop a comprehensive sentiment analysis system that empowers marketing teams to:
  - Gain deeper insights into customer sentiment and emotions.
  - Identify trends and patterns in customer feedback.
  - Customize marketing campaigns based on audience sentiment.
  - Evaluate the success of marketing strategies through sentiment metrics.

- This project will employ state-of-the-art NLP techniques and machine learning algorithms to analyze and classify textual data from various sources, including customer reviews, social media posts, and surveys.

- Natural language processing tools will be utilized to preprocess and tokenize the data. Machine learning models such as deep neural networks and support vector machines will then be trained to categorize sentiments as positive, negative, or neutral.

- To ensure the accuracy and effectiveness of sentiment analysis, we will collect data from diverse sources, including but not limited to:
  - Social media platforms (e.g., Twitter, Facebook, Instagram).
  - E-commerce websites (e.g., Amazon, eBay).
  - Customer feedback surveys.
  - Product review sites (e.g., Yelp, TripAdvisor).

Dataset Link: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

LIST of tools and software used in the process of sentiment analysis for marketing:

1.Data Collection:

- Social Media APIs: Platforms like Twitter, Facebook, and Instagram provide APIs for collecting data.
- Web Scraping Tools: Python libraries like BeautifulSoup and Scrapy for extracting data from websites.
- Survey and Feedback Tools: Tools like SurveyMonkey and Google Forms for collecting customer feedback.

2.Data Preprocessing:

- Natural Language Processing (NLP) Libraries: Python libraries such as NLTK (Natural Language Toolkit) and spaCy for text processing.
- Text Cleaning Tools: Regex for text cleaning and removal of special characters, stopwords, and irrelevant information.
- Tokenization Tools: Tools to split text into words or phrases, such as NLTK's tokenizer.

3.Sentiment Analysis Models:

- Machine Learning Libraries: Scikit-learn, TensorFlow, PyTorch for building and training sentiment analysis models.
- Pretrained Models: Models like BERT, GPT-3, VADER, and TextBlob, which are pretrained for sentiment analysis tasks.
- Sentiment Analysis APIs: Commercial APIs like IBM Watson, Google Cloud NLP, and Amazon Comprehend for prebuilt sentiment analysis.

4.Data Analysis:

- Statistical Analysis Tools: R or Python with libraries like Pandas and NumPy for statistical analysis.
- Data Visualization Tools: Matplotlib, Seaborn, Plotly, or Tableau for creating visualizations.

5.Dashboard and Reporting:

- Business Intelligence Tools: Tools like Power BI, Tableau, or Google Data Studio for creating interactive dashboards and reports.
- Custom Dashboard Development: Creating custom dashboards using web development technologies like HTML, CSS, and JavaScript.

6.Text Annotation Tools:

Labeling and Annotation Tools: Prodigy, Labelbox, and Amazon SageMaker Ground Truth for labeling data for supervised sentiment analysis.

7.Version Control and Collaboration:

- Version Control Systems:

  Git and platforms like GitHub for collaborative development and version control.
- Project Management Tools:

  Tools like Jira or Trello for project management and task tracking.

8.Cloud Services:

- Cloud Computing Platforms: AWS, Google Cloud, Microsoft Azure for scalable computing resources.
- Serverless Computing: AWS Lambda, Azure Functions for serverless data processing.

9.Database and Storage:

- Relational Databases: MySQL, PostgreSQL for structured data storage.
- NoSQL Databases: MongoDB, Cassandra for unstructured or semi-structured data storage.
- Data Warehouses: Redshift, BigQuery for analytical data storage.

10.Deployment:

- Web Application Frameworks: Django, Flask for deploying sentiment analysis applications.
- Containerization: Docker for packaging applications and models into containers.
- Serverless Deployment: AWS Lambda, Azure Functions for serverless model deployment.

11.Security and Compliance:

Data Encryption: Tools for encrypting sensitive data in transit and at rest & Compliance Tools: Tools and practices for ensuring data privacy and GDPR compliance.

Design thinking and present in form of document :

1.Empathize:

- Identify key stakeholders: Marketing teams, data analysts, decision-makers.
- Gather user stories and pain points related to sentiment analysis.
- Define the problem: Lack of real-time sentiment insights affecting marketing strategy effectiveness.

2.Define:

- Problem Statement: Develop a sentiment analysis system to understand customer sentiment and improve marketing strategies.
- Goals:
  - Real-time sentiment monitoring.
  - Enhanced customer engagement.
  - Customized marketing campaigns.
  - Improved brand reputation management.

3.Ideate:

- Brainstorm potential features and functionalities for the sentiment analysis system.
- Explore different data sources (social media, reviews, surveys).
- Consider the integration of machine learning models for sentiment classification.
- Think about user-friendly visualization and reporting options.

4.Prototype:

- Develop a prototype system with a user-friendly interface.
- Include features for data collection, preprocessing, sentiment analysis, and reporting.
- Choose NLP and machine learning tools for sentiment analysis.
- Design a simple dashboard for real-time monitoring and reporting.

5.Test:

- Share the prototype with key stakeholders for feedback.
- Conduct user testing to evaluate the ease of use.
- Ensure that the system meets user expectations.
- Gather feedback on potential improvements.

6.Implement:

- Select appropriate tools and technologies.
- Build the sentiment analysis system, integrating data sources and NLP models.
- Create a database for data storage.
- Develop real-time data processing capabilities.

7.Test:

- Conduct extensive testing, including functional, performance, and security testing.
- Verify the accuracy of sentiment analysis results.
- Ensure data privacy and compliance with regulations.
- Address any issues or bugs found during testing.

8.Deliver:

- Deploy the sentiment analysis system on a suitable platform or server.
- Train marketing teams on how to use the system.
- Provide documentation for system maintenance.
- Ensure scalability for future growth.

9.Iterate:

- Continuously monitor system performance and gather user feedback.
- Make regular updates to improve accuracy and add new features.
- Adapt to changes in customer behavior and market dynamics.
- Keep up with advancements in NLP and AI for sentiment analysis.

Design into innovation:

1.User-Centered Design:

- - Start with understanding user needs and pain points.

- Conduct user research and create detailed personas.

- Identify unique challenges faced by marketing teams in understanding customer sentiment.

2.Design Thinking for Innovation:

__- Employ the design thinking process to identify problems and generate creative solutions.

- Conduct ideation workshops involving cross-functional teams, including designers, data scientists, and marketers.

- Focus on innovation in data collection, preprocessing, and analysis methods.

3.User Interface (UI) and User Experience (UX) Design:

- Craft a user-friendly interface for sentiment analysis.

- Create intuitive data visualization and reporting tools.

- Ensure responsive design for mobile and web applications.

4.Embracing Advanced Technologies:

- Leverage machine learning and AI for sentiment analysis.

- Explore deep learning models like BERT for more accurate sentiment classification.

- Implement real-time data processing and analysis.

5.Data Sources and Integratio:

- Explore diverse data sources, including social media, customer reviews, and surveys.

- Integrate APIs for seamless data collection.

- Develop connectors for various platforms.

6.Scalability and Performance:

- Design the system to scale with the increasing data volume.

- Consider cloud-based solutions for scalability.

- Ensure optimal performance for real-time sentiment analysis.

7. Customization and Personalization:

- Allow users to customize sentiment analysis based on specific industry or product-related terms.

- Implement personalization features for individualized marketing insights

8.Ethical Considerations and Compliance:

- Build in ethical AI principles to ensure privacy and fairness.
- Address compliance requirements, including GDPR and data protection regulations.

9. Innovation Metrics:

- Define innovation KPIs related to sentiment analysis.
- Measure the impact of sentiment analysis on marketing campaigns and customer engagement.

10.Continuous Learning and Adaptation:

- Encourage a culture of continuous learning and adaptation.
- Stay updated on the latest developments in NLP and AI.
- Regularly gather feedback from users for improvements.

11. Collaboration and Cross-Functional Teams:

- Foster collaboration between marketing, data science, and design teams.
- Promote open communication and knowledge sharing.

12. Market Testing and Feedback Loops:

- Launch prototypes and minimum viable products (MVPs) for market testing.
- Create feedback loops to rapidly iterate and improve the solution.

