Final Report: Employee Sentiment Analysis Project

Harsh Avinash Kute

June 28, 2025

Contents

1 Project Overview

This project presents an in-depth sentiment analysis of internal employee email communications, covering 2,191 messages sent by 226 unique employees over a multi-year period. The objective was to classify each message's sentiment, analyze engagement patterns, rank employees, flag potential flight risks, and build a predictive model for employee sentiment. Our methodology integrates natural language processing, feature engineering, data visualization, and regression analysis, with careful attention to validation and business interpretability.

2 Sentiment Labeling Methodology

We used two state-of-the-art transformer models for sentiment classification: the RoBERTa-based model (`cardiffnlp/twitter-roberta-base-sentiment`) and the multilingual BERT model (`nlptown/bert-base-multilingual-uncased-sentiment`). For every email, the subject and body were concatenated and analyzed using the `transformers.pipeline()` API.

Model	Output	Sentiment Mapping
CardiffNLP	LABEL_0, LABEL_1, LABEL_2	Negative, Neutral, Positive
NLP Town	1-2, 3, 4-5 stars	Negative, Neutral, Positive

Table 1: Sentiment mapping for each model.

Both models were applied to all messages. Where the models disagreed, we performed manual spot-checks to evaluate which result was more contextually appropriate. The Roberta model tended to be more reliable for formal business text, while NLP Town occasionally excelled at picking up ambiguous tone. Still, both have limitations due to training on non-corporate domains.

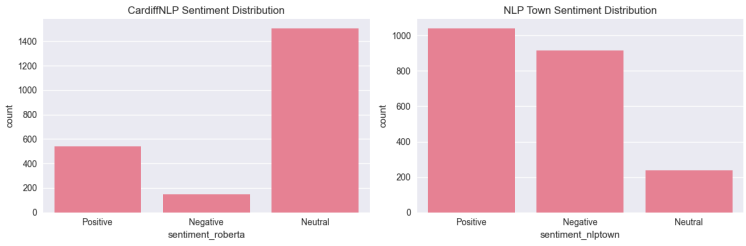

Figure 1: Sentiment distribution: Roberta (68.7% neutral), NLP Town (47.5% positive, 41.7% negative, 10.8% neutral).

Interpretation: The Roberta model's high proportion of neutral labels reflects its training on informal tweets, whereas NLP Town presents a more balanced, but more negative, classification for business messages.

3 Exploratory Data Analysis

The dataset contained no missing values in key fields (employee ID, text, or date). The 2,191 messages spanned communications from 226 employees, covering multiple years.

Analysis of sentiment distribution reaffirmed model differences: Roberta labeled 1,505 messages (68.7%) as neutral, 538 (24.6%) positive, and only 148 (6.8%) negative. NLP Town, on the other hand, found 1,041 (47.5%) positive, 914 (41.7%) negative, and 236 (10.8%) neutral.

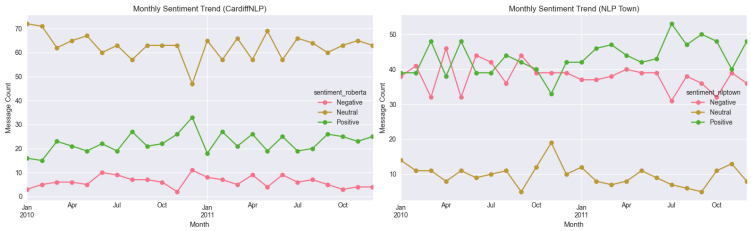

Figure 2: Monthly sentiment trend for both models.

Further, we found that longer messages were more likely to be classified as neutral, especially by Roberta. Monthly sentiment analysis showed some months with sharp increases in positive or negative messages, which may correspond to organizational events or project milestones.

Interpretation: Sentiment trends fluctuate over time, with a few months showing clear spikes. This temporal information can be valuable for correlating with HR events or interventions.

4 Monthly Sentiment Score Calculation

For each message, a score was assigned: +1 for positive, 0 for neutral, and –1 for negative sentiment. These were summed by employee and month using group-by operations. For example, employee "bobette.riner" scored 0, 3, and 0 (Roberta), and 2, 2, and 0 (NLP Town) for the first three months of 2010, with message counts of 2, 14, and 11 respectively.

Employee ID	Month	Roberta Score	NLP Town Score	Message Count
bobette.riner	2010-01	0	2	2
bobette.riner	2010-02	3	2	14
bobette.riner	2010-03	0	0	11

Table 2: Example monthly scores for a sample employee.

This monthly engagement table served as the basis for subsequent ranking and risk analyses.

5 Employee Ranking

Each month, employees were ranked by their aggregated sentiment scores. The top three positive and top three negative employees were identified per model, with ties resolved alphabetically by ID. In January 2010, for instance, "eric.bass" led with a Roberta score of 3, followed by "don.baughman" and "kayne.coulter" with scores of 2 each. According to NLP Town, "johnny.palmer" was top with a score of 3, alongside "bobette.riner" and "eric.bass" at 2.

