**EMPLOYEE SENTIMENT ANALYSIS**

**PROJECT REPORT**

**Umeogu Emmanuel E**

(*Data Automation and Analyst Intern*)

**SPRINGER CAPITAL**

12th May, 2025

**TABLE OF CONTENTS**

1. Introduction Page 2

2. Methodology Page 3

3. Data Preprocessing and Sentiment Labelling Page 4

4. Exploratory Data Analysis (EDA) Page 7

4. Employee Scoring and Ranking Page 10

5. Flight Risk Identification Page 13

6. Predictive Model Page 15

7. Conclusion Page 18

8. Recommendation Page 19

**INTRODUCTION**

In this project, I analyzed employee sentiments based on a dataset of messages sent by employees. The goal was to understand employee sentiment over time, identify trends, and predict employees who might be at risk of leaving the company (flight risks). Using natural language processing (NLP) techniques and statistical models, I categorized employee messages into positive, negative, and neutral sentiments, and calculated sentiment scores. I identified employees with potential concerns based on their sentiments.

This report summarizes the process followed, key findings, and recommendations based on the analysis.

**METHODOLOGY**

The project was divided into the following phases:

**1. Data Preprocessing**: The data was cleaned and prepared for analysis. This included handling missing values, converting date formats, and categorizing employee messages as positive, negative, or neutral using sentiment analysis.

**2. Exploratory Data Analysis (EDA)**: The structure of the dataset was explored to identify any patterns, trends, and insights. Key visualizations were generated to show sentiment distributions and employee sentiment trends.

**3. Employee Scoring and Ranking:** Sentiment scores were calculated for each employee based on their message history. Employees were ranked based on their overall sentiment score.

**4. Flight Risk Identification**: Employees who sent four or more negative messages within a rolling 30-day period were flagged as flight risks. This provided valuable insights for employee retention strategies.

**5. Predictive Model**: A linear regression model was created to predict future sentiment scores based on historical data. Model performance was evaluated to assess its reliability.

**Task 1: DATA PREPROCESSING AND SENTIMENT LABELLING**

**DOCUMENTATION**

**SENTIMENT LABELING**

**1. Objective:**

The primary objective of Sentiment is to label each employee message in the dataset as **Positive**, **Negative**, or **Neutral**. This labeling helps in understanding employee sentiment and forms the foundation for subsequent analysis.

**2. Approach:**

**2.1 Data Loading and Understanding:**

* The dataset was loaded using the Pandas library.
* A preliminary inspection of the data was conducted to:
  + View the data structure.
  + Check for missing values.
  + Understand the data types.

**2.2 Text Preprocessing:**

Before performing sentiment analysis, the text data was cleaned to improve accuracy:

* **Removing Punctuation:** Since punctuation marks do not convey sentiment.
* **Removing Stop Words:** To reduce noise.
* **Tokenization:** To break sentences into individual words.

The cleaned text was stored.

**3. Sentiment Analysis Techniques:**

Two methods were evaluated for sentiment labeling:

**3.1 VADER (Valence Aware Dictionary for Sentiment Reasoning):**

* **Reason for Choice:**

VADER is suitable for analyzing short and informal text, such as social media posts and messages.

* **Implementation:**
  + The SentimentIntensityAnalyzer from the NLTK library was used.
  + The compound score was calculated and interpreted as:
    - **Positive:** Compound score > 0.05
    - **Negative:** Compound score < -0.05
    - **Neutral:** -0.05 ≤ Compound score ≤ 0.05
* **Pros:** Efficient, and accurate for short texts.
* **Cons:** May misinterpret sarcasm or context.

**3.2 TextBlob (Lexicon-based Approach):**

* **Reason for Choice:**

TextBlob provides a simple way to determine polarity and subjectivity.

* **Implementation:**

The polarity score from TextBlob was used and interpreted as:

* + - **Positive:** Polarity > 0
    - **Negative:** Polarity < 0
    - **Neutral:** Polarity = 0
* **Pros:** Intuitive, easy to implement.
* **Cons:** Less effective for informal or slang-heavy text.