Build loading and Preprocessing the dataset:

Step 1: Import Libraries

```
import pandas as pd

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

from sklearn.model_selection import train_test_split
```

Step 2: Load the Dataset

```
# Replace 'your_dataset.csv' with the actual file path

df = pd.read_csv('your_dataset.csv')
```

Step 3: Data Inspection

```
# Display the first few rows of the dataset

print(df.head())
```

Step 4: Text Preprocessing

```
# Lowercase the text
df['text'] = df['text'].str.lower()

# Tokenize the text
df['text'] = df['text'].apply(word_tokenize)

# Remove stopwords and punctuation
stop_words = set(stopwords.words('english'))
df['text'] = df['text'].apply(lambda x: [word for word in x if word.isalnum()
and word not in stop_words])
```

Step 5: Label Encoding (if not already done)

```
X = df['text']  # Features (text data)
y = df['sentiment']  # Labels (sentiment)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Performing different activities like Feature engineering,Model Training,Evaluation for sentiment analysis for marketing:

Step 1: Feature Engineering:

- Feature engineering is the process of transforming the text data into numerical features that machine learning models can use.
- Common techniques include TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings.

Step 2: Model Training

- Now that you have transformed the text data into numerical features, you can train a sentiment analysis model.

Step 3: Evaluation

- Evaluate the model's performance on the test data using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score
- In addition to accuracy and the classification report, you can also consider other evaluation metrics like confusion matrix, ROC-AUC, and ROC curves if your sentiment analysis task involves multiple classes.

Step 4: Fine-Tuning and Advanced Models

- Depending on the results and requirements, you may want to fine-tune the model, try different algorithms, or explore more advanced techniques such as deep learning models (e.g., LSTM or BERT) for sentiment analysis.

- This might involve hyperparameter tuning, cross-validation, and larger datasets for better performance.

- Remember that the choice of model and feature engineering techniques should be based on the specific characteristics of your dataset and the sentiment analysis goals of your marketing project.

Feature selection for sentiment analysis for marketing:

- Feature selection in sentiment analysis for marketing is a crucial step in optimizing your model's performance and reducing dimensionality.
- It involves choosing the most relevant features or attributes from your data to improve model accuracy, reduce overfitting, and enhance interpretability.
- Here are some techniques and considerations for feature selection in sentiment analysis:

1.Unigrams and Bigrams

2.TF-IDF (Term Frequency-Inverse Document Frequency)

3.Feature Importance from Models

4.Sentiment Lexicons

5.Part-of-Speech Tags

6.Word Embeddings

7.Named Entity Recognition (NER)

8.Topic Modeling

9.Feature Selection Algorithms

10.Dimensionality Reduction Techniques

11.Cross-Validation

12.Domain Knowledge

13.Text-Based Features

- It's essential to experiment with different feature selection techniques and evaluate their impact on the model's performance.
- The choice of features may vary depending on the specific characteristics of your dataset and the objectives of your sentiment analysis project in marketing.
- Regularly reassess and fine-tune your feature selection approach to ensure the best results.

Advantages:

Customer Insights:

It provides valuable insights into customer opinions, emotions, and attitudes towards products and services.

Real-time Monitoring:

Allows businesses to monitor sentiment in real-time, enabling timely responses to customer feedback.

Customized Marketing:

Helps in tailoring marketing campaigns and content to match customer sentiment and preferences.

Competitive Analysis:

Enables benchmarking against competitors and identifying market trends.

Brand Reputation Management:

Supports proactive reputation management and damage control by identifying negative sentiment early.

Product Improvement:

Identifies areas for product or service improvement based on customer feedback.

Efficient Resource Allocation:

Helps in optimizing marketing budgets and resources by focusing on areas with the most significant sentiment impact.

Measurable Results:

Provides quantifiable data for assessing the success of marketing strategies.

Disadvantages:

Inaccuracy:

Sentiment analysis may not always accurately interpret context, sarcasm, or cultural nuances.

Overreliance on Automated Tools:

Relying solely on automated sentiment analysis tools can lead to incorrect assessments.

Human Bias:

Human bias may be present in the creation of sentiment analysis tools or in the interpretation of results.

Language Variability:

Variations in language and dialects can be challenging for sentiment analysis models.

Data Privacy:

Handling customer data for sentiment analysis raises privacy concerns and must comply with data protection regulations.

Cost and Resources:

Developing and maintaining sentiment analysis tools can be resource-intensive.

Complex Sentiments:

Some opinions may contain mixed or complex sentiments that are challenging to categorize.

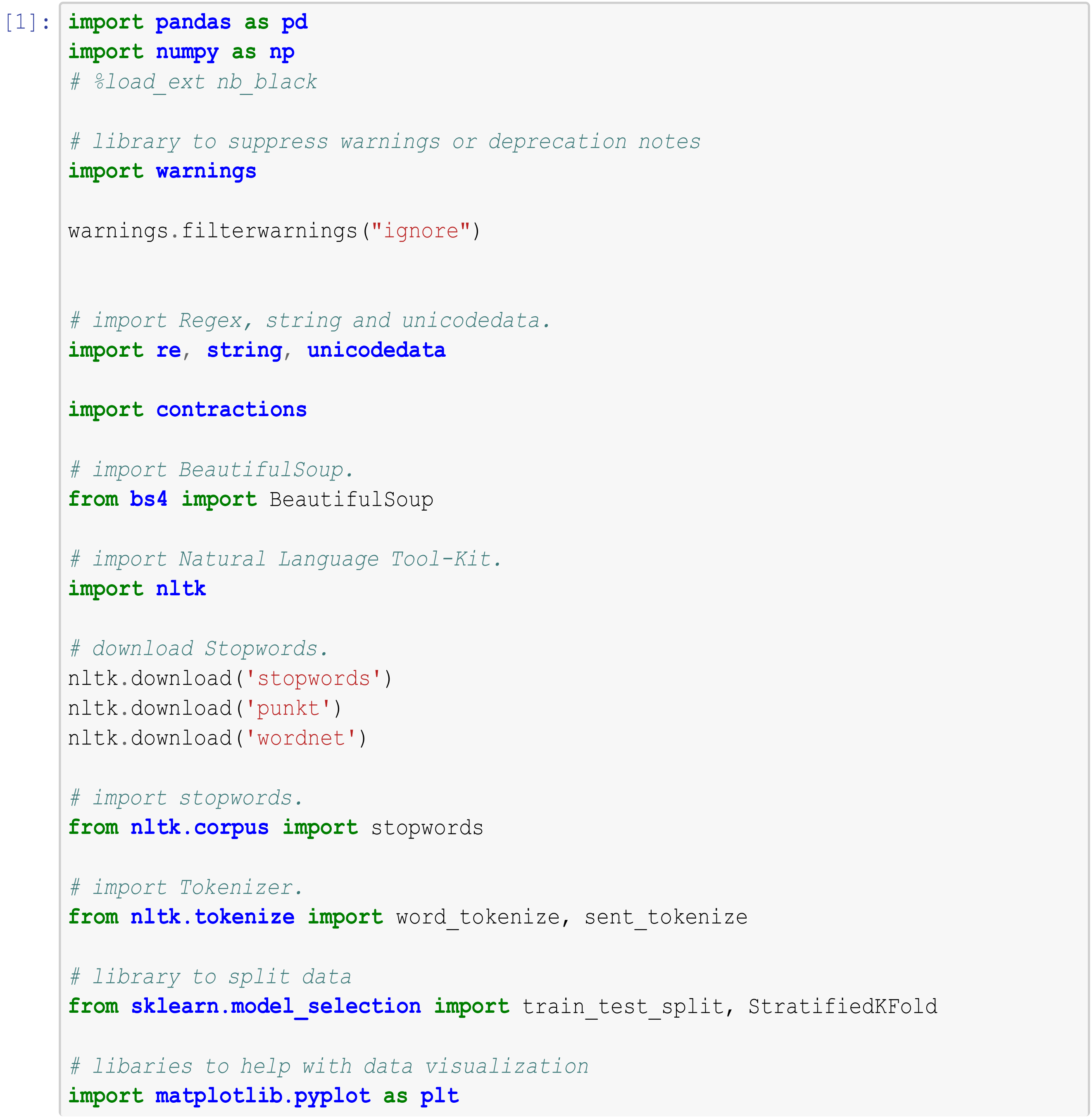
Changing Trends:

Sentiments can change rapidly, making it challenging to keep up with evolving customer attitudes.

Benefits of using Sentiment analysis for marketing:

- Customer Insights
- Real-time Monitoring
- Customized Marketing Campaigns
- Competitive Analysis
- Brand Reputation Management
- Product and Service Improvement
- Efficient Resource Allocation
- Measurable Results
- Targeted Customer Engagement
- Crisis Management
- Product Development
- Content Strategy
- Identifying Influencers
- Data-Driven Decision Making
- Trend Detection
- Improved Customer Experience
- Optimized Ad Targeting

PROGRAM:



[1]:

```
import pandas as pd
import numpy as np
# %load_ext nb_black

# library to suppress warnings or deprecation notes
import warnings

warnings.filterwarnings("ignore")


