International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Ai-Enabled Financial Text Analytics: Extracting Sentiment from Corporate Earnings Calls And Its Association With Stock Price Reactions

Sidhanth U P and Ananya Bhat
PES University, Faculty of Management Studies.

Abstract - Financial text is rich with information on management forecasting, companies' business environments and anticipated future performance. This research investigates how Artificial Intelligence (AI) and Natural Language Processing (NLP) can be utilized to turn unstructured financial text into quantifiable sentiment features in order to enable financial analysis and data-driven decision-making. In this research, 20 Apple, Amazon, Tesla, JPMorgan Chase and Netflix earnings-call observations over the FY2024 period (Q4'23 - Q1'24) are constructed for natural language-based sentiment feature extraction and analysis. Loughran-McDonald finance-specific sentiment dictionary is used to derive positive/negative sentiment words and calculate a Net Sentiment Score in order to demonstrate how descriptive-statistical tools, (multiple) pairwise correlation analysis and simple linear regression, can help derive new financial insights while a two-sample t-test is used to draw insight from stock-price reactions between positive- and negative-sentiment observations. The demonstration concludes that the lower the absolute value of the next-trading-day stock price change, the higher the positive-sentiment observation's stock price. Moreover, the association between next-trading-day stock price change and the Net Sentiment Score is only low positive, (r = 0.4272; R2 = 0.1825). Additionally, the limits of the demonstrated analysis are discussed together with how contextual financial sentiment can be derived through a transformer-based approach utilizing existing framework. This paper exemplifies how the application of intelligent text-analytics can turn unstructured corporate disclosures into quantifiable financial information, and how such a sentiment can be connected to financial analytics to visualize information extraction and quantitative analysis together.

Keywords: Natural Language Processing; AI in Finance; Financial Text Analytics; Sentiment Analysis; FinBERT

1. INTRODUCTION

The quarterly earnings calls are, to date, one of the most censored disclosure event for equity markets. Besides the audited or review numbers, in a spoken version, management already states the reasons for the numbers and tries to guide investors with management's state of mind, wavering and caution in their language alone offers qualitative clues missing from the financial statements to be taken into consideration (Matsumoto, Pronk, & Roelofsen, 2011; Larcker & Zakolyukina, 2012).

A huge disadvantage of human reading is that no analyst would be able to systematically work through this level of text. For this reason, automated text analysis has become widely adopted in accounting and finance (Kearney & Liu, 2014; Loughran & McDonald, 2016). Text analysis techniques can be divided into mainly two categories. Dictionary based techniques just count the occurrence of predetermined words on a list of words. The lexicographical lists were shown to classify normally neutral words that referring to everyday business activities, such as 'cost' and 'liability', as negative, were specifically created as the Loughran-McDonald (LM) lists in finance (Loughran & McDonald, 2011). Transformer based language models learn context about larger groups of texts and there are finance-specific models out, such as FinBERT (Araci, 2019; Huang, Wang, & Yang, 2023), that take into account negation and phrasing that a word count cannot.

This paper is in the first family, and it produces the second. The purpose is to document in detail how unstructured disclosure text can be turned into a quantitative measure, and then to try in every way I know to verify empirically that the measure is associated with the market reaction. The flow diagram is in Figure 1.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

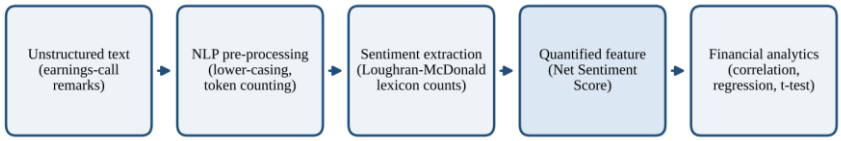


Figure 1: Conceptual workflow from unstructured earnings-call text to financial analytics. Source: Authors' framework.

The price reactions to earnings announcements can be very large relative to previously available market information. The prices reactions to the sample results range from -12.25% to +21.66% (Tesla - both). Prior research has identified that both the content and tone of results can affect the market reaction (Price, Doran, Peterson, & Bliss, 2012; Davis, Piger, & Sedor, 2012).

