Project : Employee Sentiment Analysis

NOTE : DO NOT SAVE ANY INFORMATION INTO A WORD DOCUMENT. THIS IS AN INTERNAL EVALUATION AND YOU SHOULD NOT SHARE ANY OF THE INFORMATION MENTIONED IN THIS DOCUMENT OR ELSE WERE IN PUBLIC. THIS IS NOT A TEAM PROJECT, AND IT IS EXPECTED YOU TO COMPLETE THIS ON YOUR OWN.

Dataset: test.csv (Unlabeled)

Coding Language : Python

***Use Pytorch or sklearn library for AI modeling**

***The goal of the project is to assess how good you can solve this problem statement with speed and cleverness. You can use additional libraries if you think that can simplify your solution.**

SUBMISSION (zip file) :

- Once you have completed the assignment you should make a zip file containing the "deliverables" mentioned below and email it with a zip file as attachment to me at : jbirch@glynac.ai
- The subject of the email should be "AI-project-submission".
- Include in the email your full name.
- Message me on Microsoft teams at jbirch@glynac.ai that you finished the training and would like an evaluation.

1. Project Overview

This project involves analyzing an unlabeled dataset of employee messages to assess sentiment and engagement. Your task is to work from raw data and derive insights using natural language processing (NLP) and statistical analysis techniques. The project is divided into several distinct tasks, each focusing on a different aspect of data analysis and

model development. Your final deliverables should include a well-documented codebase, a comprehensive report, visualizations, and a clear summary of findings.

2. Project Objective

The main goal is to evaluate employee sentiment and engagement by performing the following:

- **Sentiment Labeling:** Automatically label each message as Positive, Negative, or Neutral.
- **Exploratory Data Analysis (EDA):** Analyze and visualize the data to understand its structure and underlying trends.
- **Employee Score Calculation:** Compute a monthly sentiment score for each employee based on their messages.
- **Employee Ranking:** Identify and rank employees by their sentiment scores.
- **Flight Risk Identification:** A Flight risk is any employee who has sent 4 or more negative mails in a given month.
- **Predictive Modeling:** Develop a linear regression model to further analyze sentiment trends.

3. Detailed Tasks

Task 1: Sentiment Labeling

- **Objective:**

Label each employee message with one of three sentiment categories: Positive, Negative, or Neutral.

- **Requirements:**
    - Work with the provided test.csv dataset.
    - **Preferably use a large language model (LLM)** or any suitable NLP technique to determine the sentiment of each message.
    - Augment the dataset with an additional column that indicates the sentiment label for each message.
    - Document your chosen approach for labeling.

- **Notes:**

Ensure that the labeling criteria are clearly justified and reproducible.

Task 2: Exploratory Data Analysis (EDA)

- **Objective:**

Understand the structure, distribution, and trends in the dataset through thorough exploration.

- **Requirements:**
    - Examine the overall data structure (e.g., number of records, data types, missing values).
    - Investigate the distribution of sentiment labels across the dataset.
    - Analyze trends over time.
    - Explore additional patterns or anomalies that could provide insights into employee engagement.
    - Prepare visualizations (charts, graphs, tables) that effectively communicate your findings.
- **Notes:**

The EDA should form a solid foundation for later tasks by highlighting key insights and areas for further analysis.

Task 3: Employee Score Calculation

- **Objective:**

Compute a monthly sentiment score for each employee based on their messages.

- **Requirements:**
    - For each employee, assign a score to each message:
        - **Positive Message:** +1
        - **Negative Message:** –1
        - **Neutral Message:** 0 (no effect)

- Aggregate these scores on a monthly basis for each employee.
- Ensure that the score resets at the beginning of each new month.
- Clearly document your method for grouping messages by month and calculating the cumulative score.

- **Notes:**

Accuracy in the aggregation process is essential, as the resulting scores are used for subsequent ranking and risk analysis.