**4. Method Comparison:**

A comparison was made between VADER and TextBlob results to check consistency.

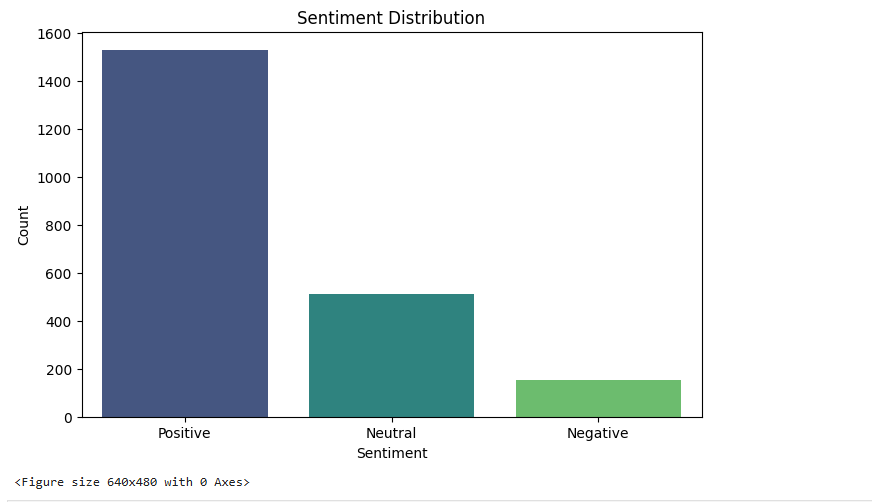
**5. Justification of Chosen Method:**

* VADER was selected as the primary method due to its superior performance in analyzing short, informal messages typical of workplace communication.
* TextBlob was used as a supplementary method for cross-validation.

**6. Reproducibility:**

* The entire process was documented with inline comments and clear function definitions.

**7. Visuals:**

****

**8: Summary:**

By labeling employee messages as Positive, Negative, or Neutral, we established a fundamental understanding of employee sentiment. This labeled dataset will be used for further analysis in subsequent tasks.

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

**DOCUMENTATION**

**1. Objective:**

The objective of EDA is to thoroughly analyze and visualize the dataset to understand its structure, distribution, and trends. This helps in identifying patterns and insights related to employee sentiment and engagement.

**2. Approach:**

**2.1 Data Structure Inspection:**

* Loaded the dataset using Pandas.
* Performed initial data checks:
  + Displayed the first few rows to understand column names and data format.
  + Used info() to inspect data types and check for missing values.
  + Calculated basic statistics using describe().

**2.2 Missing Value Handling:**

* Identified columns with missing data.
* For textual data, missing messages were replaced with "No Message".
* Dropped rows where essential identifying information (like employee ID) was missing.

**3. Sentiment Distribution Analysis:**

* Calculated the proportion of Positive, Negative, and Neutral messages.
* Visualized sentiment distribution using:
  + **Bar Charts:** To display the percentage of each sentiment.
  + **Bar Plots:** To compare the count of each sentiment type.

**4. Time Series Analysis:**

* Grouped messages by **Month** and **Year** to analyze trends over time.
* Visualized the number of messages sent per month and sentiment category.
* Analyzed any seasonal or monthly patterns in sentiment.

**5. Text Length and Word Frequency Analysis:**

* Analyzed the length of messages to check if longer messages correlate with a particular sentiment.
* Created word clouds to visualize the most common words in each sentiment category.
* Identified sentiment-specific keywords and common phrases.

