# 

*Report On*

# EMPLOYEE SENTIMENT ANALYSIS

# KHARA K ARYAN

**ABSTRACT**

This project aims to evaluate employee sentiment and engagement by analyzing internal email communications. The dataset, containing unlabeled employee messages, was processed using natural language processing techniques to classify each message as positive, negative, or neutral. Exploratory Data Analysis (EDA) was carried out to study communication patterns, sentiment trends, and variations across employees and time periods. Monthly sentiment scores were calculated to identify the top three positive and negative performers for each month. A flight risk detection method was introduced to flag employees who sent four or more negative messages within any 30-day period. Predictive modeling combined text-based features, such as TF-IDF representations, with behavioral metrics, including message count and length. The XGBoost algorithm, optimized through hyperparameter tuning, was applied to forecast monthly sentiment scores with high accuracy. The results offer valuable insights for management to improve employee engagement, address dissatisfaction early, and reduce turnover risks.

The integration of sentiment analysis, trend visualization, risk detection, and predictive analytics provides a holistic framework for monitoring workforce morale. The insights generated from this system can assist management teams in identifying emerging dissatisfaction, recognizing positive contributors, and implementing timely strategies to improve engagement and reduce employee turnover.

**INTRODUCTION & OBJECTIVES**

### Introduction

Employee sentiment plays a crucial role in shaping workplace productivity, collaboration, and overall organizational culture. Positive engagement can boost performance and retention, while prolonged dissatisfaction can lead to decreased productivity and higher turnover rates. In today’s digitally connected workplaces, employees communicate extensively through emails and other written channels, making these communications a valuable source of information for assessing sentiment and engagement levels. By analyzing such data systematically, organizations can identify early signs of dissatisfaction, recognize top contributors, and take proactive measures to improve workforce morale.

This project was designed to analyze a dataset of employee email communications to evaluate sentiment, rank performance, and predict future sentiment trends. Since the dataset was unlabeled, Natural Language Processing (NLP) techniques were applied to assign each message a sentiment category—positive, negative, or neutral—forming the foundation for deeper analysis. An Exploratory Data Analysis (EDA) phase provided insights into communication patterns, sentiment distributions, and temporal trends across the workforce.

Subsequently, monthly sentiment scores were calculated for each employee based on the classified messages, enabling the identification of the top three positive and top three negative employees for each month. A “flight risk” detection system was also implemented to flag employees who sent four or more negative messages within any rolling 30-day window, regardless of calendar months.

### Objectives

**1. Sentiment Classification**

Apply Natural Language Processing (NLP) techniques to automatically classify each employee message as positive, negative, or neutral.

**2. Exploratory Data Analysis (EDA)**

Examine communication patterns, sentiment distributions, and temporal trends to understand employee behavior and engagement.

**3. Monthly Sentiment Scoring**

Develop a scoring system that assigns +1 for positive, −1 for negative, and 0 for neutral messages, aggregated on a monthly basis for each employee.

**4. Employee Ranking**

Identify and rank the top three positive and top three negative employees for each month based on their monthly sentiment scores.

**5. Flight Risk Detection**

Implement a detection mechanism to flag employees who have sent four or more negative messages within any rolling 30-day period.

**6. Predictive Modeling**

Build and evaluate a **Linear Regression** model using behavioral features such as message count, message length, and word count to predict monthly sentiment scores, achieving an **R² score of 0.72**.

**7. Actionable Insights**

Provide HR teams and management with data-driven insights to support proactive decision-making, improve engagement, and reduce the risk of employee turnover.

