Using Language in Transcript Earnings Calls to Detect Negative Earnings Surprise

Background and Context

The Prudential Regulation Authority (PRA), a division of the Bank of England, is responsible for supervising the financial health of key institutions, including Global Systemically Important Banks (G-SIBs). A major challenge for the PRA is to identify early signs of firm-specific instability before these manifest in broader market volatility. One early warning signal is an earnings surprise, defined as the gap between actual and expected earnings per share (EPS). Negative earnings surprises (NES) are closely tied to investor uncertainty and market disruption.

This project investigates whether the language used in earnings call transcripts can provide early indicators of NES. Alongside traditional financial data, the sentiment, structure and tone of communication by senior executives may reflect internal optimism or concern before this is evident in the numbers.

Our goal is to develop a classification model to predict NES using a combination of financial indicators and linguistic features. By doing so, the PRA may be better equipped to identify at-risk firms in advance, offering opportunities for earlier intervention and ultimately contributing to the overall stability of the UK's financial system.

Project Development Process

This project aimed to create a predictive tool that can help identify whether a major financial institution is likely to experience a NES. Rather than trying to forecast actual earnings, our objective was to detect the likelihood of deviation from expectations, which could indicate risk factors not yet evident in headline financials.

To achieve this, we selected two case studies: JP Morgan (JPM) and Citigroup (Citi). These are large U.S.-based G-SIBs with extensive earnings histories and analyst expectations. While both have UK subsidiaries, we focused on their U.S. entities due to the wider availability of high-quality earnings call transcripts and financial data. JPM has experienced more frequent NES events but has generally maintained stronger and more consistent stock growth. Citi, by contrast, has faced more variable investor sentiment despite fewer NES events. This made the pair a useful comparison for identifying both consistent and subtle patterns across different communication styles and market outcomes.

Data Acquisition and Preparation

We compiled quarterly data from Q1 2015 to Q1 2025, which included:

- Earnings call transcripts
- Financial statements (income, balance sheet, cash flow)
- Macroeconomic indicators such as interest rates, CPI, GDP growth and unemployment

Most financial and macroeconomic data were collected using the Alpha Vantage API. Earnings transcripts were manually retrieved when necessary due to API limitations.

Our analysis aimed to contrast JPM and Citi's performance in light of their linguistic behaviour. JPM's frequent NES events did not hinder strong market performance, suggesting effective communication or investor confidence. Citi's more cautious communication and variable stock growth offered a contrasting case.

Data Processing and Feature Engineering

Our target variable was the earnings surprise percentage, calculated as the difference between reported and expected EPS. NES was defined as a shortfall of 5 percent or more.

Financial Data Processing:
We merged the income statement, balance sheet, cash flow and macroeconomic data into a quarterly dataset. Monthly indicators were resampled to a quarterly frequency using averages. Key preprocessing steps included:

- Lagging all features by one quarter
- Filling gaps using spline interpolation
- Removing redundant or sparsely populated features

We engineered the following financial ratios to enhance the dataset:

- Cash Ratio
- Return on Equity
- EBITDA Margin
- Debt-to-Equity Ratio
- Interest Coverage Ratio
- Cash Conversion Ratio
- Share Buyback Rate
- Dividend Payout Ratio

- Liabilities-to-Assets Ratio
- Net Income Growth
- Interest Expense Growth
- Federal Funds Rate Growth

Linguistic Data Processing:
Earnings call transcripts were processed using NLP techniques. Only dialogue from senior executives was retained, divided into Presentation and Q&A sections. We created two datasets:

- Q&A-only, to focus on unscripted responses

- Combined presentation and Q&A, averaged by quarter

Extracted features included:

- FinBERT sentiment score
- Average word and sentence length
- Vagueness rate
- Jargon count and rate
- Flesch reading ease
- Gunning Fog Index
- SMOG index
- Coleman-Liau index
- Automated readability score

Exploratory Data Analysis:

Initial analysis revealed a significant class imbalance, with NES occurring only once in Citi over the past 10 years and five times for JPM. This imbalance highlighted the need for techniques like scale_pos_weight and model evaluation beyond accuracy.

Correlation analysis showed weak or no linear relationship between FinBERT-derived sentiment scores and EPS Surprise, figure 1, challenging early assumptions that sentiment alone could be a strong predictor. Similarly, many financial metrics showed mixed correlation trends with surprise magnitude.