# import Regex, string and unicodedata.
import re, string, unicodedata

import contractions

# import BeautifulSoup.
from bs4 import BeautifulSoup

# import Natural Language Tool-Kit.
import nltk

# download Stopwords.
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')

# import stopwords.
from nltk.corpus import stopwords

# import Tokenizer.
from nltk.tokenize import word_tokenize, sent_tokenize

# library to split data
from sklearn.model_selection import train_test_split, StratifiedKFold

# libaries to help with data visualization
import matplotlib.pyplot as plt
```

```python
import seaborn as sns
import missingno as msno

# import wordcloud
import wordcloud
from wordcloud import STOPWORDS
from wordcloud import WordCloud

# remove the limit for the number of displayed columns
pd.set_option("display.max_columns", None)

# set the limit for the number of displayed rows
pd.set_option("display.max_rows", 200)

# to get diferent metric scores
from sklearn.metrics import (
    recall_score,
    accuracy_score,
    confusion_matrix,classification_report,
    f1_score,
    precision_score,
    precision_recall_fscore_support
)

# import vectorizers
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

# import rfc and cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score

# import word prepocessors
from nltk.tokenize import word_tokenize
from nltk.stem import LancasterStemmer, WordNetLemmatizer
```

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\Administrator\AppData\Roaming\nltk_data…
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\Administrator\AppData\Roaming\nltk_data…
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data] C:\Users\Administrator\AppData\Roaming\nltk_data…
[nltk_data] Package wordnet is already up-to-date!
```

[2]:
```python
df = pd.read_csv('Tweets.csv')
df.head()
```

```
[2]:          tweet_id airline_sentiment                         \
         airline_sentiment_confidence
    0  570306133677760513  neutral    1.0000
    1  570301130888122368  positive   0.3486
    2  570301083672813571  neutral    0.6837
    3  570301031407624196  negative   1.0000
    4  570300817074462722  negative   1.0000

     negativereasonnegativereason_confidence     airline \
    0      NaN   NaN   Virgin America
    1      NaN   0.0000     Virgin America 2 NaN   NaN   Virgin
    America
    3     Bad Flight  0.7033     Virgin America
    4     Can't Tell  1.0000     Virgin America
     airline_sentiment_gold    name negativereason_gold            \
        retweet_count
    0                 NaN    cairdin              NaN            0
    1                 NaN   jnardino              NaN            0

    2                 NaN yvonnalynn              NaN            0

    3                 NaN   jnardino              NaN            0

    4                 NaN   jnardino              NaN            0

                                              text tweet_coord \
    0           @VirginAmerica What @dhepburn said.     NaN

    1  @VirginAmerica plus you've added commercials t…  NaN

    2  @VirginAmerica I didn't today… Must mean I n…  NaN

    3  @VirginAmerica it's really aggressive to blast…  NaN

    4  @VirginAmerica and it's a really big bad thing…  NaN

                tweet_created tweet_location         user_timezone
    0  2015-02-24 11:35:52 -0800       NaN      Eastern Time (US &
       Canada)
    1  2015-02-24 11:15:59 -0800       NaN      Pacific Time (US &
       Canada)
    2  2015-02-24 11:15:48 -0800  Lets Play      Central Time (US &
       Canada)
    3  2015-02-24 11:15:36 -0800       NaN      Pacific Time (US &
       Canada)
    4  2015-02-24 11:14:45 -0800       NaN      Pacific Time (US &
       Canada)
```

```
[3]: texts = [[word.lower() for word in text.split()] for text in df]
     df.head()
```

```
[3]:           tweet_id airline_sentimentairline_sentiment_confidence  \
    0  570306133677760513       neutral                        1.0000
    1  570301130888122368      positive                        0.3486
    2  570301083672813571       neutral                        0.6837
    3  570301031407624196      negative                        1.0000
    4  570300817074462722      negative                        1.0000
     negativereasonnegativereason_confidence     airline \
    0          NaN                        NaN Virgin America
    1     NaN   0.0000     Virgin America 2 NaN   NaN
          Virgin America
    3     Bad Flight  0.7033     Virgin America
    4     Can't Tell  1.0000     Virgin America

                                                                  \
    airline_sentiment_gold    name negativereason_goldretweet_count
    0                 NaN    cairdin              NaN            0
    1                 NaN   jnardino              NaN            0

    2                 NaN yvonnalynn              NaN            0

    3                 NaN   jnardino              NaN            0

    4                 NaN   jnardino              NaN            0


                                             text tweet_coord \
    0           @VirginAmerica What @dhepburn said.     NaN

    1  @VirginAmerica plus you've added commercials t…  NaN

    2  @VirginAmerica I didn't today… Must mean I n…  NaN

    3  @VirginAmerica it's really aggressive to blast…  NaN

    4  @VirginAmerica and it's a really big bad thing…  NaN

                tweet_created tweet_location          user_timezone
    0  2015-02-24 11:35:52 -0800       NaN    Eastern Time (US &
       Canada)
    1  2015-02-24 11:15:59 -0800       NaN    Pacific Time (US &
       Canada)
    2  2015-02-24 11:15:48 -0800  Lets Play   Central Time (US &
       Canada)
    3  2015-02-24 11:15:36 -0800       NaN    Pacific Time (US &
       Canada)
    4  2015-02-24 11:14:45 -0800       NaN    Pacific Time (US &
       Canada)
```

```
[4]: df.info()
```

```
<class
'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to
```

```
14639 Data columns (total 15
columns):
 #   Column                      Non-Null    Dtype
                                 Count
---  ------                      ----------- -----
                                 ---
 0   tweet_id                    14640 non-  int64
                                 null
 1   airline_sentiment           14640 non-  object
                                 null
 2   airline_sentiment_confidence14640 non-  float64
                                 null
 3   negativereason              9178 non-   object
                                 null
 4   negativereason_confidence   10522 non-  float64
                                 null
 5   airline                     14640 non-  object
                                 null
 6   airline_sentiment_gold      40 non-null object
 7   name                        14640 non-  object
                                 null
 8   negativereason_gold         32 non-null object
 9   retweet_count               14640 non-  int64
                                 null
 10  text                        14640 non-  object
                                 null
 11  tweet_coord                 1019 non-   object
                                 null
 12  tweet_created               14640 non-  object
                                 null
 13  tweet_location              9907 non-   object
                                 null
 14 user_timezone              9820 non-null  object
dtypes: float64(2), int64(2),
object(11) memory usage: 1.7+ MB
```

```
[5]: df.isnull().sum()
```

```
[5]: tweet_id                        0
    airline_sentiment               0
  airline_sentiment_confidence      0
    negativereason               5462
    negativereason_confidence    4118
    airline                         0
    airline_sentiment_gold      14600
    name                            0
    negativereason_gold         14608
    retweet_count                   0
```

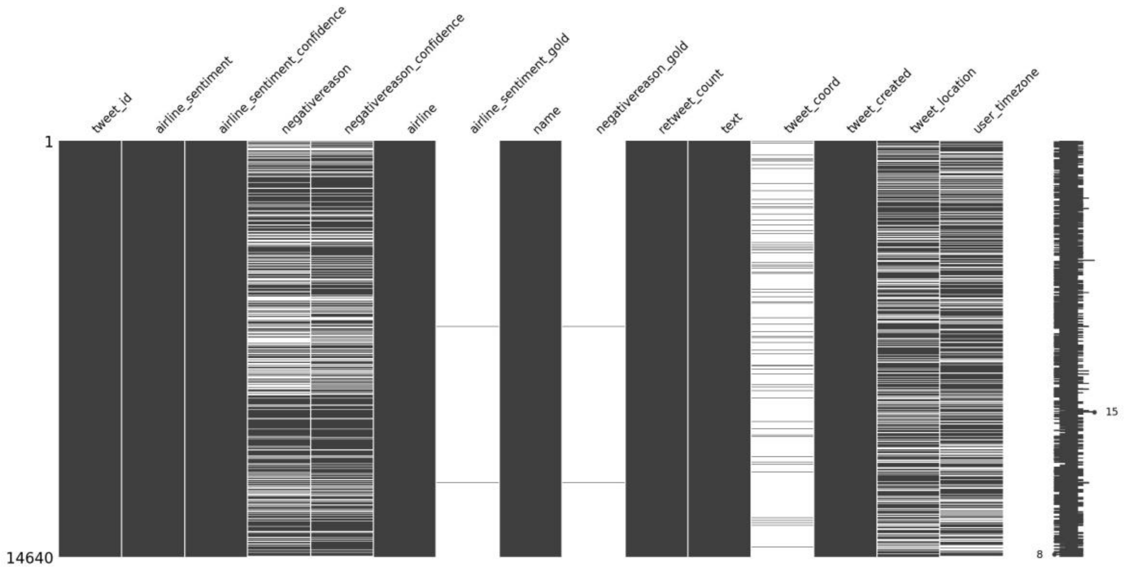
```
text                               0
tweet_coord                    13621
tweet_created                      0
tweet_location                  4733
user_timezone                   4820
dtype: int64
```

```
[6]: df.isnull().sum() / len(df) * 100
```