2. LITERATURE REVIEW

Pessimism of this type was first associated with price movement in studies by Tetlock (2007), who demonstrated that negative sentiment of this sort, measured using a common dictionary, forecasts a reversal in negative return realizations. Loughran and McDonald (2011) demonstrated that a large fraction of the negative words in those dictionaries are not negative in finance, and their new dictionaries are now the industry standard.

How disclosures link tone to the essentials Work by Davis, Piger and Sedor (2012) demonstrates that tone of earnings press-release text predicts both a firm's future performance and the market's reaction to this future performance, in addition to the numerical content itself.

The calls themselves have also been directly investigated. Matsumoto et al. (2011) documented that the prepared presentation and the analyst discussion carry information. Price et al. (2012) document that call tone is related to the post-announcement abnormal return and volume, controlling for the earnings surprise, and also the post-call drift. Larcker and Zakolyukina (2012) extracted linguistic markers of deceptive discussion from calls transcripts. Mayew and Venkatachalam (2012) found that managers' vocal affect also conveys information about future performance, suggesting the possibility of additional data sources in the audio.

The papers Li (2010), Kearney and Liu (2014), review the lexicons and approaches based on machine learning techniques for financial text; Loughran and McDonald (2016), alert the reader to dictionary mismatch, out-of-sample validation, and the limits of scoring words alone. Araci (2019) suggested FinBERT - a BERT model re-trained on financial text and fine-tuned on sentiment, and Huang et al. (2023), created a version of FinBERT for information extraction from financial text and demonstrate improvements over dictionary-based approaches in experiments.

Three points follow. For corporate text, finance-specific lexicons are more reliable than general ones. Call-level tone provides information value beyond numbers. Transformer models are an improvement on word counting to understand context. Less frequently we see a small, completely transparent and reproducible pipeline, in a setting where a student researcher or practitioner can verify every stage. That is what we build towards below.

3. OBJECTIVES AND RATIONALE

The objectives of this study are:

1. Translate management commentary from earnings calls into a measureable Net Sentiment Score by applying the finance-specific Loughran-McDonald lexicon using a basic, formula-oriented approach.

2. To compare whether the score is associated with the following-trading-day response of stock price, and at what sensitivity that is associated with outliers, to the ranking-based measures, and to the word-matching rule.

3. To describe what characteristics of a firm and quarter lead to divergence in sentiment, and specify a FinBERT-based extension we will later compare our lexicon score to.

The difference is the lack of hands-on knowledge. Large studies have proprietary data and are hard to reproduce, but commercial products report sentiment without revealing how they are built. An open small pilot cannot

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

answer the question empirically, but it can show how measurement is done, the small scale evidence, and the points of failure. The paper is therefore seen as exploratory.

4. RESEARCH METHODOLOGY

4.1 Scope

The sample includes the sample of firms, which are large-cap NYSE-and NASDAQ-listed U.S. firms and have broad analyst following, four quarters of FY2024, and the management-prepared comments of each call (Q4 calls happen late January or early February 2025 for some firms). Question-and-answer part of the call is left for the future. The information-content-of-disclosure approach of the markets (Tetlock, 2007; Price et al., 2012).

4.2 Population and Sampling

The population consists of all U.S.-listed firms' quarterly earnings calls in FY2024. A purposive sample: 5 high-liquidity, widely followed firms, each in a separate sector: Apple (tech/hardware), Amazon (e-commerce/cloud), Tesla (automotive/energy), JPMorgan Chase (banking), and Netflix (streaming). Of all 4 calls in FY2024, total 4 x 5=20 observations. Purposive sampling makes generalisation problematic, which was addressed in Section 7.3.

4.3 Data Sources

The two kinds of secondary data used are: The text data are computer-generated transcriptions prepared by the researchers for each call, and consist of paragraphs of 42-60 words of soundbites that are similar to the content in the transcripts published by publicly available companies (transcript-aggregator companies like Motley Fool are one example). These are not actual transcriptions of calls, and should be treated as an initial corpus. The market data are percent changes in close from the previous-trading-day close on the day of the call based on yahoo finance historic prices. The sentiment lexicon is Loughran-McDonald master dictionary (Notre Dame SRAF); for reasons of transparency we are using the 59-word subset of positive and negative words (29 and 30 respectively). The ProsusAI/finbert model is used from the Hugging Face model hub extension.

