AI-DRIVEN INSIGHT EXTRACTION FROM QUARTERLY EARNINGS DATA FOR STRATEGIC DECISION-MAKING

Group 3 – Pandas In Pajamas

BANK OF ENGLAND Employer Project

Contents

Background / Context

Thoughtful oversight typically depends on quantitative indicators, capital adequacy, liquidity, and funding ratios that are essential yet often lag the behavioural signals that foreshadow stress. Quarterly earnings calls provide a high-value narrative channel: executives and analysts reveal confidence, caution, and pressure dynamics that can surface before hard metrics move. The challenge is scale and linguistic complexity: reading hundreds of pages per bank across many quarters is not tractable without NLP.

This project evaluates whether narrative tone and interaction patterns from earnings calls can complement robust indicators and act as early qualitative warning signals. Using transcripts from HSBC, UBS, Deutsche Bank, Credit Suisse, and JPMorgan, we combined (i) FinBERT sentiment and dialogue features, (ii) Aspect-Based Sentiment Analysis (ABSA) with a purpose-built spin metric, (iii) temporal correlation/lead-lag checks to prudential ratios (CET1, LCR, NSFR, LDR), (iv) XGBoost modelling of language-to-KPI linkages, and (v) a Retrieval-Augmented Generation (RAG) prototype for supervisory Q&A. The objective is to augment rather than replace prudential analytics: language features triage attention, point to where risk may be building, and shorten the time to insight.

The approach aims to strengthen supervisory foresight by linking linguistic indicators directly to prudential performance, enabling earlier and more evidence-based interventions.

Project Development Process

Data preparation & integration

We scraped and parsed bank investor-relations PDFs, then standardised heterogeneous structures via regex-driven segmentation and speaker attribution. Deutsche Bank's transcripts lacked clean speaker boundaries; we engineered additional rules to recover roles reliably. To analyse Q&A dynamics at topical granularity, we added an exchange_id that flags new analyst-executive exchanges, supporting topic-aware interaction features (e.g., who addressed what, and how). The output is a structured, quarter-aligned corpus ready for sentiment, aspect, and temporal analysis.

Fundamentals and performance target

In parallel, we assembled prudential ratios (CET1, LCR, NSFR, LDR) and measured post-earnings market reaction. Each bank received a 0-9 quarterly rank (0 = strongest), combining relative fundamental strength and price-reaction performance. Two years' coverage (Q1-2022 to Q4-2024) yielded consistent patterns: Credit Suisse ranked weakest; JPMorgan strongest. This produced a compact, comparable target useful for correlation sanity-checks and supervised modelling.

Table 1. Quarterly ranking of Banks (Q1 2022 - Q4 2024) - 0-9 Ranking based on combined

quarter_label	abs_perf_1_day	abs_perf_1_week	abs_perf_1_month	abs_perf_3_months
Q1 2022	6	5	1	6
Q1 2023	4	1	2	1
Q1 2024	7	4	2	8
Q2 2022	2	0	2	3
Q2 2023	2	1	1	8

FinBERT sentiment & behavioural features

We scored sentence-level polarity with finance-tuned FinBERT and aggregated by bank-quarter-role (Executive vs Analyst). We also derived interaction features, disagreement rate and evasive rate, as proxies for tension and hedging that plain polarity can miss. We applied 3-quarter rolling means to stabilise series and reduce noise for smaller quarters. Across the pooled series, average executive sentiment exceeded analyst sentiment (0.163 vs 0.109), aligning with guidance-driven managerial optimism and analyst scepticism in stress periods. Manual spot-checks confirmed that low-sentiment clusters coincided with scrutiny of capital quality, liquidity buffers, or adverse macro commentary.