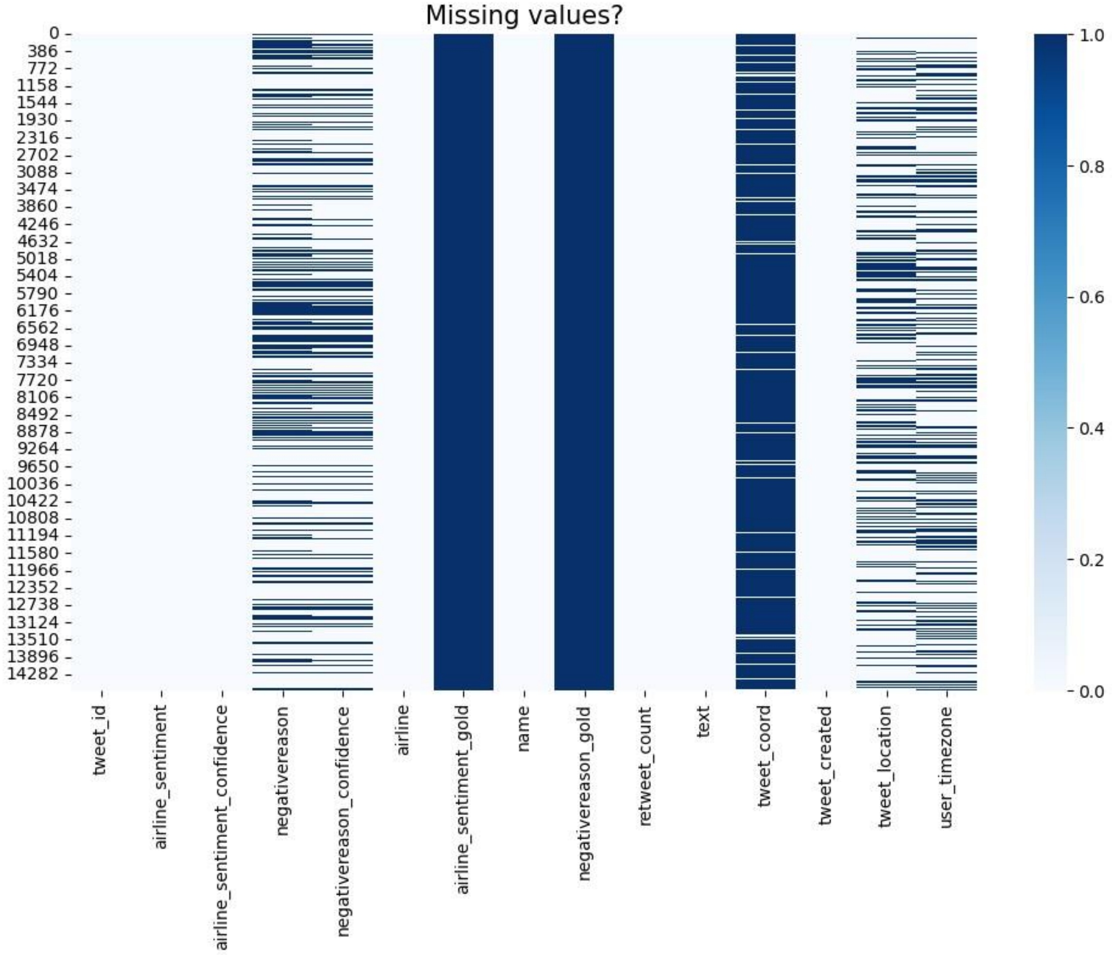
```
[6]: tweet_id                         0.000000
     airline_sentiment                0.000000
   airline_sentiment_confidence   0.000000
     negativereason                  37.308743
     negativereason_confidence       28.128415
     airline                          0.000000
     airline_sentiment_gold          99.726776
     name                             0.000000
     negativereason_gold             99.781421
     retweet_count                    0.000000
     text                             0.000000
     tweet_coord                     93.039617
     tweet_created                    0.000000
     tweet_location                  32.329235
     user_timezone                   32.923497
     dtype: float64
```

```
[7]: msno.matrix(df)
```

```
[7]: <AxesSubplot:>
```



```
[8]: plt.figure(figsize=(12,7))
     sns.heatmap(df.isnull(), cmap = "Blues")                     #Visualization
       of missing value using heatmap
     plt.title("Missing values?", fontsize = 15)
     plt.show()
```



```
[9]: print("Percentage null or na values in df")
                    ((df.isnull() | df.isna()).sum() * 100 /
                                         df.index.size).round(2)
```

```
    Percentage null or na values in df
[9]: tweet_id                        0.00
     airline_sentiment               0.00
   airline_sentiment_confidence    0.00
     negativereason                 37.31
     negativereason_confidence      28.13
     airline                         0.00
```

```
airline_sentiment_gold      99.73
name                         0.00
negativereason_gold         99.78
retweet_count                0.00
text                         0.00
tweet_coord                 93.04

tweet_created                0.00
tweet_location              32.33
user_timezone               32.92
dtype: float64
```

[10]:
```
del df["tweet_coord"]
del df["airline_sentiment_gold"]
del df["negativereason_gold"]
```

[11]:
```
df.head()
```

[11]:
```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0 570306133677760513           neutral                        1.0000
1 570301130888122368          positive                        0.3486
2 570301083672813571           neutral                        0.6837
3 570301031407624196          negative                        1.0000
4 570300817074462722          negative                        1.0000

  negativereasonnegativereason_confidence       airline       name  \
0          NaN                       NaN  Virgin           cairdin
                                          America
1          NaN                    0.0000  Virgin          jnardino
                                          America
2          NaN                       NaN  Virgin        yvonnalynn
                                          America
3     Bad Flight                  0.7033  Virgin          jnardino
                                          America
4     Can't Tell                  1.0000  Virgin          jnardino
                                          America

  retweet_count                                             text   \
0             0           @VirginAmerica What @dhepburn said.
1             0@VirginAmerica plus you've added commercials t…
2             0@VirginAmerica I didn't today… Must mean I n…
3             0@VirginAmerica it's really aggressive to blast…
4             0@VirginAmerica and it's a really big bad thing…
```

```
             tweet_created tweet_location             user_timezone
0   2015-02-24 11:35:52 -0800          NaN Eastern    Time    (US    &
                                           Canada)
1   2015-02-24 11:15:59 -0800          NaN Pacific    Time    (US    &
                                           Canada)
2   2015-02-24 11:15:48 -0800   Lets Play Central    Time    (US    &
                                           Canada)
3   2015-02-24 11:15:36 -0800          NaN Pacific    Time    (US    &
                                           Canada)
4  2015-02-24 11:14:45 -0800            NaN   Pacific Time (US & Canada)
```

[12]:
```
freq = df.groupby("negativereason").size()
```

[13]:
```
# Checking duplicates
df.duplicated().sum()
```

[13]:
```
39
```

[14]:
```
df.drop_duplicates(inplace = True)
df.duplicated().sum()
```

[14]:
```
0
```

[15]:
```
df.sample(n = 10)
```

[15]:
```
               tweet_id airline_sentimentairline_sentiment_confidence
                           \
10589569156425626329089      neutral                      1.0000
6182  568149878095753216      neutral                      0.6545
11336568196165780578304     negative                      1.0000
623   570245555064074240     negative                      1.0000
1186  569902065247322112     negative                      1.0000
2425  569213883371683840     positive                      0.6679
13299569893723091238912     negative                      1.0000
7693  569343003476819969      neutral                      0.6641
5148  569308552671707136     negative                      1.0000
11135568486436355346432     negative                      1.0000

               negativereasonnegativereason_confidence   airline \
10589                    NaN                       NaN  US
                                                        Airways
6182                     NaN                 0.0000      Southwest

11336             Can't Tell                 0.3579     US
                                                        Airways
623   Flight Booking                         0.6740        United
      Problems
1186            Late Flight                  1.0000        United
```

```
2425                  NaN                       NaN      United

13299           longlines                  0.3512     American

7693                  NaN                  0.0000        Delta

5148         Lost Luggage                  1.0000     Southwest

11135          Bad Flight                  1.0000    US
                                                     Airways

             name retweet_count \
10589observepeople           0

6182     Brian_Fox           0

11336   thefisch26           0

623  fatwmnonthemtn          0

1186   LukeXuanLiu           1

2425  PierreSchmit           0

13299elisakathleen           0

7693   dgruber1700           0

5148 scoobydoo9749           0

11135    kristenlc           0

                                            text     \
10589@usairways Does anyone know the hold times
     for…
6182 @SouthwestAir I would but you need to follow
     m…
11336@USAirways Secondary screenings, a piece of
     th…
623@united What's going on with your website? I'm…
1186@united and most frustratingly, all this delay…
2425@united gave me a smile today, with a Zero Awa…
13299@AmericanAir the most stressful morning and st…
7693                            @JetBlue flite454
5148@SouthwestAir 9 hrs in Baltimore, still not go…
11135@USAirways we bought our tickets months ago. H…
              tweet_created    tweet_location  \

10589 2015-02-21 07:27:20 -              NaN
      0800
6182  2015-02-18 12:47:41 -   NH, United
      0800                    States
```

```
11336 2015-02-18 15:51:37 -    Washington, DC
      0800
623   2015-02-24 07:35:09 -        Summit, NJ
      0800
1186  2015-02-23 08:50:15 -               NaN
      0800
2425  2015-02-21 11:15:39 -   Rixensart,
      0800                    Belgium
13299 2015-02-23 08:17:06 -        Boston, MA
      0800
7693  2015-02-21 19:48:44 -               NaN
      0800
5148  2015-02-21 17:31:50 -   Tallahassee, FL
      0800
11135 2015-02-19 11:05:03 -               NaN
      0800
                 user_timezone
10589 Eastern Time (US & Canada)
6182 Eastern Time (US & Canada)
11336Central Time (US & Canada)
623  Central Time (US & Canada)
1186     Atlantic Time (Canada)
2425                  Brussels
13299                      NaN
7693                       NaN
5148            America/Chicago
11135Eastern Time (US & Canada)
```

