Using Earnings Surprise to Assess Risk

Team 5/Risky Talk

Research Problem

Problem Statement

To help the PRA detect firm-specific instability in G-SIBs, it is essential to understand early signals that indicate a firm may be at risk before traditional financial metrics reveal distress

Objective

To predict likelihood of negative earning surprises to improve their risk assessment of individual firms

The Solution

Using two binary classification models:
Model A: financial data from the firms
Model B: financial data + linguistic metrics calculated from earnings calls

Target Variable - Earnings Surprise (the gap between actual and expected earnings per share)

J.P.Morgan

- More frequent negative earnings surprises
- Strong and consistent stock growth

CITI

- Fewer negative surprises
- Less stable performance -> reflecting ongoing restructuring and investor caution

Data Collection

Earnings Calls Transcripts
Financial Metrics

Timeline: 2015-2025

Pre-Processing

Transcripts Cleaning

- Remove headers, non-dialogue
- Segment by speaker (CEO, CFO)

Financial Data Cleaning

- Align estimates vs actuals by quarter

Feature Engineering

Trend Analysis & Visualisations

- Language vs EPS Surprise
- JPMorgan vs Citi comparisons

Exploratory Analysis

Linguistic Features

- Statistical Analysis
- Distributions
- Trends
- Quarters performance comparison

Financial Features

- Historical EPS Surprise Trends

Predictive Modeling

Models Implementation

Model Explainability

SHAP Values & Feature Importance

- Identify linguistic indicators of performance

Challenges & Refinements

Challenges & Iterations
Final Model & Limitations

Data Collection

Transcripts

Data Sources
- Alpha Vantage API
- PDF files from official websites (for missing quarters)

Bank	Year	Quarter	Period	Question_Speaker	Question	Answer_Speaker	Answer
JPMorgan	2015	Q1	2015Q1	Glenn Schorr	Just one quick clarification question on the p...	Marianne Lake	It wasn't anything particularly noteworthy in ...
JPMorgan	2015	Q1	2015Q1	Glenn Schorr	And maybe just a related question but I'm not ...	Marianne Lake	Yes, this is where it would be. I wouldn't say...
JPMorgan	2015	Q1	2015Q1	Glenn Schorr	Okay. Switching gears, in Jamie's letter, you ...	Marianne Lake	No, I mean, it is more of the same. Obviously ...
JPMorgan	2015	Q1	2015Q1	John McDonald	I was wondering on net interest income, given ...	Marianne Lake	So, again assuming for a second that rates don...
JPMorgan	2015	Q1	2015Q1	Erika Najarian	Good morning. On the CCAR, do you expect any p...	Marianne Lake	So taking your first point Erika, obviously I ...

*** Example of API output

QUESTION AND ANSWER

OPERATOR: Your first question is from the line of John McDonald with Autonomous Research. Please go ahead with your questions.

JOHN MCDONALD: Hi. Good morning.

MARK MASON: Good Morning.

JOHN MCDONALD: Mark, I wanted to ask about the net interest income. It seems like overall it came in a little bit better than you might have expected when you spoke at the Barclays Conference in September. Can you give us a little more color on what drove the improvement in the core ex-Markets NII this quarter and how that makes you feel about the setup for NII growing from here.

*** Example of PDF output

Financial Metrics

Data Sources
- Alpha Vantage API

DateEnding	reportedCurrency	grossProfit	totalRevenue	costOfRevenue	costofGoodsAndServicesSold	operatingIncome	sellin
2025-03-31	USD	18893000000	41255000000	22362000000	22362000000	5448000000	
2024-12-31	USD	17005000000	40900000000	23895000000	23895000000	4842000000	
2024-09-30	USD	17668000000	43359000000	25691000000	25691000000	4390000000	
2024-06-30	USD	17681000000	42638000000	24957000000	24957000000	4310000000	
2024-03-31	USD	18678000000	43722000000	25044000000	25044000000	4544000000	
2023-12-31	USD	15317000000	41395000000	26078000000	26078000000	-2103000000	
2023-09-30	USD	17921000000	40748000000	22827000000	22827000000	4788000000	
2023-06-30	USD	17626000000	38187000000	20561000000	20561000000	4042000000	

*** Example of API output

Data Pre-Processing

Financial Metrics

Data Merge into a quarterly dataset
(Income Statement, Balance Sheet, Cash Flow, Macroeconomic data)

↓

Data Transformation
(Lagging all features by one quarter, Spline interpolation, Removal of redundant features)