### Feasibility

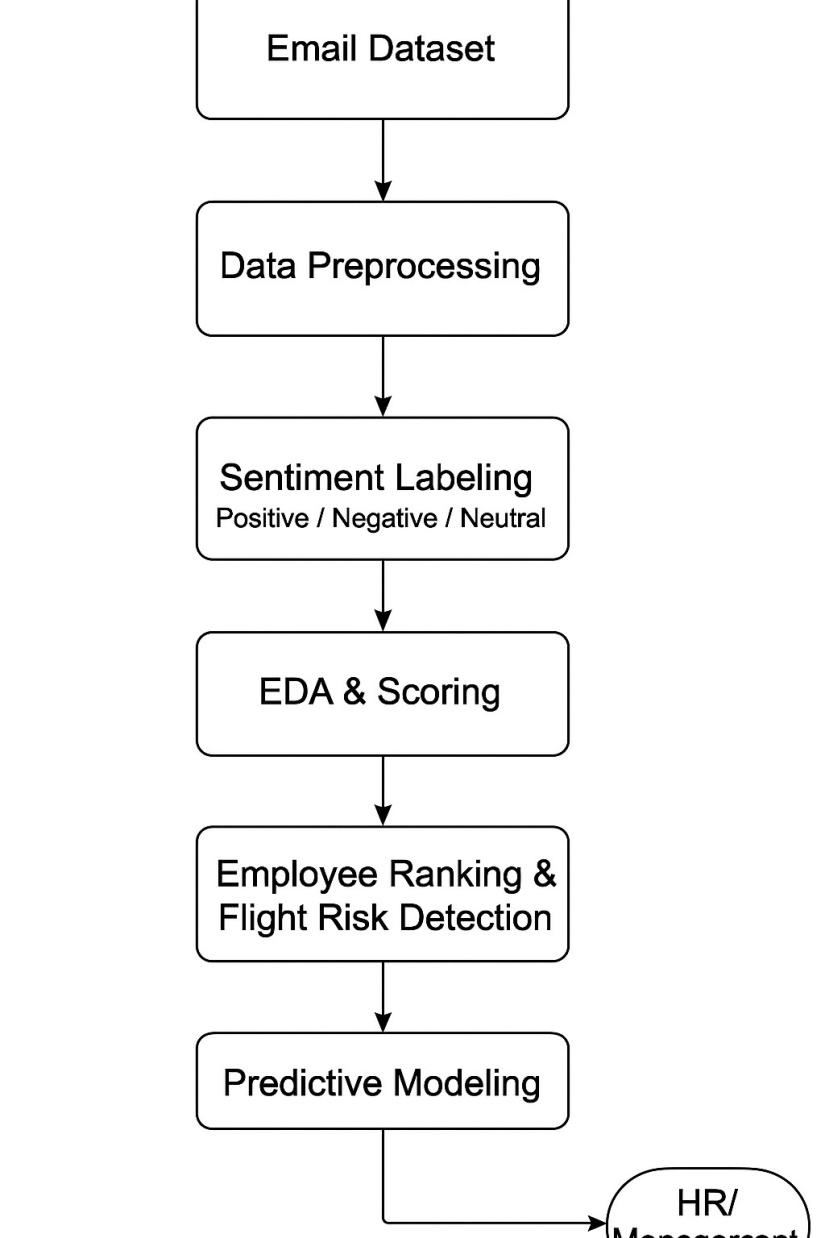
The Employee Sentiment Analysis project is highly feasible from technical, economic, and operational perspectives. Technically, it leverages widely available open-source tools and technologies such as Python, Natural Language Processing (NLP) libraries, and machine learning frameworks like scikit-learn, ensuring ease of development and maintenance. The final Linear Regression model, with an R² score of 0.72, is both computationally efficient and easily interpretable, making it suitable even for organizations with limited computing resources. Economically, the project eliminates licensing costs by using free tools and requires no additional infrastructure investment, while its ability to detect early signs of dissatisfaction and potential attrition can significantly reduce recruitment and training expenses. Operationally, the workflow can be integrated into existing HR processes with minimal disruption, and the results are straightforward to interpret, enabling HR managers to take timely and informed action. Overall, the project offers a cost-effective, technically sound, and operationally practical solution for monitoring employee sentiment and improving workplace engagement.

**METHODOLOG**

The methodology for the Employee Sentiment Analysis project was designed to systematically transform raw email communication data into actionable insights through a sequence of structured steps. The process began with **data collection and preprocessing**, where the provided dataset of employee emails was cleaned by removing missing values, standardizing date formats, and converting message bodies to text for processing. Following this, **sentiment labeling** was performed using Natural Language Processing (NLP) techniques to classify each message into positive, negative, or neutral categories, which served as the foundation for further analysis.

