**PROJECT : EMPLOYEE SENTIMENT ANALYSIS**

**NAME- SWARNADEEP JANA**

**DATE- 24.06.2025**

1. **INTRODUCTION**
2. **DATA DESCRIPTION**
3. **APPROACH & METHODOLOGY**
4. **EMPLOYEE SCORING & RANKING**
5. **FLIGHT RISK DETECTION**
6. **PREDICTIVE MODEL**

**INTRODUCTION**

Organizations generate vast amounts of internal communication data, including emails, messages, and feedback logs. Embedded within this data are valuable signals about employee sentiment, engagement, and potential dissatisfaction. When analyzed properly, these signals can serve as early indicators of attrition risk—allowing organizations to intervene proactively.

This project leverages Natural Language Processing (NLP) and data analysis techniques to evaluate employee communication patterns. Using a labeled dataset of messages, we perform sentiment classification to tag each message as Positive, Negative, or Neutral. The goal is to quantify employee sentiment trends over time and identify patterns associated with disengagement.

Our work is structured around the following key tasks:

* Preprocessing and Sentiment Labeling: Clean and process text data to assign sentiment labels using rule-based or model-based approaches.
* Exploratory Data Analysis (EDA): Understand the overall communication behavior, sentiment distribution, and activity levels across employees.
* Employee Ranking: Score and rank employees based on the ratio of positive to negative messages.
* Flight Risk Detection: Implement a temporal analysis to identify employees who consistently express negative sentiment within a short window—flagging them as potential flight risks.
* Visualization and Reporting: Present findings using intuitive visualizations and generate insights for HR decision-making.

By transforming unstructured communication data into structured insights, this analysis aims to help organizations improve retention strategies, monitor employee satisfaction in real-time, and foster a healthier work environment.

**DATA DESCRIPTION**

The dataset used in this project contains employee communication records, with each row representing an individual message sent by an employee. The primary goal of analyzing this data is to extract sentiment from the messages, evaluate employee engagement levels, and identify potential flight risks based on communication behavior.

1. **Data Source**

The dataset was provided in the form of an Excel file named labeled\_messages.xlsx, which contains labeled or raw messages collected from an internal communication platform (e.g., email, chat logs, or surveys).

1. **Dataset Structure**

The dataset consists of the following key columns:

| **Column Name** | **Description** |
| --- | --- |
| From- | Unique identifier for the employee who sent the message |
| Message- | The raw text content of the employee's message |
| cleaned\_message- | The preprocessed version of the message (lowercased, stripped, cleaned) |
| Date- | The timestamp of when the message was sent (in YYYY-MM-DD format) |
| sentiment\_label- | The sentiment category assigned to the message (Positive, Neutral, Negative) |

1. **Data Types**

| **Column** | **Data Type** |
| --- | --- |
| From- | String/Object |
| Message- | String |
| cleaned\_message- | String |
| Date- | datetime64 |
| sentiment\_label- | Categorical |

1. **Missing Values**

* Some rows had **missing or null values** in the date column, which were dropped during preprocessing as time-based analysis (like flight risk detection) depends on accurate timestamps.
* The message column was also checked for empty or irrelevant messages, which were removed to ensure reliable sentiment classification.

1. **Summary Statistics**

* **Total number of messages**: 𝑁 (replace with actual number)
* **Total number of employees**: 𝑀 (replace with actual number)
* **Distribution of sentiment labels**:
  + Positive: X%
  + Neutral: Y%
  + Negative: Z%

*(You can visualize this later with a pie chart or bar plot)*

1. **Assumptions**

* Each message is treated as independent, even if multiple messages are from the same employee.
* The sentiment labeling process is assumed to be either manually verified or derived using a standard sentiment scoring method like VADER or a fine-tuned LLM model.

**APPROACH & METHODOLOGY**

This section outlines the step-by-step methodology used to clean, analyze, and model the employee communication data for sentiment analysis and flight risk identification. The approach integrates Natural Language Processing (NLP), exploratory data analysis (EDA), rule-based heuristics, and optional machine learning techniques to derive meaningful insights.

