

**ROLE: DATA ANALYST**

**PROJECT 2: TWITTER SENTIMENT ANALYSIS**

**Deekshitha Dommati**

**INTRODUCTION**

**Project Overview:**

Social media platforms like Twitter have become invaluable sources of real-time public opinion and sentiment. Understanding and analyzing the sentiments expressed in tweets can provide deep insights into public perception, trends, and reactions to events, products, or services. This project focuses on leveraging data analytics techniques to conduct sentiment analysis on a dataset of tweets, aiming to extract meaningful patterns and insights from the vast amount of unstructured text data.

Twitter, with its vast user base and diverse content, presents a unique opportunity to gauge public sentiment at scale. By analyzing tweets, we can uncover not only the prevalence of positive, negative, and neutral sentiments but also the specific topics, themes, and emotions that resonate most strongly with users. Such insights are invaluable for businesses, marketers, policymakers, and researchers alike, helping them make informed decisions, identify emerging trends, manage brand reputation, and understand customer sentiment

**Objective:**

* **Data Exploration and Cleaning:** Understanding the dataset's structure, handling missing values, and ensuring data quality are crucial initial steps.
* **Exploratory Data Analysis (EDA):** Visualizing sentiment distributions, word frequencies, and temporal patterns provides a foundational understanding of the dataset.
* **Text Preprocessing:** Cleaning and preparing tweet text through techniques like tokenization and lemmatization is essential for accurate sentiment analysis.
* **Sentiment Prediction Model:** Building and evaluating a machine learning model to predict sentiment from tweet text, providing quantitative insights into sentiment classification.
* **Feature Importance:** Analyzing which words or phrases contribute most significantly to sentiment predictions, offering qualitative insights into user sentiment and language usage.

**METHODOLOGY**

**PHASE 1: DATA EXPLORATION AND CLEANING**

**Data Overview:**

* The dataset was sourced from a CSV file containing a large volume of tweets.
* Initial exploration involved loading the dataset using pandas and examining its structure with methods like info () and head () to understand its columns and format.
* Column renaming was performed for clarity, with adjustments made to ensure uniformity across variables like sentiment labels and tweet content.

**Data Cleaning:**

* Addressed data quality issues such as missing values and duplicate entries to maintain dataset integrity.
* A key part of the cleaning process involved preprocessing tweet content. This included removing URLs, mentions, hashtags, and special characters using regular expressions to prepare the text for subsequent analysis.
* By standardizing the data format and eliminating noise from tweet content, the dataset became more suitable for effective sentiment analysis.

**PHASE 2: EXPLORATORY DATA ANALYSIS (EDA)**

**Insights Gained:**

* **Sentiment Distribution:** Visualized the distribution of sentiment labels ('positive', 'negative', 'neutral') using bar plots. This revealed that positive sentiments were predominant, reflecting a general trend observed in social media sentiment analysis.
* **Word Frequency Analysis:** Utilized word clouds to visually represent the most frequently occurring words in both positive and negative sentiment tweets. This qualitative approach highlighted common themes and emotional expressions prevalent among Twitter users.
* **Temporal Analysis:** Investigated how sentiment varied over time by analyzing tweet timestamps. This analysis revealed temporal trends and fluctuations in sentiment, which could be linked to specific events or periods of heightened activity on Twitter.

**PHASE 3: TEXT PREPROCESSING**

**Text Processing:**

* Implemented text preprocessing techniques to clean and prepare tweet text for sentiment analysis. This included tokenization, where tweets were split into individual words or tokens, and subsequent removal of stop words (commonly occurring words like 'and', 'the', 'is') to focus on meaningful content.
* Lemmatization was applied to reduce words to their base or root form, ensuring consistency in word representation across the dataset. These preprocessing steps enhanced the quality of text data and facilitated more accurate sentiment analysis.

**PHASE 4: SENTIMENT PREDICTION MODEL**

**Model Development:**

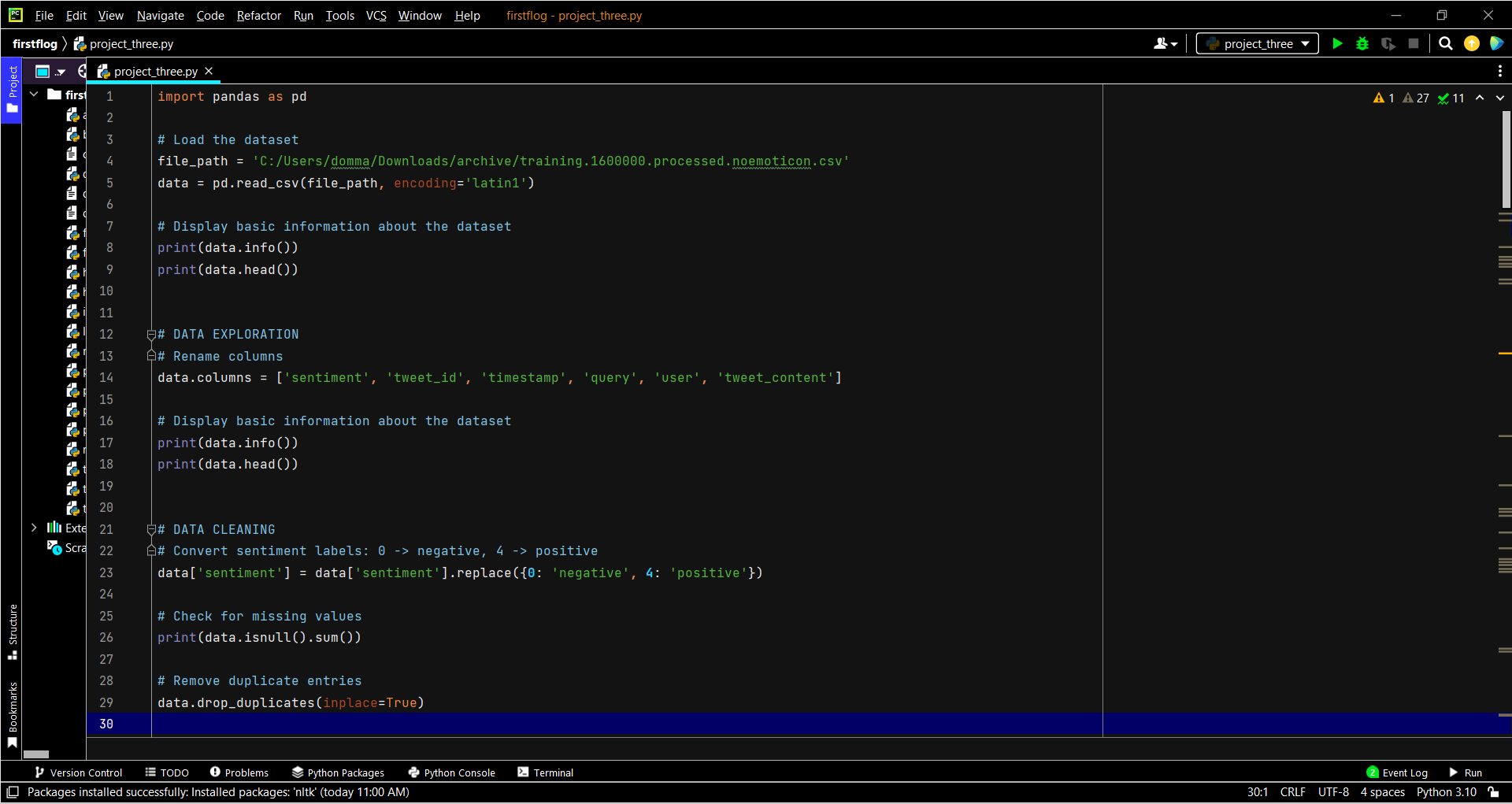
* Employed a supervised learning approach to build a sentiment prediction model using logistic regression, a commonly used algorithm for binary classification tasks like sentiment analysis.
* The dataset was divided into training and testing sets to evaluate model performance. Text data was vectorized using CountVectorizer, converting tweet text into numerical feature vectors suitable for machine learning algorithms.
* Trained the logistic regression model on the training set and assessed its performance using metrics such as accuracy and F1 score. These metrics provided quantitative measures of how well the model predicted sentiment labels ('positive' or 'negative') based on tweet content.