Task 4: Employee Ranking

- **Objective:**

Generate ranked lists of employees based on their monthly sentiment scores.

- **Requirements:**
  - Create two distinct lists:
    - **Top Three Positive Employees:** The three employees with the highest positive scores in a given month.
    - **Top Three Negative Employees:** The three employees with the lowest (most negative) scores in each month.
  - Sort them first in descending order and then in alphabetical order.
  - Ensure that the ranking is clearly derived from the monthly scores calculated in Task 3.
  - Present the rankings in a clear and organized format (e.g., tables or charts).
- **Notes:**

Your report should include a brief discussion of how these rankings were determined.

Task 5: Flight Risk Identification

- **Objective:**

Identify employees who are at risk of leaving based on their monthly sentiment scores.

- **Requirements:**

    - A Flight risk is any employee who has sent 4 or more negative mails in the span of 30 days (irrespective of the score).
    - The 30-day period is rolling count of days, irrepective of months.
    - Extract a list of these employees
    - Ensure that this flagging process is robust.
- **Notes:**

This task is critical for identifying potential issues in employee engagement and retention.

Task 6: Predictive Modeling

- **Objective:**

Develop a linear regression model to analyze sentiment trends and predict sentiment scores using a variety of independent variables that may influence sentiment scores.

Select independent variables (features) you believe may influence sentiment scores. You are encouraged to use:

- **Requirements:**
    - Select appropriate features from the dataset that may influence sentiment scores (e.g., message frequency in a month, message length, average message length, word count).
    - Split the data into training and testing sets to evaluate model performance.
    - Develop a linear regression model and validate its effectiveness using suitable metrics.
    - Interpret the model results and discuss the significance of the findings.
- **Notes:**

The predictive modeling task is aimed at exploring how sentiment trends can be quantified and forecasted. Ensure that your model evaluation is clearly documented.

4. Deliverables

Upload the following in a GitHub repo and share the link to that repo.

- **Code Submission (ipynb, py or other files):**
    - The main file where **all the processes are done should be in an ipynb file**. All other files (if exists) should be supporting the code implemented in the ipynb file.
    - A complete codebase implementing all tasks (including ipynb or other files).
    - Well-commented and structured code to facilitate understanding and reproducibility.
    - Include your observations, titles and describing comment about the steps you are taking. This should be such that the ipynb file can be read along with the comments etc and be understood your process.
    - Give title to each section, provide commentary and observations.
- **Final Report (doc / docx file):**
    - A detailed document that includes:
        - A description of your approach and methodology.
        - Key findings from the EDA.
        - Explanation of the employee scoring and ranking processes.
        - Flight risk identification criteria and outcomes.
        - Overview and evaluation of the predictive model.
    - Include visualizations and tables to support your conclusions.
- **Visualizations (image in a 'visualization' folder):**
    - Charts and graphs summarizing the EDA.
    - Visual representation of employee rankings and flight risk analysis.
    - Plots or graphs that illustrate model performance metrics.
- **Readme file with Summary (README.md file) :**
    - A concise summary that highlights:
        - The top three positive and negative employees.
        - The list of employees flagged as flight risks.
        - Key insights and recommendations based on your analysis.

5. Additional Guidelines

- **Documentation:**

Maintain clear documentation of your process, decisions, and any assumptions made during the project.

- **Clarity and Organization:**

Ensure that the final deliverables are organized and easy to follow. Use headings, subheadings, and bullet points where appropriate.

- **Testing and Validation:**

Validate your methods at each stage of the project. This includes verifying data integrity, ensuring correct calculation of scores, and proper model evaluation.

- **Reproducibility:**

The project should be structured so that another team member can reproduce your results from raw data to final outputs.

This detailed problem statement is intended to guide you through the project requirements without providing specific solution approaches. It is designed to ensure clarity on what needs to be achieved while leaving the method of implementation open for you to design and document. Good luck, and we look forward to reviewing your innovative solutions!