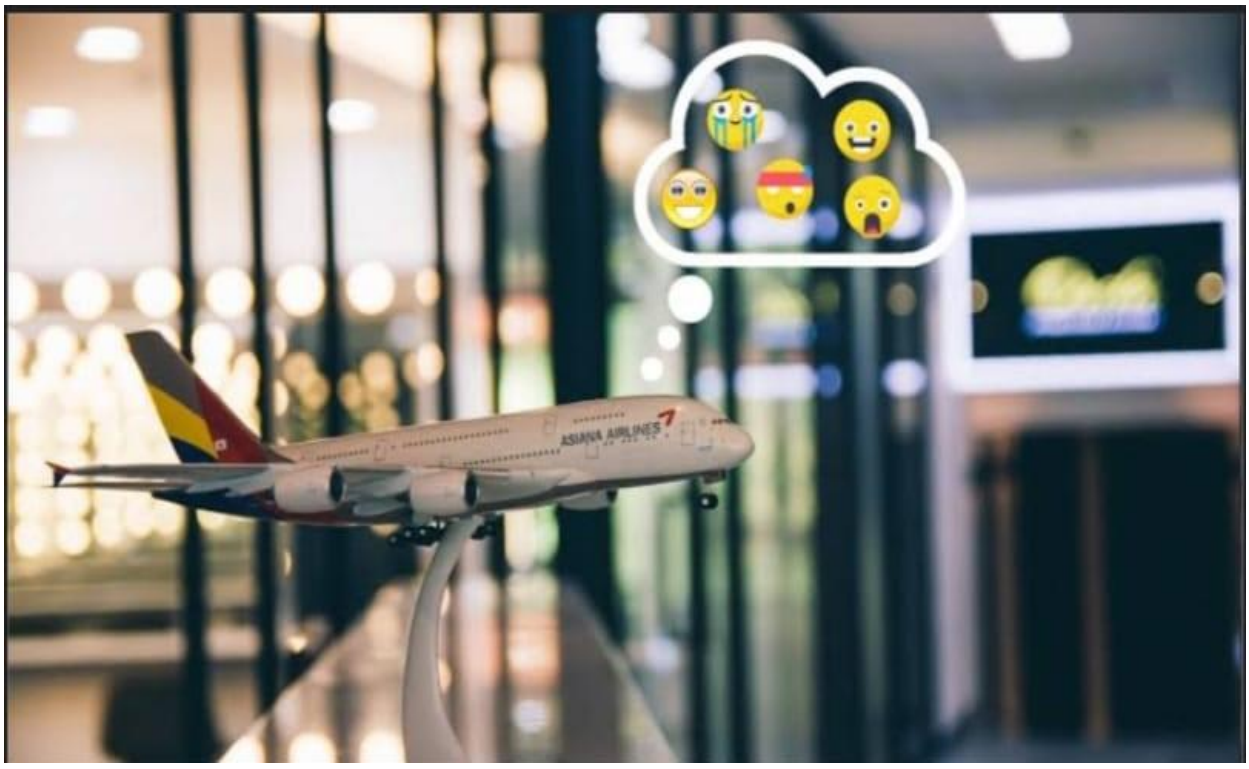
PHASE 4 PROJECT SUBMISSION

PROJECT TITLE:

SENTIMENT ANALYSIS FOR MARKETING

DEVELOPMENT PART 2

TOPIC: Continue building the Sentiment Analysis for marketing model using python for feature engineering, model training and evaluation



INTRODUCTION:

- Sentiment analysis is a powerful tool in the realm of US airlines marketing, enabling companies to gain valuable insights into how customers perceive their services.

- In a highly competitive industry where customer experience is paramount, understanding the sentiment behind customer feedback is critical.

- Sentiment analysis involves the use of natural language processing and machine learning techniques to classify and quantify customer comments and reviews, helping airlines discern whether the sentiment is positive, negative, or neutral.

- This analysis goes beyond merely gauging customer satisfaction; it provides airlines with the ability to pinpoint specific pain points and areas of excellence, offering actionable insights for marketing campaigns, service improvements, and crisis management.

- By harnessing the sentiments expressed in social media, customer reviews, and other textual data, US airlines can tailor their marketing strategies to enhance customer satisfaction, strengthen brand loyalty, and stay ahead in a dynamic and competitive industry.

- Furthermore, sentiment analysis equips airlines with the tools to detect and respond to emerging trends, ensuring that their marketing efforts remain relevant and responsive to evolving customer sentiments.