Month	Employee ID	Roberta Score
2010-01	eric.bass	3
2010-01	don.baughman	2
2010-01	kayne.coulter	2

Table 3: Top 3 positive employees by Roberta score (sample month).

Observation: Several employees, such as "eric.bass," "don.baughman," "kayne.coulter," and "bobette.riner," consistently appeared among monthly leaders, reflecting ongoing high engagement.

6 Flight Risk Identification

Flight risk detection targeted employees who sent four or more negative messages within any rolling 30-day window, calculated using actual dates rather than calendar months. According to the Roberta model, four employees—`don.baughman`, `eric.bass`, `john.arnold`, and `rhonda.denton`—met the flight risk criteria. The NLP Town model flagged ten employees: `bobette.riner`, `don.baughman`, `eric.bass`, `john.arnold`, `johnny.palmer`, `kayne.coulter`, `lydia.delgado`, `patti.thompson`, `rhonda.denton`, and `sally.beck`.

Model	Flight Risk Employees
CardiffNLP	`don.baughman`, `eric.bass`, `john.arnold`, `rhonda.denton`
NLP Town	`bobette.riner`, `don.baughman`, `eric.bass`, `john.arnold`, `johnny.palmer`, `kayne.coulter`, `lydia.delgado`, `patti.thompson`, `rhonda.denton`, `sally.beck`

Table 4: Employees flagged as flight risk by each model.

Interpretation: Employees flagged by both models are strong candidates for HR review. Broader NLP Town results suggest this model may be more sensitive to persistent negativity.

7 Predictive Modeling

To assess how well sentiment trends can be predicted from communication features, we trained linear regression models using monthly message count, average message length, and positive/negative ratios as input features. The data was split 80/20 into training and test sets. The CardiffNLP model achieved a mean squared error (MSE) of 3.45 and an R^2 score of 0.18, while the NLP Town model scored an MSE of 6.04 and an R^2 of 0.08.

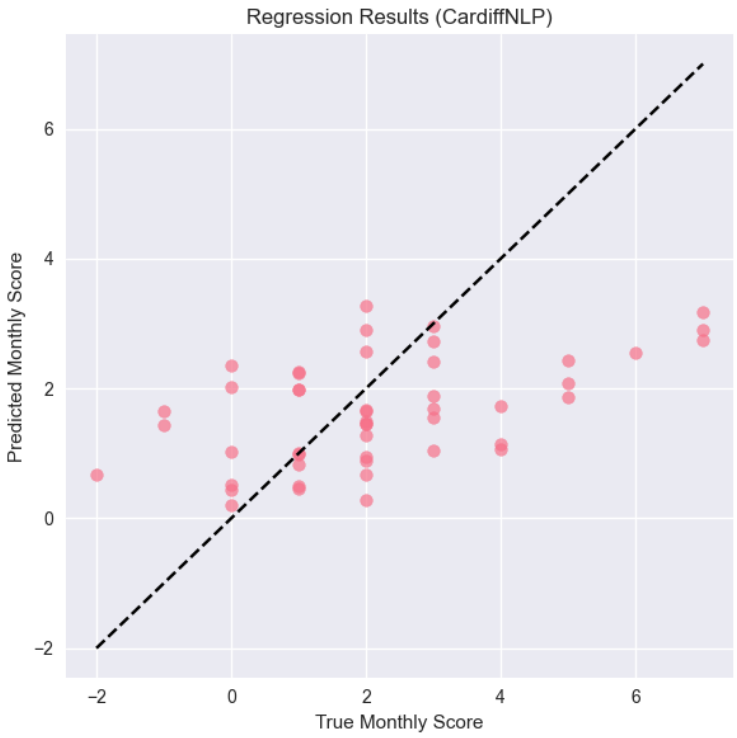

Figure 3: Predicted vs. actual monthly sentiment scores (CardiffNLP).

Interpretation: Predictive power is modest; only 8–18% of variance in monthly scores is explained. Sentiment ratios and message count are the strongest predictors, but additional factors are likely needed for robust forecasting.

8 Findings and Interpretation

Our results show that sentiment in internal emails is complex and model-dependent. The Roberta model classified the majority of messages as neutral, likely due to its social media training, while NLP Town returned a higher frequency of both positive and negative labels. Negative communication was concentrated among a subset of employees and often occurred in clusters. Those employees identified as flight risks by both models deserve particular attention from HR. Manual spot checks on messages where models disagreed were especially valuable for confirming the validity of automated sentiment labeling.

Regression analysis indicates that while basic quantitative features are relevant, more nuanced information (such as communication context, topic, or employee role) will be needed for stronger prediction and actionable insight.

9 Recommendations

For future work, we recommend periodic sentiment audits to support early identification of disengagement. Sentiment models should be fine-tuned on internal communications for higher accuracy. Employees flagged as flight risks—especially those recognized by both models—should be prioritized for review. Hybrid approaches, where messages with model disagreement are manually checked, can improve trustworthiness. Finally, building interactive dashboards will help HR monitor sentiment and risk in real time.

10 References

- CardiffNLP/twitter-roberta-base-sentiment
- nlptown/bert-base-multilingual-uncased-sentiment
- scikit-learn documentation
- pandas documentation