# **Employee Sentiment Analysis**

Nishigandha V Wankhade

LLM Assessment Project

Springer Capital

May 26, 2025

## **Project Overview**

This project involves analyzing an unlabeled dataset (test.csv) of employee messages to assess sentiment and engagement. The workflow includes working on raw data and derives insights such as sentiment labeling, exploratory data analysis, employee score and ranking, flight risk detection, and predictive modeling using natural language processing (NLP) and statistical analysis techniques.   
This report summarizes all tasks and outcomes based on the dataset and implemented models.

1. **Project Objective**

The primary goal of this project is to assess employee sentiment and engagement using data analysis and machine learning techniques. The process begins with sentiment labeling, where each employee message is automatically categorized as Positive, Negative, or Neutral using natural language processing tools. Next, exploratory data analysis (EDA) is conducted to understand the structure of the dataset, identify patterns, and examine communication trends over time.

Following this, a sentiment score is assigned to each message based on its label and aggregated monthly for each employee. These monthly scores are then used for employee ranking, highlighting the top three most positive and negatively engaged employees within a particular time period.

The project also includes flight risk identification, where employees who have sent four or more negative messages within any 30-day rolling window are flagged as potential disengagement risks. Lastly, predictive modeling is performed using linear regression and advanced algorithms to forecast future sentiment scores based on historical data. This comprehensive workflow provides valuable insights into employee behavior and supports proactive decision-making to enhance workplace engagement.

1. **Detailed Tasks**

**Task 1: Sentiment Labeling**

The sentiment labeling task involved cleaning and preprocessing raw employee messages using techniques like case normalization, stop word removal, punctuation and URL stripping, and lemmatization. It involves assigning sentiment labels - Positive, Negative, or Neutral, to each employee message in the dataset. This was achieved using the VADER (Valence Aware Dictionary and sEntiment Reasoner) model from the Natural Language Toolkit (NLTK) and a transformer-based model (cardiffnlp/twitter-roberta-base-sentiment).

VADER is a lexicon and rule-based natural language processing (NLP) technique, not a large language model (LLM). It is specifically designed for sentiment analysis on short texts such as social media posts or email messages and performs well without requiring large-scale training data (Hutto & Gilbert, 2014).

In this project, VADER calculates a compound sentiment score for each message, ranging from -1 (most negative) to +1 (most positive). Messages with a compound score ≥ 0.5 were labeled as *Positive*, those with ≤ -0.05 as *Negative*, and those in between as *Neutral*. This logic was applied to the message body column (body) using Python and the pandas library, and the results were saved in a new column called 'sentiment'.

**Task 2: Exploratory Data Analysis (EDA)**

The Exploratory Data Analysis (EDA) phase focused on preparing and understanding the dataset before applying sentiment analysis and modeling techniques. The analysis began with checking the overall data structure, confirming the number of records and ensuring that no missing values were present. The 'date' column was converted to datetime format, and records were sorted chronologically to enable trend analysis. As part of text preprocessing, the message body ('body') and subject fields were cleaned by converting text to lowercase, removing email signatures, URLs, punctuation, and special characters. Stop words were excluded, and words were lemmatized to standardize different forms of the same root word (e.g., "running" to "run"). Time-based features such as month, weekday, and weekend flag were extracted from the date column to support temporal analysis. Additionally, sender email addresses were normalized to lowercase and stripped of extra spaces to prevent duplication or inconsistencies. Finally, duplicate messages were removed by checking for repeated combinations of message body, sender, and timestamp. This clean and enriched dataset provided a solid foundation for downstream sentiment analysis, pattern detection, and modeling tasks.

The sentiment analysis was conducted using three different models: VADER (a rule-based sentiment analyzer from NLTK), BERT (a transformer-based LLM using the cardiffnlp/twitter-roberta-base-sentiment model), and DistilBERT (a lightweight distilled version of BERT fine-tuned for sentiment classification).

