Employee Sentiment Analysis

Submitted by:
Soham Mohanlal Tare
(sohamt2004@gmail.com)
https://github.com/SohamTare/Employee-Sentiment-Analysis

Objective

The objective of this project is to examine employee communications in order to understand sentiment patterns, measure engagement levels, rank employees based on sentiment behavior, identify potential flight risks, and build a predictive model to analyze factors influencing sentiment over time.

1. Introduction

Employee sentiment is a key indicator of organizational health, directly impacting productivity, engagement, and retention. Gaining insights into how employees feel through their everyday communication can help organizations identify dissatisfaction early and take corrective action.

This project analyzes an unlabeled dataset of employee email messages to automatically determine sentiment, study communication trends, calculate sentiment-based engagement scores, rank employees, detect potential flight risks, and develop a predictive model to better understand sentiment dynamics.

2. Dataset Description

The dataset contains internal employee email communications with the following primary attributes:

- **from**: Unique employee identifier (email address)
- **subject**: Subject line of the email
- **body**: Main content of the message
- **date**: Timestamp indicating when the message was sent

To support time-based analysis, additional features such as **year** and **month** were derived from the date field.

3. Sentiment Labeling Methodology

As the dataset did not contain predefined sentiment labels, sentiment classification was performed automatically using the **VADER (Valence Aware Dictionary and sEntiment Reasoner)** sentiment analysis tool.

Sentiment Classification Criteria:

- **Positive**: Compound score ≥ 0.05
- **Negative**: Compound score ≤ -0.05
- **Neutral**: Compound score between -0.05 and 0.05

Sentiment Distribution:

- **Positive**: 1545 messages
- **Neutral**: 382 messages
- **Negative**: 264 messages

The distribution shows that most employee communications are positive in nature, while a smaller yet meaningful portion reflects negative sentiment that requires attention.

(Sentiment distribution and trend visualizations can be included to support this analysis.)

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain insights into communication behavior and sentiment patterns across the organization.

Key observations from the analysis include:

- Employee sentiment varies over time, reflecting fluctuations in morale.
- Negative messages are generally longer, suggesting stronger emotional expression.
- A limited number of employees contribute a disproportionately high number of messages, indicating higher engagement levels.

Multiple visualizations were generated, including sentiment trends over time and message length comparisons, to validate these findings.

5. Employee Sentiment Scoring

Each email message was assigned a numerical score based on its sentiment classification:

- **Positive**: +1
- **Negative**: -1
- **Neutral**: 0

These scores were aggregated on a **monthly basis for each employee**, with scores resetting at the start of every new month. The resulting monthly sentiment scores were used for ranking and further analysis.

6. Employee Ranking

Employees were ranked each month using their aggregated sentiment scores:

- **Top 3 Positive Employees**: Employees with the highest sentiment scores
- **Top 3 Negative Employees**: Employees with the lowest sentiment scores

Ranking was performed by sentiment score first, followed by alphabetical ordering to ensure consistency and fairness. This ranking highlights both highly engaged employees and those potentially experiencing dissatisfaction.

7. Flight Risk Identification

An employee was flagged as a **flight risk** if they sent **four or more negative messages within any rolling 30-day period**, regardless of calendar months.

This rolling window approach enables early detection of sustained negative sentiment patterns. Employees meeting this criterion were identified and listed for further review and possible intervention.

8. Predictive Modeling

A **Linear Regression model** was developed to analyze the factors influencing monthly sentiment scores.

Features Used:

- Monthly message count
- Average message length
- Total message length
- Number of negative messages

Model Insights:

- The number of negative messages has the strongest negative influence on sentiment score.
- Higher communication frequency often aligns with improved sentiment.
- Message length can indicate emotional intensity in communication.

While the model is not intended for precise prediction, it provides valuable insights into sentiment-driving factors.

9. Key Findings and Recommendations

Key Findings:

- Overall employee sentiment is largely positive.
- A small group of employees consistently exhibits negative sentiment patterns.
- Communication frequency and negativity play a major role in sentiment scores.

Recommendations:

- Proactively monitor employees identified as flight risks.
- Encourage open dialogue to address recurring negative sentiment.
- Use sentiment analysis as an ongoing feedback and engagement monitoring tool.

10. Conclusion

This project demonstrates the effective use of NLP techniques and statistical analysis to derive meaningful insights from raw employee communication data. The solution provides a scalable and reproducible approach for sentiment monitoring, engagement evaluation, and early risk detection, supporting data-driven organizational decision-making.