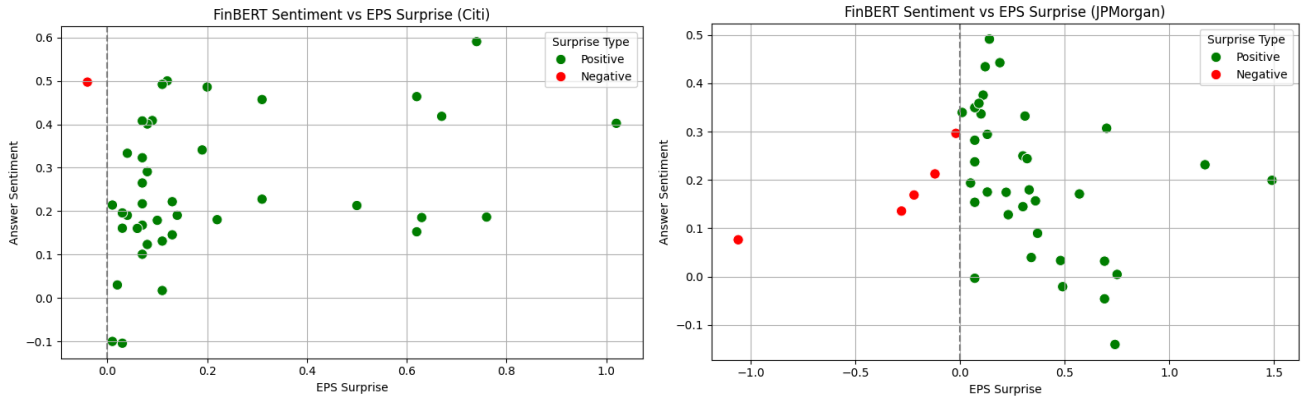


Figure 1. Correlation of Sentiment vs EPS Surprise for Citi (left) and JPM (right).

We also examined linguistic metrics over time (e.g. jargon rate, word length, readability) and looked for a correlation between them and EPS surprise, figure 2. However, some features like vagueness and sentence complexity varied in quarters that preceded NES, suggesting potential indirect signals worth testing in classification models.

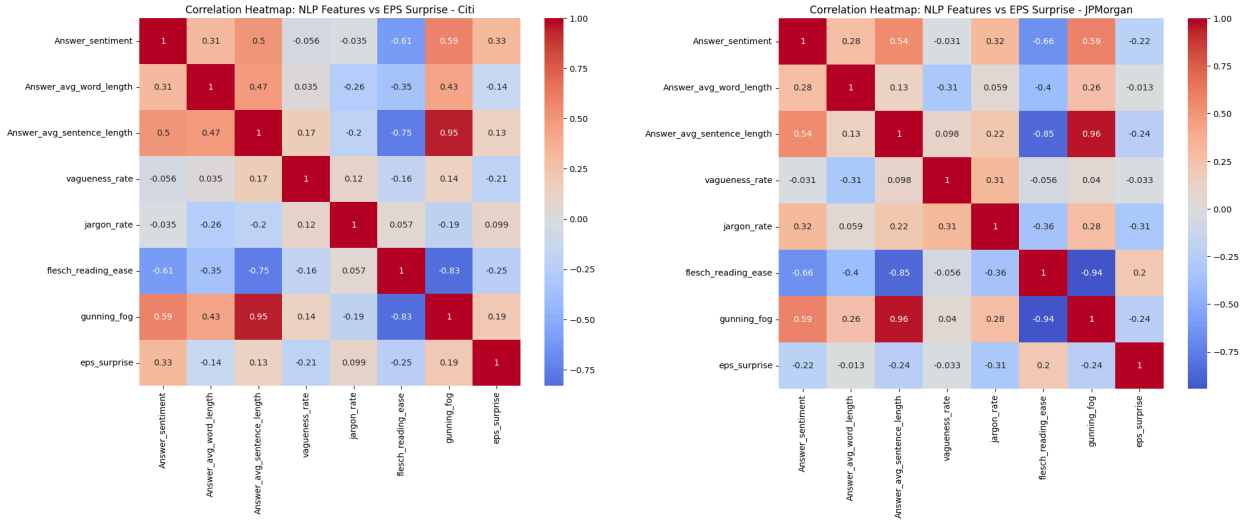


Figure 2. Correlation of linguistic metrics vs EPS Surprise for Citi (left) and JPM (right).

These findings motivated a shift away from regression toward a classification approach, integrating both financial and linguistic features to detect better subtle signals associated with NES events.

Model Selection and Baseline

We chose a binary outcome variable (EPS surprise < 5%). This is because we are most interested in when a bank suffers low EPS relative to expectations (as this is a sign of possible distress). The threshold was raised from 0% to 5% to address the class imbalance problem.

Random Forest and XGBoost classifiers were chosen as models as we believed the relationships to be non-linear, and these models do not require extensive feature selection. Moreover, they are more robust and easier to train than deep learning models on small datasets.

We compared performance to two baselines: a naive forecast and a random baseline, where the class balance is assumed to be known beforehand and predictions are randomly selected according to this ratio.

Model Tuning and Evaluation

We trained 10 model versions (not all shown in the table below) by varying: model type, hyperparameter tuning, reducing features, and, most importantly, including NLP features and compared them to the baseline.