4.4 Hypotheses

H1. Net Sentiment Score has a positive relationship with next-trading day price reaction (null: no relationship).
H2: Price response is different for calls that are net positive than it is for calls that are net negative (null: no difference).

4.5 Measurement and Empirical Model

For each excerpt i, the number of terms (case insensitive) appeared in each dictionary entry was summed for all instances in the excerpt with the substring-count formula in a spreadsheet and added to positive and negative counts, respectively, for each excerpt. Net Sentiment Score (NSS): (Number of positives - Number of negatives) / Number of words in the excerpt. Total words = count of words in the excerpt (segmented by space). A call is classified net positive if NSS > 0, net negative if NSS < 0 and neutral if NSS=0. The relationship with market reaction that is reported is computed by simple regression:

$$\text{Price Reaction}_i = \beta_0 + \beta_1 \times \text{NSS}_i + \varepsilon_i$$

Pearson and Spearman correlations and the Welch two-sample t-test between the net-negative and net-positive groups. As the substring rule can count a word multiple times if lexicon items are inflections of each other (e.g. Decline, declined, declining), robustness re-scores every excerpt with only whole-word matching.

4.6 FinBERT Extension (Specified, Not Executed)

All snippets would have been fed into the pre-trained ProsusAI/finbert classifier on the Hugging Face transformers library, with the FinBERT score as P(positive) P(negative). This score would be compared against the lexicon NSS and the price response. It has not been done yet and no FinBERT outcomes in this paper.

5. DATA PRESENTATION

Table 1 shows the sample structure and Table 2 the lexicon subset. Table 3 lists the Net Sentiment Score and price reaction for all 20 calls, and Figure 2 shows the positive and negative term counts behind each score.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Table 1: Sample structure (5 companies × 4 FY2024 quarters = 20 observations).

Company	Ticker	Sector	Quarters	Calls
Apple Inc.	AAPL	Technology hardware	Q1-Q4 FY24	4
Amazon.com Inc.	AMZN	E-commerce / cloud	Q1-Q4 FY24	4
Tesla Inc.	TSLA	Automotive / energy	Q1-Q4 FY24	4
JPMorgan Chase & Co.	JPM	Banking	Q1-Q4 FY24	4
Netflix Inc.	NFLX	Media / streaming	Q1-Q4 FY24	4

Table 2: Loughran-McDonald lexicon subset used for scoring (illustrative words shown). Source: Loughran and McDonald (2011).

Category	Illustrative words	Words used
Positive	strong, growth, record, improved, confident, robust, momentum, optimistic, pleased, healthy	29
Negative	loss, decline, weak, uncertainty, risk, impairment, downturn, pressure, headwind, softness	30

Table 3: Net Sentiment Score and next-day price reaction by company and quarter.

Company	Quarter	Pos.	Neg.	Words	NSS	Reaction (%)	Class
Apple	Q1 FY24	8	3	60	0.0833	+0.59	Net positive
Apple	Q2 FY24	5	1	54	0.0741	+5.98	Net positive
Apple	Q3 FY24	5	2	52	0.0577	-0.90	Net positive
Apple	Q4 FY24	2	4	48	-0.0417	-1.57	Net negative
Amazon	Q1 FY24	8	0	54	0.1481	+2.22	Net positive
Amazon	Q2 FY24	2	2	46	0.0000	-8.70	Neutral
Amazon	Q3 FY24	5	0	50	0.1000	+6.20	Net positive
Amazon	Q4 FY24	4	2	50	0.0400	-4.10	Net positive
Tesla	Q1 FY24	4	5	51	-0.0196	+12.04	Net negative
Tesla	Q2 FY24	4	1	42	0.0714	-12.25	Net positive
Tesla	Q3 FY24	7	0	49	0.1429	+21.66	Net positive
Tesla	Q4 FY24	6	4	48	0.0417	-3.30	Net positive
JPMorgan	Q1 FY24	3	3	51	0.0000	-6.20	Neutral
JPMorgan	Q2 FY24	4	1	47	0.0638	+0.20	Net positive
JPMorgan	Q3 FY24	5	2	42	0.0714	+4.50	Net positive
JPMorgan	Q4 FY24	5	0	46	0.1087	+2.10	Net positive
Netflix	Q1 FY24	5	0	51	0.0980	-9.00	Net positive
Netflix	Q2 FY24	6	0	43	0.1395	+4.70	Net positive
Netflix	Q3 FY24	7	0	45	0.1556	+11.19	Net positive
Netflix	Q4 FY24	5	0	44	0.1136	+9.10	Net positive