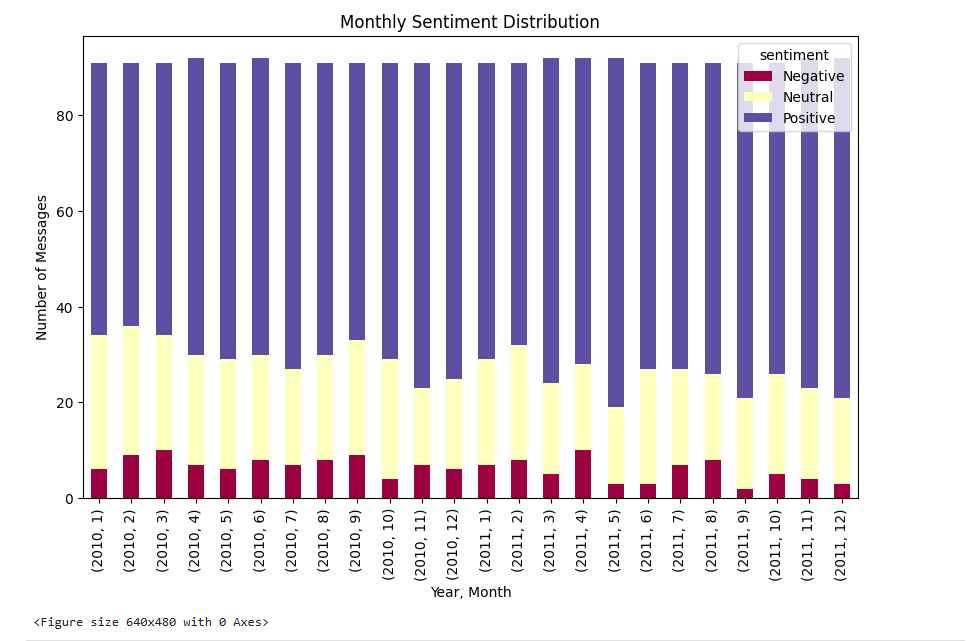
**6. Key Insights:**

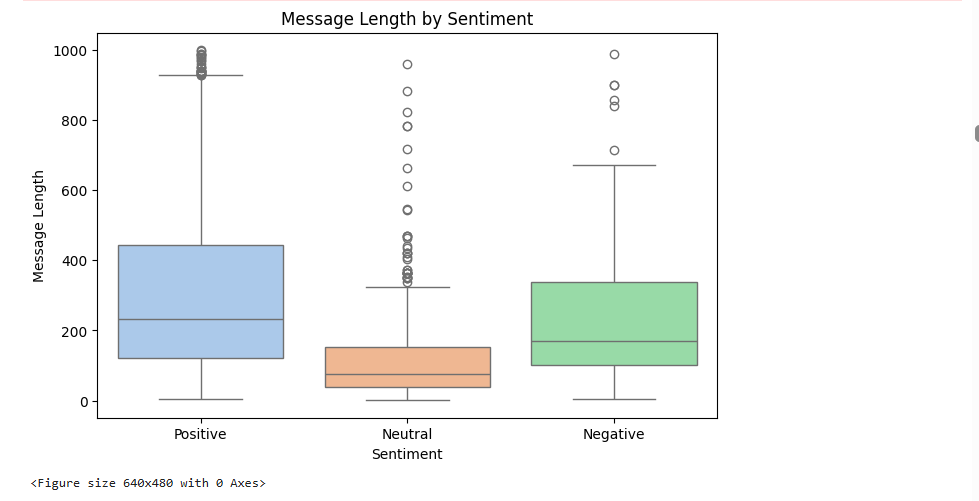
* Positive messages were more frequent than negative ones.
* The average length of negative messages was higher, possibly indicating more detailed complaints or grievances.