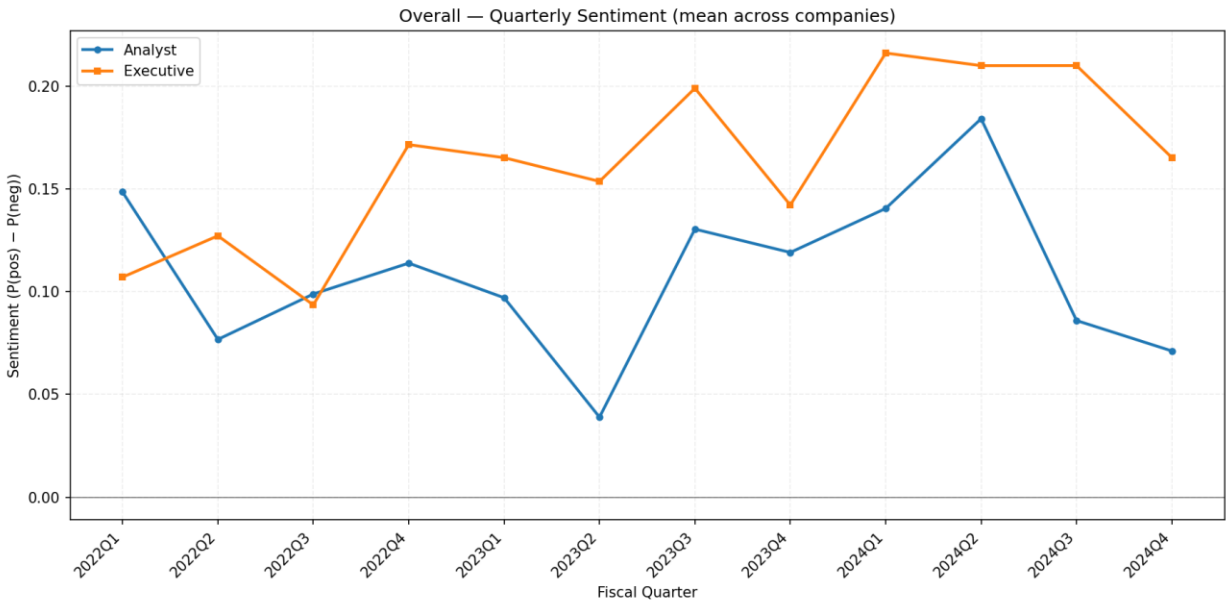

Figure 1. Executive vs Analyst Mean Sentiment (2022-2024) - FinBERT sentiment comparison, showing rolling 3-Quarter means

Aspect-Based Sentiment Analysis (ABSA) & spin detection

Traditional sentiment analysis fails to capture obfuscation because it treats all positive sentiment equally, whether expressed with confidence or hedged with uncertainty. The spin detection module addresses this limitation by constructing a composite metric that quantifies the degree to which positive language is undermined by linguistic markers of uncertainty or ambiguity.

We implemented ABSA to capture sentiment by aspect. We used yangheng/deberta-v3-base-absa-v1.1, which jointly identifies the aspect and sentiment in context important in financial transcripts where sentiment cues and targets may be separated by intervening text.

Table 2. Comparison of candidate sentiment models evaluated for ABSA and financial-text suitability

Model	Strengths	Weaknesses
yangheng/deberta-v3-base-absa-v1.1 (~90.9% F1, ~80–81% F1 on ABSA tasks)	• **Multi-aspect ABSA (24 aspect classification grid)** • **ABSA-specific fine tuning** • **Financial domain adaptation**	• Higher computational cost
ProsusAI/FinBERT (~86–88% accuracy on finance sentiment tasks)	• Financial language training • Regulatory compliance focus	• **General sentiment only** • **No aspect extraction**
amphora/FinABSA (~87% accuracy on finance ABSA test split)	• Trained on financial ABSA dataset (SEntFiN) • Domain-specific sentiment mapping • Handles finance jargon well	• Finance domain • **Usually single aspect–sentiment pair**

This analytical framework provides granular, quantifiable insights into bank earnings call communications by detecting sentiment manipulation ("spin") in management guidance, cross-sentiment patterns revealing mixed messaging within statements, bank-to-bank comparative risk signals across quarters, and speaker-level discrepancies between management and analyst sentiment.

To detect spin, positive language co-occurring with uncertainty, we engineered a composite metric:

Spin = 0.5 × sentiment entropy + 0.4 × hedging indicator + 0.1 × positive bias.