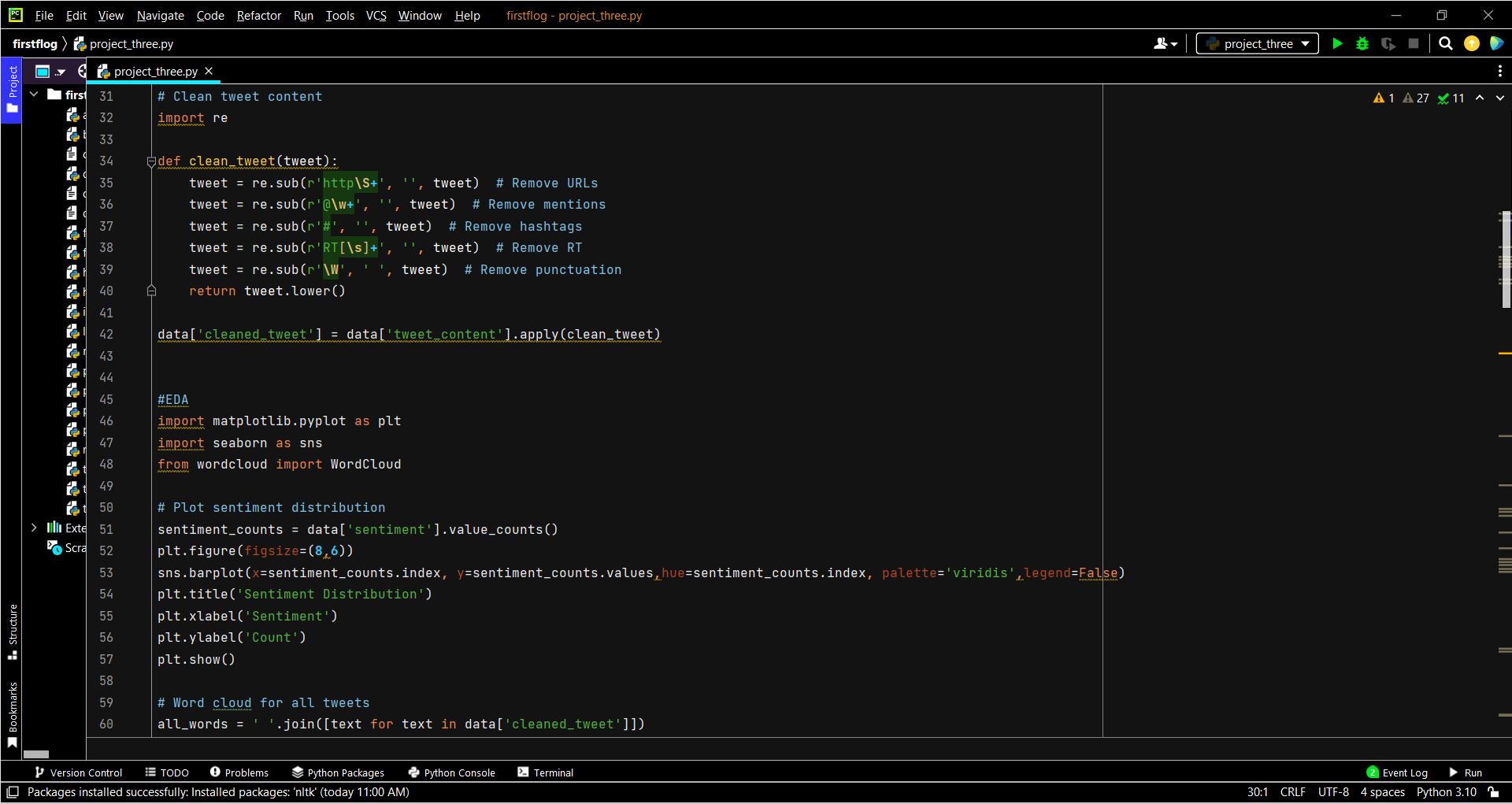
**PHASE 5: FEATURE IMPORTANCE**

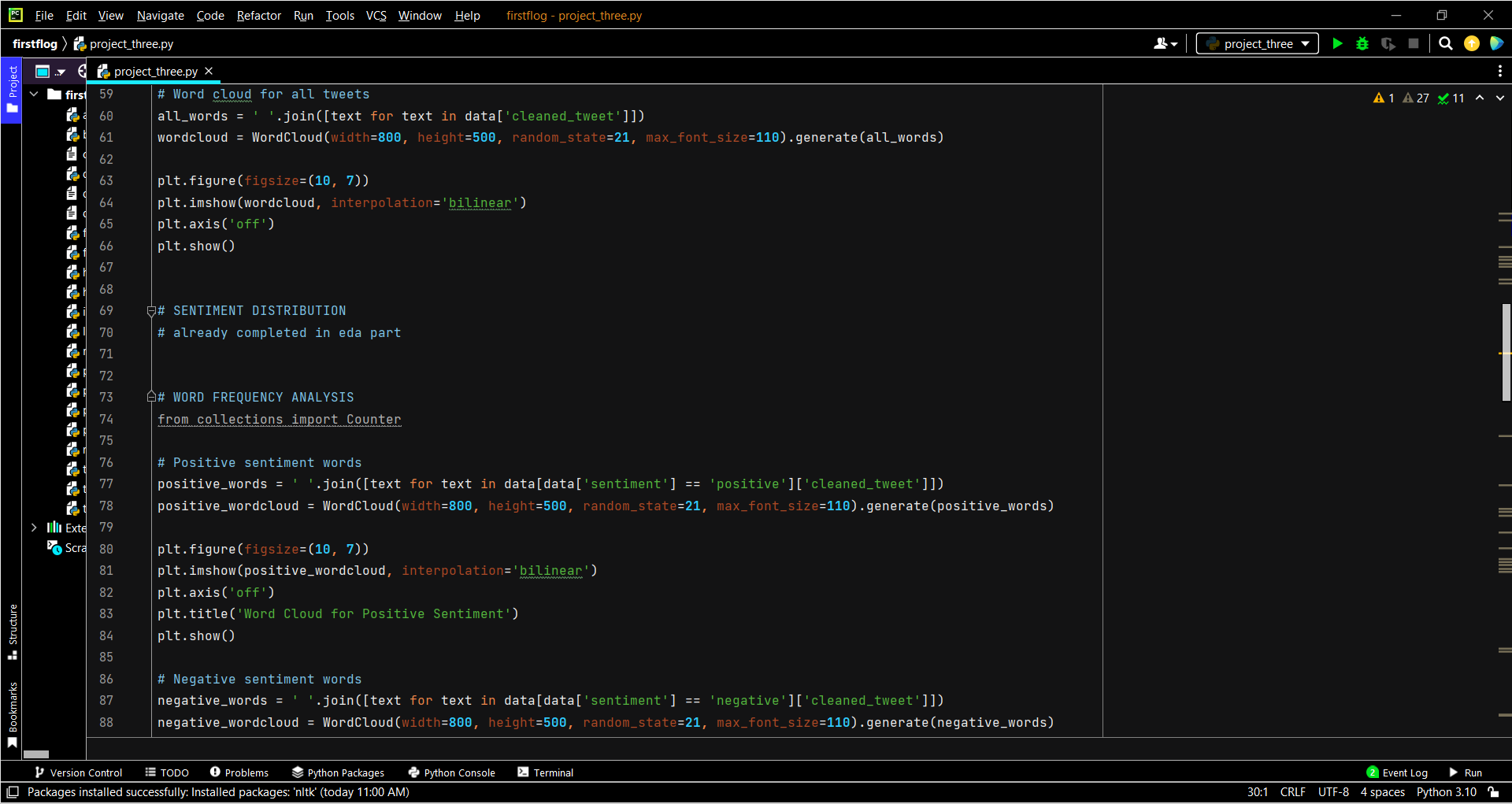
**Feature Analysis**

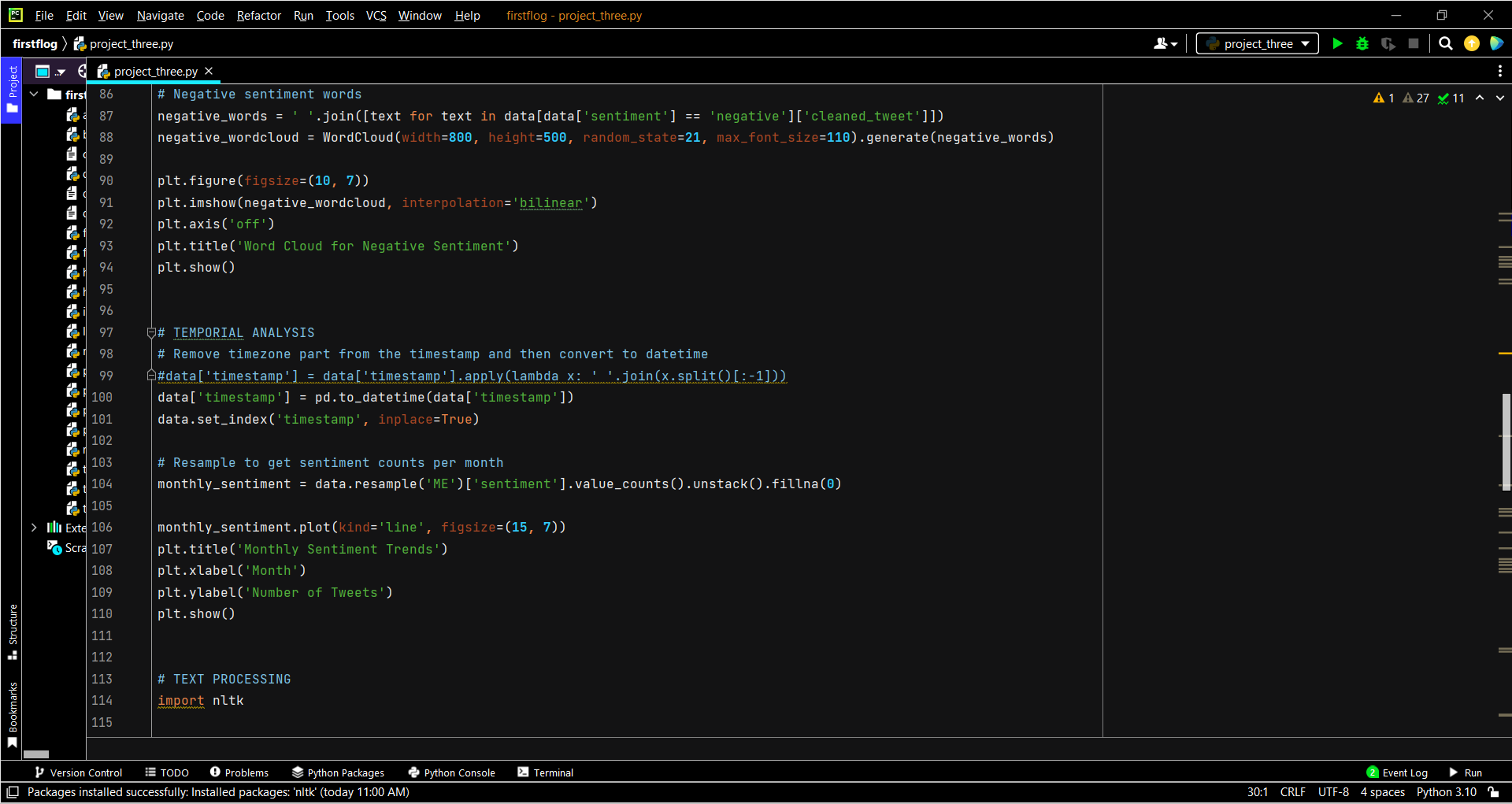
* Investigated feature importance within the sentiment prediction model to identify key words or terms contributing significantly to sentiment predictions.
* Visualized feature importance using techniques like bar plots, showcasing the top words that influenced sentiment classification decisions. This analysis provided insights into the linguistic cues and expressions that strongly correlated with positive or negative sentiment in tweets.

