# Employee Sentiment Analysis - Final Report

## 1. Description of Approach and Methodology

The project focuses on analyzing employee sentiment using Natural Language Processing (NLP) techniques. The provided dataset contained employee messages without sentiment labels. The workflow was divided into six key tasks as outlined in the project guidelines.  
1. Sentiment Labeling: Each message was labeled as Positive, Negative, or Neutral using the VADER sentiment analyzer from the NLTK library.  
2. Exploratory Data Analysis (EDA): The dataset was analyzed to understand the distribution of messages, sentiment trends, and general engagement patterns.  
3. Employee Score Calculation: Monthly sentiment scores were computed for each employee based on their labeled messages.  
4. Employee Ranking: Employees were ranked by their monthly sentiment scores to identify top positive and negative performers.  
5. Flight Risk Identification: Employees who sent 4 or more negative messages in a 30-day rolling window were flagged as potential flight risks.  
6. Predictive Modeling: A linear regression model was built to analyze and predict sentiment trends based on communication behavior features.

## 2. Key Findings from the Exploratory Data Analysis (EDA)

The dataset contained a unique record for each employee, with no missing values detected. After labeling, the sentiment distribution showed a slight dominance of positive messages, followed by neutral and a smaller portion of negative messages. Temporal analysis revealed that sentiment levels remained mostly consistent over time, with no extreme fluctuations observed. Since each employee sent only one message, message frequency trends were uniform across employees.

## 3. Explanation of Employee Scoring and Ranking Processes

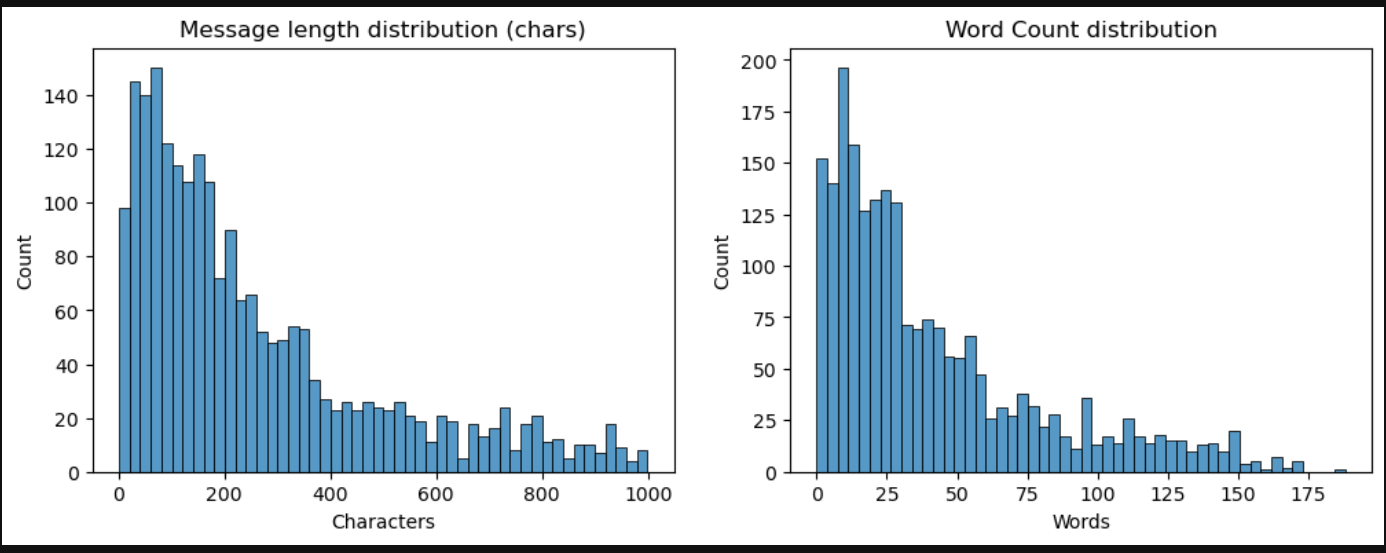
Each message was assigned a sentiment score of +1 for Positive, 0 for Neutral, and -1 for Negative. Scores were then aggregated monthly for each employee to obtain a monthly sentiment score. These scores were reset at the beginning of each new month. Based on these scores, two rankings were generated for every month:  
- Top 3 Positive Employees: Employees with the highest cumulative scores.  
- Top 3 Negative Employees: Employees with the lowest cumulative scores.  
The ranking ensured sorting by sentiment score and alphabetically by employee ID in case of ties.