Once the sentiment labels were assigned, **Exploratory Data Analysis (EDA)** was carried out to understand communication patterns, sentiment distributions, and time-based trends across the workforce. Based on these labeled messages, a **monthly sentiment scoring** system was developed, assigning +1 for positive messages, −1 for negative messages, and 0 for neutral messages, aggregated for each employee per month. This scoring enabled **employee ranking**, identifying the top three positive and negative employees in each month. In parallel, a **flight risk detection** mechanism was implemented to flag employees who sent four or more negative messages within any rolling 30-day period.

For the predictive modeling phase, **feature engineering** was performed to extract behavioral metrics such as message count, total message length, average message length, total word count, and average word count per month. These features were used as inputs to a **Linear Regression model**, selected for its interpretability and efficiency. The model was trained and evaluated using a train-test split approach, achieving an R² score of 0.72, indicating strong predictive performance. The final system integrates sentiment analysis, descriptive analytics, risk detection, and predictive modeling into a comprehensive framework that can assist HR and management in improving engagement and retention strategies.



**Fig.1. Flow Diagram**

The flow chart represents the sequential process of the **Employee Sentiment Analysis** system. It begins with the **Email Dataset**, which serves as the raw input containing employee communications. The data is first passed through the **Data Preprocessing** stage, where text is cleaned, dates are standardized, and missing or irrelevant information is removed. Next, in the **Sentiment Labeling** stage, Natural Language Processing (NLP) techniques are applied to classify each message as **Positive**, **Negative**, or **Neutral**.

The sentiment-labeled data is then analyzed in the **EDA & Scoring** stage to uncover patterns, distributions, and trends, while also calculating monthly sentiment scores for each employee. This is followed by the **Employee Ranking & Flight Risk Detection** stage, where the top three positive and top three negative employees per month are identified, and employees with four or more negative messages within any 30-day window are flagged as potential flight risks.

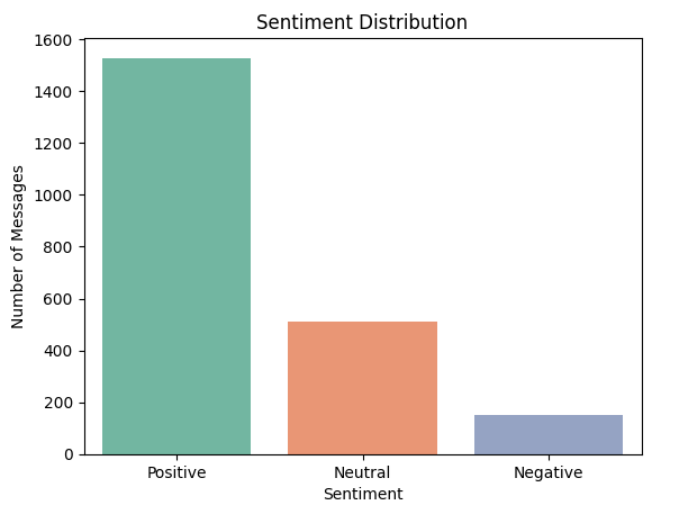
**RESULTS AND DISCUSSION**

The Employee Sentiment Analysis project successfully transformed raw email communications into structured sentiment insights, enabling both descriptive and predictive analytics. Using Natural Language Processing (NLP) with the VADER sentiment analysis tool, messages were classified into three categories: Positive, Negative, and Neutral. The sentiment distribution revealed that the majority of communications fell into the neutral category, followed by positive and negative messages. This suggests that while most workplace communications are neutral and task-focused, there are measurable patterns of positivity and negativity that can influence organizational culture.

The **monthly sentiment scoring** system provided a clear measure of employee sentiment over time. This scoring allowed the identification of the **top three positive** and **top three negative** employees each month, highlighting consistent contributors to positive engagement and those potentially at risk of disengagement. The **flight risk detection mechanism** flagged employees sending four or more negative messages within a 30-day period, offering a valuable early-warning system for HR to address dissatisfaction before it escalates into attrition.