[16]:
```
df.describe().T
```

[16]:
```
                                  count         mean          std  \
tweet_id                       14601.0 5.692156e+177.782706e+14
airline_sentiment_confidence14601.08.999022e-011.629654e-01
negativereason_confidence      10501.0 6.375749e-013.303735e-01
retweet_count                  14601.0 8.280255e-027.467231e-01
                                        min          25%          50%  \
tweet_id                       5.675883e+175.685581e+175.694720e+17
airline_sentiment_confidence3.350000e-016.923000e-011.000000e+00
negativereason_confidence      0.000000e+003.605000e-016.705000e-01
retweet_count                  0.000000e+000.000000e+000.000000e+00
                                        75%          max
```

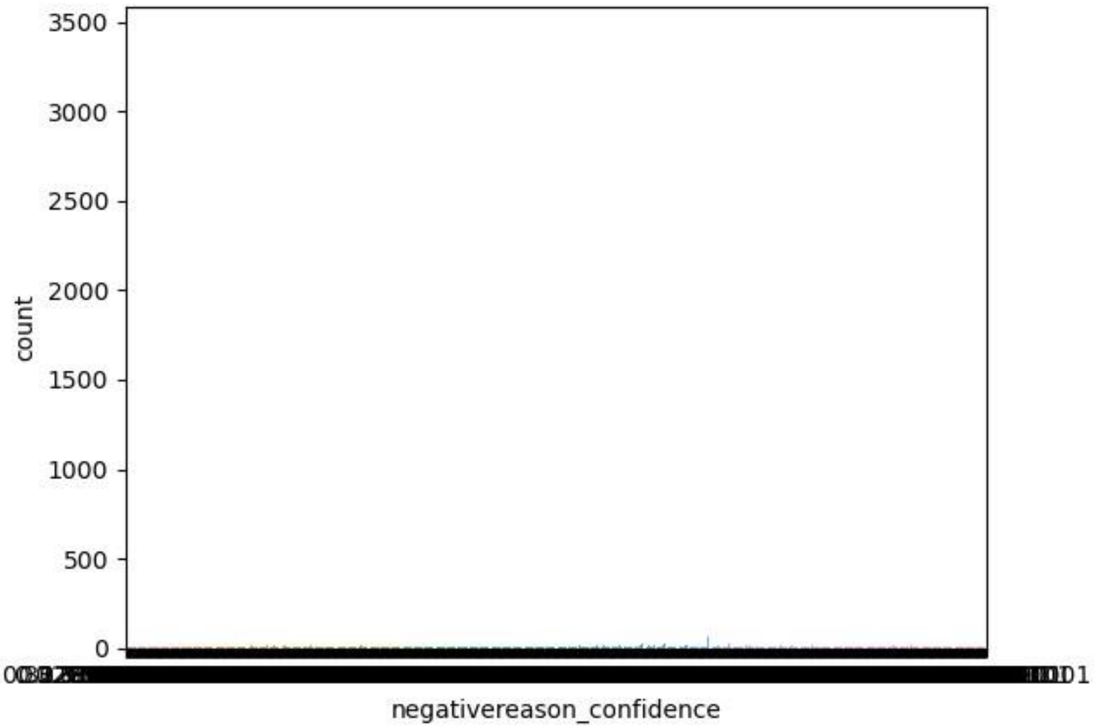
```
tweet_id         5.698884e+17         5.703106e+17
airline_sentiment_confidence         1.000000e+00
1.000000e+00            negativereason_confidence
1.000000e+00      1.000000e+00      retweet_count
0.000000e+00 4.400000e+01
```

```
[17]: df.nunique()
```

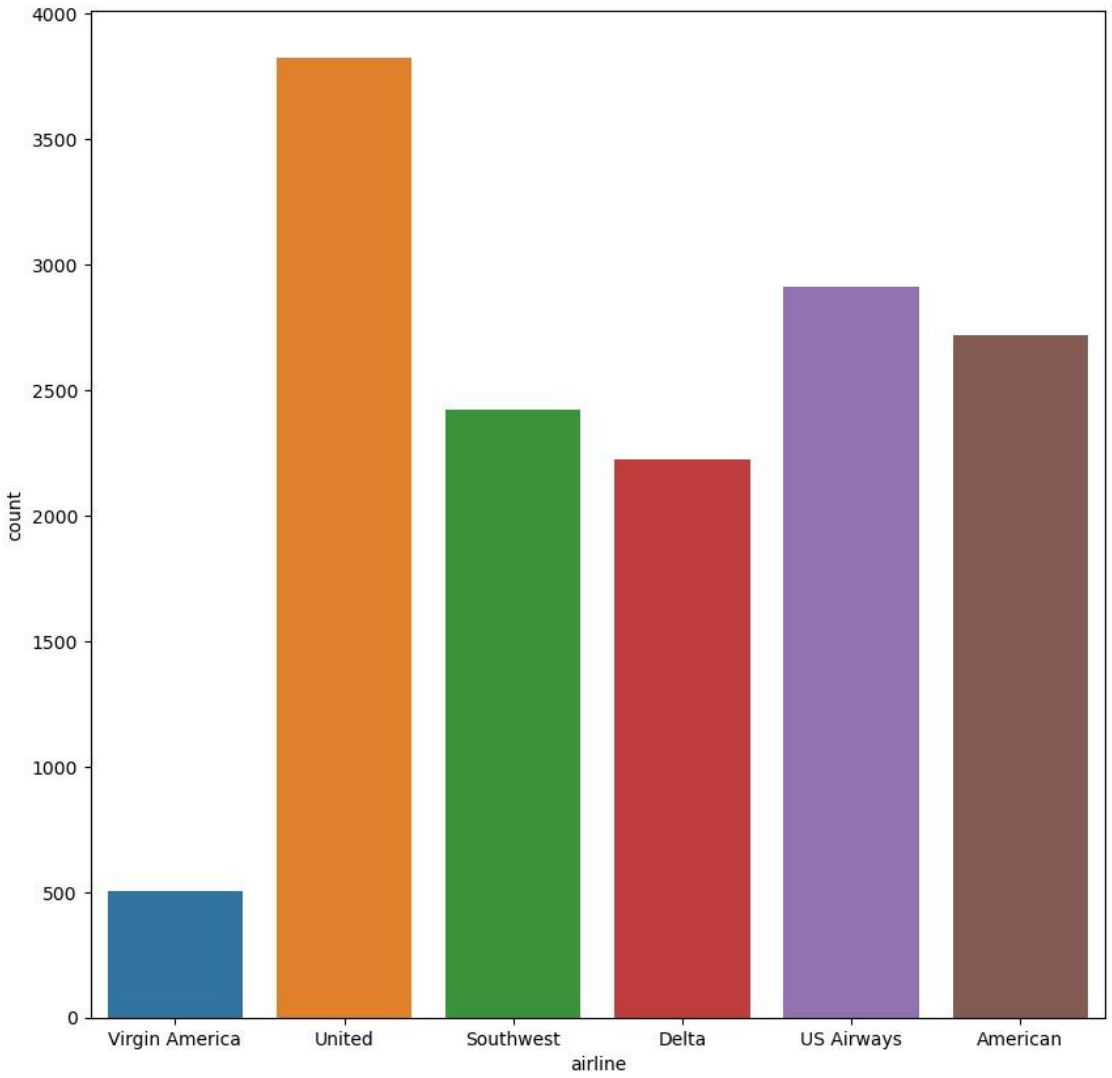
```
[17]: tweet_id                      14485
     airline_sentiment                 3
    airline_sentiment_confidence  1023
     negativereason                   10
    negativereason_confidence     1410
      airline                          6
      name                          7701
     retweet_count                    18
      text                         14427
     tweet_created                 14247
     tweet_location                 3081
     user_timezone                    85
```

```
dtype: int64
```

```
[18]: ax = sns.countplot(x = "negativereason_confidence", data = df)
```



```
[19]: plt.figure(figsize = (10, 10))
      ax = sns.countplot(x = "airline", data = df)
```



```
[20]:  import plotly.graph_objects as go
      crosstab_sentiments=pd.crosstab(df.airline, df.negativereason)
      companies=list(crosstab_sentiments.index)

      fig = go.Figure(data=[ go.Bar(name=col_name, x=companies,
          y=list(crosstab_sentiments[col_name]))
      for col_name in list(crosstab_sentiments.columns)])
      # Change the bar mode
      fig.update_layout(barmode='stack', title='Sentiment
                        distribution per company',
                    yaxis=dict(title='Sentiment
                    distribution'),
                  xaxis=dict(title='Companies'))
fig.show()
```

