SENTIMENTAL ANALYSIS FOR MARKETING USING AI

BATCH MEMBER

Name :Priyanka.S,

Reg no:410121104033,

Adhi College of Engineering and Technology

Phase 3 submission document

Project title: Sentimental analysis for marketing

Phase 3: Development part 1

Topic: Start building the sentiment analysis solution by loading dataset and preprocessing the data.

Sentimental analysis for marketing

Introduction:

Sentiment analysis is a marketing tool that helps you examine the way people interact with a brand online. This method is more comprehensive than traditional online marketing tracking, which measures the number of online interactions that customers have with a brand, like comments and shares. Using sentiment analysis, you can label individual interactions as positive, negative or neutral. Once you've figured out how to determine and track these labels, you can use this new data set for a variety of marketing purposes, including your online strategy.

Sentimental analysis is an extremely useful tool to have since higher numbers of interactions don't always equate to better results. For example, if you were to receive 10 replies on a social post and all of them were positive, your post likely had a more compelling effect on your audience than if you receive 100 replies with only 10 of them being positive. The primary purpose of sentiment analysis is to respond to commentary more constructively.

Dataset link:

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

Necessary step to follow:

Import Necessary Libraries:
Start by importing the required libraries, such as Pandas, NumPy, and Natural Language Processing (NLP) libraries like NLTK or spaCy.

Load and Explore the Dataset:
Load your sentiment dataset (e.g., CSV file) using Pandas. You should have a column with text data and another with labels (positive/negative).

Text Preprocessing:
Preprocess the text data. This includes:
- Lowercasing the text.
- Removing punctuation and special characters.
- Tokenization (splitting text into words or tokens).
- Removing stop words (common words like "the," "and," "is" that do not contribute much to sentiment).

Text Vectorization:
You need to convert text data into numerical form. Common techniques include:
- Bag of Words (BoW): Count the frequency of each word in the text.

- TF-IDF (Term Frequency-Inverse Document Frequency): Weigh words based on their importance in the document and across the corpus.
- Word Embeddings (e.g., Word2Vec or GloVe): Represent words in a dense vector space.

Split the Dataset:
Divide your dataset into training and testing sets to evaluate the model's performance.

Select a Machine Learning Model:
Choose a machine learning algorithm for sentiment analysis, like Logistic Regression, Naive Bayes, or a neural network.

Train and Evaluate the Model:
Train the model using the training data and evaluate its performance using the testing data. Common metrics include accuracy, precision, recall, and F1-score.

Make Predictions:
Once the model is trained, you can use it to make sentiment predictions for new text data.

Program:

```
  # import contractions library.
!pip install contractions missingno wordcloud
```

In [13]:
linkcode

```
# Import necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as msno
```

```
import warnings
warnings.filterwarnings(action='ignore')

# Import NLTK and download required resources
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokeniz
e
from nltk.stem import LancasterStemmer, WordNetLemmat
izer

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

# Import other libraries
import re
import string
import unicodedata
import contractions
from sklearn.feature_extraction.text import CountVect
orizer, TfidfVectorizer
import wordcloud
import train_test_split, StratifiedKFold
from sklearn.svm import LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
from sklearn.metrics import (
    recall_score,
    accuracy_score,
    confusion_matrix,
    classification_report,
    f1_score,
    precision_score,
```

```
    precision_recall_fscore_support
)

# Set options for displaying data
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", 200)
```

```
df = pd.read_csv('Tweets.csv')
df.head()
```

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added com	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & C

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
											mercials t...				anada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast t...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

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

In [16]:
linkcode

```
df.head()
```

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn	NaN	2015-02-24 11:35:52 -	NaN	Eastern Time (US &

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
											said.		0800		Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)

tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone	
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

```
df.info()
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 14640 entries, 0 to 14639

Data columns (total 15 columns):
```

```
 #   Column                         Non-Null Count  Dt
ype

---  ------                         --------------  --
---

 0   tweet_id                       14640 non-null  in
t64

 1   airline_sentiment              14640 non-null  ob
ject

 2   airline_sentiment_confidence  14640 non-null  fl
oat64

 3   negativereason                 9178 non-null   ob
ject

 4   negativereason_confidence      10522 non-null  fl
oat64

 5   airline                        14640 non-null  ob
ject

 6   airline_sentiment_gold         40 non-null     ob
ject

 7   name                           14640 non-null  ob
ject

 8   negativereason_gold            32 non-null     ob
ject
```