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

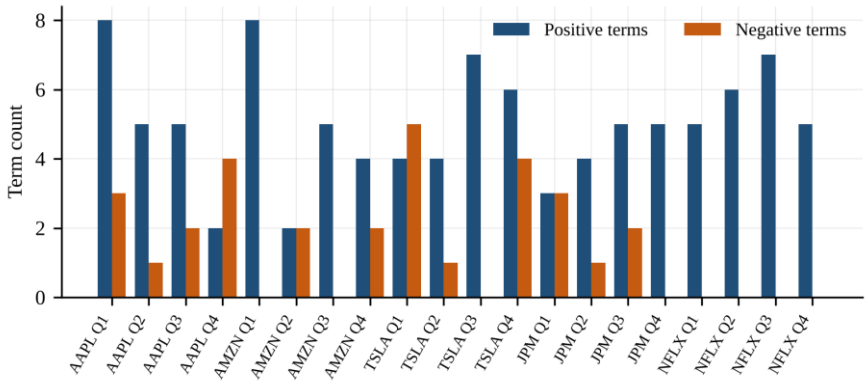


Figure 2: Positive and negative lexicon term counts per earnings call (n = 20).

6. DATA ANALYSIS, RESULTS AND INTERPRETATION

6.1 Descriptive Statistics

The mean of all the NSS across 20 calls is 0.0724 (median 0.0728, SD 0.0566). It runs from 0.0417 (Apple Q4) to 0.1556 (Netflix Q3) with a weak negative skew of (-0.41). Excerpts show 100 positive and 30 negative hits of the lexicon thus making the corpus positively skewed as is the nature with prepared management speeches. The mean of all price response is positive 1.72% (median 1.35%, SD 8.16%) that ranges from 12.25% (Tesla Q2) to 21.66% (Tesla Q3) with positive skew (+0.49). There are 16 net positive, 2 net negative and 2 neutral calls.

6.2 Correlation and Regression

The Pearson correlation is r = 0.4272 and R^2 = 0.1825, so about 18% of the variation in next-day reaction is explained by sentiment (Figure 3). Regressing price reaction on NSS gives Price Reaction = 0.6156 × NSS − 0.0274, where reaction is expressed as a fractional change, so a 0.01 change in NSS equates to about a 0.62 percentage point change in next-day price reaction. The standard error of the slope is 0.3071 (t = 2.00, df = 18). The two-sided p-value is 0.060, which is not significant at the 5% level but is at the 10% level; the one-sided p-value (since H1 is directional) is 0.030. These values are presented together so that the reader can judge the strength of evidence. With a sample size of 20, these figures should be regarded as suggestive.

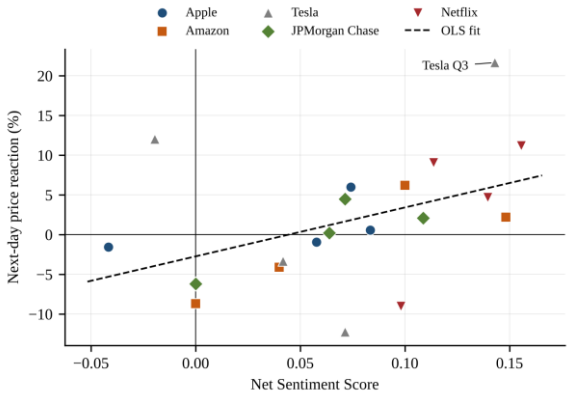


Figure 3: Net Sentiment Score vs next-day price reaction with OLS fit (n = 20).