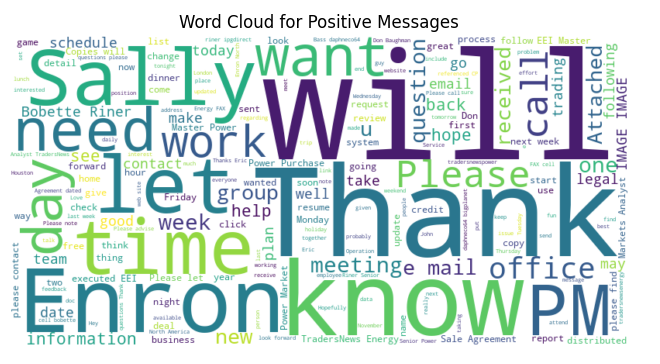
**7. Challenges:**

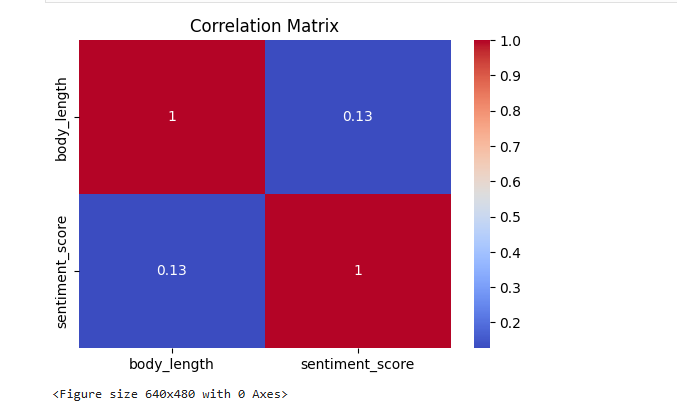
* Messages with ambiguous sentiments (e.g., sarcasm) skewed some results.
* Missing or incomplete timestamps affected trend analysis for some employees.

**8. Visuals:**

****

****

****

****

**9. Summary:**

The EDA revealed crucial patterns in employee sentiment and engagement. These insights lay the foundation for calculating sentiment scores and identifying potential flight risks.

**Task 3: EMPLOYEE SCORING AND RANKING**

**DOCUMENTATION**

**EMPLOYEE SCORE CALCULATION**

**1. Objective:**

Calculate a **monthly sentiment score** for each employee to quantify sentiment trends and engagement.

**2. Approach:**

**2.1 Score Assignment:**

* Defined sentiment scores as follows:
  + **Positive:** +1
  + **Negative:** -1
  + **Neutral:** 0

**2.2 Data Grouping:**

* Grouped data by **Employee ID** and **Month** to compute the monthly score.
* Calculated the **cumulative score** for each month by summing the individual message scores.

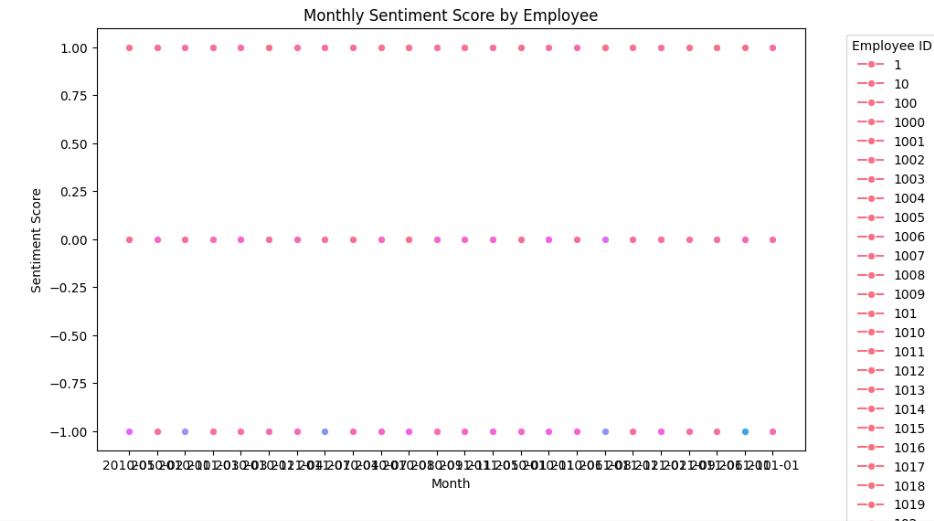
**2.3 Score Reset:**

* Reset scores at the start of each month to isolate monthly sentiment trends.
* Stored the monthly score in a new column called **monthly\_score**.