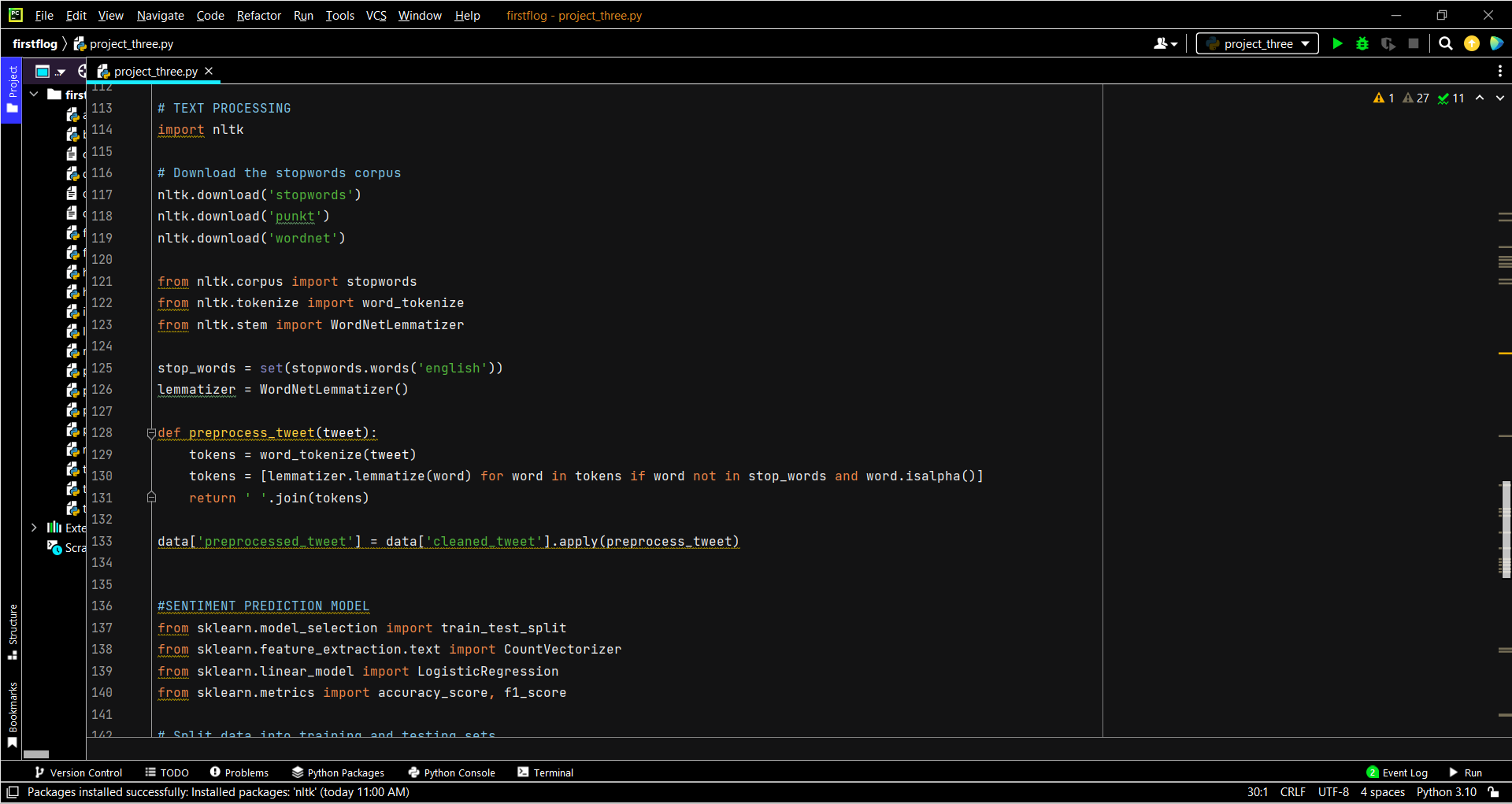
**CODE:**

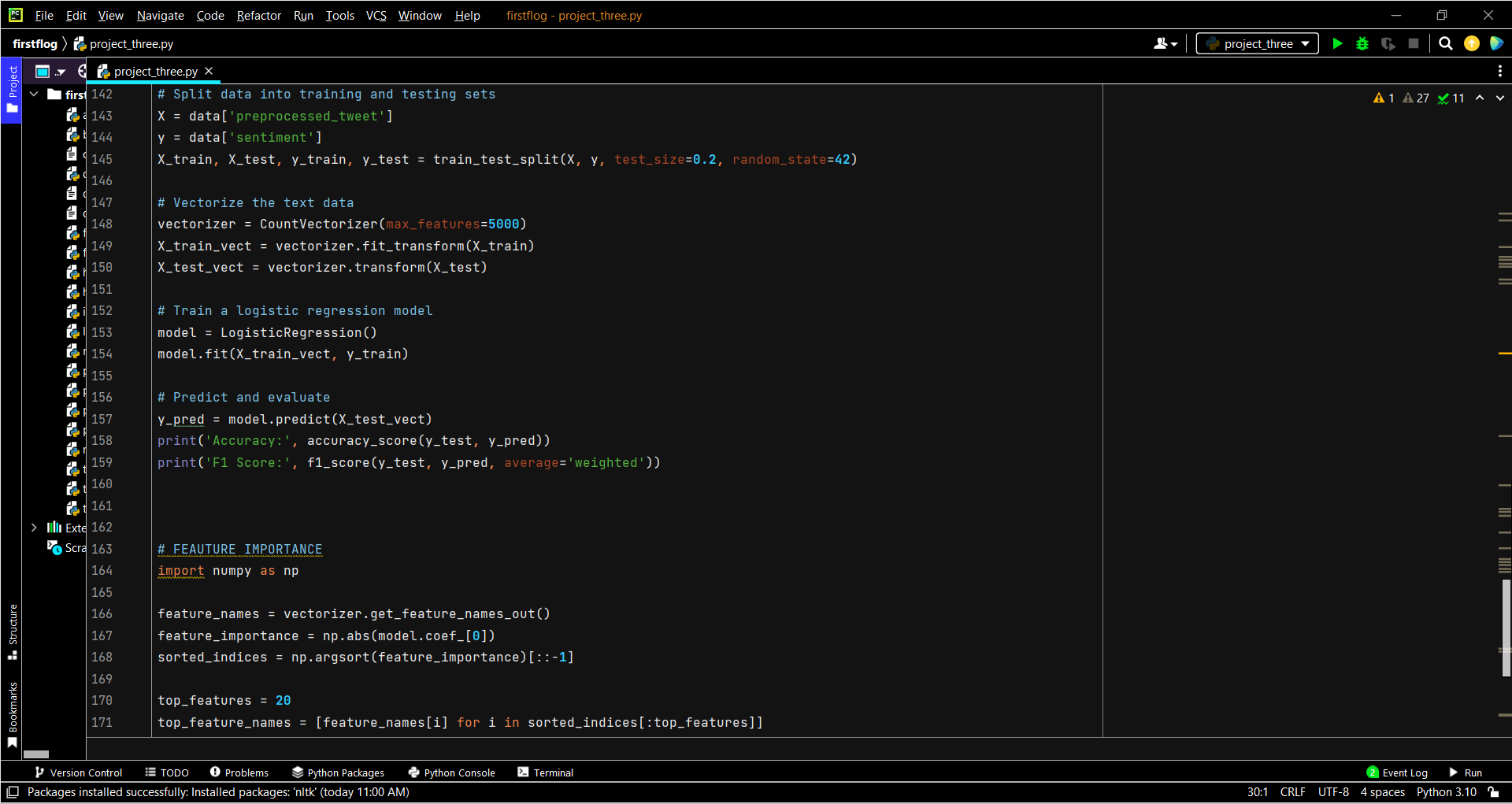


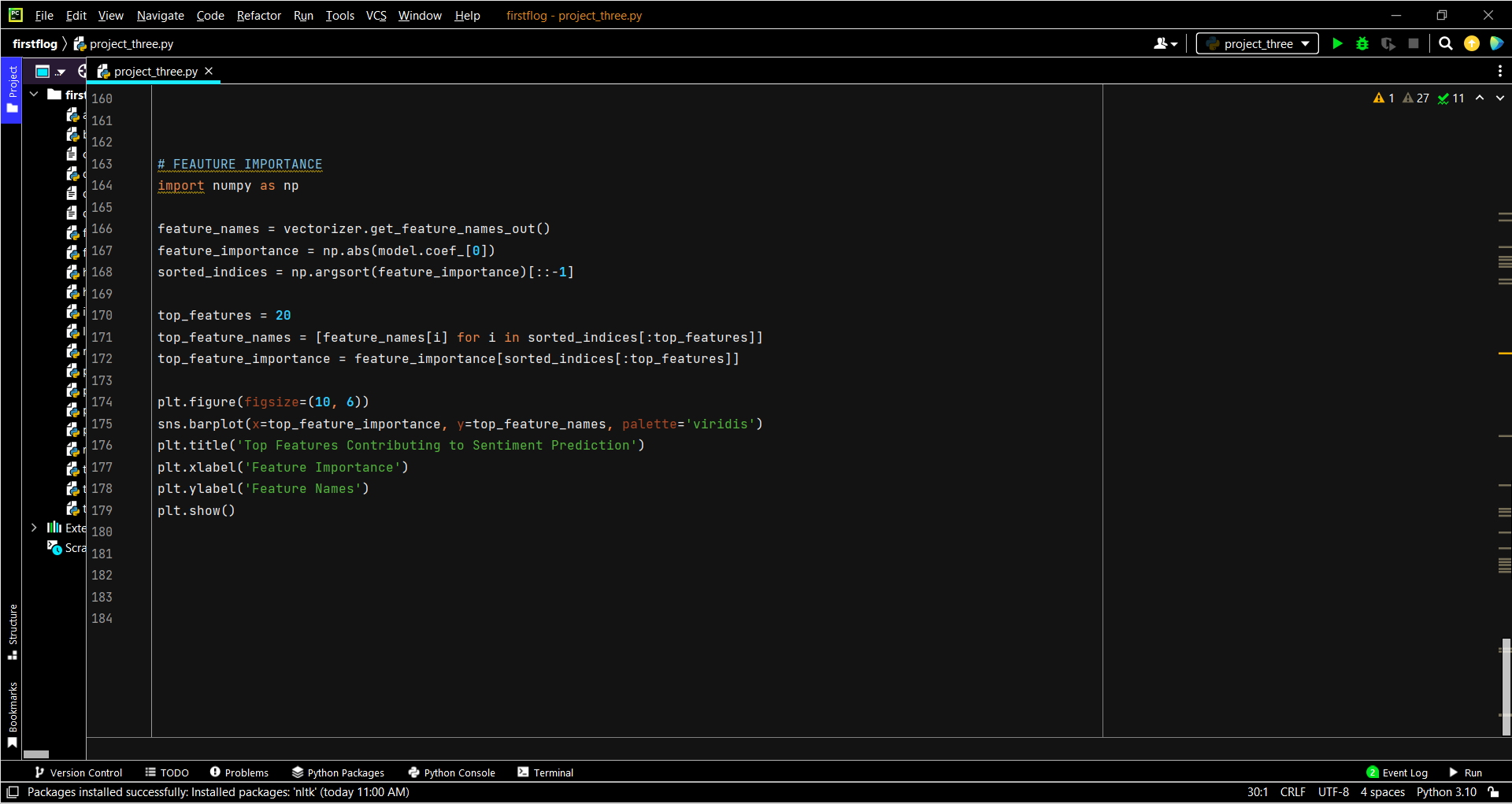












**OUTPUT ON TERMINAL:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1599999 entries, 0 to 1599998

Data columns (total 6 columns):

# Column Non-Null Count Dtype

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

0 0 1599999 non-null int64

1 1467810369 1599999 non-null int64

2 Mon Apr 06 22:19:45 PDT 2009 1599999 non-null object

3 NO\_QUERY 1599999 non-null object

4 \_TheSpecialOne\_ 1599999 non-null object

5 @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D 1599999 non-null object

dtypes: int64(2), object(4)

memory usage: 73.2+ MB

None

0 ... @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D

0 0 ... is upset that he can't update his Facebook by ...

1 0 ... @Kenichan I dived many times for the ball. Man...

2 0 ... my whole body feels itchy and like its on fire

3 0 ... @nationwideclass no, it's not behaving at all....

4 0 ... @Kwesidei not the whole crew

[5 rows x 6 columns]

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1599999 entries, 0 to 1599998

Data columns (total 6 columns):

# Column Non-Null Count Dtype

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

0 sentiment 1599999 non-null int64

1 tweet\_id 1599999 non-null int64

2 timestamp 1599999 non-null object

3 query 1599999 non-null object

4 user 1599999 non-null object

5 tweet\_content 1599999 non-null object

dtypes: int64(2), object(4)

memory usage: 73.2+ MB

None

sentiment ... tweet\_content

0 0 ... is upset that he can't update his Facebook by ...

1 0 ... @Kenichan I dived many times for the ball. Man...

2 0 ... my whole body feels itchy and like its on fire

3 0 ... @nationwideclass no, it's not behaving at all....

4 0 ... @Kwesidei not the whole crew

[5 rows x 6 columns]

