Sentiment Analysis and Visualization in Social Media Data

Objective:

To analyze and visualize sentiment patterns in social media data to understand public opinion and attitudes towards specific topics or brands using the Twitter Sentiment Analysis Dataset.

Dataset:

- **Training Data:** `twitter_training.csv`
- **Validation Data:** `twitter_validation.csv`

Data Preprocessing:

1. **Loading the Data:**
    - Loaded the training and validation datasets.
    - Renamed columns for clarity.
2. **Handling Missing Values:**
    - I checked for missing values.
    - Handled missing data by removing records with missing text values.
3. **Tokenization and Text Preprocessing:**
    - Tokenized the text and removed stopwords and punctuation.
    - Lemmatized the tokens to normalize the text.

Exploratory Data Analysis (EDA):

1. **Distribution of IDs and Game Types:**
    - Plotted bar charts show the distribution of IDs and game types in the dataset.
2. **Sentiment Distribution Across Games:**
    - I visualized the sentiment distribution for different games using bar plots and cat plots.
3. **Missing data visualization:**
    - Identified and visualized missing data using heatmaps.

Sentiment Distribution:

1. **Pie Charts:**
    - Plotted pie charts show the distribution of sentiments in both training and validation datasets.

Entity Sentiment Analysis:

1. **Sentiment Distribution for Top Entities:**
    - Grouped data by game and sentiment to analyze sentiment distribution for the top 10 games and entities.
    - Created stacked bar charts for both training and validation datasets.

Common Words Analysis:

1. **Most Common Words:**
    - extracted and displayed the most common words for each sentiment (positive, negative, or neutral) in the training data.
2. **Word Clouds:**
    - Generated word clouds for visual representation of common words in each sentiment category.

TF-IDF Vectorization:

1. **TF-IDF Vectorization:**
    - Combined text columns for TF-IDF vectorization.
    - Transformed the text data into TF-IDF vectors.

Model Training and Evaluation:

1. **Logistic Regression Classifier:**
    - Trained a logistic regression classifier on the training data.
    - Evaluated the classifier on the validation data.
2. **Performance Metrics:**

Classification Report:

	precision	recall	f1-score	support
Irrelevant	0.82	0.73	0.77	171
Negative	0.79	0.88	0.83	266
Neutral	0.86	0.77	0.81	285
Positive	0.80	0.86	0.83	277
accuracy			0.82	999
macro avg	0.82	0.81	0.81	999
weighted avg	0.82	0.82	0.82	999

Confusion Matrix:

Actual/Predicted	Positive	Negative	Neutral	Irrelevant
Positive	217	12	7	14
Negative	8	181	6	15
Neutral	22	16	153	20
Irrelevant	11	9	14	161

Visualization of Model Performance:

1. **Confusion Matrix Heatmap:**
    - Generated a heatmap for the confusion matrix to visually inspect the model performance.

Sentiment Co-occurrence Analysis:

1. **Pivot Table and Heatmap:**
    - I created a pivot table and normalized the sentiment distribution to analyze co-occurrences.

    - I plotted a heatmap to visualize the sentiment co-occurrence by game/entity.

Conclusion:

The analysis provides a detailed understanding of sentiment patterns across different games and entities in social media data. The visualizations and model performance metrics indicate that the approach effectively captures public opinion and attitudes towards specific topics or brands.

Overall Results:

- Successfully preprocessed and analyzed the Twitter Sentiment Analysis Dataset.
- Generated comprehensive visualizations to understand sentiment distribution and patterns.
- I trained and evaluated a logistic regression classifier, achieving an accuracy of 82%.
- Identified common words and their sentiment associations through word clouds.
- Provided actionable insights and recommendations based on the analysis.