Employee Sentiment Analysis – Project Report

Candidate Name: Ansh

Project Title: Employee Sentiment Analysis

Dataset: test.csv

Submission Context: Internal Individual Evaluation

Technologies Used: Python, Pandas, NumPy, TextBlob, scikit-learn, Matplotlib, Seaborn, datetime, NLTK

# Introduction

This project aims to analyze employee communication data, extracted from email content, to evaluate overall sentiment and engagement within an organization. The analysis uses Natural Language Processing (NLP) techniques and statistical modeling to understand trends in employee behavior, identify potential disengagement (flight risk), and model sentiment variations over time. The dataset provided was unlabeled, which added an additional challenge of implementing an effective rule-based sentiment classifier as a foundation for all further tasks.

# Sentiment Labeling

To begin the analysis, the sentiment of each message had to be labeled as Positive, Negative, or Neutral. Since the dataset was unlabeled, I implemented a rule-based sentiment classification using the TextBlob library, which computes sentiment polarity scores for text. A threshold-based approach was adopted where messages with polarity greater than 0.1 were marked as Positive, polarity less than -0.1 as Negative, and messages with polarity between -0.1 and 0.1 were labeled Neutral. This automated method allowed for efficient labeling without requiring manual annotation.

# Exploratory Data Analysis (EDA)

Following the labeling, a comprehensive exploratory data analysis was conducted to gain an understanding of the data distribution, structure, and hidden trends. Initially, data types, missing values, and message lengths were examined. The distribution of sentiment categories revealed that the majority of messages fell into the Neutral category, followed by Positive and Negative. This suggests a predominance of formal or information-centric communication.  
  
A time-series analysis was also conducted by converting the date column to datetime format and extracting month-wise data. This allowed for the construction of monthly sentiment trend graphs, which provided valuable insights into fluctuations in employee mood and engagement over time. Visualizations created using Matplotlib and Seaborn were used to represent the sentiment distribution, monthly trends, and employee message volume. These visualizations serve as the foundation for all subsequent modeling and ranking efforts.

# Monthly Sentiment Score Calculation

To quantify employee sentiment over time, a scoring mechanism was implemented. Each message was assigned a score based on its sentiment label: +1 for Positive, 0 for Neutral, and -1 for Negative. These scores were then aggregated on a per-employee, per-month basis using Pandas groupby operations. This resulted in a structured dataset containing cumulative sentiment scores for each employee per month. This method enabled the identification of performance trends and mood swings, facilitating employee monitoring and decision-making at the organizational level.

# Employee Ranking

Using the monthly sentiment scores, two types of rankings were computed for each month. The first was a list of the top three employees with the highest cumulative sentiment scores, representing the most positively engaged individuals. The second list included the three employees with the lowest sentiment scores, indicating potential dissatisfaction or disengagement. These rankings were sorted first by score in descending order and then alphabetically to ensure consistent and reproducible outputs. This mechanism provided a fair and interpretable method for recognizing performance extremes within the workforce.

# Flight Risk Identification

One of the critical parts of this project was to flag employees who could be at risk of leaving the organization. For this, a flight risk was defined as any employee who sends four or more negative emails within a rolling 30-day window. This rule was implemented using a rolling count of negative messages per employee, irrespective of month boundaries. The implementation involved creating a filtered dataframe containing only negative messages, then applying a rolling window with grouping by employee and time. Employees meeting this criterion were flagged as potential flight risks, allowing for early intervention by HR.

# Predictive Modeling: Linear Regression

In the final task, a linear regression model was developed to predict future sentiment trends. The goal was to forecast sentiment scores based on time and message frequency features. For each employee, the dataset included a time index (e.g., month number), message volume, and cumulative sentiment score. The data was split into training and testing sets. A linear regression model from scikit-learn was trained on the data and evaluated using metrics such as Mean Absolute Error (MAE) and R² score. While the results showed moderate accuracy, the model demonstrated the feasibility of predicting workforce sentiment using basic numerical features.

# Visualizations

Multiple visualizations were created and saved in the ‘visualization’ folder. These include: Sentiment distribution histograms and pie charts, Monthly sentiment trend lines, Bar plots showing top and bottom ranked employees, Line chart of rolling negative message count to identify flight risks, and Regression line plotting predicted vs actual sentiment scores. These visualizations serve as both exploratory tools and presentation-ready summaries of insights derived from the dataset.

# Key Findings and Conclusion

The sentiment analysis showed that most employee communication remains neutral, indicating a formal communication culture. However, periodic drops in sentiment scores hint at potential workplace dissatisfaction during certain months. The top-ranked employees were consistent in maintaining positive sentiment, making them valuable morale contributors.  
  
On the other hand, the flight risk analysis was effective in identifying employees who might be disengaging or dissatisfied, based on repeated negative communications. The ranking system helped distinguish both high-performing and struggling employees. The predictive model, while not highly precise, showed that trends in sentiment can be modeled based on time and activity.  
  
In conclusion, this project successfully applied rule-based sentiment analysis, data aggregation, ranking mechanisms, risk detection, and machine learning to uncover meaningful insights from unstructured employee communications. These insights can inform HR decisions, guide organizational engagement strategies, and serve as a prototype for more sophisticated AI-based workforce analytics in the future.