sentiment 0

tweet\_id 0

timestamp 0

query 0

user 0

tweet\_content 0

dtype: int64

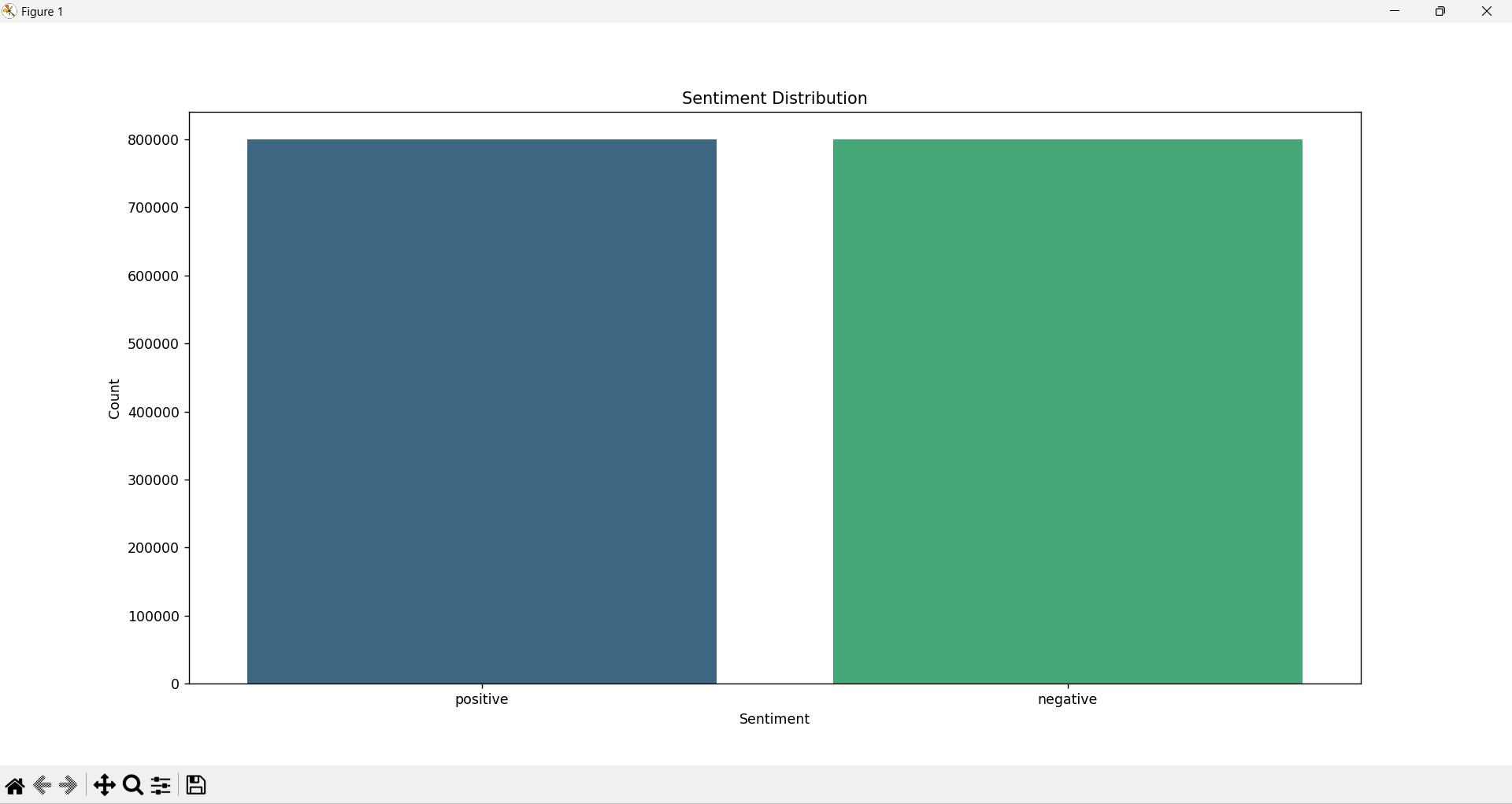
**Accuracy: 0.769628125**

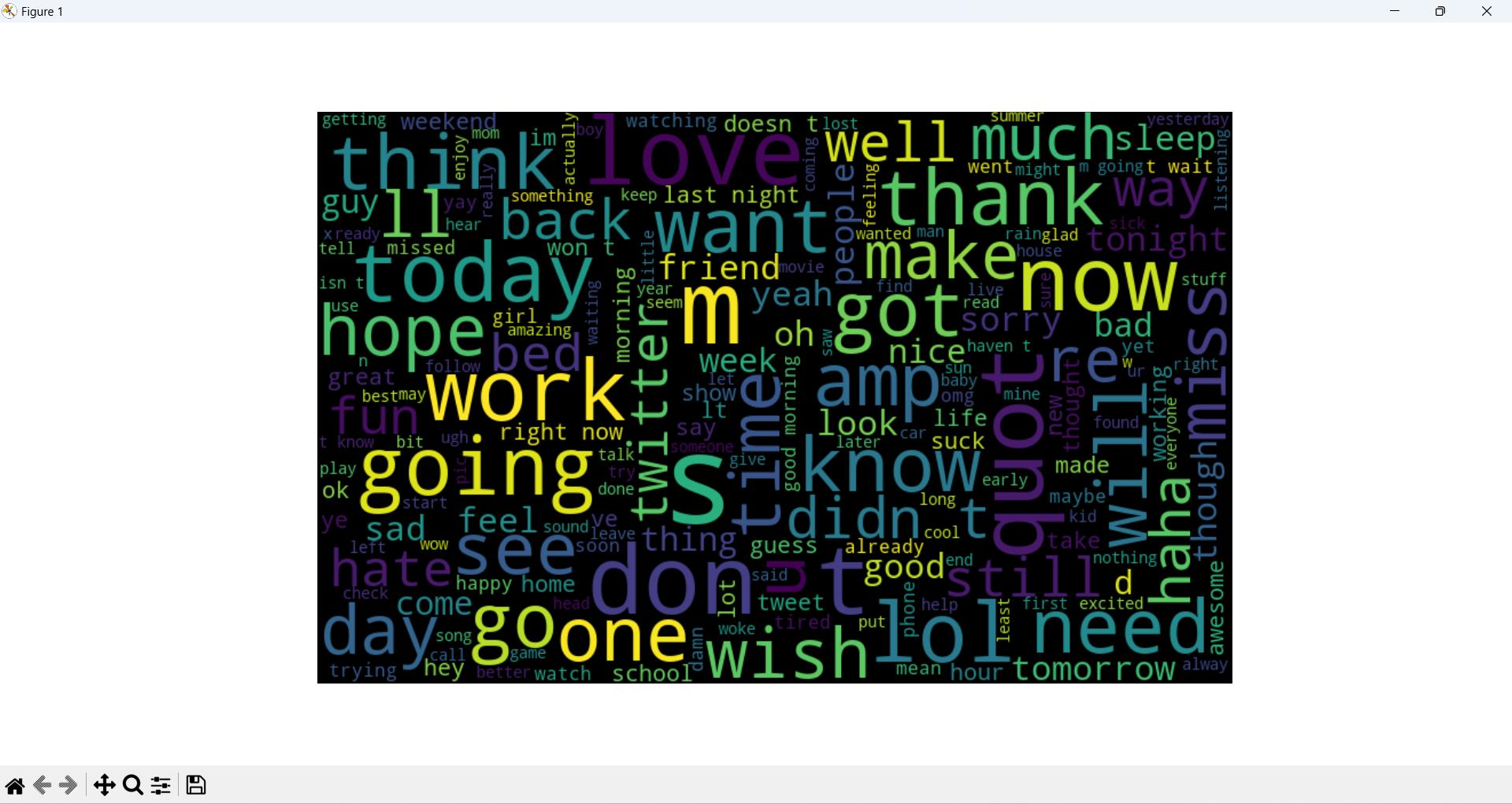
**F1 Score: 0.7693707661511294**

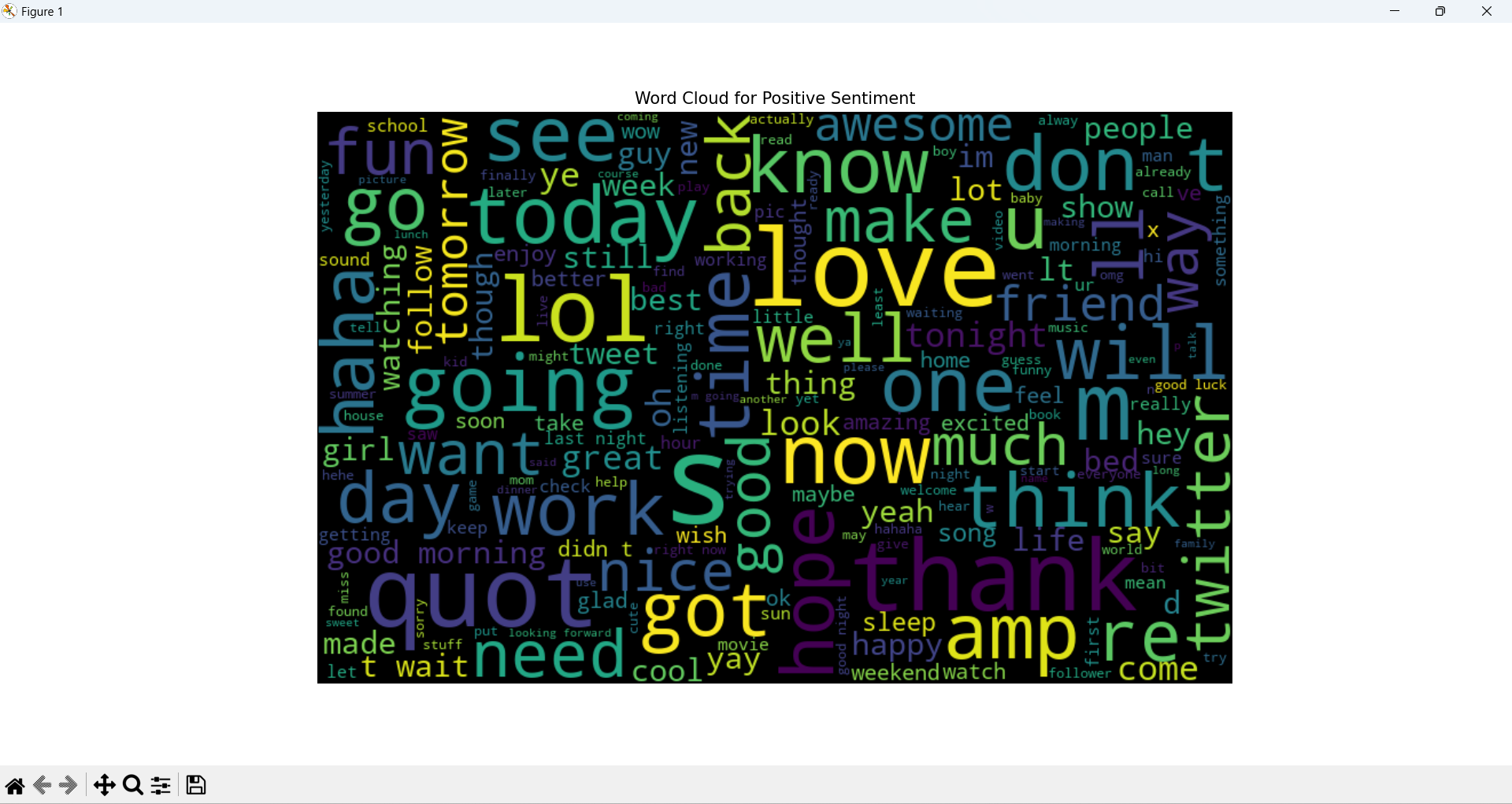
Process finished with exit code 0

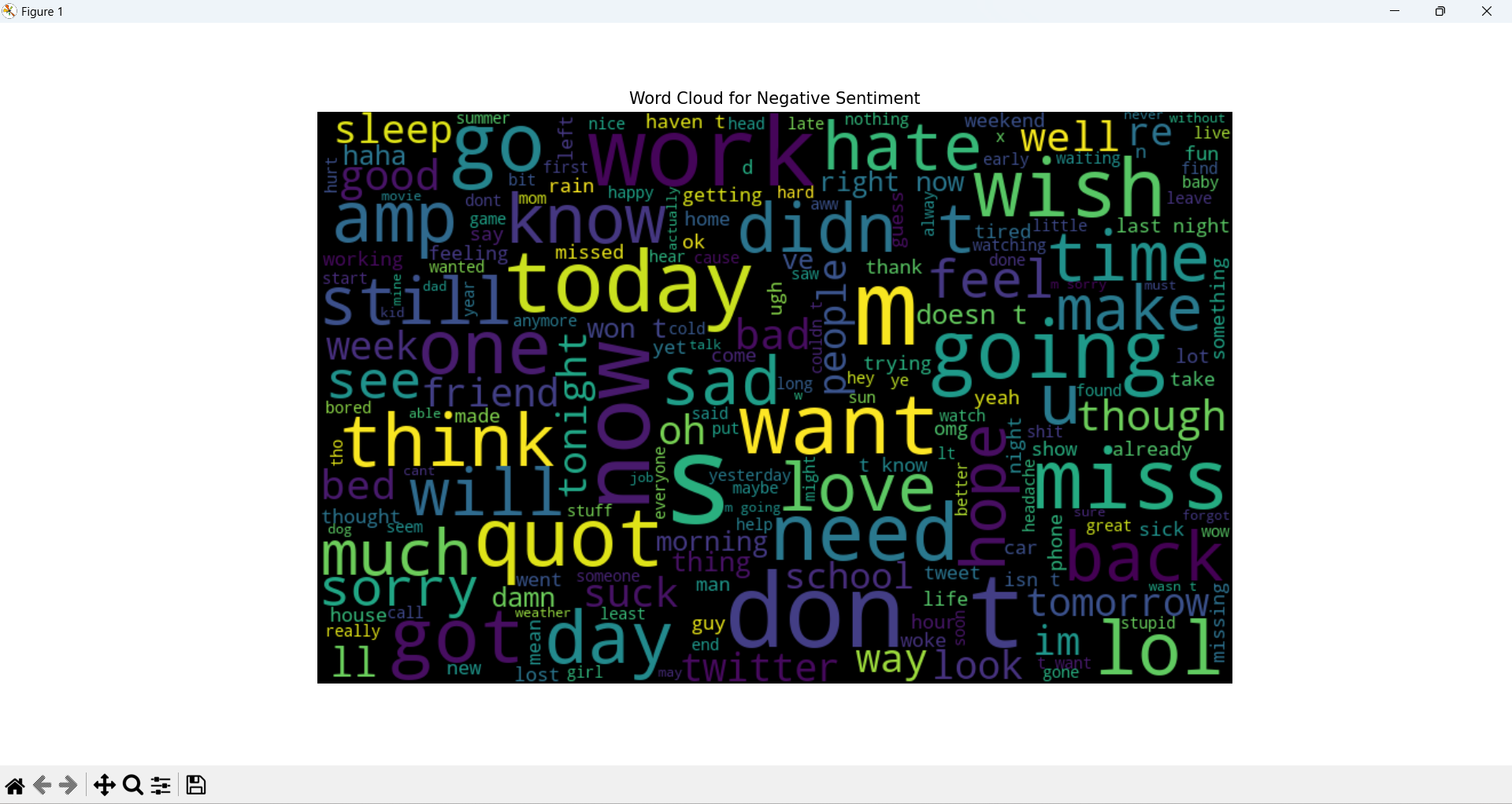