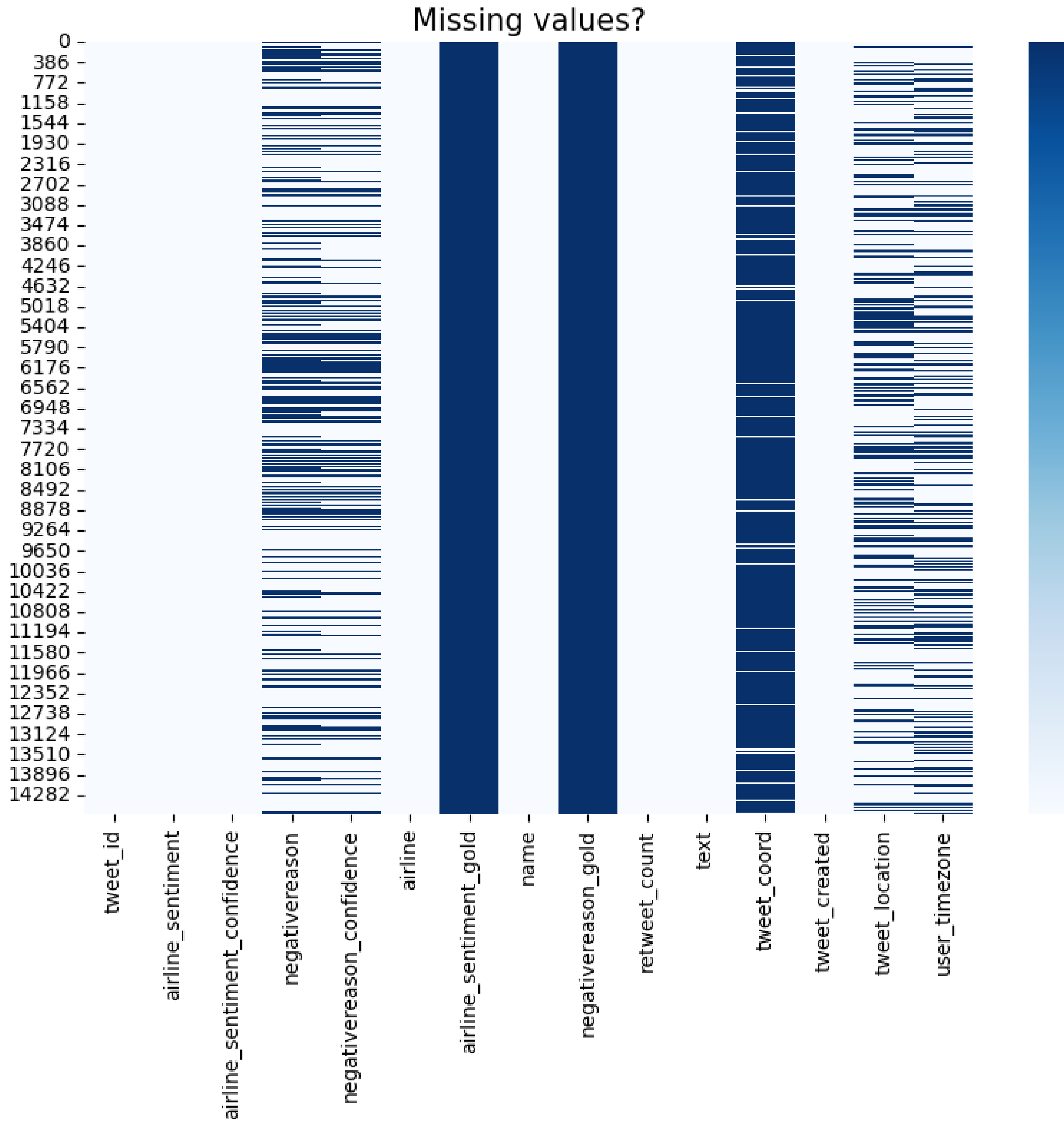
```
 9   retweet_count                    14640 non-null  in
t64

 10  text                             14640 non-null  ob
ject

 11  tweet_coord                      1019 non-null   ob
ject

 12  tweet_created                    14640 non-null  ob
ject

 13  tweet_location                   9907 non-null   ob
ject

 14  user_timezone                    9820 non-null   ob
ject

dtypes: float64(2), int64(2), object(11)

memory usage: 1.7+ MB
```

```
df.isnull().sum()
```

```
Out[3]:
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
```

```
#Visualization of missing value using heatmap
plt.figure(figsize=(10,7))
sns.heatmap(df.isnull(), cmap = "Blues")
plt.title("Missing values?", fontsize = 15)
plt.show()
```



Missing values?
0
386
772
1158
1544
1930
2316
2702
3088
3474
3860
4246
4632
5018
5404
5790
6176
6562
6948
7334
7720
8106
8492
8878
9264
9650
10036
10422
10808
11194
11580
11966
12352
12738
13124
13510
13896
14282
tweet_id
airline_sentiment
airline_sentiment_confidence
negativereason
negativereason_confidence
airline
airline_sentiment_gold
name
negativereason_gold
retweet_count
text
tweet_coord
tweet_created
tweet_location
user_timezone

Conclusion:

In conclusion, loading and processing datasets for sentiment analysis in marketing is a crucial step in harnessing valuable insights from customer feedback. Effective handling of data allows marketers to gain a deeper understanding of consumer sentiment, enabling them to make informed decisions and create targeted strategies to enhance their brand's reputation and customer satisfaction.