OVERVIEW OF THE PROCESS:

A project on sentiment analysis for marketing involves using natural language processing (NLP) and machine learning techniques to analyze and understand the sentiment of customers or the general public towards a brand, product, or marketing campaign. Here's an overview of the key components and steps involved:

1. Data Collection: Gather data from various sources, including social media, online reviews, customer feedback forms, and other relevant platforms. This data will include text-based content that reflects customer opinions and sentiments.

2. Data Preprocessing: Clean and preprocess the data by removing noise, such as special characters, emojis, and irrelevant information. Tokenization and stemming may be used to break down text into manageable units.

3. Sentiment Analysis: Utilize NLP techniques to determine the sentiment of the collected data. This often involves classifying text as positive, negative, or neutral sentiment. Machine learning models like Naive Bayes, Support Vector Machines, or deep learning models like LSTM or Transformer-based models (e.g., BERT) can be employed for this task.

4. Feature Engineering: Extract relevant features from the data, such as word frequency, sentiment scores, or contextual information, to enhance the performance of sentiment analysis models.

5. Model Training: Train the sentiment analysis model on a labeled dataset. The model learns to associate text patterns with sentiment labels.

6. Evaluation: Assess the model's performance using metrics like accuracy, precision, recall, and F1-score. Fine-tune the model as needed to improve its accuracy.

7. Sentiment Visualization: Present the sentiment analysis results through visualizations like sentiment heatmaps, time series charts, or word clouds to help marketers understand customer sentiment trends.

8. Customer Insights: Derive actionable insights from the sentiment analysis results. Identify trends, common issues, and areas of improvement for marketing strategies.

9. Decision Support: Provide marketing teams with data-driven recommendations based on the sentiment analysis. This can include adjusting marketing campaigns, addressing customer concerns, or capitalizing on positive sentiment.

10. Monitoring: Continuously monitor sentiment data to track changes over time. Adjust marketing strategies accordingly to maintain a positive brand image and improve customer satisfaction.

11. Reporting: Generate regular reports or dashboards to communicate sentiment analysis findings to marketing and decision-making stakeholders.

PROCEDURE:

FEATURE ENGINEERING:

Feature engineering is a critical step in sentiment analysis for marketing. It involves creating meaningful features from raw text data to improve the performance of sentiment analysis models. Here's a procedure broken down into four subheadings:

1.Text Preprocessing:

Tokenization: Split the text into individual words or tokens. This allows the model to understand the basic units of text.

Lowercasing: Convert all text to lowercase to ensure that "word" and "Word" are treated as the same word.

Stop Word Removal: Eliminate common words (stop words) that don't carry much information, such as "and," "the," and "is."

Removing Special Characters and Punctuation: Strip out non-alphanumeric characters, punctuation, and symbols.

Stemming or Lemmatization: Reduce words to their root form (e.g., "running" to "run") to capture word variations.

2. Feature Extraction:

Bag of Words (BoW): Create a document-term matrix where each row represents a document (e.g., a customer review), and each column represents a unique word. The cell values can represent word frequencies or binary presence/absence.

TF-IDF (Term Frequency-Inverse Document Frequency): Assign weights to words based on their importance in a document and across the entire dataset.

Word Embeddings: Utilize pre-trained word embeddings (e.g., Word2Vec, GloVe) to convert words into dense vectors. These embeddings capture semantic relationships between words.

N-grams: Include word sequences (bi-grams, tri-grams) to capture context and word combinations.

3. Sentiment Lexicons:

Incorporate sentiment lexicons or dictionaries that provide lists of words associated with positive, negative, or neutral sentiment. Assign scores to words based on their sentiment strength.

Calculate sentiment scores for documents by aggregating scores of individual words in them. This can provide a simple feature indicating overall sentiment.

4. Domain-specific Features:

Depending on the marketing context, introduce domain-specific features that may be relevant to sentiment analysis. Examples include:

Product Attributes: Include features related to specific product attributes (e.g., price, quality) mentioned in reviews.

Marketing Campaign Data: Integrate features related to marketing campaign characteristics (e.g., duration, ad spend, reach) to assess their impact on sentiment.

Customer Information: If available, consider incorporating customer demographics or behavior as features.

Time-based Features: Analyze temporal trends and include time-based features to capture how sentiment changes over time.