[21]:
```
crosstab_neg_reasons = pd.crosstab(df["airline"], df["negativereason"])
companies = list(crosstab_neg_reasons.index)

fig = go.Figure(data = [
    go.Bar(name = col_name, x = companies, y =
  list(crosstab_neg_reasons[col_name]))
for col_name in list(crosstab_neg_reasons.columns)])

fig.update_layout(barmode = "stack",
                  title = "Negative Reasons Distribution per Company",
                  yaxis = dict(title = "Negative reasons Distribution"),
                  xaxis = dict(title = "Companies"))
fig.show()
```

[22]:
```
labels = list(crosstab_neg_reasons.columns)
values = [crosstab_neg_reasons[col_name].sum() for col_name in labels]

# Use `hole` to create a donut-like pie chart
fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=.3)])
fig.update_layout(title='Overall distribution for negative reasons')
fig.show()
```

[23]:
```
df.drop(df.loc[df["airline_sentiment"] == "neutral"].index, inplace = True)
```

[24]:
```
data = df[["airline_sentiment", "text"]]
data.head()
```

[24]:
```
  airline_sentiment                                               text
1          positive @VirginAmerica    plus    you've    added
                    commercials t…
3          negative @VirginAmerica it's really aggressive to
                    blast…
4          negative @VirginAmerica and it's a really big bad
                    thing…
5          negative @VirginAmerica seriously would pay $30 a
                    fligh…
6          positive @VirginAmerica yes, nearly every time I
                    fly VX…
```

[25]:
```
X = df["text"]
y = df["airline_sentiment"]
X
```

```
[25]: 1    @VirginAmerica plus you've added
     commercials t… 3 @VirginAmerica it's really
     aggressive to blast…
     4       @VirginAmerica and it's a really big bad thing…
     5       @VirginAmerica seriously would pay $30 a fligh…
     6       @VirginAmerica yes, nearly every time I fly VX…
                                     …
     14633         @AmericanAir my flight was Cancelled Flightled…
     14634         @AmericanAir right on cue with the delays
     14635         @AmericanAir thank you we got on a different f…
     14636         @AmericanAir leaving over 20 minutes Late Flig…
     14638 @AmericanAir you have my money, you change my …
     Name: text, Length: 11510, dtype: object
```

```
[26]:
                                                                  y
```

```
[26]: 1    positive 3
     negative
     4       negative
     5       negative
     6       positive
               …
     14633   negative
     14634   negative
     14635   positive
     14636   negative
     14638   negative
     Name: airline_sentiment, Length: 11510, dtype: object
```

```
[27]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
= 0.2,↩
      ↪random_state = 42) print(X_train.shape,
     X_test.shape, y_train.shape, y_test.shape)
```

```
     (9208,) (2302,) (9208,) (2302,)
```

```
[28]: tfidf = TfidfVectorizer(stop_words="english")
```

```
[29]: tfidf.fit(y_train)
```

```
[29]: TfidfVectorizer(stop_words='english')
```

```
[30]: print(tfidf.get_feature_names_out())
```

```
     ['negative' 'positive']
```

[31]:
```
print(tfidf.vocabulary_)
```

```
{'negative': 0, 'positive': 1}
```

[32]:
```
print(df)
```

```
              tweet_id airline_sentimentairline_sentiment_confidence
                               \
1     570301130888122368      positive                        0.3486
3     570301031407624196      negative                        1.0000
4     570300817074462722      negative                        1.0000

5    570300767074181121      negative                          1.0000
6    570300616901320704      positive                          0.6745
…                …              …                        …
14633569587705937600512      negative                          1.0000
14634569587691626622976      negative                          0.6684
14635569587686496825344      positive                          0.3487
14636569587371693355008      negative                          1.0000
14638569587188687634433      negative                          1.0000

             negativereasonnegativereason_confidence      airline\
1                     NaN                  0.0000 Virgin
                                                  America
3              Bad Flight                  0.7033 Virgin
                                                  America
4              Can't Tell                  1.0000 Virgin
                                                  America
5              Can't Tell                  0.6842 Virgin
                                                  America
6                     NaN                  0.0000 Virgin
                                                  America
…                   …                       …          …

14633     Cancelled                        1.0000     American
          Flight
14634        Late Flight                   0.6684     American

14635                 NaN                  0.0000     American

14636Customer Service                      1.0000     American
     Issue
14638Customer Service                      0.6659     American
     Issue

              name retweet_count  \
1          jnardino           0

3          jnardino           0
```

```
4         jnardino          0

5         jnardino          0

6       cjmcginnis          0

…             …          …

14633RussellsWriting          0

14634 GolfWithWoody          0

14635KristenReenders          0

14636     itsropes          0

14638    SraJackson          0

                                                 text  \
1    @VirginAmerica plus you've added commercials
     t…
3    @VirginAmerica it's really aggressive to
     blast…
4    @VirginAmerica and it's a really big bad
     thing…
5    @VirginAmerica seriously would pay $30 a
     fligh…
6    @VirginAmerica yes, nearly every time I fly
     VX…
…                                             …

14633@AmericanAir my flight was Cancelled
     Flightled…
14634      @AmericanAir right on cue with the
           delays
14635@AmericanAir thank you we got on a different
     f…
14636@AmericanAir leaving over 20 minutes Late
     Flig…
14638@AmericanAir you have my money, you change my
     …
            tweet_created   tweet_location       user_timezone

1    2015-02-24 11:15:59 -              NaN  Pacific  Time  (US  &
     0800                                    Canada)
3    2015-02-24 11:15:36 -              NaN  Pacific  Time  (US  &
     0800                                    Canada)
4    2015-02-24 11:14:45 -              NaN  Pacific  Time  (US  &
     0800                                    Canada)
5    2015-02-24 11:14:33 -              NaN  Pacific  Time  (US  &
     0800                                    Canada)
```

```
6    2015-02-24 11:13:57 -   San Francisco CA Pacific  Time  (US  &
     0800                                     Canada)
…                      …             …                 …
146332015-02-22 12:01:06 -       Los Angeles              Arizona
     0800
146342015-02-22 12:01:02 -               NaN                Quito
     0800
146352015-02-22 12:01:01 -               NaN                  NaN
     0800
146362015-02-22 11:59:46 -             Texas                  NaN
     0800
146382015-02-22 11:59:02 -        New Jersey  Eastern  Time  (US  &
     0800                                     Canada)
[11510 rows x 12 columns]
```

```
[33]: data[data["airline_sentiment"] ==
"negative"]["text"]
```

