**Employee Sentiment Analysis**

**Introduction:**

This project aims to analyze the sentiment of email communications using natural language processing techniques. The dataset contains email content, and the primary goal is to understand the emotional tone of each message and uncover communication patterns within an organization.

Using tools like TextBlob, the project labels each email as Positive, Negative, or Neutral based on the text's polarity score. This sentiment labeling forms the foundation for further insights such as:

* Detecting employee satisfaction or discontent
* Identifying potential flight risks
* Ranking communication behavior
* Enhancing HR analytics with data-driven insights

In addition to sentiment labeling, the project also performs exploratory data analysis (EDA) to visualize trends and distributions across departments, message types, and time periods. The steps that have been taken for this project along with their explanation are given below:

**1. Sentiment Loading:**

**A screen shot of a computer code

AI-generated content may be incorrect.**

This code uses the TextBlob library to perform sentiment analysis on the text data from an email dataset. The function get\_sentiment\_textblob(text) calculates the polarity score of a given text (ranging from -1 to +1), where positive values indicate positive sentiment, negative values indicate negative sentiment, and values near zero are considered neutral. Based on the score, the function returns 'Positive', 'Negative', or 'Neutral'. The .apply() method is then used to apply this function to each entry in the 'body' column of the DataFrame, storing the results in a new column called 'Sentiment'.

**Observations:**

This sentiment labeling approach provides a quick and interpretable way to categorize emails by emotional tone. From the sample results, we observe that emails with greetings and appreciation are marked as Positive, those with unfortunate or negative context are labeled Negative, and informational or neutral content like meeting reminders are tagged as Neutral. This method is useful for detecting trends in communication and can support further analysis such as identifying employee mood, team morale, or potential HR concerns.

**2. Exploratory Data Analysis**

Summary of the dataset:

A screenshot of a computer

AI-generated content may be incorrect.

Insights from some visualizations:

A graph with blue squares

AI-generated content may be incorrect.

* Positive messages dominate the dataset, with the highest count (around 1150). This indicates that most email communications carry a constructive or appreciative tone, reflecting a generally professional or supportive workplace environment.
* Neutral messages are the second most common, with about 820 entries. These typically include routine messages such as reminders, calendar invites, or factual updates without emotional language.
* Negative messages are the least frequent (around 230), suggesting fewer emails express dissatisfaction, criticism, or negative sentiment. This could either indicate a healthy communication culture or that sensitive/negative feedback is conveyed via other channels.

A graph with different colored lines

AI-generated content may be incorrect.

* Positive sentiment remains consistently high throughout the timeline (2010–2011), showing a stable and healthy tone in workplace communication. Peaks in February 2010, June 2010, and early 2011 indicate periods of increased positivity, possibly linked to successful events or announcements.
* Neutral sentiment fluctuates more significantly, especially in mid-2010 and early 2011, which may indicate operational updates, scheduling, or system notifications increasing in volume during those months.
* Negative sentiment is consistently the lowest across all months, with occasional small spikes (e.g., October 2010 and February 2011). These could reflect isolated incidents or business challenges that triggered more critical emails.

**3. Employee Score Calculation**

**A computer code with text

AI-generated content may be incorrect.**

This code assigns numeric values to sentiment labels (Positive = 1, Neutral = 0, Negative = -1) and calculates a Monthly Sentiment Score for each employee based on their email communications. By mapping sentiments to scores and grouping the data by sender and month, it generates a new dataset that reflects the overall emotional tone expressed by each employee over time.

These scores help reveal communication patterns—employees with consistently high scores tend to express positive sentiment, while declining or negative scores may indicate dissatisfaction or stress. This enables HR or management to track morale trends, flag potential issues early, and better understand internal communication dynamics.

**4. Employee Ranking**

**A screenshot of a computer

AI-generated content may be incorrect.**

This code identifies the top 3 most positive and negative communicators (employees) each month by sorting them based on their Monthly\_Sentiment\_Score. The top positive employees are sorted in descending order of score, highlighting those who frequently use optimistic and encouraging language. In contrast, the bottom-ranking (most negative) employees are sorted in ascending order, identifying those who tend to use negative or critical tone in communication.

By tagging these with Rank\_Type and combining them into a unified employee\_ranking dataset, this approach enables monthly tracking of employee sentiment leadership and concern. It helps management recognize and appreciate consistently positive communicators while also providing early warnings about employees whose tone may signal disengagement, stress, or issues requiring attention.

**5. Flight Risk Identification**

**A screenshot of a computer code

AI-generated content may be incorrect.**

This code identifies potential flight-risk employees by analyzing the frequency of negative emails they send. It begins by filtering messages with a 'Negative' sentiment and ensuring the date column is in proper datetime format. The messages are then sorted by sender and date to prepare for time-based analysis. For each employee, the code checks if there are four or more negative messages sent within any 30-day period. If such a pattern is found, the employee is flagged as a potential risk, and their name is added to a set of at-risk individuals.

The final output is a DataFrame listing all employees who meet this criteria, sorted alphabetically. This approach provides an effective early-warning system for identifying employees who may be experiencing dissatisfaction or disengagement. It enables HR or management to intervene proactively before potential resignations, promoting employee well-being and retention.

**6. Predictive Modeling**

**A computer screen shot of a program

AI-generated content may be incorrect.**

This code builds a linear regression model to predict an employee’s monthly sentiment score based on two features: the number of messages sent in that month (Message\_Count) and a numeric representation of time (Month\_Num). It first prepares the feature set by merging message counts with the previously calculated sentiment scores and then converts the month to an ordinal value for time progression. After defining features (X) and target (y), it splits the dataset into training and test sets.

A linear regression model is trained to learn the relationship between messaging behavior and sentiment. The model's predictions are evaluated using Mean Squared Error (MSE) and R² Score, which measure prediction accuracy and model fit. Additionally, the coefficients are extracted to interpret the influence of each feature on sentiment scores. This approach enables organizations to forecast sentiment trends based on communication frequency and time, offering predictive insight into employee tone or engagement.

**Conclusion:**

This project successfully demonstrates how sentiment analysis combined with exploratory data analysis and predictive modeling can provide meaningful insights into employee communication patterns. By classifying messages, scoring sentiment, identifying top communicators, flagging flight risks, and predicting sentiment trends, the approach offers a data-driven foundation for improving organizational awareness, employee engagement, and HR decision-making.