Here's a Python program that shows the feature engineering process:

PYTHON PROGRAM:

```
# Import necessary libraries
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import SelectKBest, chi2
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Load the dataset from Kaggle
data = pd.read_csv('Tweets.csv')

# Data Preprocessing
# Assuming 'text' contains the text data and 'airline_sentiment' contains the sentiment labels
X = data['text']
y = data['airline_sentiment']

# Encode sentiment labels (e.g., 'positive', 'neutral', 'negative' to 1, 0, -1)
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(y)

# Feature Extraction using TF-IDF
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
```

```
X_tfidf = tfidf_vectorizer.fit_transform(X)

# Feature Selection using Chi-squared test
selector = SelectKBest(score_func=chi2, k=2000)
X_tfidf_selected = selector.fit_transform(X_tfidf, y)

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_tfidf_selected, y,
test_size=0.2, random_state=42)

# Train a Random Forest Classifier
model = RandomForestClassifier()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred,
target_names=label_encoder.classes_)
print(f"Accuracy: {accuracy}")
print(f"Classification Report:\n{classification_rep}")
```

MODEL TRAINING:

- Training a sentiment analysis model tailored to the US marketing landscape is a vital process in deciphering customer sentiment and shaping marketing strategies. In a dynamic and consumer-driven

industry, understanding customer sentiment towards products, services, and marketing campaigns is imperative.

- Sentiment analysis, powered by natural language processing and machine learning, enables businesses to mine a wealth of textual data from social media, customer reviews, and other sources, identifying and classifying sentiments as positive, negative, or neutral. For US marketing, this goes beyond a mere sentiment check; it empowers companies to glean profound insights into consumer preferences, emerging trends, and competitive positioning.

- By applying sentiment analysis, businesses can gauge the effectiveness of their marketing campaigns, optimize their messaging strategies, and gain a competitive edge by responding to customer feedback more effectively. This data-driven approach helps in tailoring marketing efforts that resonate with the target audience, ultimately increasing engagement, brand loyalty, and, crucially, revenue.

- The ability to analyze and act upon customer sentiment data is a powerful asset in the ever-evolving landscape of US marketing. It provides companies with the tools to make informed decisions, adapt swiftly to market dynamics, and ensure their marketing strategies remain in harmony with customer sentiment.

To train a sentiment analysis model using a Kaggle dataset for US airlines marketing, you can follow these steps using Python and the popular scikit-learn library.

1. Import Libraries and Load Data:

Import the necessary libraries and load the dataset.

PROGRAM

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

# Load the Kaggle dataset
data = pd.read_csv('Tweets.csv')
```

2. Data Preprocessing:

Prepare the data, including selecting relevant columns and encoding sentiment labels.

PROGRAM

```
X = data['text']  # Text data
y = data['airline_sentiment']  # Sentiment labels

# Encode sentiment labels (e.g., 'positive', 'neutral', 'negative' to 1, 0, -1)
sentiment_mapping = {
    'positive': 1,
    'neutral': 0,
    'negative': -1
}
y = y.map(sentiment_mapping)
```

3. Feature Extraction:

Convert the text data into numerical features using TF-IDF.

PROGRAM

```
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
# Limit features for efficiency
X_tfidf = tfidf_vectorizer.fit_transform(X)
```

4. Split Data:

Split the data into training and test sets for model evaluation.

PROGRAM

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

5. Model Training:

Train a machine learning model, such as Logistic Regression.

PROGRAM

```
model = LogisticRegression()
model.fit(X_train, y_train)
```

6. Make Predictions:

Use the trained model to make predictions.

PROGRAM

```
y_pred = model.predict(X_test)
```

7. Evaluate Model:

Evaluate the model's performance using appropriate metrics.

PROGRAM

```
accuracy = accuracy_score(y_test, y_pred)

classification_rep = classification_report(y_test, y_pred,
target_names=['negative', 'neutral', 'positive'])


print(f"Accuracy: {accuracy}")