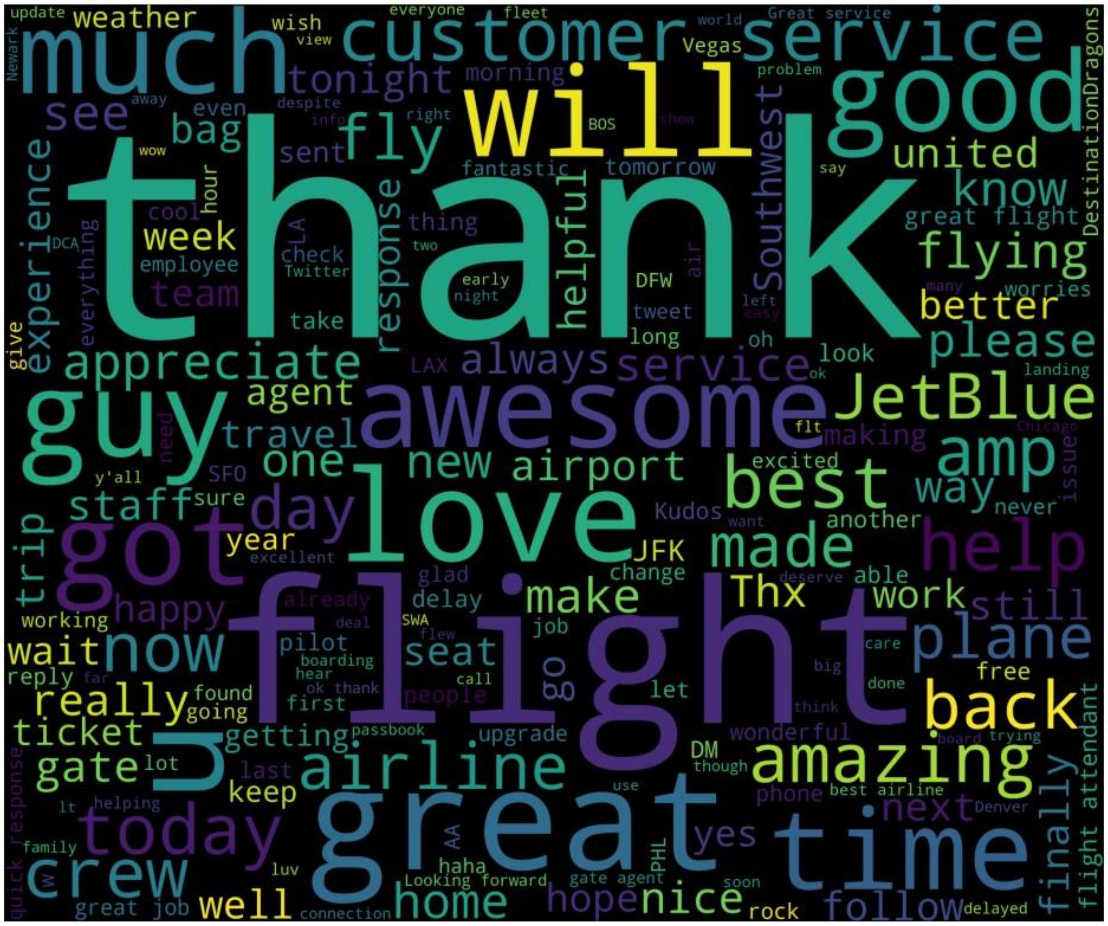
```
[33]: 3    @VirginAmerica it's really aggressive to
      blast… 4   @VirginAmerica and it's a really big
      bad thing… 5     @VirginAmerica seriously would
      pay $30 a fligh… 15   @VirginAmerica SFO-PDX
      schedule is still MIA.
      17     @VirginAmericaI flew from NYC to SFO last we…
                                   …
      14631@AmericanAir thx for nothing on getting us out…
      14633         @AmericanAir my flight was Cancelled Flightled…
      14634         @AmericanAir right on cue with the delays
      14636@AmericanAir leaving over 20 minutes Late Flig…
      14638 @AmericanAir you have my money, you change my …
      Name: text, Length: 9157, dtype: object
```

```
[34]: count_vect = CountVectorizer(stop_words="english")
      neg_matrix = count_vect.
       ↪fit_transform(data[data["airline_sentiment"]=="negative"]["te
      xt"]) freqs = zip(count_vect.get_feature_names_out(),
      neg_matrix.sum(axis=0). ↪tolist()[0])
      # Sort from largest to smallest
      print(sorted(freqs, key=lambda x: -x[1])[:100])
```

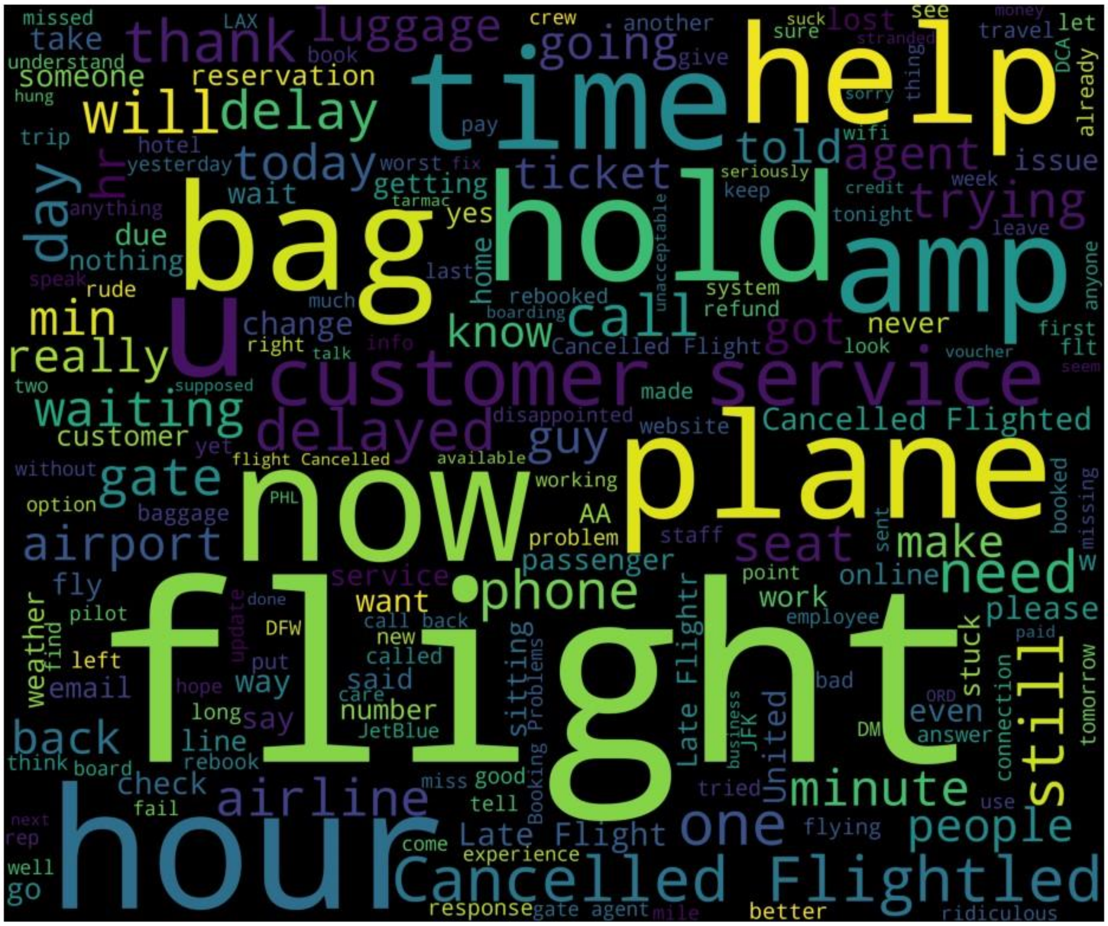
```
[('flight', 2937), ('united', 2899), ('usairways', 2375),
('americanair', 2089),
('southwestair', 1214), ('jetblue', 1051), ('cancelled', 921),
('service', 746),
('hours', 646), ('just', 622), ('help', 618), ('hold', 611),
('customer', 609),
('time', 596), ('plane', 530), ('delayed', 505), ('amp', 503),
('hour', 452),
('flightled', 445), ('http', 436), ('flights', 419), ('bag', 415),
('gate',
410), ('ve', 398), ('don', 388), ('late', 377), ('need', 373),
('phone', 367),
('waiting', 341), ('thanks', 315), ('got', 298), ('airline', 294),
('like',
291), ('trying', 288), ('delay', 272), ('wait', 272), ('today', 269),
('minutes', 266), ('day', 251), ('going', 249), ('bags', 245),
('luggage', 245),
('told', 245), ('airport', 244), ('people', 242), ('worst', 241),
('fly', 237), ('really', 236), ('did', 227), ('guys', 224),
('weather', 224), ('lost', 221),
('agent', 218), ('hrs', 217), ('way', 212), ('make', 211), ('change',
210),
('seat', 208), ('flighted', 205), ('want', 205), ('check', 204),
('know', 201),
('days', 200), ('home', 194), ('virginamerica', 191), ('baggage',
190),
('getting', 181), ('sitting', 179), ('ticket', 176), ('tomorrow',
176), ('let',
174), ('min', 171), ('customers', 169), ('flying', 168), ('line',
164),
('email', 163), ('online', 163), ('experience', 162), ('didn', 161),
('stuck',
160), ('work', 159), ('bad', 157), ('number', 156), ('won', 156),
('said', 155),
('seats', 154), ('30', 153), ('10', 150), ('problems', 150),
('times', 150),
('crew', 149), ('flightr', 148), ('doesn', 146), ('good', 145),
('ll', 144),
('aa', 143), ('travel', 142), ('yes', 142), ('response', 139),
('miss', 137)]
```