* ***VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analyzer***
* **Distribution:**
* Neutral: Above 1000 messages
* Positive: Approximately 900 messages
* Negative: Fewer than 200 messages

**Figure 1**

*Sentiment Distribution using VADER Sentiment Analyzer*

A graph of different colored rectangular shapes

AI-generated content may be incorrect.

* **Observation**:  
   VADER tends to classify most messages as neutral or slightly positive, with a relatively small number marked as negative as shown in Figure 1. This is consistent with its rule-based approach, which works well for social media-style, short-text sentiment but may under detect nuanced negative tones.
* ***BERT(Bidirectional Encoder Representations from Transformers) Sentiment Analyzer (Roberta-based)***
* **Distribution**:
* Neutral: More than 1400 messages
* Positive: Around 600 messages
* Negative: Less than 200 messages

**Figure 2**

*Sentiment Distribution using BERT Sentiment Analyzer*

A graph of different colored squares

AI-generated content may be incorrect.

* **Observation**:  
   The BERT-based sentiment analysis in Figure 2. shows a stronger bias toward neutral classification, likely due to its training on Twitter data where neutrality is common. Positive messages are significantly fewer compared to VADER, and negative messages remain the lowest. This suggests BERT captures a more conservative and context-aware interpretation of sentiment.
* **DistilBERT Sentiment Analyzer**
* **Distribution**:
* Negative: Approximately 1200 messages
* Positive: Around 1000 messages

**Figure 3**

*Sentiment Distribution using DistilBERT Sentiment Analyzer*

A graph of different colored squares

AI-generated content may be incorrect.

* **Observation:**  
   DistilBERT outputs only two sentiment classes, Positive and Negative as shown in Figure 3. because the model (distilbert-base-uncased-finetuned-sst-2-english) was fine-tuned on the SST-2 (Stanford Sentiment Treebank) dataset, which is a binary classification dataset. As such, there is no "Neutral" class in its output. Interestingly, DistilBERT classifies a much larger portion of messages as negative, indicating a sharper distinction in polarity compared to the other models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **VADAR** | **BERT (Roberta)** | **DistilBERT** |
| **Sentiment Classes** | Positive, Neutral, Negative | Positive, Neutral, Negative | Positive, Negative |
| **Most Frequent Sentiments** | Neutral | Neutral | Negative |
| **Least Frequent Sentiments** | Negative | Negative | N/A (No Neutral) |
| **Model Type** | Rule-based (lexicon) | Transformers (LLM) | Distilled Transformer |
| **Text Sensitivity** | Low–Moderate | High | Moderate-High |
| **Token Limit** | N/A | 512 | 512 |

**Table 1:** *Comparison of VADER, BERT and DistilBERT* *Sentiment Analyzers*

* **VADER** is quick and effective for straightforward sentiment classification, especially in domains like emails or short reviews, but may underperform in capturing deep contextual sentiment.
* **BERT** offers a more refined and neutral-leaning sentiment classification, benefiting from its pretraining on large contextual datasets but may require more resources.
* **DistilBERT**, while faster and lighter, shows a tendency toward binary polarization, which may be suitable for use cases requiring clear positive/negative sentiment boundaries but lacks the nuance of neutral categorization.

After evaluating the performance of the various sentiment analyzers, the results from the BERT sentiment analyzer have been found to be the most suitable and will be used as the basis for continuing the remainder of the project.

**Identifying Monthly Sentiment Trend:**

Figure 4 displays a Monthly Sentiment Trend of employee messages categorized into Negative, Neutral, and Positive sentiments over time from January 2010 to December 2011. It shows that Neutral messages consistently dominate, maintaining the highest volume each month. Positive messages remain the second most frequent, with a slight upward trend, while Negative messages are the least frequent, staying relatively low and stable. Overall, this indicates a workplace communication environment where neutral and positive sentiments prevail, with minimal negativity over the observed period.

**Figure 4**

*Monthly Sentiment Trend*

A graph of different colored lines

AI-generated content may be incorrect.