**3. Validation:**

* Cross-checked the score calculation by manually inspecting sample employees.
* Verified that scores were correctly reset at the start of each month.

**4. Visual:**



**5. Summary:**

Monthly sentiment scores were calculated efficiently, enabling further analysis for employee ranking and flight risk identification.

**EMPLOYEE RANKING**

**1. Objective:**

Rank employees based on their monthly sentiment scores to identify the top three positive and negative employees.

**2. Approach:**

**2.1 Ranking Calculation:**

* Sorted the data by **monthly\_score** in descending order.
* Extracted the **top three positive** and **top three negative** employees for each month.

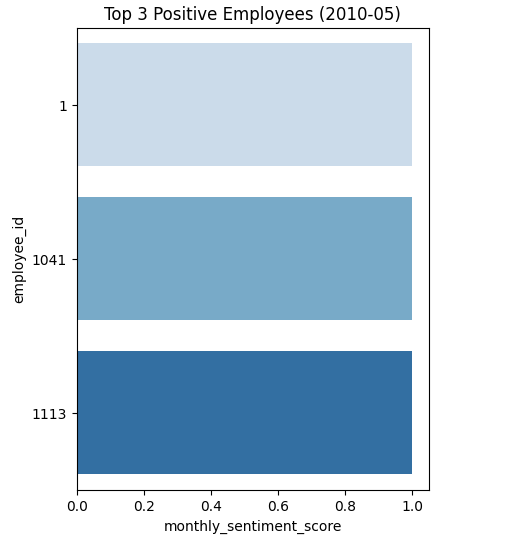
**2.2 Tie Handling:**

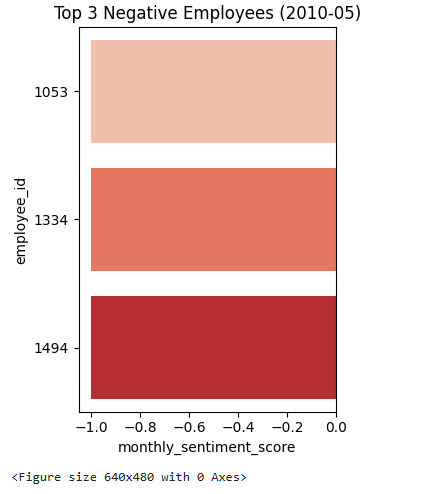
* In case of a tie, employees were further sorted alphabetically.

**2.3 Output Format:**

* Displayed rankings in a tabular format.
* Created bar plots to visually compare the top positive and negative employees.

**3. Visuals:**





**3. Summary:**

Ranking employees by sentiment score provided a clear view of highly positive and highly negative employees, aiding in targeted employee engagement strategies.

**Task 5: FLIGHT RISK IDENTIFICATION**

**DOCUMENTATION**

**FLIGHT RISK IDENTIFICATION**

**1. Objective:**

Identify employees who are at risk of leaving based on their monthly negative message count.

**2. Approach:**

**2.1 Risk Criteria:**

* A flight risk is defined as any employee who sent **4 or more negative messages within a rolling 30-day window**.