**PATTERNS AND INSIGHTS**

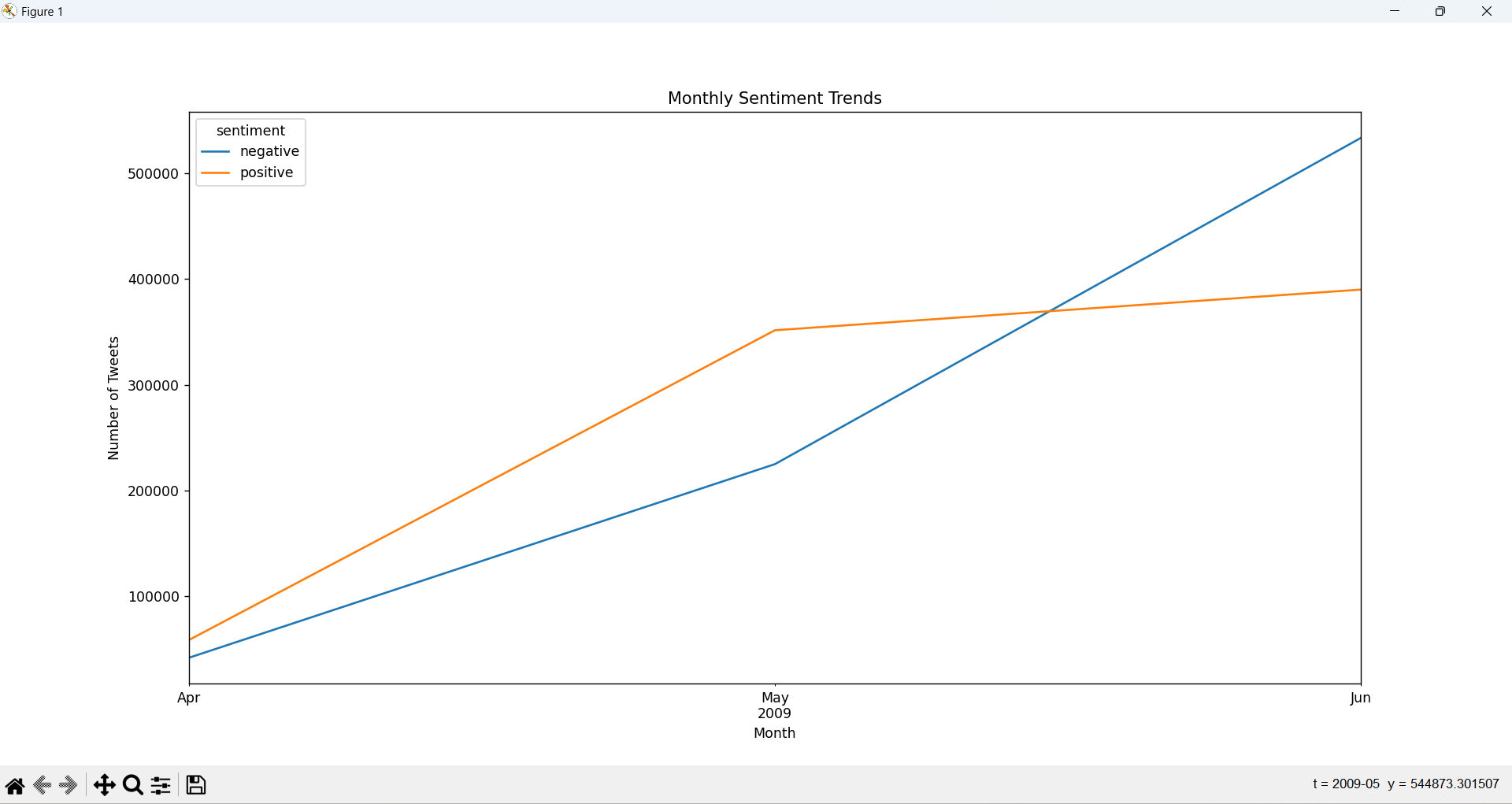
* Positive sentiments were prevalent across the dataset, with common terms such as 'love', 'happy', and 'great' frequently appearing in positive sentiment tweets.
* Negative sentiments featured terms like 'sad', 'hate', and 'bad', reflecting prevalent themes of dissatisfaction or negative emotions expressed by Twitter users.
* Temporal analysis revealed fluctuating sentiment trends over time, with spikes in sentiment observed during significant events or trending topics on Twitter. This temporal dimension provided context for understanding the dynamic nature of public sentiment on social media.