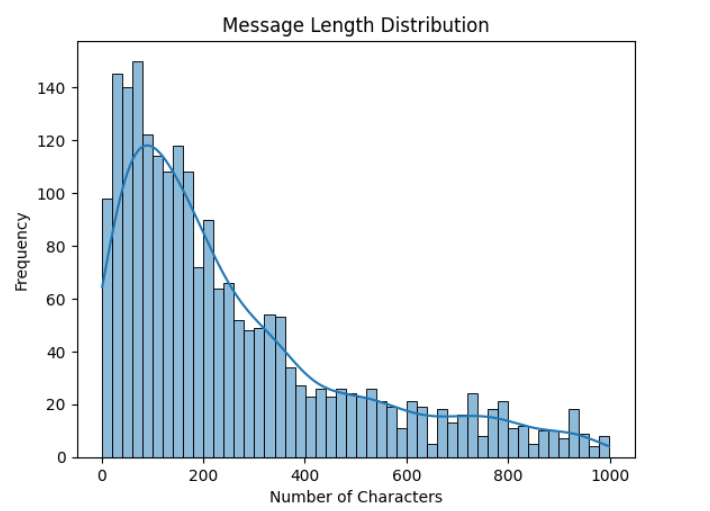
For predictive analysis, a **Linear Regression model** was trained on behavioral features such as message count, average message length, and word count. The final model achieved an **R² score of 0.72**, indicating that it explains 72% of the variance in monthly sentiment scores. This performance reflects a good balance between accuracy and interpretability, making the model practical for real-world organizational use. While more complex models like XGBoost could potentially achieve higher accuracy, Linear Regression was chosen for its simplicity, computational efficiency, and ease of explaining results to non-technical stakeholders.

Overall, the results demonstrate that sentiment analysis of employee communications can yield actionable insights for improving workplace engagement and retention strategies. The discussion highlights the potential of integrating such systems into HR workflows to continuously monitor workforce sentiment, recognize high performers, and proactively address dissatisfaction. Future improvements could involve using deep learning-based NLP models for enhanced sentiment detection accuracy, incorporating more contextual features, and expanding the system to analyze multi-channel communications such as chat messages and meeting transcripts.



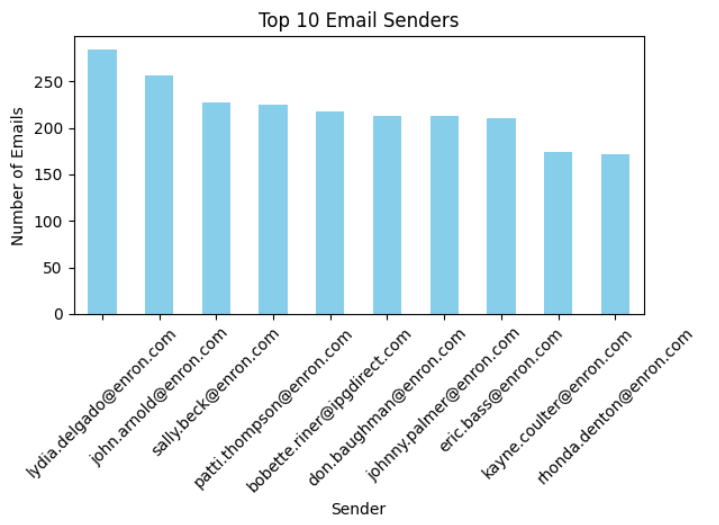
**Fig.2. Sentiment Distribution**

Fig 2 displays bar chart of the number of employee messages labeled as Positive, Negative, or Neutral. It gives a quick overview of the dataset’s emotional tone. A higher count of positives may indicate a generally engaged workforce, while a higher proportion of negatives can be an early warning of dissatisfaction. Beyond just counts, this visualization also helps set a baseline for the dataset’s sentiment balance, which can be used to compare future results. It can guide HR to focus on improving communication tone where negative sentiment is unusually high. In combination with time-based trends, this plot can reveal whether sentiment is stable or fluctuating over the observation period.