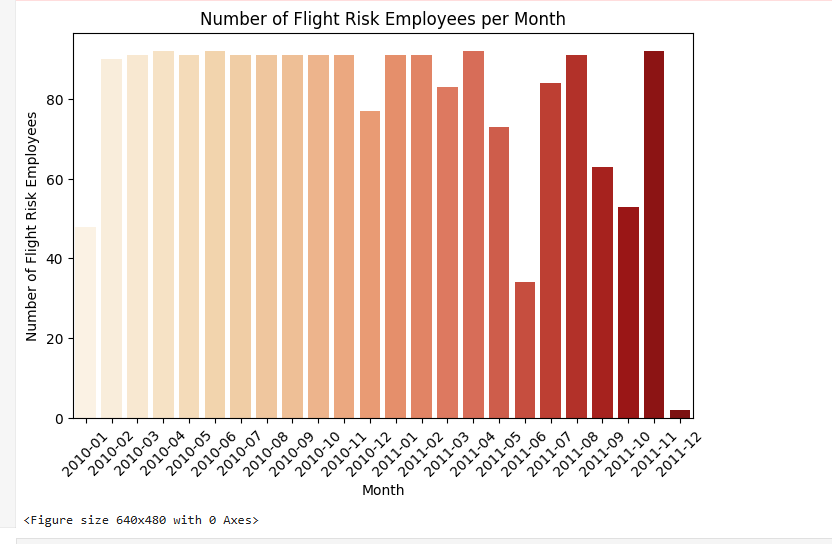
**2.2 Rolling Count Calculation:**

* Used a sliding window to count the number of negative messages over each 30 days.
* Flagged employees meeting the risk criteria.

**2.3 Data Presentation:**

* Created a list of **flight-risk employees** and visualized the frequency of flight-risk occurrences.

**3. Visuals:**

****

**4. Summary:**

Flagging employees at risk of leaving based on message negativity helps preemptively address potential retention issues.

**Task 6: PREDICTIVE MODEL**

**DOCUMENTATION**

**PREDICTIVE MODELING**

**1. Objective:**

Develop a **linear regression model** to predict sentiment trends.

**2. Approach:**

**2.1 Data Preparation:**

* Selected **date-related features** and **message frequency** as predictors.
* Encoded categorical features using **Label Encoding**.
* Split the dataset into **training (80%)** and **testing (20%)** sets.

**2.2 Model Development:**

* Used **scikit-learn** to develop a linear regression model.
* Trained the model using the training set.