- *Entropy* (normalised Shannon entropy of the model's class probabilities) captures ambiguity;
- *Hedging* flags modal and cautious terms ("may", "might", "we believe", "monitoring", "headwinds");
- *Positive bias* measures net positive probability beyond negative.
- Weights are theory-driven (ambiguity and hedging dominate true obfuscation); future work can use a grid search to optimise via correlation to adverse prudential movements. This produces reproducible, auditable aspect-level time series for both tone and framing (spin).

```
HEDGE_TERMS = ["may", "might", "could", "should", "potentially",
               "we believe", "we expect", "cautiously", "monitoring", ...]

def spin_features_row(text, pneg, pneu, ppos):
    # Linguistic hedging markers
    hedges = len(HEDGE_RE.findall(text))

    # Sentiment uncertainty (Shannon entropy)
    ent = sentiment_entropy(pneg, pneu, ppos)

    # Sentiment bias
    pos_minus_neg = ppos - pneg

    # Composite spin score
    return {
        "spin_score": 0.5*ent + 0.4*(hedges>0) + 0.1*max(0.0, pos_minus_neg)
    }
```

Temporal linkage to prudential metrics

We aligned FinBERT and ABSA series with CET1, LCR, NSFR, and LDR, testing contemporaneous and 3-quarter rolling windows. At the pooled level:

- LDR showed a positive link with sentiment (~+0.11): more positive tone during active lending.
- NSFR was mildly positive (~+0.04), consistent with confident tone in stable funding environments.
- LCR was weakly negative (~-0.03): more cautious language when liquidity buffers are emphasised.
- CET1 was near-neutral (~+0.01).

Aspect-level analysis revealed consistent patterns across banks: NSFR correlated positively with many aspects (mean ~+0.23), while CET1 tended to correlate negatively (mean ~-0.22). Interpretation: funding stability aligns with confident managerial tone; phases of capital build or capital caution coincide with more guarded communication. LCR hovered near zero; LDR was mildly negative by aspect (likely reflecting cautious messaging during tighter lending regimes).

Aspect	CET1	NSFR	LCR	LDR	Interpretation
Capital	-0.22	+0.23	+0.01	-0.12	Cautious tone when capital is strengthened; positive tone when funding and lending conditions improve.
Liquidity	-0.19	+0.24	+0.05	-0.06	More conservative tone during high-liquidity buffers; optimism during active lending periods.
Digital Strategy	-0.24	+0.21	+0.03	-0.09	Defensive language around technology investments during capital strain.
Economic Environment	-0.23	+0.20	+0.04	-0.08	Negative tone in macroeconomic discussion aligns with capital tightening.
Regulation	-0.21	+0.23	+0.04	-0.08	Positive tone in periods of stable regulatory compliance and funding security.
Expenses	-0.22	+0.23	-0.01	-0.09	Cost-control discussions become more cautious when capital constraints increase.
Fees / Revenue	-0.22	+0.21	+0.03	-0.09	Constructive tone when revenue stability supports stronger funding ratios.
Loans / Credit Risk	-0.22	+0.21	+0.02	-0.07	Loan-quality discussions become negative when capital tightening dominates.

Table 3. Aspect-Based Sentiment (ABSA) Correlations with Prudential Fundamentals - Pooled Sample

At the bank level, examples reinforce the pooled view: HSBC showed strong negative correlations between tone and CET1 on *economic environment* (-0.684) and *digital strategy* (-0.669), consistent with defensive/guarded language in capital-tightening phases; UBS showed positive tone-NSFR correlations for *capital* (+0.623) and *deposits* (+0.614), matching confident tone amid funding stability. Visual inspection of role-delta plots indicated that rising analyst pessimism and higher spin often preceded prudential tightening by roughly one quarter.

Table 4. High-Magnitude Aspect-Fudamental Correlations at the Bank Level

Bank	Aspect	Fundamental	Correlation (r)	Observations (n)	Interpretation
HSBC	Economic Environment	CET1	-0.684	10	Strongly negative relationship: cautious tone in macroeconomic discussions coincides with stronger capital positions.
HSBC	Digital Strategy	CET1	-0.669	10	Defensive tone during technology and strategy discussions reflects managerial caution under capital tightening.
UBS	Capital	NSFR	+0.623	11	Positive association: confident tone aligns with stable long-term funding conditions.
UBS	Deposits	NSFR	+0.614	11	Optimistic tone in deposit and funding discussions during periods of balance-sheet stability.

