**VRS College of Engineering and Technology**

**Arasur.**

# **Sentiment analysis for marketing**

Presented by,

S.Latha,

K.Akilandeshwari,

M.Mathina,

R.Thamizharasi.

**Sentiment analysis for marketing**

# **Introduction**:

Sentiment analysis, also known as opinion mining, is a natural language processing technique used to determine and extract subjective information from text data. It involves identifying and categorizing opinions, sentiments, and emotions expressed within a piece of text, such as social media posts, customer reviews, or survey responses. The primary goal of sentiment analysis is to understand the attitudes, opinions, and feelings of individuals or groups regarding a particular topic, product, service, or event.

# **Abstract:**

Sentiment analysis, a prominent field in natural language processing, involves the automated extraction and categorization of sentiments expressed in text data. This analysis, also known as opinion mining, discerns emotions, attitudes, and opinions as positive, negative, or neutral.. By uncovering valuable insights from textual data, businesses and researchers can make data-driven decisions, enhance customer experiences, and optimize products and services.

# **Dataset:**

**Data Source URL:**[https://www.kaggle.com/datasets/tweets dataset/market-based-sentiment-analysis](https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis)

# **Keywords:**

* Sentiment Analysis
* Twitter-Specific Terms
* NLP Techniques
* Machine Learning Algorithms
* Twitter Airline US Dataset
* Sentiment Classes
* Evaluation Metrics
* Data Preprocessing
* Industry-Specific Terms
* Social Media Analysis
* Ethical Considerations

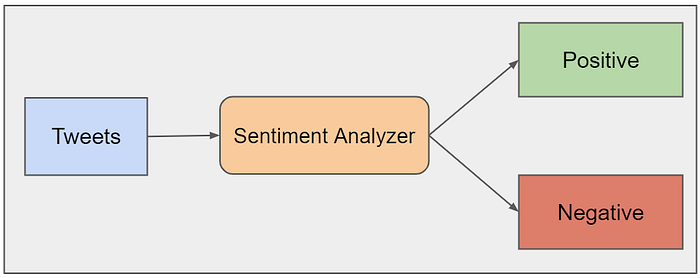
# **Design Thinking:**

Design thinking is a human-centered, iterative problem-solving methodology that prioritizes empathy, creativity, and collaboration to address complex challenges. Rooted in understanding the users' needs, it encourages innovative solutions that resonate with real-world requirements. Here's a breakdown of the key stages and principles of design thinking



l

# **Twitter Sentiment Analysis Dataset:**



**Step 1:**

Install pandas and numpy



**Step 2:**

CountVectorizer is a feature extraction technique used in natural language processing (NLP) and text mining to convert a collection of text documents into a numerical feature matrix.

of a word in the given text.



Cccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccccc

**Step 3:**

WordNetLemmatizer is a part of the NLTK (Natural Language Toolkit) library in Python. It is used for lemmatizing words, which means reducing a word to its base or root form.



**Step 4:**

Upload dataset file(twitter airlines US)



**Step 6:**

Load the dataset

**Step 7:**

Display the frames



**Step 8:**

Install nltk (natural processing language)



**Step 9:**

To drop unnecessary columns from a DataFrame in Python, you can use the drop() method provided by pandas, a powerful data manipulation library. Here's how you can do it.



airline\_sentiment text

0 neutral @VirginAmerica What @dhepburn said.

1 positive @VirginAmerica plus you've added commercials t...

2 neutral @VirginAmerica I didn't today... Must mean I n...

3 negative @VirginAmerica it's really aggressive to blast...

4 negative @VirginAmerica and it's a really big bad thing...

... ... ...

120 negative @VirginAmerica Use another browser! 2015 &amp;...

121 negative @VirginAmerica And now the flight Flight Booki...

122 negative @VirginAmerica I like the customer service but...

123 positive @VirginAmerica thanks to your outstanding NYC-...

