**FINAL REPORT**

**Approach and Methodology**

**Introduction**

The primary aim of this analysis is to understand, quantify, and forecast employee sentiment trends through a combination of natural language processing, feature engineering, and predictive modeling. By leveraging message-level data and behavioral indicators, the goal is to build a robust analytical framework that not only classifies sentiment but also identifies patterns that may signal disengagement or flight risk.

Specifically, the analysis seeks to:

* **Data Ingestion**: test dataset was loaded from a excel file.
* **Text Preprocessing**: The dataset was cleaned by removing special characters, converting text to lowercase, and trimming extra spaces.
* **Label employee messages**: Labeled sentiment categories (Positive (0.08), Negative (-0.08) and Neutral) using TextBlob, enabling structured sentiment tracking across time and individuals.
* **Explore the dataset thoroughly**: Uncover structural insights, sentiment distributions, temporal trends, and anomalies that may reflect shifts in employee engagement.
* **Quantify sentiment dynamics**: Computing monthly sentiment scores per employee, enabling comparative analysis and trend detection.
* **Identify flight-risk employees**: Used a rolling 30-day window of negative message activity, providing actionable intelligence for retention strategies.
* **Develop a predictive model**: A linear regression was built to estimate sentiment scores based on linguistic and behavioral features such as message length, word count, sentiment ratio etc.
* **Evaluate model performance**: MAE and R2 to evaluate the performance of the model.

Ultimately, this analysis aims to transform unstructured communication data into meaningful insights that support proactive human resource interventions, foster a healthier workplace culture, and enhance strategic decision-making around employee engagement and retention.

**The exploratory data analysis highlighted the following**:

* **Missing Values**: The dataset was mostly complete except for the “**body”** feature that showed 35 rows of missing data and the rows were dropped.

Missing values:

Subject 0

body 35

date 0

from 0

sentiment 0

dtype: int64

* Summary of the descriptive statistics from the dataset
* **Duplicates**: The dataset had no significant duplication.
* **Sentiment Distribution**: Visualization showed most messages were 1152 Positive, 769 Neutral and 235 negative emails.
* **Message Length Analysis**: Compared message length across sentiment categories. Negative messages tended to be longer, suggesting detailed complaints or emotional expression.
* **Outliers Removal:** Interquartile range technique was used to reduce longer negative messages which pose as outliers.
* **Monthly Sentiment Trend:** Sentiment counts were grouped by month and a line plot showed fluctuations in sentiment volume over time. Negative sentiment spikes aligned with specific months, possibly indicating stress periods or organizational changes. Positive sentiment increased during celebratory periods example holidays, bonuses

**Employee Scoring and Ranking**

The employee engagement levels by assigning sentiment-based scores to individual messages and aggregating them monthly. This scoring system enabled the identification of employees with consistently positive contributions as well as those exhibiting signs of disengagement or dissatisfaction. The rankings derived from these scores serve as a foundation for performance monitoring, recognition, and early intervention strategies.

* **Score Label and score**:
  + POSITIVE = +1
  + NEGATIVE = -1
  + NEUTRAL = 0

Two distinct rankings were generated for each month:

**Top Three Positive Employees**: Employees with the highest cumulative sentiment scores. Sorted by score (descending), then alphabetically to break ties.

Top Three Positive Employees:

from year\_month monthly\_sentiment\_score

112 johnny.palmer@enron.com 2011-06 14

222 sally.beck@enron.com 2010-08 14

185 patti.thompson@enron.com 2011-07 13

**Top Three Negative Employees**: Employees with the lowest cumulative sentiment scores. Sorted by score (ascending), then alphabetically.