Hyperparameter tuning was optimised for recall, as we are interested in capturing as many predictions for high-risk EPS misses.

Cross-validation of recall and precision metrics was also conducted to see the variance in each metric. This was high due to the small sample sizes for training (which affects model robustness) and validation (which increases variance).

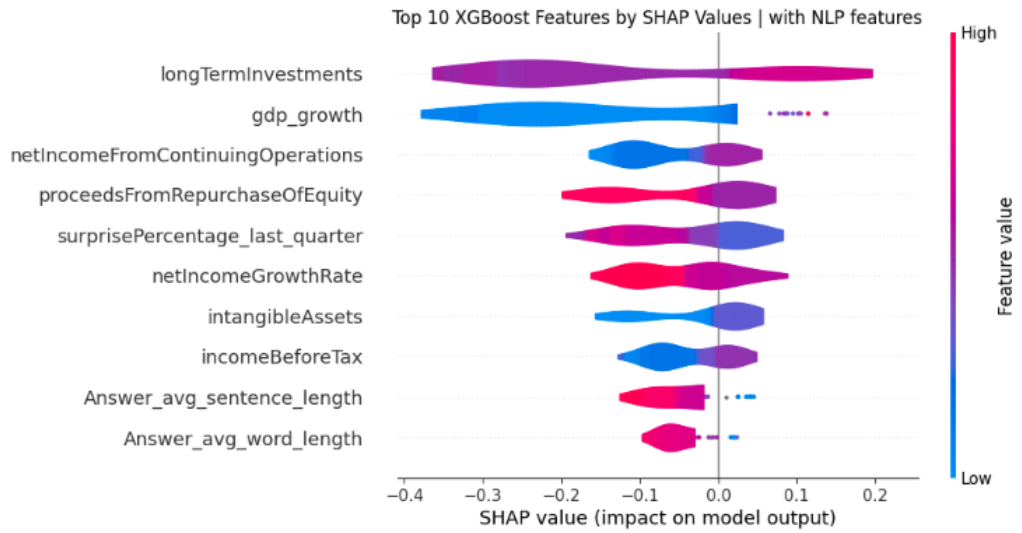
Fig 3- Negative EPS surprise model evaluation

Model Grouping	Model	Recall	Precision
Baseline	Random Baseline*	0.50	0.30
	Naive Prediction Baseline**	0.12	0.5
Financial Features Only	Random Forest	0.25	1.0
	Random Forest - Tuned	0.0	0.0
	XGBoost	0.60	0.38
	XGBoost - Tuned	0.75	0.40
	XGBoost - Tuned (Top 10 features only)	0.75	0.40
Financial + NLP Features	XGBoost	0.38	0.43
	XGBoost - Tuned	0.75	0.43

The best model shows promising recall and precision. However, the cross-validation variance for recall and precision was high due to the small sample sizes, and the NLP features had minimal impact on model performance.

This is reflected in the SHAP plot, where 8/10 of the top features are financial. Using longer sentences and longer words is associated with a lower risk of negative EPS surprise, which may be because CEOs/CFOs can talk more eloquently when they have high confidence in the business for the subsequent quarter.

Fig 4– SHAP plot of feature importances for best model version



Challenges and Iterations

Challenges for the prediction model included class imbalance and low sample size. The former was addressed by adjusting the threshold for negative surprise to make the classes more balanced, and the latter by examining cross-validation. Ideally, we would have more data, which could solve both problems (high imbalance is less of an issue with a large sample size, and larger training datasets will lead to more robust models)

Final Model and Current Limitations

Our tuned XGBoost model using only financial features achieves better performance (recall: 0.75, precision: 0.40) than the baseline (recall: 0.50, precision: 0.30), though this was with a small sample size (n=52).

After integrating linguistic features (e.g., sentiment, vagueness, readability), recall was unchanged, and precision increased slightly to 0.43. This may be due to noise in linguistic signals and limited training examples. Additionally, firm-specific language styles introduce variability that hinders generalisability across institutions.

With a larger dataset across more firms and quarters, we expect improved model robustness and insight into how communication patterns relate to financial outcomes.

Results

We framed the prediction task as a binary classification problem: detecting NES, which is defined as actual EPS falling short of estimates by 5% or more. This threshold was chosen to reduce class imbalance while capturing potential firm-level financial stress signals.

We selected XGBoost because it can model non-linear relationships and handle small datasets without extensive preprocessing. The best-performing model was a tuned XGBoost classifier using only financial features (recall: 0.75, precision: 0.40). When linguistic features were added (sentiment, vagueness, readability), recall remained unchanged while precision rose slightly to 0.43. SHAP analysis revealed that 8 out of the top 10 predictors were financial, though longer words and sentences were weakly associated with lower risk, possibly reflecting speaker confidence.

High variance in cross-validation metrics reflects the limited sample size. Expanding the dataset and including more firms would likely improve model stability and reveal stronger language-risk relationships.