**Key Findings from the Exploratory Data Analysis (EDA):**

* The dataset included email communications of employees, with sentiment labels (Positive, Neutral, Negative) assigned using advanced sentiment analysis models like BERT.
* Initial visualizations and frequency counts revealed patterns in communication sentiment over time.
* Certain employees showed consistently negative sentiment patterns, while others had more balanced or positive communication styles.
* Seasonal and monthly sentiment trends were noticeable, providing insights into employee morale and communication behaviors.

**Task 3: Employee Score Calculation**

In this task, a monthly sentiment score was calculated for each employee based on the messages they sent. Each message was first assigned a numerical score according to its sentiment: +1 for positive messages, -1 for negative messages, and 0 for neutral messages. These scores reflect the tone or emotional weight of each communication. To analyze trends over time, the messages were grouped by the sender (employee) and by the month in which they were sent. The scores were then summed for each employee within each month, resulting in a monthly sentiment score. This approach provides a clear view of how an employee’s communication tone changes over time. These scores are crucial for further analysis, including employee ranking and identifying potential flight risks or dissatisfaction trends.

**Figure 5**

*Monthly Employee Sentiment Score*

A screenshot of a number list

AI-generated content may be incorrect.

**Task 4: Employee Ranking**

In this task, employees were ranked based on their monthly sentiment scores, which were calculated in the previous step, as shown in Figure 5. For each month, two separate lists were generated: one for the top three employees with the highest positive scores and another for the bottom three employees with the most negative scores. To ensure fairness and clarity, the data was first sorted by month, followed by the sentiment score in descending order, and then alphabetically by employee name. This sorting method ensures that when multiple employees have the same score, their names appear in a consistent and organized way.

These rankings provide valuable insight into communication patterns, helping identify employees who consistently contribute positively to workplace communication, as well as those whose messaging trends may signal dissatisfaction or disengagement.

**Figure 6** *Monthly Review of**Top 3 Positive and Negative Sentiment Employees*

|  |  |
| --- | --- |
| A screenshot of a computer  AI-generated content may be incorrect. | A screenshot of a computer  AI-generated content may be incorrect. |
|  | A graph with a red bar  AI-generated content may be incorrect. |

**Task 5: Flight Risk Identification**

In this task, the goal was to find employees who might be thinking about leaving the company. To do this, we looked for employees who sent 4 or more negative emails within any 30-day period. It doesn’t matter which month the emails were in; we checked every rolling 30-day window.

Before conducting any advanced analysis, a preliminary review was performed using a pivot table in Excel (Figure 7) to identify potential employees considering leaving the company. This involved manually checking for individuals who had sent four or more negative emails within any 30-day period. The highlighted values in the pivot table indicate employees who may be at risk of leaving, based on this initial assessment.

**Figure 7**

*Initial Flight Risk Assessment using Pivot Table in Excel*

A screenshot of a calendar

AI-generated content may be incorrect.

After performing advanced analysis, it was observed that approximately 80% of the initial assumptions made using the Excel pivot table were accurate in identifying employees who were potentially at risk of leaving the company as shown in Figure 8.

**Figure 8**

*Flight Risk Assessment Results After Advanced Analysis*

A number and date on a white background

AI-generated content may be incorrect.

To further validate the findings, a focused analysis was conducted on a specific employee, “bobette.ringer.” The results confirmed the accuracy of the previous analysis, as it was found that this employee had indeed sent exactly four negative messages as shown in Figure 9, during the period from July 15, 2010, to August 14, 2010, matching the earlier prediction (Figure 8).

**Figure 9**

*Cross Validating Flight Risk Assessment Results by Focusing on a Single Employee*

A screenshot of a computer

AI-generated content may be incorrect.

An employee was flagged as a "flight risk" if they sent four or more negative emails within any rolling 30-day period, regardless of calendar month boundaries. To identify such cases, a rolling window analysis was employed, systematically scanning each employee's communication timeline to detect these patterns. This approach enabled the creation of a refined list of potentially at-risk employees based on their negative sentiment frequency.