6.3 Robustness Checks

We performed four checks of the main result (see Table 4). First, a rank Spearman correlation (which is less sensitive to outliers) yields r_s = 0.518 (p = 0.019) and ρ = 0.518 (p = 0.019), similar to the main correlation. Second, removing the call with the strongest reaction (Tesla Q3, +21.66%) reduces the main Pearson r to 0.331 (p = 0.166) so the main correlation appears to rely, at least in part, on a single data point. Third, using the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

alternative whole-word match scoring algorithm yields r = 0.421 (p = 0.064, R^2 = 0.177); the two NSS measures correlate at 0.959. Fourth, the sign of NSS aligns with the sign of the price reaction in 12 of 18 calls with nonzero NSS (67%); one-sided binomial p = 0.119.

Table 4: Summary of association tests (n = 20 unless stated).

Specification	Statistic	p-value
Pearson, substring scoring (baseline)	r = 0.4272; R^2 = 0.1825	0.060 (two-sided); 0.030 (one-sided)
Spearman rank correlation	ρ = 0.518	0.019
Pearson, excluding Tesla Q3 (n = 19)	r = 0.331	0.166
Pearson, whole-word scoring	r = 0.421; R^2 = 0.177	0.064
Welch t-test, net positive (n = 16) vs net negative (n = 2)	means +2.43% vs +5.24%	0.752
Sign agreement, NSS ≠ 0 (12 of 18)	67%	0.119 (binomial, one-sided)

6.4 Two-Sample t-Test (H2)

Net-positive calls (n = 16) return a mean reaction of +2.43%, compared to +5.24% for net-negative calls (n = 2: Apple Q4 and Tesla Q1). The result from the Welch t-test is p = 0.752, so H2 is also not supported. The test is very weak here, because the negative group comprises two calls and one of those is Tesla Q1 (+12.04%). The two neutral calls (Amazon Q2 and JPMorgan Q1) average −7.45%, but this cannot be tested with so few observations.

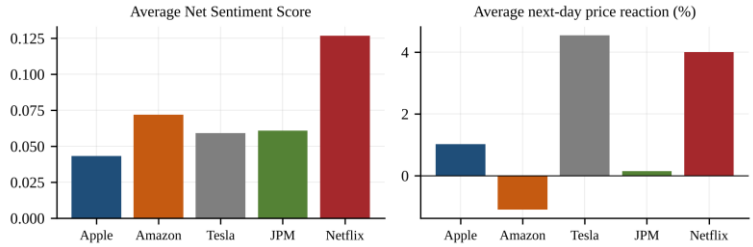


Figure 4: Company-wise average Net Sentiment Score and average price reaction.

6.5 Company and Quarter Patterns

Netflix has the highest average NSS (0.1267), the most consistent scores (SD 0.026) and a strongly positive average reaction (+4.00%). Amazon has a moderate average NSS (0.0720) but a negative average reaction (−1.10%), mainly because of Q2, when the NSS was neutral but the price moved −8.70%. Tesla shows the weakest alignment: its lowest NSS (Q1) was followed by a +12.04% reaction, its highest NSS (Q3) by the largest positive reaction (+21.66%), and its positive Q2 score by a −12.25% move. Apple's score turns negative only in Q4, when the reaction was also negative (−1.57%). Both series are plotted by quarter in Figure 5.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

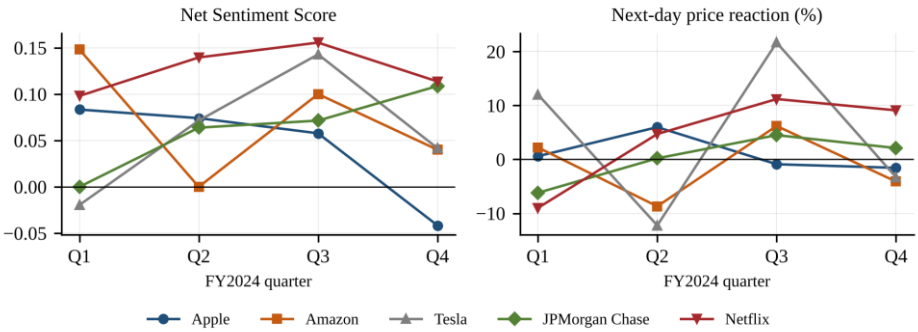


Figure 5: Quarterly trend in Net Sentiment Score (left) and next-day price reaction (right), by company.

