**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 (0.05), Negative (-0.05), 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**: Cross-validation and grid search were used to ensure reliability and generalizability, and interpret feature importance to understand key drivers of sentiment. RSME 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

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

85 john.arnold@enron.com 2011-02 12

112 johnny.palmer@enron.com 2011-06 12

**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

221 sally.beck@enron.com 2010-07 -2

229 sally.beck@enron.com 2011-03 -2

102 johnny.palmer@enron.com 2010-07 -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:

* word\_count - Number of words in the message
* avg\_word\_length - Average word length
* exclaim\_count - Number of exclamation marks
* pos\_term\_count - Count of positive terms (e.g., "great", "thank", "happy")
* neg\_term\_count - Count of negative terms (e.g., "issue", "delay", "frustrated")
* msg\_count\_month - Number of messages sent by employee in the month
* sentiment\_ratio - Ratio of positive to negative messages per employee per month
* body\_length - Total character count of the message

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

**Target**: sentiment\_score

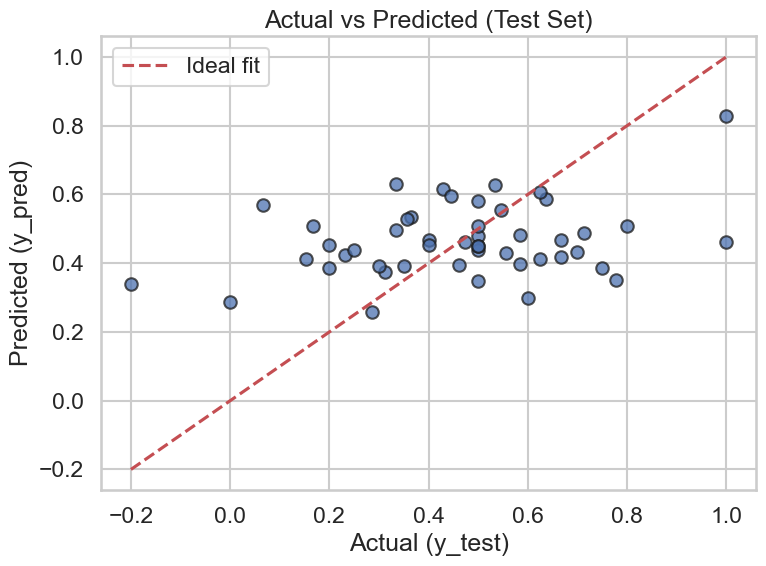
Model Approach:

* Model Type: ElasticNet regression (Linear Regression)
* Pipeleine: StandardScalar + ElasticNet
* Hyperparameter Tuning: GridSearchCV with 5-fold cross-validation
* Parameter Grid: regressor\_\_alpha: [0.01, 0.1, 1.0, 10.0], regressor\_\_l1\_ratio': [0.2, 0.5, 0.8]

Training method: 80/20 Train-Test split

**Evaluation**:

* R² Score: Assessed how well sentiment explained variation in the target. The R² is 0.175, which shows that the model explains 17.5 % of the variance in the sentiment trends.
* Root Mean Squared Error: Measured prediction accuracy. The RMSE is 0.616, this indicates that the prediction deviated only by 0.6 points from the actual value.
* Model explains only 17.3% of the variance in sentiment scores. Predictive power is limited, but slightly better than a naive mean-prediction baseline.



**Feature Coefficient:**

|  |  |  |
| --- | --- | --- |
| pos\_term\_count | 0.198043 | Strong positive influence; more positive terms higher sentiment score |
| word\_count | 0.102084 | Longer messages (in word count) tend to be more positive |
| avg\_word\_length | 0.078698 | Use of longer words correlates with more positive sentiment |
| sentiment\_ratio | 0.037660 | Higher ratio of positive to negative messages boosts sentiment score |
| body\_length | 0.014427 | Slight positive effect from overall message length |
| exclaim\_count | 0.000000 | No measurable impact from exclamation marks |
| neg\_term\_count | -0.000315 | Slight negative influence from negative terms |
| msg\_count\_month | -0.024943 | Higher monthly message volume slightly correlates with lower sentiment |

**Reflection on Challenges in Analysis and Prediction**

During this sentiment prediction task, several challenges were encountered that affected the performance of the linear regression model. One major challenge was **feature representation**. The numerical features extracted from text (such as average length, word counts, and sentiment ratios) may not have been sufficient to fully capture the subtle nuances of human sentiment. Text data is inherently **high-dimensional, context-dependent, and ambiguous**, which makes it difficult for simple linear models to detect meaningful patterns.

Another challenge was **model limitations**. Linear regression and its variants (Ridge, Lasso, ElasticNet) assume a linear relationship between features and target values. However, sentiment is often influenced by **non-linear interactions** in language, such as sarcasm, negation, or cultural expressions, which are not well captured by these models. This could explain the relatively low R² scores observed.

In addition, **data quality and variability** posed obstacles. Sentiment labels are often noisy because human annotation is subjective, and text messages may include slang, abbreviations, or misspellings that distort feature extraction. Missing values, infinite values, and imbalanced data distributions also required preprocessing steps, which may have reduced the richness of the dataset.

Finally, **hyperparameter tuning** and model evaluation were non-trivial. Selecting an appropriate regularisation strength (alpha) was essential, but the model still struggled to generalize effectively. This suggests that while cross-validation helped, the chosen set of models may not have been the best fit for the problem.