Employee Sentiment Analysis Project Final Report

Author: Varsana Ilango

Date: June 17, 2025

Introduction: This project implements an employee sentiment analysis utilizing internal email communications. The primary goals of this analysis are to understand employee engagement patterns, identify potential employee flight risks, and build predictive models to forecast future sentiment trends.

Methodology: The dataset was employee emails with fields like Subject, Body, Form, Data, and etc. Preprocessing was done to handle the missing data and remove the invalid entries. Then the date fields were converted to a datetime format and data was tokenized and cleaned. For sentiment labeling, the TextBlob library was used to calculate polarity scores for each message. Based on these polarity scores, each message was categorized as positive if its polarity was greater than 0.05, negative if it was below -0.05, or neutral.

EDA: The exploratory data analysis provided insights into the structure and dynamics of sentiment. The distribution of sentiments revealed that positive messages were the most common in the dataset. Monthly trends were visualized using line plots, which showed that peaks in positive sentiment corresponded to specific months. Negative sentiments were distributed throughout the dataset. In addition to numerical trends, qualitative patterns in language were identified using word clouds. Positive communications frequently contained words such as “thanks,” “appreciate,” and “good,” while negative messages were often characterized by terms like “problem,” “concern,” and “issue.”

Employee Scoring and Ranking: A scoring system was developed in which each positive message contributed a score of +1, each neutral message contributed a score of 0, and each negative message -1 point. Scores were aggregated on a monthly basis for each employee, which provided a clear view of individual sentiment trends over time. Employees were ranked both positively and negatively based on their monthly scores. Those with the highest scores for a given month were identified as top positive contributors, while those with the lowest scores were classified as top negative contributors for that month. Examples of positive score employees include kayne.coulter@enron.com and bobette.riner@ipgdirect.com. Examples of negative scores include employees such as rhonda.denton@enron.com and johnny.palmer@enron.com.

Flight Risk Identification: A critical part of this analysis involved identifying employees who might be at risk of disengagement or departure. An employee was flagged as a potential flight risk if they sent four or more negative emails within any rolling 30-day period, regardless of the

specific month. The analysis identified several employees who met this criterion. The total number of flagged employees was 8. These findings suggest that frequent negative communication within a short time frame is a meaningful signal of potential disengagement or dissatisfaction.

Predictive Modeling: To explore the possibility of forecasting future sentiment trends, a predictive model using linear regression was developed. The feature used in this model was time, represented by UNIX timestamps, while the target variable was the aggregated sentiment score for each employee. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The values for these metrics were MAE: 2.643, RMSE: 3.148, R^2: 0.000.

Visualizations: These include bar charts that summarize the overall distribution of sentiments, line plots depicting sentiment trends over time, and word clouds illustrating the most common terms found in positive, negative, and neutral communications. All visual materials are included in the /visualization/ directory provided with this report.