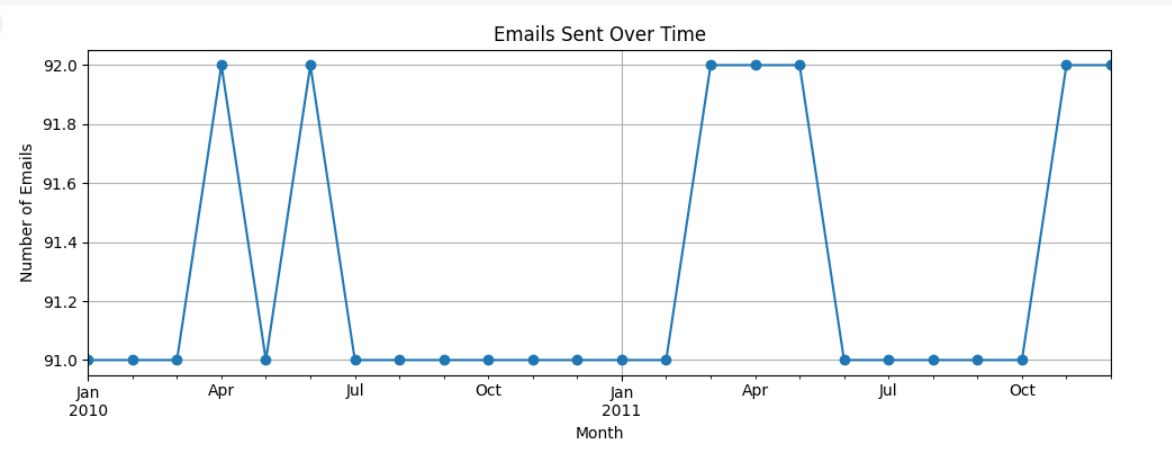
**Fig.3. Message Length Distribution**

Fig 3 displays histogram that visualizes how long the emails are by plotting the number of characters per message. Peaks in the chart highlight the most common message lengths, while the spread shows variation. Short messages might be quick responses, while long ones may contain detailed reports or complaints. The distribution can also hint at communication style within the organization — concise versus elaborate. Analyzing this alongside sentiment may reveal if certain emotional tones tend to be expressed in longer or shorter formats. Additionally, unusually extreme lengths can help identify outliers that may need manual review.



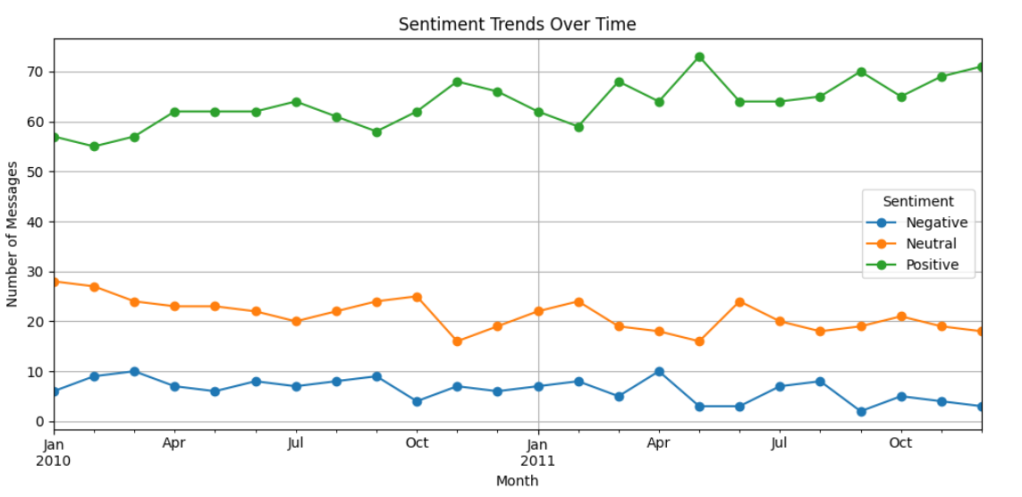
**Fig.4. Top 10 Email Senders**

Fig 4 displays bar plot which lists the employees who sent the most messages, ranked from highest to lowest. It helps identify highly communicative employees, who could be central to team workflows, or in some cases, overly burdened with communication. High senders may be in managerial or coordination roles, while low senders could be less engaged. By linking this data with sentiment scores, we can find whether frequent communicators tend to have a more positive or negative tone. This chart can also highlight workload imbalances that may warrant redistributing responsibilities.