↓

Features Engineering

Transcripts

Data Split into Presentations and Q&A

↓

Data Cleaning
(Lowercasing, Punctuation normalisation, Whitespace handling)

↓

Features Engineering

*** Stopwords were not removed to preserve linguistic structures relevant for feature extraction
*** All metrics were calculated per intervention in the presentation and per answer in the Q&A

Features Engineering

Financial Metrics

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

Transcripts

Sentiment
Emotional tone using FinBert

Word and Sentence Length
Syntactic structure and complexity

Vagueness
Count - clarity and precision in speech
Rate - vagueness level relative to text length

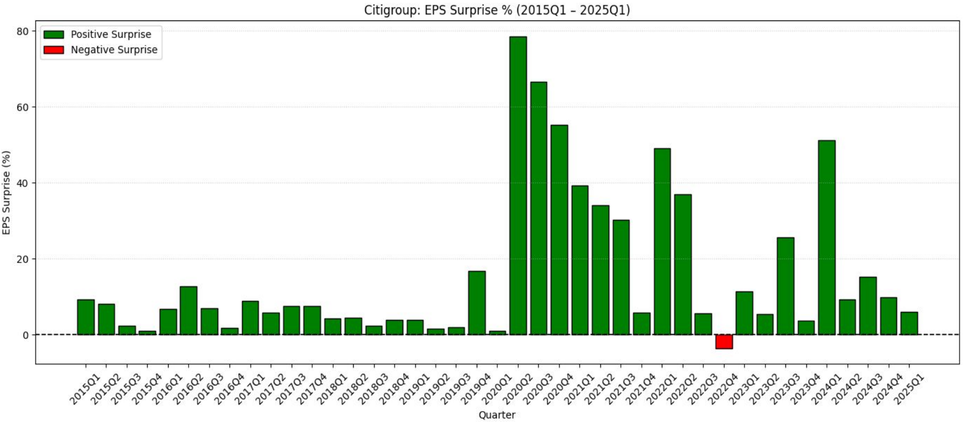
Jargon
Count - use of specialised language
Rate - jargon usage relative to text length

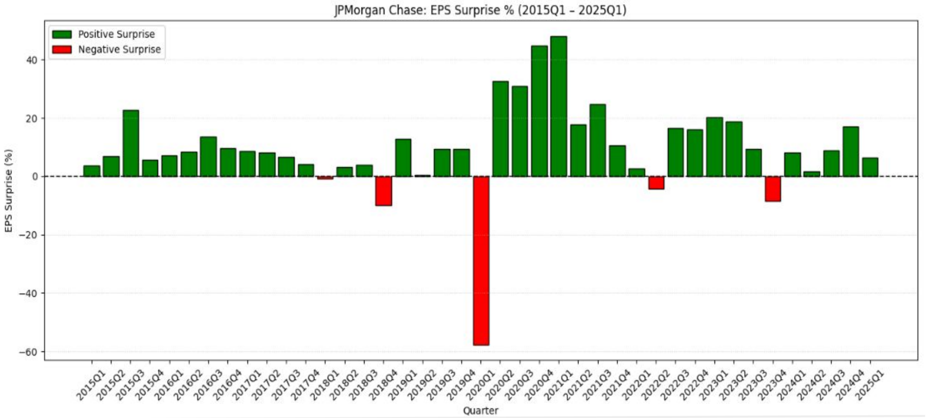
Readability
Flesch reading ease
Gunning fog
Smog index
Coleman Liau
Automated Readability