124 positive @VirginAmerica you have the absolute best team...

125 rows × 2 columns

**Step 10:**

The preprocess\_tweet function performs various preprocessing steps on each tweet. df['text'] refers to the column containing the original tweets, and df ['preprocessed\_text'] is the column where preprocessed tweets will be stored.

# Function to preprocess the text

def preprocess\_text(text):

    # Remove punctuations and numbers

    text = re.sub('[^a-zA-Z]', ' ', text)

    # Single character removal

    text = re.sub(r'\s+[a-zA-Z]\s+', ' ', text)

    # Removing multiple spaces

    text = re.sub(r'\s+', ' ', text)

    # Converting to Lowercase

    text = text.lower()

    # Lemmatization

    #text = text.split()

    #lemmatizer = WordNetLemmatizer()

    #text = [lemmatizer.lemmatize(word) for word in text if not word in set(stopwords.words('english'))]

    #text = ' '.join(text)

    return text

# Apply the preprocessing to the 'text' column

df['text'] = df['text'].apply(preprocess\_text)

# Display the first 5 rows of the dataframe after preprocessing

df.head()

airline\_sentiment text

0 neutral virginamerica what dhepburn said

1 positive virginamerica plus you ve added commercials t...

2 neutral virginamerica didn today must mean need to ta...

3 negative virginamerica it really aggressive to blast o...

4 negative virginamerica and it a really big bad thing a...

**Step 11:**

Splitting the data:

Split your data into two parts training set and testing set. The training set is used to train your machine learning model, and the testing set is used to evaluate its performance. Use the training data (X\_train and y\_train) to train your sentimental analysis model.



**Step 12:**

Feature exctraction:

Feature extraction in sentiment analysis for marketing involves identifying and extracting relevant information from textual data to understand customer sentiments effectively. Key techniques include recognizing specific product aspects, detecting emotions, extracting opinion words, identifying industry-specific keywords, categorizing feedback, analyzing temporal patterns, and understanding social interactions.



**Step 13:**

Model Training:

Feed the preprocessed and feature-extracted data into the selected model. During training, the model learns the patterns and relationships between features and sentiment labels from the training data.Use a separate set of data (validation or test set) to evaluate the model's performance.



**Step 14:**

Classification report, confusion matrix, accuracy score



Classification Report:

precision recall f1-score support

negative 0.79 0.95 0.86 1889

neutral 0.65 0.41 0.50 580

positive 0.80 0.50 0.62 459

accuracy 0.77 2928

macro avg 0.75 0.62 0.66 2928

weighted avg 0.76 0.77 0.75 2928

Confusion Matrix:

[[1799 65 25]

[ 312 235 33]

[ 169 60 230]]

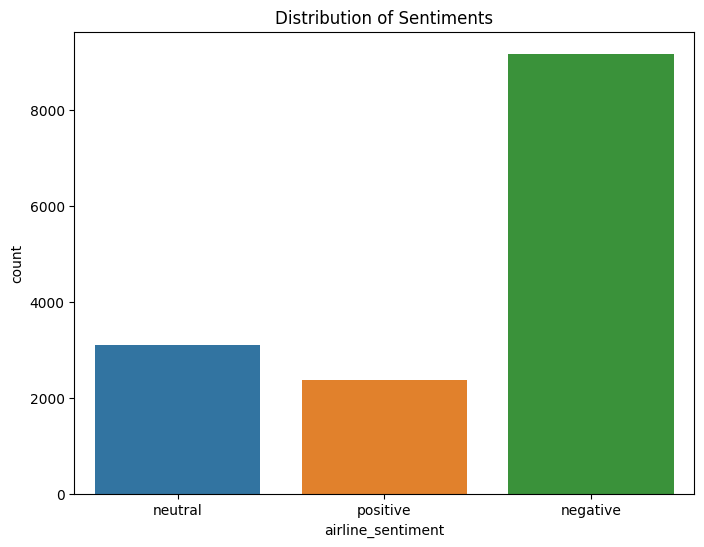
Accuracy Score:

0.773224043715847

**Distribution of sentiment:**

To visualize the distribution of sentiments in the Twitter airline dataset, you can use a bar chart or any other appropriate visualization method. In this code, the value\_counts() function is used to count the occurrences of each sentiment class in the 'sentiment' column of the dataset. The resulting counts are then plotted as a bar chart using matplotlib.