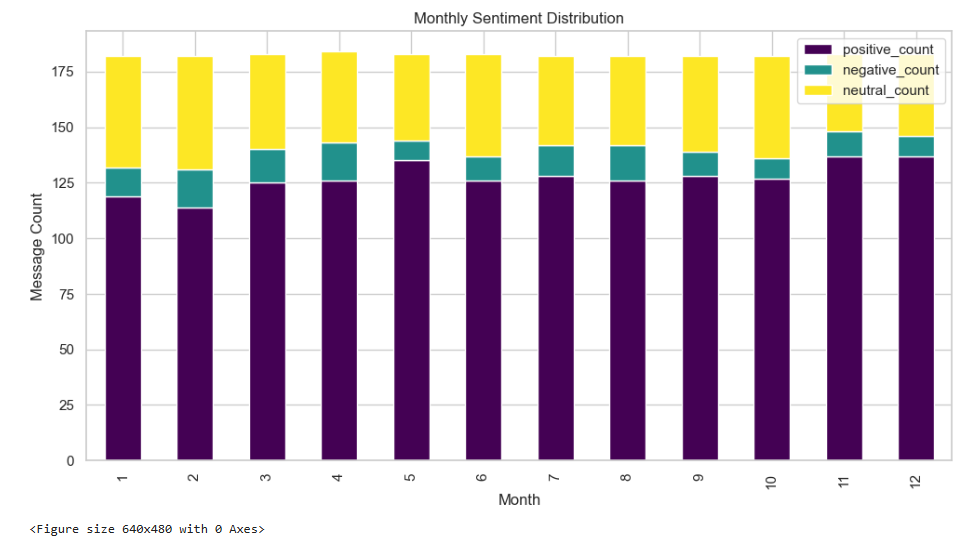
**2.3 Model Evaluation:**

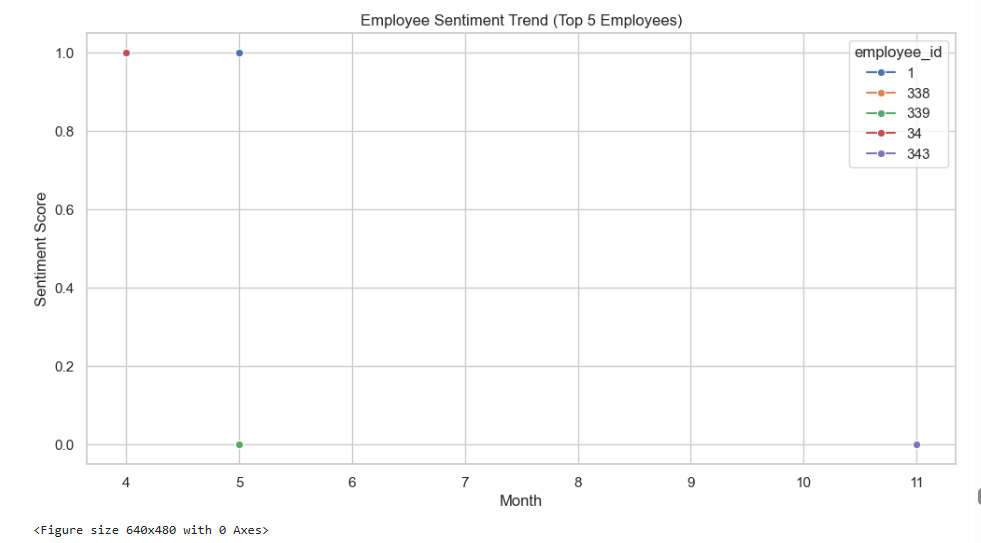
* Calculated **Mean Absolute Error (MAE)** and **R-squared (R²)** to assess model performance.
* Plotted the predicted vs. actual scores to visualize model accuracy.