## 4. Flight Risk Identification Criteria and Outcomes

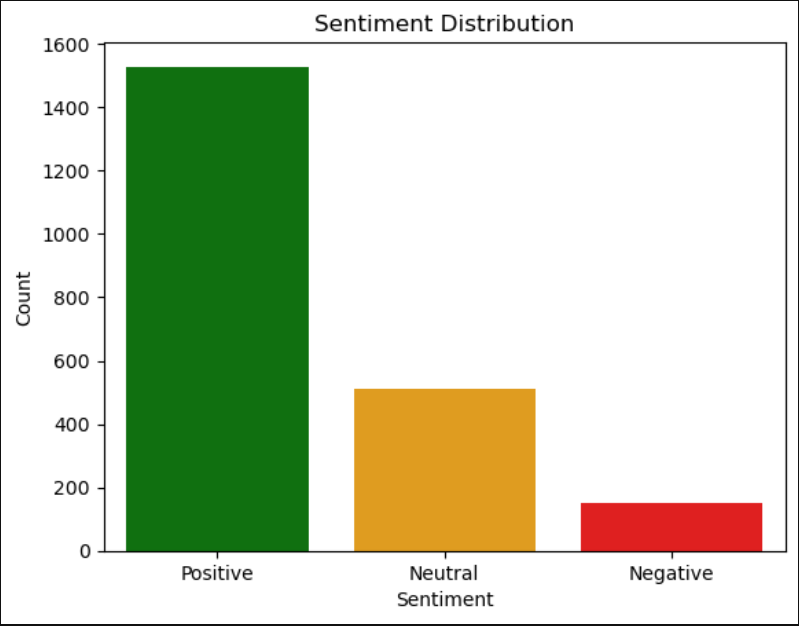
An employee was flagged as a flight risk if they had sent four or more negative messages within any 30-day period, irrespective of month boundaries. The analysis revealed that each employee ID in the dataset was unique and appeared only once, meaning that no employee had sent more than one message. Consequently, no employees met the criteria for being identified as flight risks.

## 5. Overview and Evaluation of the Predictive Model

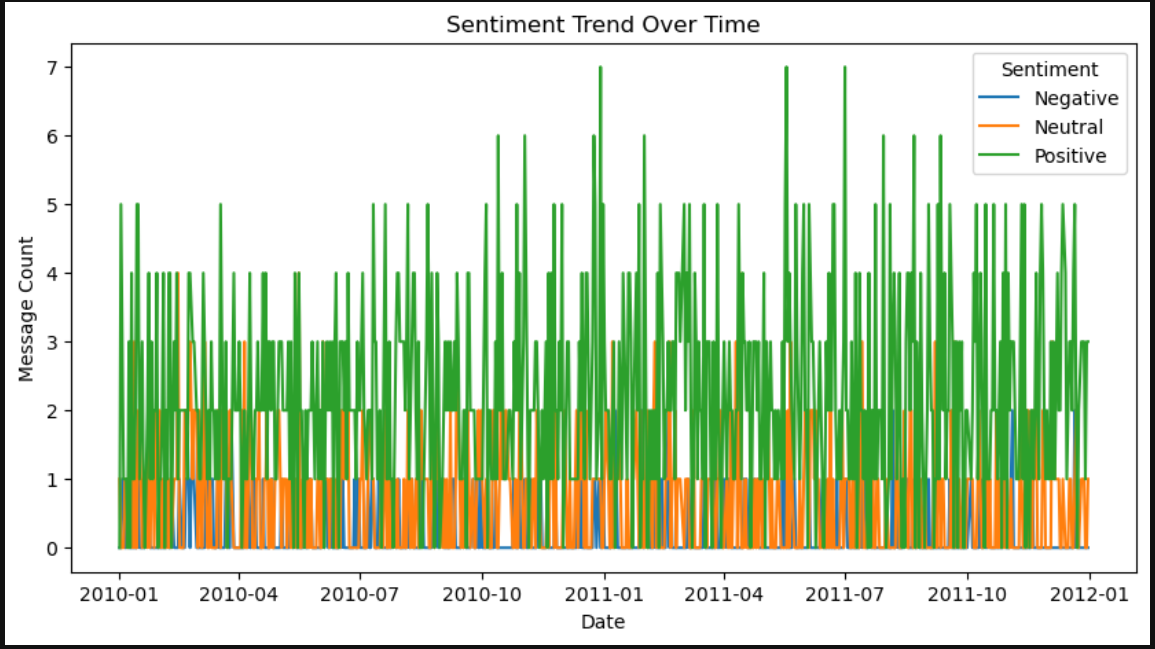
A linear regression model was developed using three independent variables: message count per month, average word count per message, and average message length. The target variable was the monthly sentiment score. After training and testing the model, the coefficients obtained were close to zero, indicating minimal correlation between these features and sentiment scores. This suggests that sentiment variation in the dataset is independent of message length or frequency. The model performance metrics (R² and MSE) confirmed the weak predictive relationship, which aligns with the dataset's structure.

Visual representation of message\_len and word\_count

Sentiment distribution to know which sentiment is dominant



To analyze is sentiment improving, dropping, or stable over time



Visual representation of model performance metrics