**Task 6: Predictive Modeling**

Linear regression and Random Forest regression models were built to predict monthly sentiment scores. Features included message count and sentiment class counts. Initially Linear regression had poor results (R² = -0.12) as compared to but Random Forest (R² = 0.91). After feature engineering, Linear regression (R² = 1.00) outperformed Random Forest regression model (R² = 0.91). Evaluation used metrics like mean squared error (MSE), mean absolute error (MAE), and R² score, along with residual plots and prediction scatter plots.

* **Predictive Modeling done by Linear Regression before Feature Engineering:**

A black text on a white background

AI-generated content may be incorrect.

**Figure 10**

*Linear Regression Before Feature Engineering*

A graph with blue dots and red lines

AI-generated content may be incorrect.

The current performance metrics of the linear regression model: MSE = 4.55, MAE = 1.60, and R² = -0.12, suggest that the model is not accurately capturing the patterns or trends in the employee sentiment score data. An MSE of 4.55 indicates that the squared prediction errors are moderately high, and the MAE of 1.60 means that, on average, the model's predictions deviate by 1.6 points from the actual monthly scores, which may be significant depending on the typical score range. Most critically, the negative R² score (-0.12) highlights that the model performs worse than a simple baseline model that always predicts the mean sentiment score. This implies that the current feature(s) used in this case, are not sufficient to explain the variability in sentiment scores, and the model is either underfitting or missing key predictors such as message frequency, monthly score, etc.

**Recommendations to Improve model performance:**

Add more features to the model like message count per employee per month, monthly score, and sentiment count for each sentiment label.

* **Predictive Modeling done by Linear Regression after Feature Engineering:**



Model Evaluation Summary:

MSE = 0.00 Models predictions exactly match the actual values (no squared error).

MAE = 0.00 There is no average absolute error, the model is 100% accurate.

R² = 1.00 The model explains 100% of the variance in the target variable.

Figure 11 shows, model built perfectly predicts the current data and it is a well-fitted model. These results are a strong indicator that the approach used to calculate and structure the sentiment scores was effective and consistent, making the data highly predictable.

**Figure 11**

*Linear Regression After Feature Engineering*

A graph with a line and a red dotted line

AI-generated content may be incorrect.

However, it’s also important to critically evaluate whether this perfect accuracy might be due to data leakage, overfitting, or an overly simplistic relationship (e.g., just one numeric feature driving the entire prediction). Nevertheless, from a business and analytical standpoint, the results are promising, as they suggest that employee sentiment trends can be reliably forecasted, which can aid HR or management teams in proactive engagement and risk mitigation strategies.

* **Predictive Modeling done by Random Forest Regression**

A white background with black text

AI-generated content may be incorrect.

Model Evaluation Summary:

MSE = 0.78 Low average squared error, which is very good!

MAE = 0.55 On average, predictions are only ±0.55 units off

R² Score = 0.91 Model explains 91% of the variance in sentiment scores, which is outstanding!

In terms of the overall project results, this shows that advanced machine learning models like Random Forest are much more suited for predicting employee sentiment trends than earlier linear models but not better than improved linear regression model which was done after feature engineering. It validates the quality of the sentiment scoring pipeline and suggests that meaningful forecasting of employee behavior and morale is possible when using the right modeling techniques. This finding enhances the credibility and practical applicability of the entire Employee Sentiment Analysis project.

The added features like message\_count, positive\_count, negative\_count, and neutral\_count clearly improved the model's ability to capture meaningful patterns as shown in Figure 11 and Figure 12.

**Figure 12**

*Random Forest Regression After Feature Engineering*

A graph with green dots and a line

AI-generated content may be incorrect.

A Random Forest method also helps in identifying that “Positive” sentiments plays more crucial role among other features in the predictive modeling.

**Figure 13**

*Feature Importance Using Random Forest Method*

A graph with a bar graph

AI-generated content may be incorrect.