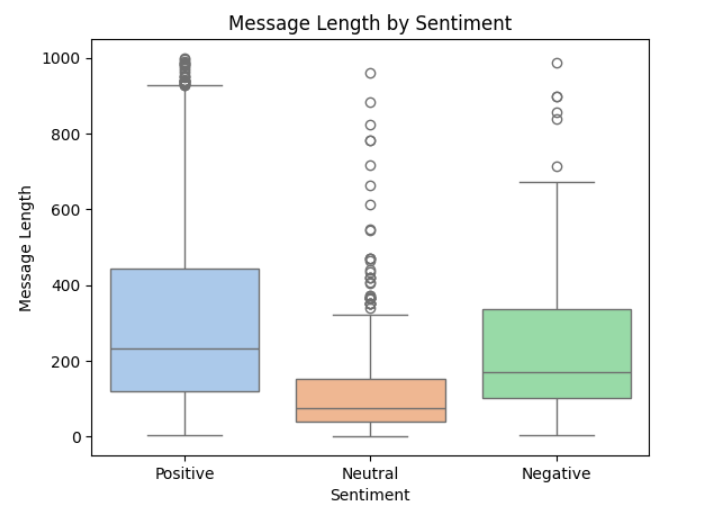
**Fig.5. Email Sent Over Time (Time Series)**

Fig 5 displays time-series plot which shows the number of emails sent each month. It reveals seasonal trends or spikes in activity, which can correlate with events like deadlines, company announcements, or crises. A rising trend may suggest increasing collaboration, while sudden drops could point to disengagement or reduced workload. Overlaying sentiment data on this chart can uncover whether busy months coincide with higher stress levels. This insight can help in resource planning and preventing burnout during peak communication periods.



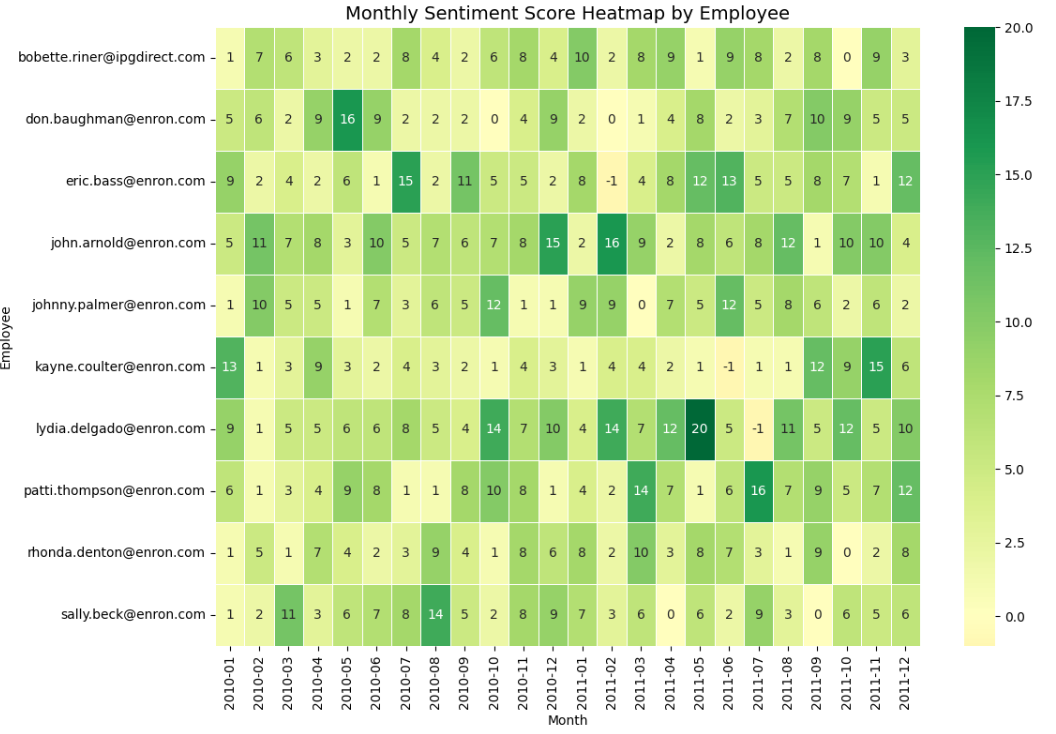
**Fig.6. Sentiment Trends Over Time(multi line chart)**

Fig 6 is a multi-line chart that displays Positive, Neutral, and Negative message counts per month. It helps spot patterns — for example, an increase in negative sentiment in certain months may indicate specific organizational stress points. Tracking these changes over time allows HR to measure the effectiveness of interventions aimed at improving morale. Sharp fluctuations might also indicate external influences such as policy changes, leadership shifts, or industry events. A stable positive trend can be a sign of healthy organizational culture.