Top Three Negative Employees:

from year\_month monthly\_sentiment\_score

229 sally.beck@enron.com 2011-03 -3

131 kayne.coulter@enron.com 2011-01 -1

183 patti.thompson@enron.com 2011-05 -1

**Flight Risk Identification Criteria and Outcomes**

The essence of Flight Risk Identification was to proactively detect potential employee disengagement by flagging individuals who exhibit sustained negative communication behavior. By applying a rolling 30-day window to sentiment-labeled messages, we identified employees who may be at risk of attrition, enabling timely intervention and support

* **Criteria**: Employees with consistently negative sentiment scores (total score ≥ 4) in the span of 30 days span were flagged as high risk.
* **Outcome**: 10 individuals showed patterns of negative communication and were shortlisted for review.

employee\_id

0 eric.bass@enron.com

1 johnny.palmer@enron.com

2 lydia.delgado@enron.com

3 patti.thompson@enron.com

4 sally.beck@enron.com

5 john.arnold@enron.com

6 bobette.riner@ipgdirect.com

7 [rhonda.denton@enron.com](mailto:rhonda.denton@enron.com)

8 [kayne.coulter@enron.com](mailto:kayne.coulter@enron.com)

9 don.baughman@enron.com

**Predictive Model overview and Evaluation**

Predictive modeling task was to quantify sentiment trends and forecast sentiment scores using structured features derived from unstructured message data. By modeling sentiment as a continuous score, the analysis enables nuanced tracking of employee engagement and supports early detection of dissatisfaction or morale shifts.

**Feature Engineering**: A diverse set of features was extracted to capture both linguistic tone and behavioral patterns:

* max\_len\_words - The maximum number of words in a single message by a user
* active\_days - The number of days a user has been active (sending at least one message).
* pos\_ratio - Ratio (proportion) of messages with positive sentiment relative to total messages.
* neg\_ratio - Ratio (proportion) of messages with negative sentiment relative to total messages.
* lag1\_neg\_inter - Interaction between previous time-step behavior and negative sentiment (a lagged negative ratio feature).
* active\_days \* pos\_ratio - Users who are both active and positive show stronger performance.
* active\_days \* neg\_ratio - Users who are active but negative reduce performance.
* max\_len\_words \* active\_days - Users who are active and write longer messages contribute positively.
* max\_len\_words \* neg\_ratio - Long messages that are negative slightly reduce outcomes.
* max\_len\_words \* pos\_ratio - Long messages with positive sentiment slightly increase outcomes.
* max\_len\_words \* lag1\_neg\_inter - max\_len\_words \* lag1\_neg\_inter
* active\_days \* lag1\_neg\_inter - Active users with past negative behavior.
* pos\_ratio \* neg\_ratio - Users expressing both positive and negative sentiment.
* pos\_ratio \* lag1\_neg\_inter - Positive communication interacting with past negativity.
* neg\_ratio \* lag1\_neg\_inter - Current negativity combined with a history of negativity.

**LassoCV** for feature selection, as it is well-suited to high-dimensional data and automatically performs variable shrinkage to reduce redundancy. Features with **non-zero coefficients** after Lasso regularization were retained.

Selected features after LassoCV: ['max\_len\_words', 'active\_days', 'pos\_ratio', 'neg\_ratio', 'lag1\_neg\_inter']

Missing values and infinities were handled using imputation and replacement strategies to ensure model stability.

**Target**: sentiment\_score

Model Approach:

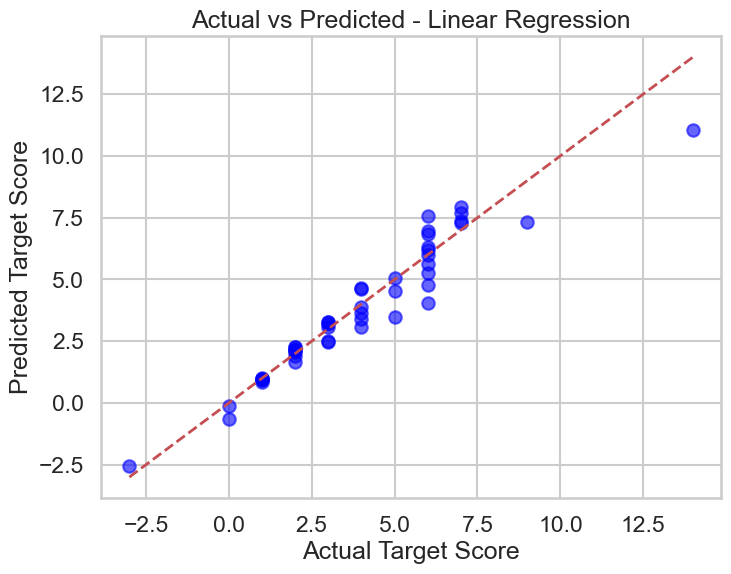
* Model Type: Linear Regression

**Feature Standardization**: (scaler, StandardScaler()). Normalizes all features to have **mean = 0** and **standard deviation = 1.**

Training method: 80/20 Train-Test split

**Evaluation**:

* R² Score: Assessed how well sentiment explained variation in the target. The model explains **92.1% of the variance** in sentiment scores. This is a very strong indicator of predictive accuracy, especially considering the complexity and noise inherent in text-derived features.
* MAE: On average, predictions differ from actual sentiment scores by **~0.53 units**. MAE 0.525 given that the **target score ranges from -2 to 15** (a 17‑unit span), this error represents only **~3.1% of the possible range.** This reflects **high predictive accuracy with low practical error.**



**Feature Coefficient:**

|  |  |  |
| --- | --- | --- |
| active\_days | 2.318322 | Strongest predictor. More active days per sender/month strongly **increase sentiment score.** |
| pos\_ratio | 2.050234 | Very strong positive influence.  Senders with a higher share of positive messages strongly drive the sentiment score upward. |
| active\_days pos\_ratio | 1.230477 | Interaction effect: being both active and positive massively boosts score. Captures synergy between volume of activity and positivity. |
| neg\_ratio | -1.036169 | Strong negative influence.  Higher proportion of negative messages drags scores down significantly. |
| active\_days neg\_ratio | -0.728769 | Interaction effect: frequent activity combined with more negativity strongly **depresses score.** |
| max\_len\_words active\_days | 0.225608 | Longer messages combined with more active days boost score a bit. |
| max\_len\_words | 0.159957 | On its own, writing more verbose messages adds small positive impact. |
| max\_len\_words lag1\_neg\_inter | -0.049260 | suggests that detailed communication does not fully offset past negative influence. |

**Reflection on Challenges in Analysis and Prediction**

**Data Quality & Preparation**

One of the first challenges encountered was dealing with **real‑world email data**, which is inherently messy and incomplete. Dates had to be cleaned, missing fields in subjects and bodies had to be handled, 1and categorical sentiment labels needed to be mapped numerically. In addition:

* **Missing or sparse data** (e.g., certain senders having few messages in a month) created problems for monthly aggregation.
* **Imbalanced sentiment distribution** (more Positive/Neutral than Negative) increased the risk that the regression model would underlearn negative patterns.

Careful preprocessing was crucial — filling missing text with combined fields, filtering invalid dates, and handling divisions by zero in ratios avoided biased or broken features.

**Linear regression assumes**:

* **Linearity** between predictors and target (difficult with human text data).
* **Independence** of errors (challenged by temporal correlations in lag features).
* **Normality and constant variance of errors** (often violated in real data).

To address these, scaling and polynomial interaction terms were added. Still, interpreting coefficients under potential violations required caution. Linear regression offered transparency, but with fragile assumptions compared to more robust models (e.g., tree ensembles).

**Performance Evaluation Challenges**

Key metrics used: **R², and MAE**

* R² ≈ 0.92 signaled strong explanatory power.
* But with target scores ranging from -3 to 14, even a small MAE (~0.53) is only ~3.1% of the range, which is excellent but easy to misinterpret without normalization.

Percentage‑based interpretation of errors relative to the target range provided clarity. Without this normalization step, the raw values may have seemed larger than they actually were.