[35]:

```
new_df = data[data["airline_sentiment"] == "positive"] words = "
".join(new_df["text"]) cleaned_word = " ".join([word for word in
words.split() if "http" not in word↪
 ↪and not word.startswith("@") and word != "RT"]) wordcloud =
WordCloud(stopwords = STOPWORDS, background_color = "black", width
= 3000, height = 2500). ↪generate(cleaned_word) plt.figure(figsize
= (12, 12)) plt.imshow(wordcloud) plt.axis("off") plt.show()
```



```
[36]: new_df = data[data["airline_sentiment"] == "negative"] words = " 
".join(new_df["text"]) cleaned_word = " ".join([word for word in 
words.split() if "http" not in word
 and not word.startswith("@") and word != "RT"]) wordcloud = 
WordCloud(stopwords = STOPWORDS, background_color = "black", width 
= 3000, height = 2500). generate(cleaned_word) plt.figure(figsize 
= (12, 12)) plt.imshow(wordcloud) plt.axis("off") plt.show()
```



```
[37]: data.drop(data.loc[data["airline_sentiment"] == "neutral"].index, inplace =
        True)
```

```
[38]: from sklearn.preprocessing import LabelEncoder
      le = LabelEncoder()

      le.fit(data["airline_sentiment"])
      data["airline_sentiment_encoded"] = le.transform(data["airline_sentiment"])
      data.head()
```

```
[38]:   airline_sentiment                                              text  \
      1          positive @VirginAmerica    plus    you've    added
                          commercials t…
      3          negative @VirginAmerica it's really aggressive to
                          blast…
      4          negative @VirginAmerica and it's a really big bad
                          thing…
```

```
5         negative @VirginAmerica seriously would pay $30 a
                   fligh…
6         positive @VirginAmerica yes, nearly every time I
                   fly VX…
   airline_sentiment_encoded
1                         1
3                         0
4                         0
5                         0
6                         1
```

```
[39]: def tweet_to_words(tweet): letters_only =
         re.sub("[^a-zA-Z]", " ", tweet) words =
         letters_only.lower().split()
         stops = set(stopwords.words("english"))
         meaningful_words = [w for w in words if not w
         in stops] return(" ".join( meaningful_words ))
```

```
[40]: nltk.download("stopwords") data["clean_tweet"] =
      data["text"].apply(lambda x: tweet_to_words(x))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\Administrator\AppData\Roaming\nltk_data…
[nltk_data] Package stopwords is already up-to-date!
```

```
[41]: data.info()
```

```
<class
'pandas.core.frame.DataFrame'>
Int64Index: 11510 entries, 1 to
14638 Data columns (total 4
columns):
#   Column                  Non-Null CountDtype
--- ------                  --------------------
0  airline_sentiment        11510 non-nullobject
1   text                    11510 non-nullobject
2  airline_sentiment_encoded   11510 non-int32
null
3   clean_tweet             11510 non-nullobject
dtypes: int32(1), object(3)
memory usage: 404.6+ KB
```

```
[42]: X = data["clean_tweet"]
      y = data["airline_sentiment"]
```

```
[43]: print(X.shape, y.shape)
```

```
(11510,) (11510,)
```

```
[44]: X_train, X_test, y_train, y_test = train_test_split(X, y,
      random_state = 42) print(X_train.shape, X_test.shape, y_train.shape,
      y_test.shape)
```

```
(8632,) (2878,) (8632,) (2878,)
```

```
[45]: vect = CountVectorizer()
      vect.fit(X_train)
```

```
[45]: CountVectorizer()
```

```
[46]: X_train_dtm = vect.transform(X_train)
      X_test_dtm = vect.transform(X_test)
```

```
[47]: vect_tunned = CountVectorizer(stop_words = "english", ngram_range = (1, 2),□
        ↪min_df = 0.1, max_df = 0.7, max_features = 100)
      vect_tunned
```

```
[47]: CountVectorizer(max_df=0.7, max_features=100, min_df=0.1,
                      ngram_range=(1, 2), stop_words='english')
```

```
[48]: from sklearn.svm import SVC model =
      SVC(kernel = "linear", random_state = 10)
      model.fit(X_train_dtm, y_train) pred =
      model.predict(X_test_dtm) print("Accuracy Score: ",
      accuracy_score(y_test, pred) * 100)
```

```
Accuracy Score:90.7574704656011
```

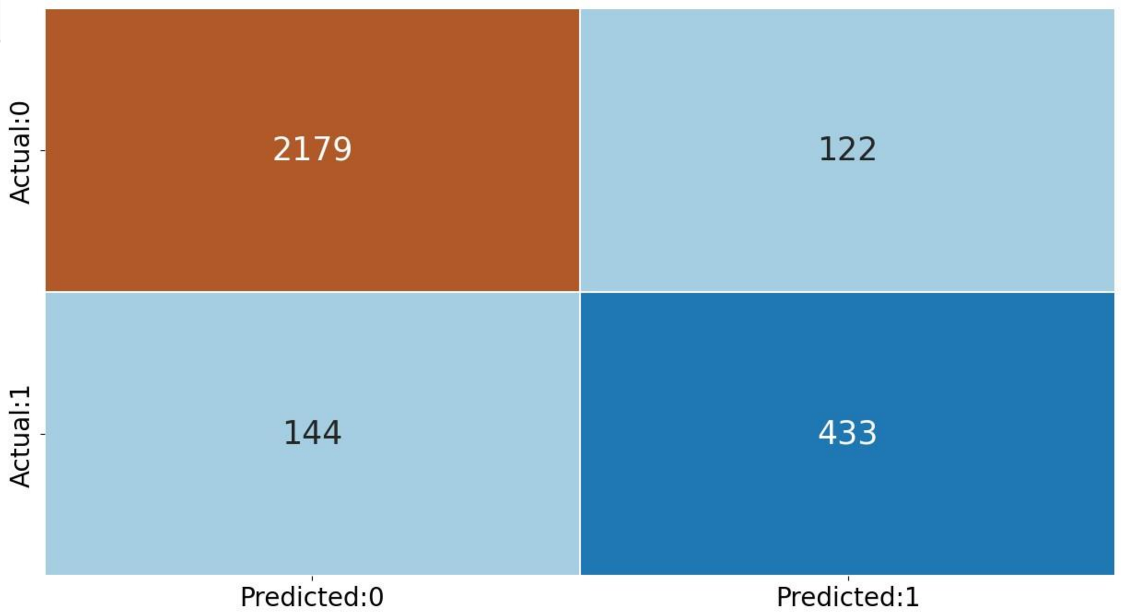
```
[49]: print("Confusion Matrix\n\n", confusion_matrix(y_test, pred))
```

```
Confusion Matrix

 [[2179 122]
 [ 144 433]]
```

```
[50]: #defining the size of the canvas
      plt.rcParams['figure.figsize'] = [15,8]
      #confusion matrix to DataFrame
      conf_matrix = pd.DataFrame(data = confusion_matrix(y_test,
        pred),columns =□ ↪['Predicted:0','Predicted:1',], index =
        ['Actual:0','Actual:1',])
      #plotting the confusion matrix sns.heatmap(conf_matrix, annot =
      True, fmt = 'd', cmap = 'Paired', cbar =□
```

```
 ↪False,linewidths = 0.1, annot_kws =
{'size':25}) plt.xticks(fontsize = 20)
plt.yticks(fontsize = 20) plt.show()
```



```
[51]: print(classification_report(y_test, pred))
```

```
              precision    recall  f1-score   support

    negative       0.94      0.95      0.94      2301
    positive       0.78      0.75      0.77       577

    accuracy                           0.91      2878
   macro avg       0.86      0.85      0.85      2878
weighted               0.91      0.91      0.91      2878
avg
```

```
[ ]:
```

CONCLUSION:

- The overall project output for sentiment analysis in marketing encompasses a multifaceted approach to extracting valuable insights from textual data.
- It begins with the collection and preprocessing of data from sources like social media, customer reviews, and surveys.
- Sentiment analysis results provide a granular understanding of sentiment scores for individual data points, which are then aggregated to reveal trends over time or across different categories and products.
- Visualizations and reports present these insights in an easily digestible format, aiding marketing teams in comprehending sentiment dynamics.
- Additionally, competitive analysis assesses how a brand's sentiment stacks up against competitors.
- Key findings and actionable recommendations arise from the analysis, offering strategic insights for enhancing customer satisfaction, addressing negative sentiment, and leveraging positive sentiment.
- Furthermore, documentation and training materials facilitate the effective utilization of sentiment analysis insights, while a continuous monitoring plan ensures adaptability to evolving sentiment.
- This holistic approach equips marketing teams to make data-driven decisions and refine their strategies for improved customer engagement and brand success.
- Overall, sentiment analysis is a valuable tool that can be used to improve marketing in a variety of ways.
- By understanding customer sentiment and developing actionable recommendations, marketers can create more effective campaigns, develop better products and services, and improve the customer experience.
- As sentiment analysis tools become more sophisticated and accessible, it is likely to become an even more essential tool for marketers.