Final Report: Employee Sentiment Analysis

# Introduction

This project focuses on uncovering actionable insights from internal employee communications through advanced **Natural Language Processing (NLP)** and **Machine Learning (ML)** techniques. The primary data source is a collection of internal emails, including subjects and bodies of messages, which are processed to assess employee sentiment, engagement levels, and potential flight risks.

The core objective of this analysis is to support **Human Resource (HR)** and **People Analytics** teams in proactively identifying employee dissatisfaction or disengagement that may lead to attrition. The project follows a structured methodology comprising the following key stages:

* **Sentiment Labelling**: Each email is evaluated using the VADER Sentiment Analyzer to classify content as Positive, Neutral, or Negative based on sentiment scores. This helps in quantifying the emotional tone of communication over time.
* **Exploratory Data Analysis (EDA)**: Patterns and trends are visualized and statistically examined to understand overall sentiment distribution, frequent senders, and departments with high negative sentiment scores.
* **Employee Scoring and Ranking**: Each employee is assigned a sentiment score based on the cumulative sentiment of their emails. This score is used to rank employees, allowing HR to identify top performers or employees showing early signs of disengagement.
* **Flight Risk Detection**: Employees who send four or more negative emails within a 30-day window are flagged as potential flight risks. This temporal rule-based approach enables HR to act pre-emptively.
* **Predictive Modelling**: A machine learning model (Random Forest Regressor) is trained using TF-IDF features extracted from email text to predict sentiment intensity. This provides a scalable method to evaluate new communications for early warning signals.

# 2. Methodology

## 2.1 Data Loading & Preprocessing

- The dataset was loaded from a .csv file containing email metadata (subject, body, sender, and date).  
- Subject and body were combined into a new full\_text column to form a meaningful text representation.  
- Missing values in text and dates were handled appropriately.  
- Employee names were extracted from the from field, and date fields were parsed to datetime.

## 2.2 Sentiment Analysis

- Used VADER (Valence Aware Dictionary and sEntiment Reasoner) from the vaderSentiment library.  
- Generated four sentiment scores for each email: positive, negative, neutral, and compound.  
- Labeled sentiment based on the compound score:  
 - ≥ 0.05 → Positive  
 - ≤ -0.05 → Negative  
 - Between → Neutral

## 2.3 Sentiment Scoring

- Each sentiment label was assigned a score:  
 - Positive → +1  
 - Neutral → 0  
 - Negative → -1  
- Scores were aggregated per employee to evaluate general tone and communication behavior.

# 3. Exploratory Data Analysis (EDA)

## 3.1 Email Volume Overview

- Over 13,000 emails analyzed.  
- Average email length: ~220 characters.  
- Communication concentrated around specific business dates.

## 3.2 Sentiment Distribution

|  |  |
| --- | --- |
| Sentiment | Percentage |
| Positive | ~75% |
| Neutral | ~20% |
| Negative | ~5% |

## 3.3 Top Negative Email Senders

- A small group of employees had consistently negative tone across multiple emails.  
- These individuals were flagged for further analysis as potential disengagement risks.

# 4. Employee Scoring & Ranking

## 4.1 Scoring Logic

- For each employee:  
 - Total Sentiment Score = Sum of all individual email sentiment scores.  
 - Normalized Score = Total Sentiment Score / Total Emails Sent

## 4.2 Categories Based on Scores

- Highly Engaged: Score ≥ +0.5  
- Moderately Engaged: Score between 0 and +0.5  
- Slightly Negative: Score between -0.3 and 0  
- At Risk: Score ≤ -0.3

## 4.3 Sample Results

|  |  |  |  |
| --- | --- | --- | --- |
| Employee | Total Emails | Score | Status |
| kayne | 62 | +14 | Highly Engaged |
| eric | 49 | +9 | Engaged |
| lydia | 34 | +9 | Engaged |

# 5. Flight Risk Identification

## 5.1 Criteria

- Employees who sent ≥4 negative emails within 30 consecutive days were flagged as flight risks.

## 5.2 Identified Employees

|  |  |  |  |
| --- | --- | --- | --- |
| Employee | Period Start | Period End | Negative Emails |
| Bobette | 05-03-2011 | 04-04-2011 | 4 |
| Eric Bass | 22-04-2011 | 22-05-2011 | 4 |

# 6. Predictive Modeling

## 6.1 Objective

Predict an email's sentiment compound score using the textual content.

## 6.2 Feature Engineering

Used TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer to extract top 500 features from email text.

## 6.3 Model Used

Random Forest Regressor from scikit-learn. Model was trained on 80% of the dataset and tested on 20%.

## 6.4 Evaluation Metrics

|  |  |
| --- | --- |
| Metric | Value |
| R² Score | 0.6128 |
| Mean Squared Error | 0.6998 |
|  |  |

## 6.5 Sample Output Columns

Employee, Original Text, Actual Sentiment, Predicted Sentiment

# 7. Conclusion

Email sentiment analysis provides valuable insights into employee morale and communication tone.  
Predictive modeling can help anticipate sentiment and potential issues proactively.  
Integrating this framework into HR analytics enables early detection of burnout or dissatisfaction.