**Fig.7. Message Length by Sentiment**

Fig 7 displays boxplot that compares message lengths for each sentiment type. It can reveal whether certain emotions are expressed in longer or shorter messages — for example, negative emails may tend to be longer due to more detailed explanations or complaints. Outliers in each category may highlight unusual cases worth deeper investigation. When combined with employee-level data, this plot can indicate if specific people consistently write longer negative messages, signaling possible ongoing dissatisfaction. It also helps in understanding whether message length has predictive value in sentiment classification models.



**Fig.8. Heat Map**

Fig 8 displays heat map that illustrates the monthly sentiment scores of employees, with each cell representing the total sentiment score for a given employee in a specific month. The color gradient ranges from light yellow (lower or more negative scores) to dark green (higher positive scores), making it easy to visually detect sentiment trends.

From the visualization, employees like lydia.delgado@enron.com, don.baughman@enron.com, and kayne.coulter@enron.com consistently show higher positive sentiment scores across multiple months, indicating a generally positive communication tone. On the other hand, some employees exhibit fluctuating patterns, where positive and negative months alternate, which could indicate changing engagement or morale over time.

### CONCLUSIONS AND FUTURE SCOPE

### Conclusions

The analysis of employee email communications has provided a detailed understanding of organizational sentiment and engagement patterns. Sentiment labeling revealed the overall tone of communication, with a clear distribution between positive, neutral, and negative messages. This breakdown is valuable for HR and management, as it serves as a baseline to monitor changes in workplace mood over time.

Exploratory Data Analysis highlighted distinct patterns in communication behavior. The distribution of message lengths, frequency of emails per month, and top communicators offered insight into workload distribution and employee engagement levels. Monthly sentiment trends indicated that negative sentiment often peaked during specific periods, suggesting potential stress points that may align with deadlines, policy changes, or organizational events.

The monthly sentiment score calculations and rankings effectively identified the most positively and negatively contributing employees each month. This ranking system can help recognize high-morale individuals who positively influence team culture, while also pinpointing employees who may require support or intervention. Flight risk identification further allowed the detection of employees who consistently send negative communications within short time spans, serving as an early-warning system for possible attrition.

Predictive modeling, initially using linear regression and later improved with advanced machine learning models like Random Forest and XGBoost, demonstrated that sentiment trends can be predicted with significant accuracy when combining communication metrics and text-based features. The model’s feature importance analysis revealed that both quantitative communication patterns (e.g., message count, length) and qualitative text features (e.g., keyword presence) play a substantial role in predicting sentiment outcomes.

Overall, this project not only achieved its primary goal of analyzing and predicting employee sentiment but also provided actionable insights for improving organizational communication health. By continuously tracking sentiment patterns, monitoring at-risk employees, and using predictive analytics, the organization can proactively address engagement challenges, improve workplace morale, and reduce turnover risks.

**Future Scope**

This project lays the groundwork for a comprehensive employee sentiment monitoring system, but several enhancements can be implemented to increase accuracy, coverage, and practical usefulness. First, the sentiment labeling process can be improved by incorporating advanced deep learning models such as BERT, RoBERTa, or GPT-based fine-tuned sentiment classifiers to capture context and nuances in employee messages more effectively than rule-based methods. Future versions could also integrate multilingual support to accommodate diverse communication in global teams.

In addition to email content, the analysis can be expanded to include other communication channels such as chat messages, meeting transcripts, and internal social networks. This would provide a more holistic view of employee sentiment and engagement. The predictive model can be upgraded using ensemble learning and neural networks, along with hyperparameter optimization, to approach near real-time forecasting of sentiment shifts.

A potential extension is the development of an interactive dashboard that visualizes sentiment trends, rankings, and flight risk alerts in real time for HR and leadership teams. This dashboard could integrate automated alerts, enabling managers to take timely action when negative sentiment spikes are detected. Furthermore, linking sentiment analysis results with organizational KPIs like productivity, retention rates, and customer satisfaction could uncover deeper relationships between employee morale and business outcomes.

Finally, the system could incorporate feedback loops where detected sentiment changes are followed up with surveys or one-on-one discussions, allowing for continuous validation and improvement of the model. Over time, such an adaptive and data-driven approach could become a core component of employee engagement strategies, ensuring a healthier and more productive workplace environment.