Predictive modelling (XGBoost)

To quantify explanatory power, we merged transcript-derived features (role sentiment, sentiment deltas, disagreement/evasive, spin, aspect summaries) with bank-quarter metadata, yielding 5,909 observations across 27 variables. We trained XGBoost regressors (80/20 time-aware split) on two targets: executive sentiment (linguistic outcome) and number of exchanges (structural outcome).

Executive sentiment: $R^2 \approx 0.96$, MAE = 0.007, RMSE = 0.008, near-perfect tracking, confirming that our feature set sufficiently encodes managerial tone.

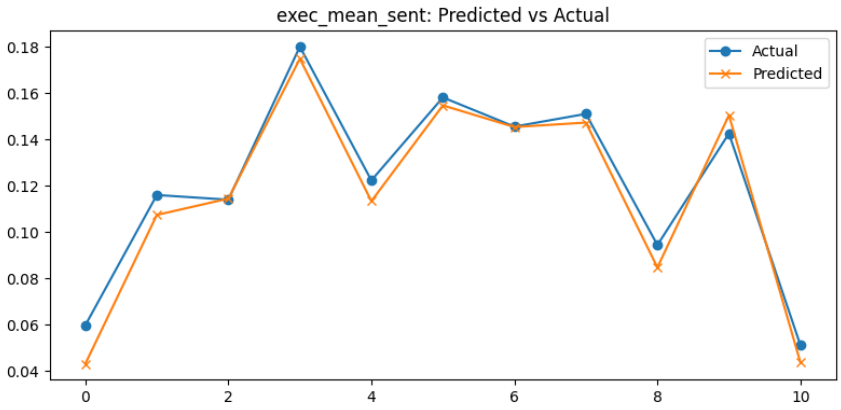

Figure 2. XGBoost performance for merged data for executive sentiment

Exchanges: $R^2 \approx 0.48$, with correct timing of major peaks but mild magnitude overestimation in calmer periods - unsurprising given context variables not captured in text alone (e.g., earnings surprises, incident newsflow).

Across discrepancy-type KPIs, performance varied: e.g., discrepancy_magnitude ($R^2 = 0.979$) and relative_to_bank_avg ($R^2 = 0.960$) were strong; aggregated/lagged summaries were weaker (e.g., bank_quarter_avg_discrepancy $R^2 = 0.425$), indicating language excels for "nowcasting" and short-horizon diagnostics more than longer-lag forecasting. Feature importance ranked role sentiment highly, with operator roles minimal.

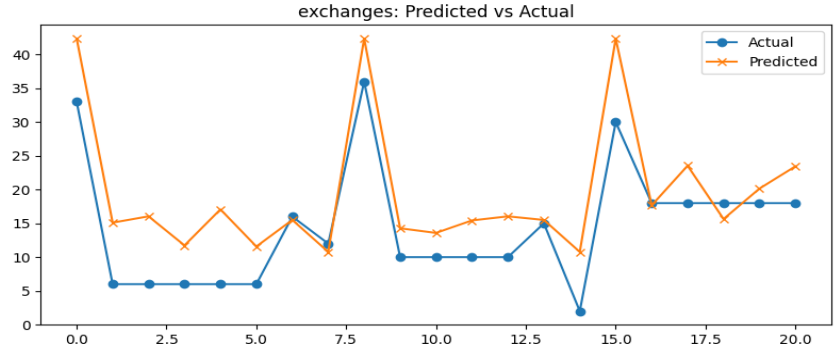

Figure 3. XGBoost performance for merged data for executive sentiment - Predicted vs Actual

XGBoost was selected over linear and deep-learning baselines because it handled non-linear relationships effectively on a small quarterly dataset while remaining interpretable

Retrieval Augmented Generation (RAG) - Chatbot Development

Retrieval Augmented Generation (RAG) was used to analyse unstructured PDF transcripts, combined with the quantitative average bank score (from fundamental analysis), sentiment, disagreement and avoidance scores. Transcripts were chunked (800 tokens, 100 overlap), embedded via all-mpnet-base-v2, stored in Chroma, and queried through LangChain with Phi-4 Mini-Instruct suitable for Q&A and Conversational RAG. Text-summaries were added to each bank/quarter-combination from the CSV files, embedded and merged. Metadata (bank, quarter, source) were added to the Chroma database for precise retrieval. Early tests correctly surfaced linguistic shifts preceding Credit Suisse's 2022 stress, illustrating diagnostic value. Similar findings can be found in UBS's 2024 transcripts, showing repeated evasive phrasing and avoidance of numerical detail, consistent with financial risk.