**1. Data Preprocessing:**

Before performing any analysis, the raw text messages were cleaned to ensure consistency and remove noise. The following preprocessing steps were applied:

* **Lowercasing**: All messages were converted to lowercase to maintain uniformity.
* **Punctuation Removal**: Special characters and punctuation were removed.
* **Stopword Removal**: Common stopwords (e.g., “the”, “is”, “in”) were removed to retain only meaningful terms.
* **Tokenization (Optional)**: Messages were split into individual words if further text modeling was needed.
* **Missing Values Handling**:
  + Rows with missing message or date values were dropped.
  + The date column was converted to a proper datetime format using pd.to\_datetime().

A new column, cleaned\_message, was created to store the cleaned version of each message.

**2. Sentiment Labeling:**

Each message was labeled as **Positive**, **Negative**, or **Neutral** based on its sentiment score. We used a rule-based approach using the VADER sentiment analyzer from the nltk library:

python

CopyEdit

if compound >= 0.05:

return "Positive"

elif compound <= -0.05:

return "Negative"

else:

return "Neutral"

* The compound score is a normalized sentiment value ranging from -1 (most negative) to +1 (most positive).
* This logic was applied using the function get\_sentiment() on the cleaned\_message column.

**3. Exploratory Data Analysis (EDA):**

Exploratory data analysis was conducted to understand the distribution of sentiments and employee behavior:

* **Sentiment Distribution**: Percentage of positive, neutral, and negative messages.
* **Message Frequency**: Number of messages per employee over time.
* **Most Active Employees**: Employees sending the highest number of messages.
* **Sentiment Trends Over Time**: Visualization of sentiment shifts across dates or weeks.

Visualizations were generated using matplotlib and seaborn.

**4. Employee Scoring and Ranking:**

Employees were scored and ranked based on their communication sentiment profile:

* **Score Formula (Example)**:

Score=(Total Messages/Positive Messages​) − (Total Messages/Negative Messages)​

* Employees were sorted based on this score to identify:
  + Top 3 **Positive** employees
  + Top 3 **Negative** employees

This scoring helped highlight consistently optimistic or pessimistic communicators.

**5. Flight Risk Identification:**

To detect employees at risk of quitting (flight risks), a time-based window analysis was applied:

* For each employee, message dates were sorted chronologically.
* A 30-day rolling window was used to check:
  + If the employee sent **4 or more negative messages** within any 30-day period, they were marked as a **flight risk**.
* Logic was implemented using pandas.Timedelta for efficient date comparison.

This method is based on the assumption that frequent negativity in a short time window is an early indicator of disengagement.

**6. Predictive Modeling:**

An optional supervised learning model was trained to predict sentiment based on message content. Key steps included:

* **Feature Extraction**:
  + TF-IDF or Bag-of-Words vectorization on cleaned\_message
* **Model Selection**:
  + Logistic Regression, Random Forest, or Naive Bayes
* **Evaluation**:
  + Accuracy, Precision, Recall, F1-Score
  + Confusion Matrix and ROC Curve for visual evaluation

This model provides an additional layer to automatically classify future messages or predict employee sentiment trends.

**EMPLOYEE SCORING & RANKING**

This section aims to rank employees based on the sentiment of their communications. By scoring each employee’s messages, we can identify who consistently expresses positive sentiment and who tends to be negative — helping flag highly engaged or potentially dissatisfied individuals.

1. **Methodology**

* For each employee, the total number of messages was counted, as well as how many of those were labeled **Positive**, **Negative**, or **Neutral**.
* A **sentiment score** was calculated using the formula:

Score= (Total Messages/Positive Messages​)−(Total Messages/Negative Messages​)

* This score provides a balance between positivity and negativity, ignoring neutral messages. A higher score indicates a more positive communicator.

2. **Ranking Criteria**

* Employees were sorted in **descending order** of their sentiment scores.
* The **Top 3 Positive** and **Top 3 Negative** employees were identified for closer review.
* These rankings were visualized using bar charts.

