**Sentiment Analysis of Employee Emails**

**Objective**

The primary goal of this project is to analyze the sentiment of employee emails over time, identify patterns in communication, and flag employees who may be at flight risk due to persistently negative sentiment.

**Dataset**: A cleaned version of email data.

**Steps Taken**:

* Parsed date column to extract **year**, **month**, **day**, and **weekday**.
* Removed duplicates and null entries.

**Sentiment Scoring**

* Applied **pre-trained sentiment analysis model** to calculate a **sentiment score** for each email.
* Scores range between 0 (negative) and 1 (positive).

**Result**: Positive sentiments are the highest followed by neutral and negative sentiments. This shows there is a positive working environment and achievement of goals.

**Ranking of Employees**

**Aggregating Sentiment Scores**

* Grouped the data by employee (from) and month.
* Computed the **total sentiment score per employee per month**.
  + This score reflects the **net positivity/negativity** of emails sent by an employee in a given month.
* Merged the aggregated monthly scores back into the main DataFrame for reference.

Defined a function to select:

* **Top 3 Positive Employees**: Highest monthly sentiment scores (most positive).
* **Top 3 Negative Employees**: Lowest monthly sentiment scores (most negative).

Handled tie-breaking using **employee names** in ascending order.

Iterated through each unique month.

For each month:

* Extracted a **distinct list** of employees and their monthly sentiment scores.
* Applied the ranking function to get:
  + Top 3 positive employees
  + Top 3 negative employees

**Flight Risk Employees**

* Extracted only those emails labelled with ‘Negative’ sentiment from the dataset.
* Created a separate Data Frame (negative\_df) containing only negative sentiment records.
* Sorted the filtered data by:

from (employee who sent the email)

date (email timestamp)

* Ensured chronological order of emails for each employee to facilitate rolling window analysis.
* For each employee:

Iterated through their negative emails.

For each email, defined a **30-day window** starting from the date of that email.

Counted the number of negative emails sent within that window.

If the **count of negative emails within any 30-day period ≥ 4**, the employee was flagged as **flight risk**.

**Predictive Modelling**

* Implemented a **Linear Regression** model to predict sentiment scores based on:
  + Day of the month
  + Weekday
  + Month number
  + Year
  + Email message count
* **Model Evaluation**:
  + Mean Squared Error (MSE): 0.4402
  + R² Score: -0.0071 → Indicates **poor predictive power**, suggesting the need for more complex models or additional features.
* The **year** had the highest impact on sentiment trends, likely due to organizational changes or evolving communication culture.
* Day, month had minimal impact individually.

**I have included the detailed explanation of codes and interpretation in markdown cells.**