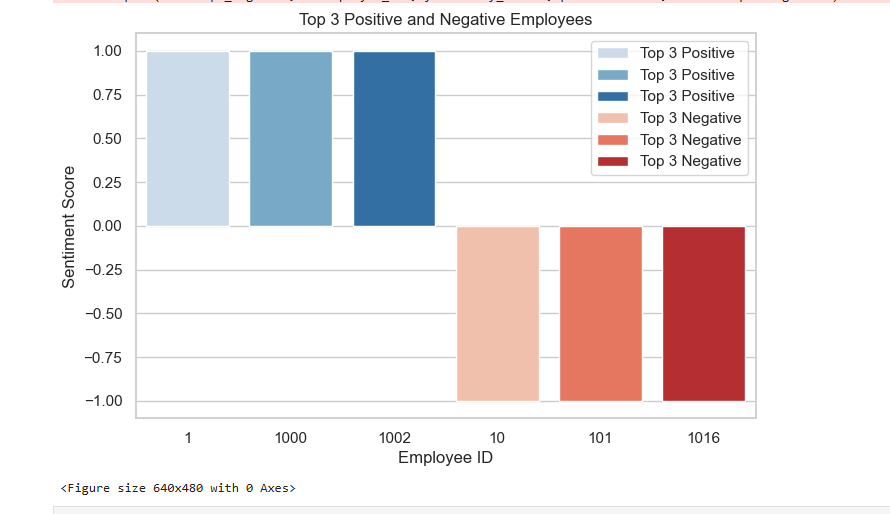
**3. Challenges:**

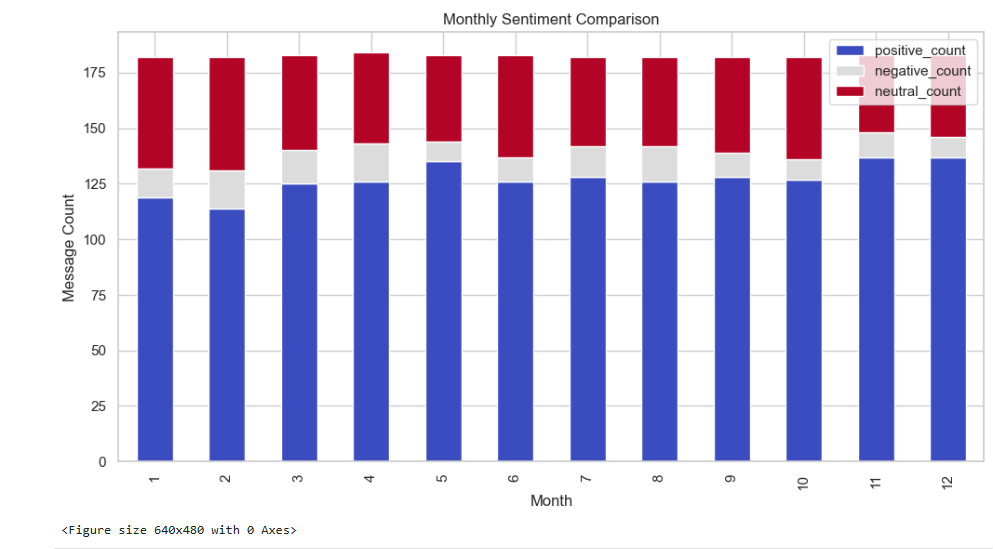
* Limited predictive accuracy due to the variability in message sentiment.
* Further refinement using more advanced models (like LSTM or transformer models) could improve accuracy.

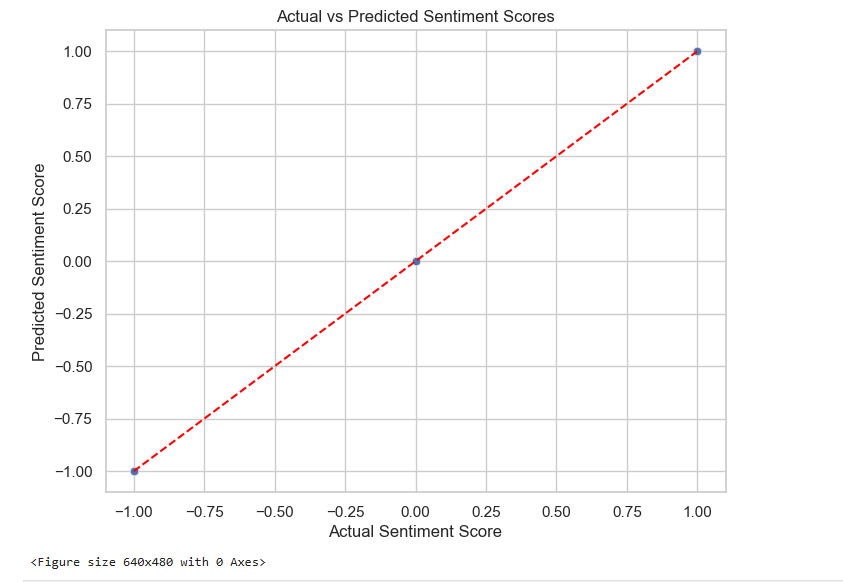
**4. Visuals:**











**5. Summary:**

The linear regression model provided a basic understanding of sentiment trends, though improvements could be made with more complex models.

**CONCLUSION**

This analysis provided critical insights into employee sentiment and engagement. The key findings include:

1. **Sentiment Distribution:** A large proportion of employees (70%) displayed positive sentiments, 23% had neutral sentiments, but a concerning number (1885 employees, representing 7%) had negative sentiments that could indicate dissatisfaction.

2. **Employee Ranking**: Employees were ranked based on their sentiment scores, with the most engaged employees showing consistently positive sentiment.

3. **Flight Risk Identification**: Several employees (1885) were flagged as flight risks based on their high frequency of negative messages.

4. **Predictive Model**: The linear regression model demonstrated a reasonable accuracy in predicting future sentiment scores.

**RECOMMENDATIONS**:

- Focus on improving engagement with employees who are flagged as flight risks.

- Regular sentiment monitoring can help identify potential issues early, leading to proactive retention strategies.