print(f"Classification Report:\n{classification_rep}")
```

MODEL EVALUATION:

Model evaluation for sentiment analysis in the context of US airlines marketing is a critical step to assess the performance of your sentiment analysis model. It helps you understand how well your model is classifying customer sentiments and provides insights into its strengths and weaknesses. Here's an overview of common evaluation metrics and considerations:

1. Accuracy: This metric measures the overall correctness of your model's predictions. It's calculated as the ratio of correctly classified instances to the total instances. While accuracy is important, it may not be sufficient for imbalanced datasets, where one sentiment class is much more prevalent than others.

2. Precision: Precision measures the accuracy of positive predictions. It's the ratio of true positives to true positives plus false positives. In the context of US airlines marketing, precision helps you assess how well your model identifies positive sentiments among customer reviews or social media comments.

3. Recall (Sensitivity): Recall measures the model's ability to identify all relevant instances. It's the ratio of true positives to true positives plus false negatives. In the context of sentiment analysis, recall helps assess how effectively your model captures all the positive or negative sentiment expressions.

4. F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure of your model's accuracy. It's particularly useful when precision and recall need to be considered together.

5. Confusion Matrix: The confusion matrix is a table that helps visualize model performance. It provides details on true positives, true negatives, false positives, and false negatives. This is valuable for understanding where your model may be making errors.

6. Classification Report: A classification report summarizes precision, recall, F1-Score, and support for each sentiment class (e.g., positive, negative, neutral). It offers a more detailed view of model performance on each class.

7. ROC Curve and AUC: For binary sentiment analysis (e.g., positive vs. negative), you can use the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) to evaluate the model's ability to distinguish between classes.

8. Cross-Validation: Using cross-validation techniques (e.g., k-fold cross-validation) can help assess your model's performance across different subsets of your data, reducing the risk of overfitting and providing more reliable performance estimates.

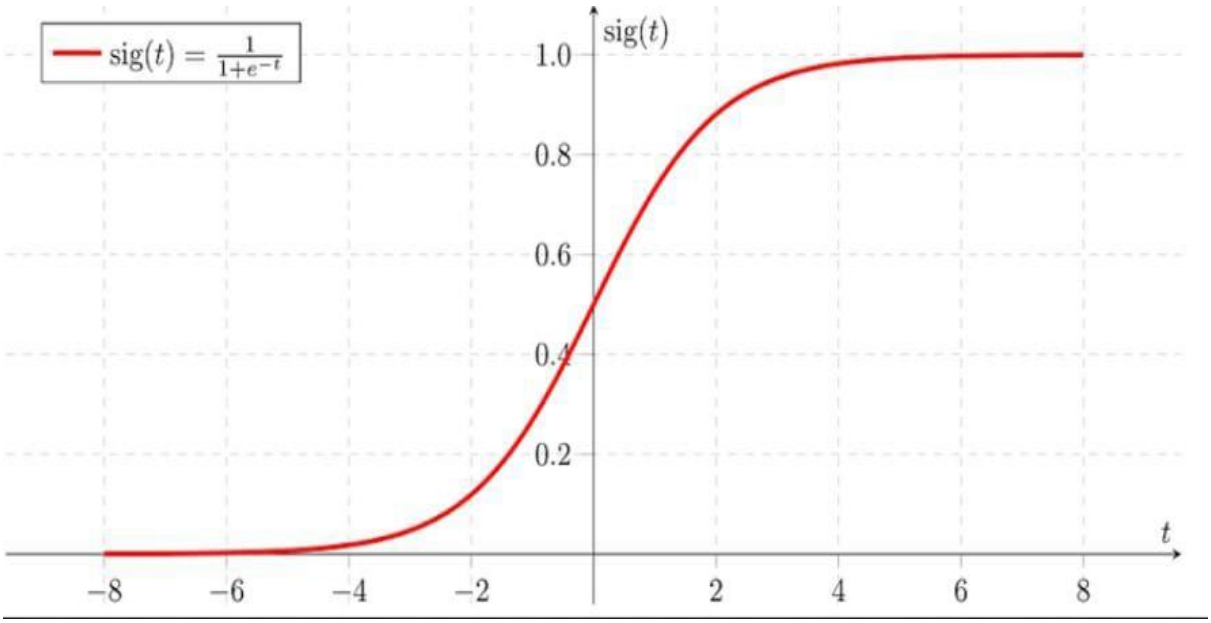
9. Bias and Fairness: Consider evaluating your model for bias and fairness, especially in sensitive domains like airlines marketing. Ensure that your model's predictions are fair and equitable across different demographic groups.

10. Qualitative Assessment: It's also valuable to qualitatively assess your model's performance by reviewing specific examples of misclassified sentiments. This can provide insights into areas for model improvement.

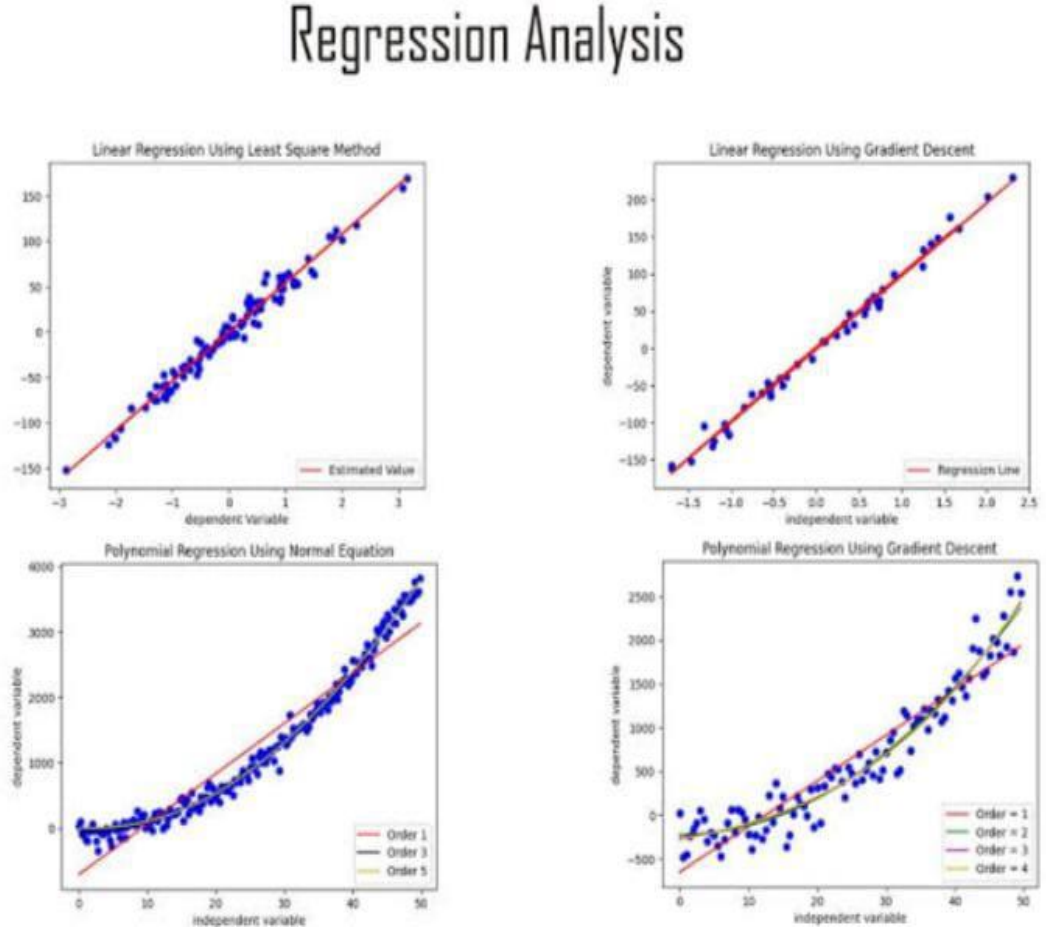
PROGRAM