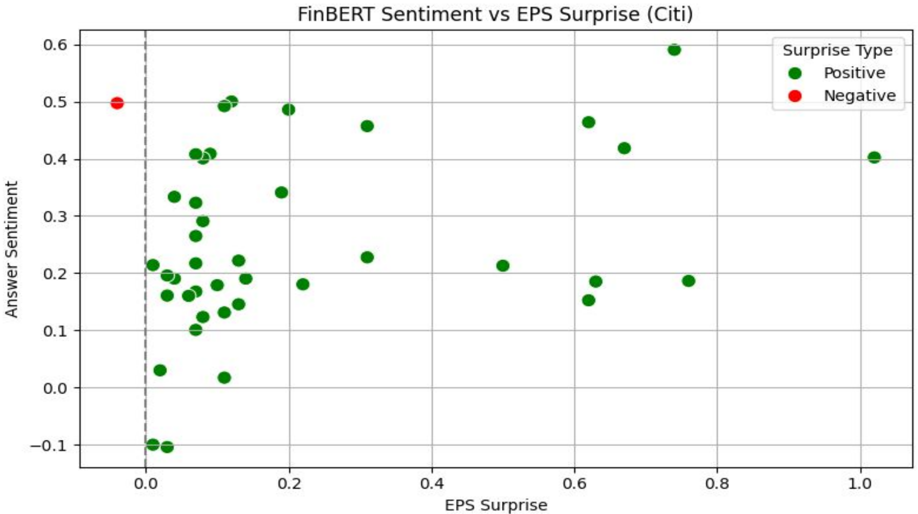
Exploratory Analysis

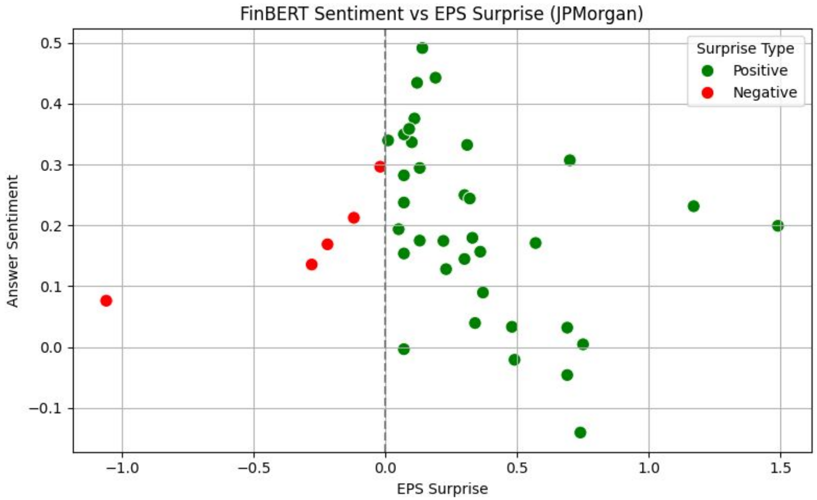
Earnings Surprise Trends vs Sentiment Analysis

Negative Earnings Surprise is relatively rare.





