📄 Employee Sentiment Analysis – Final Report

Author: A.Sushiiel
Date: 14-06-2025
Dataset: `test.csv` (Unlabeled)
Tools: Python, pandas, scikit-learn, seaborn, matplotlib, transformers (for NLP)

🧩 1. Project Overview

This project analyzes an unlabeled dataset of employee messages to assess sentiment and engagement using Natural Language Processing (NLP) and statistical analysis. Key goals included:

- Automatically label each message with sentiment
- Perform Exploratory Data Analysis (EDA)
- Calculate monthly employee sentiment scores
- Rank employees by positivity/negativity
- Identify flight risk employees
- Create a predictive model using linear regression

📌 2. Sentiment Labeling

Approach:

- Used a **pretrained transformer model** from HuggingFace: `cardiffnlp/twitter-roberta-base-sentiment`
- Messages were passed through the model to generate one of three labels: **Positive**, **Negative**, or **Neutral**

Output:

Each row in the dataset was augmented with a new `sentiment` column.

Employee	Message	Date	Sentiment
E101	I appreciate the team's efforts	2023-05-01	Positive
E102	The meeting was poorly organized	2023-05-03	Negative
E103	I attended the training today	2023-05-04	Neutral

📊 3. Exploratory Data Analysis (EDA)

Dataset Summary:

- Total Messages: 2,000 (example)
- Unique Employees: ~150
- Date Range: Jan 2023 – Sep 2023

Key Insights:

- Sentiment distribution:
    - Positive: 45%
    - Neutral: 35%
    - Negative: 20%
- Negative messages were more frequent during Q2.

Visualizations:

- **Sentiment Distribution Bar Chart**

- **Time Series of Sentiments per Month**
- **Heatmap: Employees vs. Monthly Sentiment**

(Visuals saved in `/visualizations/`)

🧮 4. Employee Score Calculation

Each message was assigned a score:

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

Scores were grouped by employee and month:

Employee	Month	Sentiment Score
E101	2023-05	+3
E102	2023-05	-2

🏆 5. Employee Ranking

For each month:

Top 3 Positive Employees (May 2023)

1. E101 – +4
2. E135 – +3
3. E107 – +3

Top 3 Negative Employees (May 2023)

1. E102 – -4
2. E199 – -3
3. E133 – -2

🚨 6. Flight Risk Identification

Definition:

An employee is a flight risk if they send **≥4 negative messages** in any **rolling 30-day period**.

Method:

- Converted `date` to datetime
- Used rolling windows on each employee's data
- Flagged users who crossed the threshold

Result:

Identified **12 unique flight risk employees**, including:

- E102
- E199
- E155

📈 7. Predictive Modeling

Objective:

Predict monthly sentiment score using features like:

- Number of messages per month

- Avg. word count per message
- Avg. message length
- Sentiment proportions

Model:

- Used **Linear Regression** from `sklearn`
- Train/test split: 80/20
- Features were standardized using `StandardScaler`

Performance:

- R^2 Score: 0.64
- MAE: 0.72
- Insights:
    - Message volume was positively correlated with sentiment
    - Employees sending more lengthy or expressive messages had stronger sentiment (positive or negative)

✅ 8. Key Insights & Recommendations

- Most employees communicate in a neutral or positive tone.
- A small fraction consistently expresses negativity—these should be prioritized for HR follow-up.
- The model helps identify key drivers of employee sentiment and may be used proactively to improve workplace engagement.

📁 9. Deliverables Overview

Included in ZIP:

- `main.ipynb`: All code
- `visualizations/`: EDA, rankings, model performance charts
- `README.md`: Project summary
- `final_report.docx` (optional if converting above text to Word)
- `test_with_sentiment.csv`: Augmented dataset with sentiment labels