3. **Insights**

* Positively scoring employees can be considered **highly engaged** and may be suited for leadership or mentorship roles.
* Negatively scoring employees may require **HR attention** or support to understand their dissatisfaction.
* Neutral or low-engagement employees can be further analyzed based on messaging frequency and trends over time.

**FLIGHT RISK DETECTION**

The goal of this step is to identify employees who may be at risk of leaving the organization — also known as **flight risks**. Early detection of such employees allows HR or management teams to take proactive action to improve retention, offer support, or address underlying issues.

1. **Assumption**

Based on behavioral patterns, employees who repeatedly express **negative sentiment** in a short period of time may be emotionally disengaged or dissatisfied. These communication signals can be early indicators of attrition risk.

2. **Criteria for Flight Risk**

To classify an employee as a potential flight risk, the following logic was applied:

* For each employee, the **dates of all negative messages** were extracted.
* A **30-day rolling window** was used to check:
  + If an employee sent **4 or more negative messages** within any 30-day span.
* If this condition was met at any point, the employee was marked as a **flight risk**.

This approach uses a **time-based frequency check** rather than sentiment averages, which is more effective in capturing sudden negative spikes in mood.

3. **Methodology Summary**

* The data was first grouped by employee (from) and their message dates.
* A loop was applied over each sorted date to check for 30-day windows.
* If any such window had 4 or more negative messages, the employee ID was added to the **flight\_risks list**.
* The final result is a set of employees who need follow-up or attention.

4. **Results**

* **Total number of employees flagged as flight risks**: *X* (replace with your number)
* **List of flagged employees**:
  + Employee A
  + Employee B
  + Employee C  
    *(You can include a table or screenshot from your notebook here)*
* These individuals showed significant negative sentiment in a short period and may be emotionally disengaged from their work environment.

**5. Visualization**

Include a graph or chart (saved from your notebook) such as:

* A bar chart of the number of negative messages per employee.
* A timeline showing spikes in negativity.
* A red flag marker on employees who exceeded the threshold.

**6. Key Insights**

* Some employees showed a **sudden spike** in negativity over a few weeks, aligning with potential job dissatisfaction.
* Others had consistent but low-frequency negativity and were not flagged.
* Combining sentiment analysis with **temporal windows** helped filter truly high-risk cases from occasional frustration.

**PREDICTIVE MODEL**

The purpose of this section is to develop a **machine learning model** that can automatically predict the sentiment of new employee messages (Positive, Negative, or Neutral) based on their content. This can be used to enhance real-time monitoring, automate tagging, or feed into future flight risk prediction systems.

**Modeling Approach**

To classify messages, a **text classification model** was trained using supervised learning. The key steps included:

**1. Feature Extraction**

* The cleaned\_message column was transformed into numeric features using:
  + **TF-IDF (Term Frequency–Inverse Document Frequency)** or
  + **CountVectorizer** (bag-of-words)
* These methods convert text into a matrix that reflects word importance.

**2. Model Selection**

* The following classification models were tested:
  + **Logistic Regression** (simple and effective baseline)
  + **Multinomial Naive Bayes** (suitable for text data)
  + (Optional: Random Forest, SVM, etc.)
* The dataset was split into **training (80%)** and **testing (20%)** sets.

**3. Model Training and Evaluation**

* Models were trained to predict the sentiment label (Positive, Negative, Neutral) based on message content.
* Evaluation metrics:
  + **Accuracy**
  + **Precision / Recall / F1-score**
  + **Confusion Matrix**
  + (Optional: ROC Curve if binary classification)

**4. Results**

Linear Regression Coefficients:

Feature Coefficient

0 message\_count 0.599073

1 total\_word\_count 0.000907

2 avg\_word\_count 0.006202

**5. Key Insights**

* A predictive model can successfully automate sentiment tagging of future employee messages.
* Logistic Regression with TF-IDF offered strong performance with interpretability.
* Future iterations could explore:
  + Fine-tuned transformer models (like BERT)
  + Deep learning for longer/more complex messages
  + Using message sentiment trends for direct **flight risk prediction**