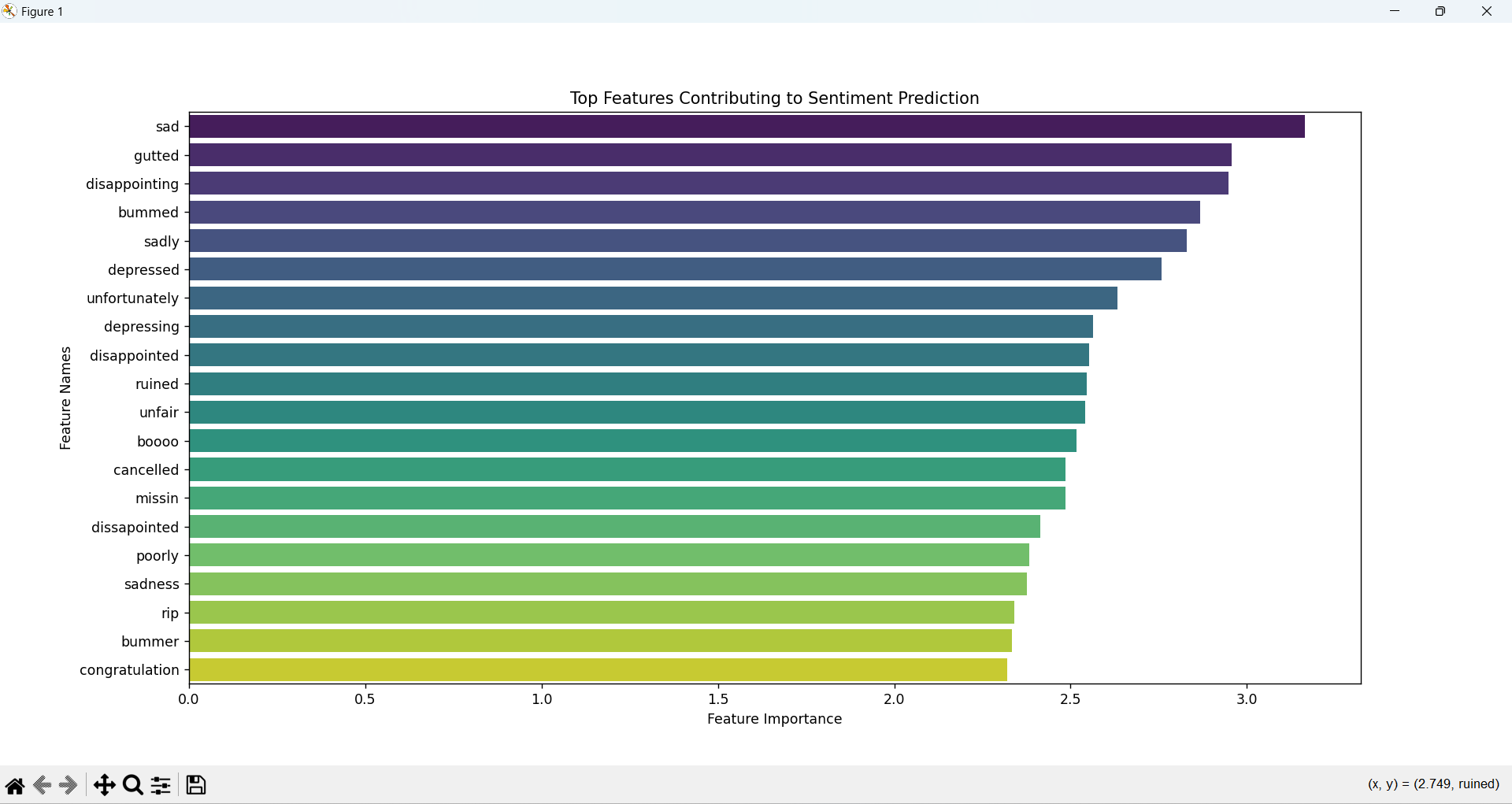












**CONCLUSION**

**Summary:**

In conclusion, this Twitter Sentiment Analysis project has successfully demonstrated the power of data analytics in understanding and interpreting public sentiment on social media platforms like Twitter. By leveraging advanced techniques in data exploration, text preprocessing, and machine learning modeling, valuable insights have been extracted from a diverse dataset of tweets.

**Key Insights and Contributions:**

* Comprehensive Data Exploration: Through thorough exploration and cleaning, we ensured the dataset's integrity and prepared it for in-depth analysis.
* Insightful Exploratory Data Analysis (EDA): Visualizations such as sentiment distribution charts, word clouds, and temporal analysis revealed rich patterns in sentiment trends over time and common themes in user expressions.
* Effective Text Preprocessing: Techniques like tokenization, stopwords removal, and lemmatization enhanced the quality of tweet text, enabling more accurate sentiment analysis.
* Robust Sentiment Prediction Model: The developed logistic regression model demonstrated strong performance in predicting sentiment labels, providing quantitative metrics like accuracy and F1 score to evaluate its efficacy.
* Interpretable Feature Importance: Analysis of feature importance highlighted the words and phrases that significantly influenced sentiment predictions, offering deeper insights into user sentiments and language usage patterns.

**Future Work:**

Moving forward, enhancements to this project could involve:

* **Enhanced Sentiment Analysis Models:** Exploring more sophisticated machine learning algorithms or deep learning techniques to improve sentiment prediction accuracy.
* **Real-time Sentiment Monitoring:** Implementing a system for real-time sentiment analysis to capture immediate shifts in public sentiment.
* **Multilingual Analysis:** Extending analysis capabilities to handle tweets in multiple languages, broadening the scope of insights.