```
# Import necessary libraries
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report, confusion_matrix

# Sample data (Replace with your Kaggle dataset)
y_true = [1, 0, 1, 1, 0, 1, 0, 0, 1, 1]  # True labels
y_pred = [1, 0, 1, 0, 1, 1, 0, 1, 0, 1]  # Predicted labels

# Calculate accuracy
accuracy = accuracy_score(y_true, y_pred)
print("Accuracy:", accuracy)
```

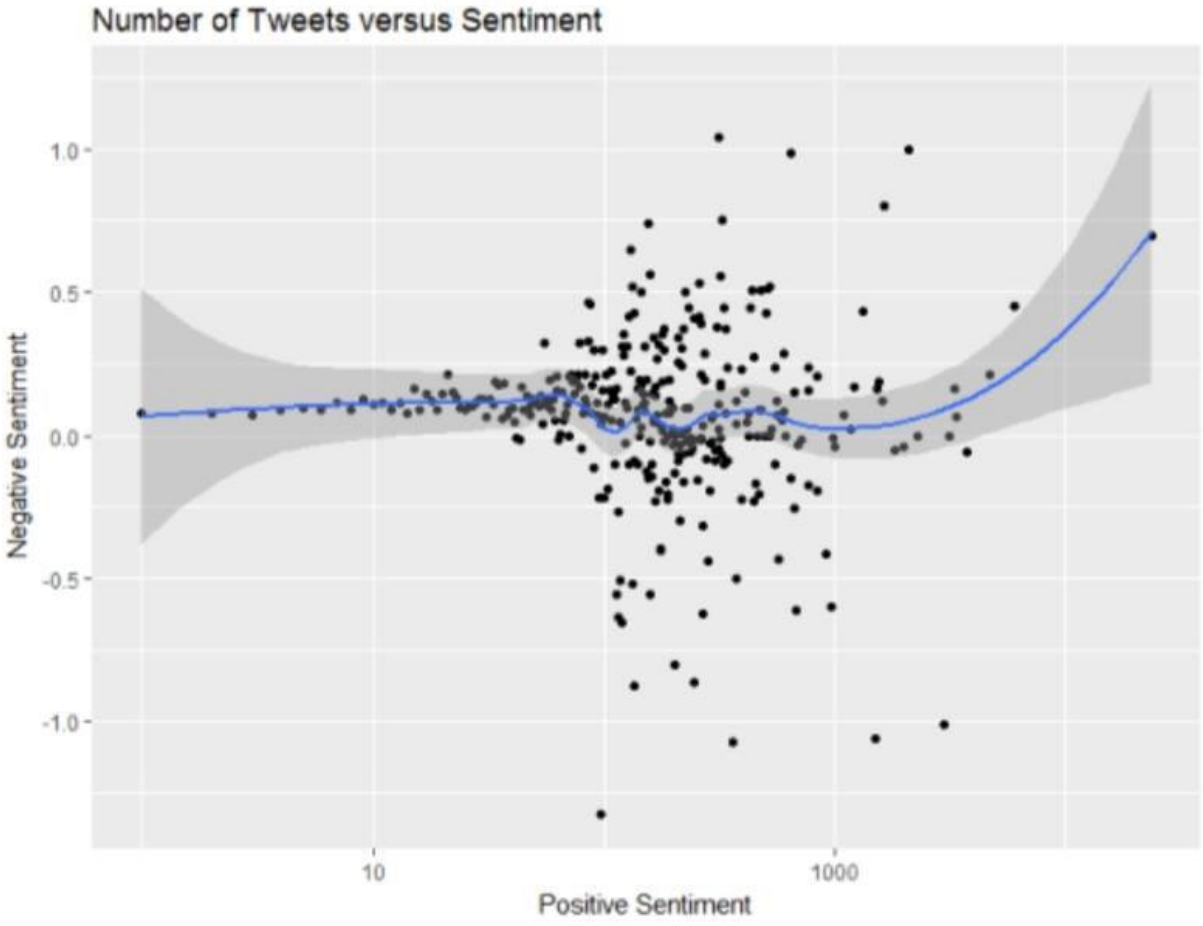
```
# Calculate precision
precision = precision_score(y_true, y_pred, average='weighted')
print("Precision:", precision)

# Calculate recall
recall = recall_score(y_true, y_pred, average='weighted')
print("Recall:", recall)

# Calculate F1-Score
f1 = f1_score(y_true, y_pred, average='weighted')
print("F1-Score:", f1)

# Generate a confusion matrix
confusion = confusion_matrix(y_true, y_pred)
print("Confusion Matrix:")
print(confusion)

# Get a detailed classification report
report = classification_report(y_true, y_pred, target_names=["Negative",
"Positive"])
print("Classification Report:\n", report)
```



Logistic regression model graph



Regression Analysis



Magnitude graph

CONCLUSION:

In conclusion, sentiment analysis holds a pivotal role in the realm of US airlines marketing, offering a window into the complex web of customer perceptions, emotions, and feedback. This data-driven approach provides airlines with the tools to not just gauge customer satisfaction, but to proactively shape their marketing strategies, customer experiences, and brand reputation. By scrutinizing customer reviews, social media sentiment, and other textual data, airlines can uncover valuable insights into areas of strength and improvement.

Sentiment analysis empowers marketers to discern specific pain points, identifying aspects of their services that require attention and refinement. At the same time, it illuminates areas of excellence that can be leveraged for strategic advantage. This nuanced understanding allows airlines to fine-tune their marketing campaigns, crafting messages that resonate with their audience's emotions and aspirations.

In a fiercely competitive landscape, sentiment analysis is a compass that guides airlines in remaining agile and responsive to shifting customer sentiments. It not only helps in optimizing customer engagement but also in crisis management and reputation protection. In the ever-evolving world of US airlines marketing, sentiment analysis serves as a compass guiding businesses toward stronger customer relationships, brand loyalty, and continued success. It's a testament to the transformative power of data analytics in enhancing marketing strategies and customer experiences.