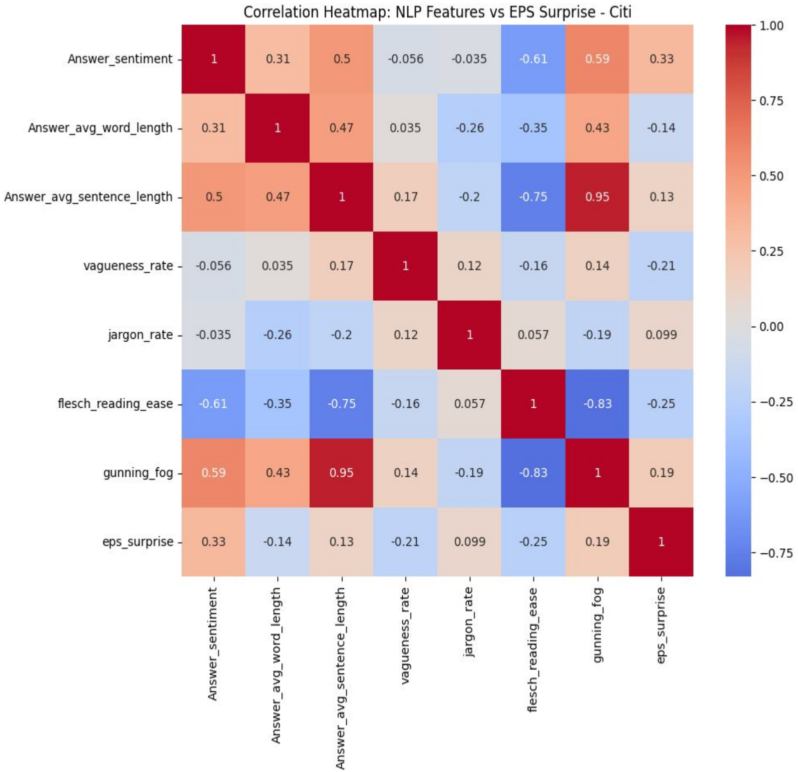


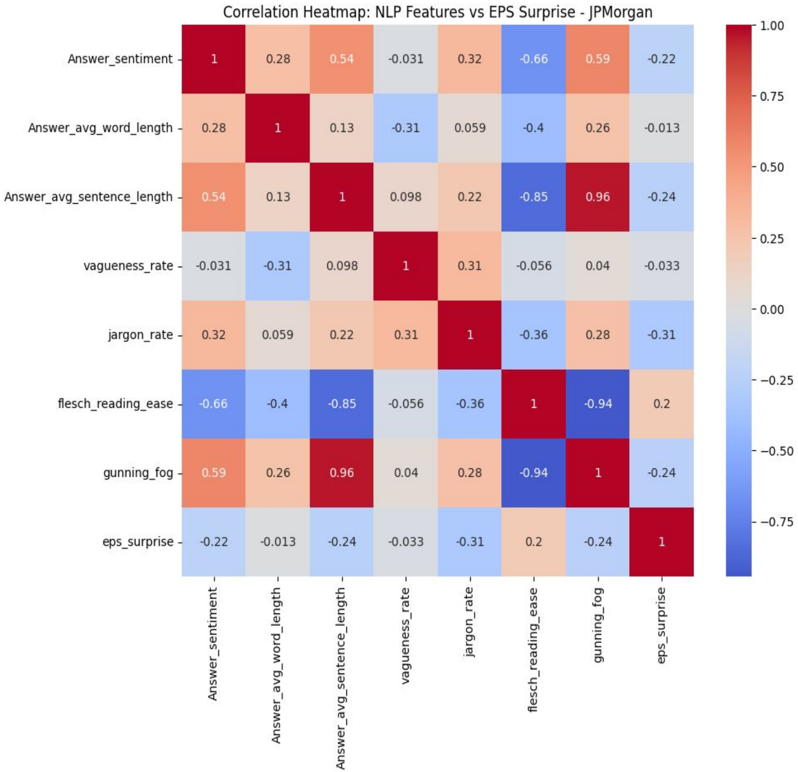


Exploratory Analysis

Earnings Surprise Trends vs Calculated Linguistic Metrics

There is poor correlation between earnings surprise and linguistic metrics





Using features from financial statements we have created a binary classification model which predicts EPS beat vs expectations

Select outcome variable

EPS surprise <+5% was chosen as EPS surprise <0% led to a severe class imbalance which we could not model effectively given the small sample size

Create model features

Secondary research was conducted to inform feature engineering (ratios and growth rates) to **create features which are likely to predict EPS** performance. These were derived from income, balance sheet and cashflow statement variables

Select model features

Initially, all features were kept (as we were using model types which are robust to high feature count).

Subsequently we cut down the feature list to see if it affected recall and precision

Select model types

Tree-based models (Random Forest and XGBoost) were chosen as they allow for non-linear relationships, show good performance on low sample sizes and are robust to overfitting even with many features.

Tune models

Models were tuned using random search and optimised for recall as we are most interested in catching as many high risk banks as possible

Evaluate and repeat

Models were evaluated using recall and precision of the positive class and compared to baselines (random and naive prediction)

Cross validation was also conducted

Based on the best model, we can predict low/negative EPS performance with ~60% recall and ~45% precision

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

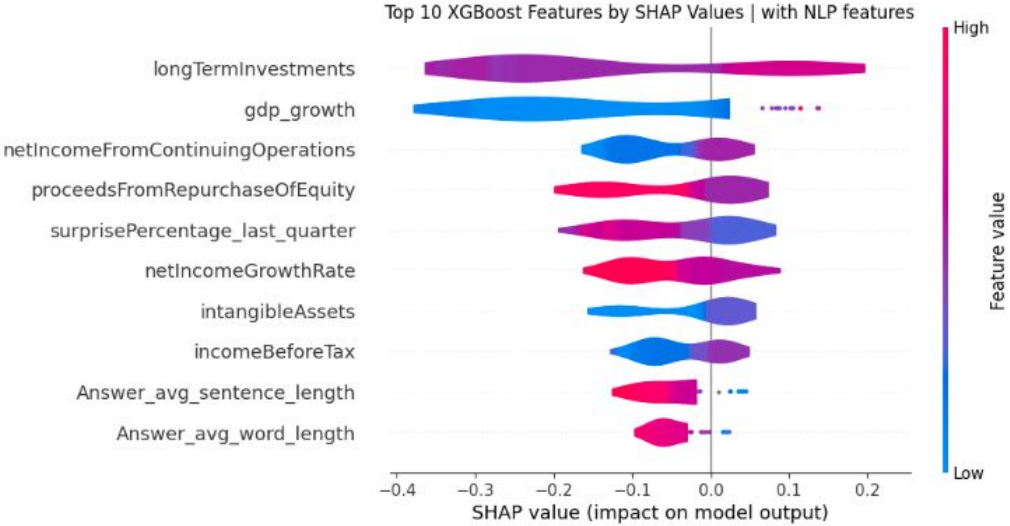
Whilst the initial results of the best model look promising we should note:

1) The NLP features only increased precision by a small amount
2) A very small dataset was used to train and test this model (n=52 total) and as a result the cross-validation scores for recall and precision have high variance (which is higher when we include the linguistic features)

*Random baseline based on 50:50 prediction of positive class vs an actual positive class % of 30%

** Naive prediction baseline based on using last period's observed value as the prediction for the current period

Financial features are more important than NLP features. More sophisticated language is associated with lower risk of EPS miss



Commentary

Risks of negative EPS surprise are increased if:

- Higher long term investments (perhaps because banks who invest more in the future are distressed currently / are less focused on the subsequent quarter)
- Lower GDP growth
- Lower net income from operations
- Lower proceeds from repurchase of equity (lower share buybacks represent less confidence and also more shares to dilute earnings for EPS in the next quarter)

NLP features have lower importance:

- Longer sentences and word lengths are associated with lower likelihood of negative EPS surprise - perhaps as it indicates confidence from the CEO/CFO

Challenges, Refinements & Final Model

Challenges & Iterations

- Imbalanced classes (few NES)
- Small sample size (n = 52)
- Adjusted threshold for class balance
- Cross-validation for robustness

Final Model & Limitations

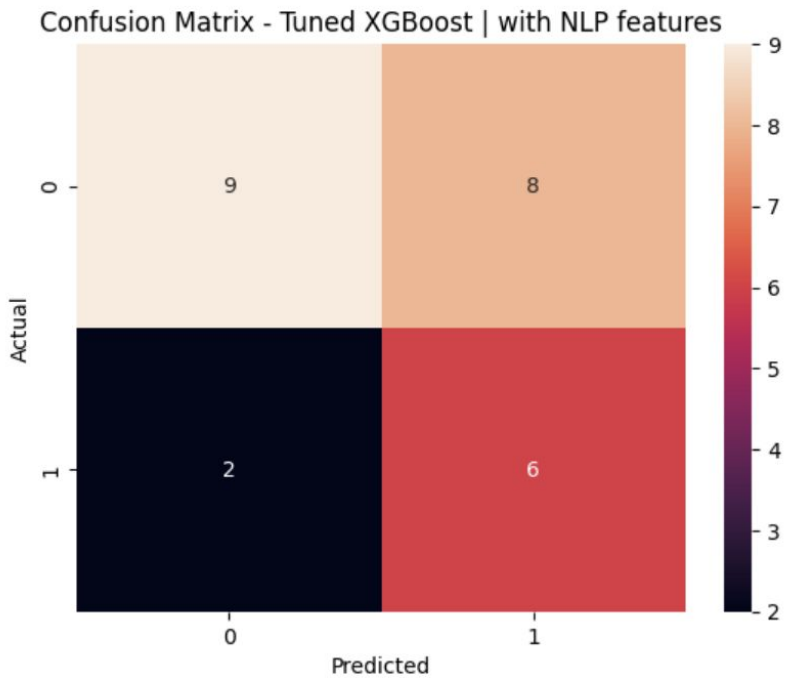
XGBoost (financial only):

Recall: **0.75** Precision: **0.40**

with NLP features:

Recall: **0.75** Precision: **0.43**

➔ Modest improvement, more data may unlock further gains
➔ Firm-specific language limits generalisation



	precision	recall	f1-score	support
0	0.82	0.53	0.64	17
1	0.43	0.75	0.55	8
accuracy			0.60	25
macro avg	0.62	0.64	0.59	25
weighted avg	0.69	0.60	0.61	25

Conclusion

- **Tuned XGBoost mode**l, focused on predicting negative earnings surprises (NES), demonstrated that financial features remain the strongest predictors, with recall of 0.75 and precision of 0.40.

- Incorporating **linguistic metrics,** such as sentiment, vagueness, and readability, resulted in a modest precision improvement (to 0.43), but did not substantially enhance recall.

- **SHAP analysis** confirmed the dominance of financial variables, though some linguistic signals (longer words, sentences) weakly indicated lower risk.

- Model performance was constrained by the small sample size, as reflected by high variance in cross-validation.

Future work: Expanding the dataset and broadening firm coverage are essential to improve model robustness and to uncover deeper links between financial language and risk.