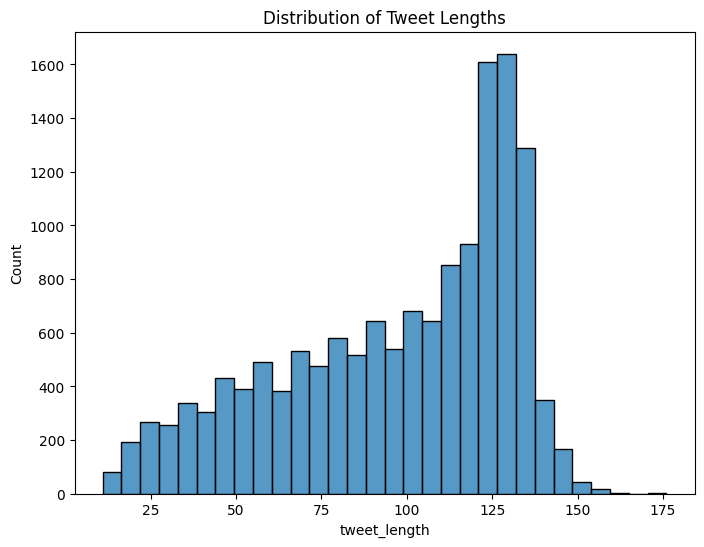




**Histogram of tweet lengths:**

Load your dataset containing a column named 'text'. Calculate the length of each tweet in the 'text' column.Plot a histogram with tweet lengths on the x-axis and frequency on the y-axis.

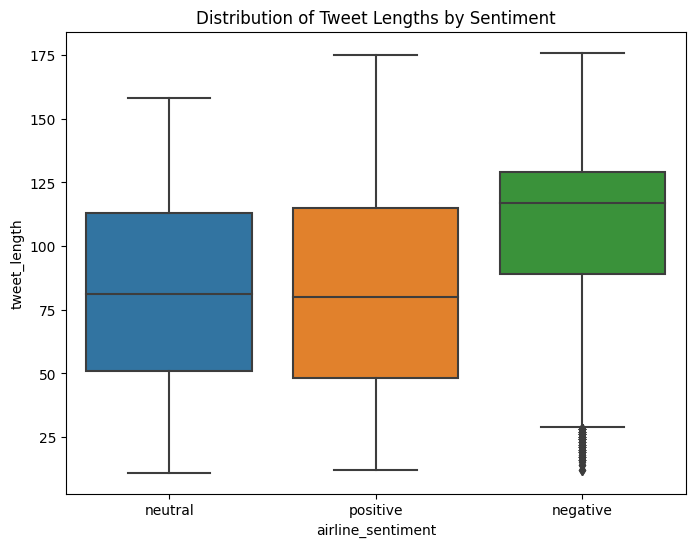




**Boxplot of tweet lengths**:

The boxplot shows the range and distribution of tweet lengths, allowing for a quick understanding of the variation in text lengths within the dataset. Use the 'tweet\_length' column to create a boxplot, displaying the distribution of tweet lengths. Compute the length of each tweet and store it in a new column, like 'tweet\_length'.





**Confusion matrix:**

Train a machine learning model on the training data. the confusion matrix is calculated using the test data and then visualized as a heatmap using the seaborn library. The rows of the confusion matrix represent the actual classes (negative, neutral, positive), and the columns represent the predicted classes.