Layer	**Component**	**Justification**
Embedding Model	Sentence-transformers/all-mpnet-base-v2 (HuggingFace)	Top-performing sentence embedding model for capturing semantic similarity, suitable for Q&A and chatbot RAG. Balance between quality and efficiency.
Vector Database + Retriever	Chroma	Compatibility with RAG and LangChain
LLM (Generator)	Phi-4-mini-instruct	Used for its efficiency and strong mathematical and logical reasoning performance for complex Q&A and multi-step problem solving
Orchestration Layer	LangChain RetrievalQA or ConversationalRetrievalChain.	RetrievalQA enables semantic search and comparative Q&A within one PDF. allow interactive exploration of the transcripts and comparison with quantitative indicators and sentiment/avoidance/disagreement-scores.

Validation, iterations & challenges

- **Validation:** manual spot-reads on extremes; cross-checks for FinBERT bias; role-segregated aggregation; rolling windows to ensure minimum observations.
- **Iterations/decisions:** added exchange_id to ground interaction analysis; standardised aspect list for cross-bank comparability; tuned rolling window (3Q) to balance stability vs responsiveness; selected XGBoost for non-linear, mixed-type features.
- **Challenges:** heterogeneous PDFs (esp. Deutsche Bank speaker attribution), short quarterly panels limiting lag inference power, institution-specific disclosure styles that can shift vocabulary and sentiment baselines.
- Pipeline components were developed collaboratively across the team, and then integrated and jointly validated through a shared notebook.

Ethics, governance & sustainability (CLO3)

- **Data ethics & privacy:** only public investor transcripts; no personal data.
- **Explainability:** ABSA aspect scores and the spin metric create an auditable trail; RAG returns source text to support supervisory review.
- **Bias & fairness:** checked FinBERT outputs for systematic polarity drift unrelated to context; triangulated with ABSA and manual reads to reduce model artefact risk.
- **Sustainability:** pipeline-first approach (reproducible, maintainable); modular components (embedding/model swaps without re-architecting).

Results

Executives were consistently more optimistic than analysts (0.163 vs 0.109 pooled means). Role-delta and spin-delta series showed that analyst pessimism and rising spin often preceded prudential tightening by ~1Q.

Sentiment related positively to LDR (+0.11) and modestly to NSFR (+0.04), but was slightly negative vs LCR (-0.03) and neutral to CET1 (+0.01). Aspect-level analysis strengthened interpretation: NSFR was broadly positive across aspects (mean ~+0.23), while CET1 was negative (mean ~ -0.22), indicating guarded tone in capital-conservative phases.

HSBC: strong negative CET1 links (economic environment -0.684; digital strategy -0.669) consistent with cautious language during capital build. UBS: positive tone-NSFR relations (+0.623 for capital; +0.614 for deposits) reflecting confidence in stable funding periods.

XGBoost explained executive sentiment with $R^2 \approx 0.96$, and exchanges with $R^2 \approx 0.48$, capturing turning points but overestimating magnitude in quiet regimes, evidence that language features are powerful for nowcasting decision-support, with diminishing returns at longer lags or for structurally noisy KPIs.

Discussion & Conclusion

Language in earnings calls adds real supervisory value: when funding is stable, tone is more confident; when capital is being conserved, tone turns guarded, and these patterns often lead prudential movements. Combining FinBERT, ABSA (with a spin metric), and temporal linkage produced consistent, interpretable signals across banks and aspects, while XGBoost confirmed that linguistic features encode managerial tone strongly and provide useful nowcast power for interaction intensity.

The RAG prototype operationalises this by returning explainable, cited passages, letting supervisors interrogate text directly and tie narrative signals to metrics; early Credit Suisse tests illustrate practical utility. Constraints remain (short panel, heterogeneous disclosure), but the pipeline is reproducible, auditable, and extensible. Priority next steps are expanding coverage, stress-testing the spin metric against prudential outcomes, and introducing uncertainty calibration. Overall, narrative analytics complement prudential measures and can improve early detection of emerging stress.