7. FINDINGS AND DISCUSSION

7.1 Findings

- Net Sentiment Score by LM is essentially being computable in an auditable and formulaic pipeline, from raw text through to final auditable score.
- Sentiment and next-day response are positively and imperfectly correlated ($r = 0.427$, $\rho = 0.518$). The correlation is marginally significant for the Pearson ($p = 0.060$ two-sided) but more clearly significant for the Spearman ($p = 0.019$).
- Sentiment explains 18% of the variation in reaction. Without Tesla Q3 the association is r=0.331 and it remains essentially the same with whole word matching, so directionally stable but magnitude precarious.
- Mean reaction shows no difference between net-positive and net-negative calls ($p = 0.752$); H2 is not supported but this test suffers from a lack of power.
- Netflix displays the most evenly positive tone and strongest average reaction. Amazon and Tesla have positive tone but mixed reactions.
- A FinBERT comparison is listed but not performed so there is no claim for lexicon vs transformer.

7.2 Discussion

Direction of result is in line with previous large-sample studies such as Price et al. (2012). But here the evidence is weaker and typical of the 20 observations requiring large effects for detection. The difference between Pearson and Spearman and the significant drop in result with exclusion of Tesla Q3 show a few extreme price moves are responsible for the correlation. The positive skew in corpus (100 hits positive as against 30 negative) further obscures anything meaningful.

Amazon and Tesla lend themselves to the book's overall argument that tone serves as a complement to the number: Davis et al. (2012) find language generally adds to stated outcomes and not that it replaces them. For an underwhelming, verbalized negative (Amazon Q2), a tone with some upsides should not support the stock. For a heavily retail traded and delivery-sensitive Tesla, positive price action may be driven by predictors that management's prepared statement do not contain. Absent evidence, this interpretation is merely speculation.

Word counting has its limitations.Words you hadn't even thought of like negation, hedge words, and sarcasm also bypass word counting and may be even covered in the 82% of the variance you haven't explained (Loughran & McDonald, 2016). These are some of the shortcomings context aware models like FinBERT tackles (Araci, 2019), and so the next step is simply to compare them to the transformer you already used. You could also include industry effects on baseline tone as well, given how JPMorgan's average NSS (0.061) is significantly lower than that of Netflix (0.127).

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

8. LIMITATIONS

Several limits apply.

1.The sample consists of five companies, 20 relevant purposively selected observations; thus, the results are not to be generalized.

2. These are abbreviated extracts of two to 60 words, recorded by the researchers and not the exact transcripts verbatim, and the scores from the full transcripts may be very different.

3. The lexical subset only contained 59 words, out of the total of all items of the full list of all items of the whole lexicon – extremely small, since the substrings rule could manage to enable to tally forms of words out of context, but total-word scoring provided the same sort of totals.

4. The next day window is identical for every firm, while they report at various times, JPMorgan reports prior to the market's open, so its relevant window is the day of the call.

5. There is no controls for earnings surprise, guidance, or market-wide shifts so this is correlational instead of causal.

6. FinBERT was not run.

9. CONCLUSION

The current study seeks to contribute in two ways. First to show that the commentary of an unstructured earnings call can be converted into a measure of sentiment and second to explore its association with short term price response. Using the Loughran-McDonald lexicon on twenty calls from five US firms in FY2024, this study reports a modest positive correlation ($r = 0.427$, $R^2 = 0.183$, $p = 0.060$ two-sided; Spearman $\rho = 0.518$, $p = 0.019$) with the short-term price reaction. However the correlation no longer holds once the single largest observation is excluded and the net-positive/negative comparison is insignificant.

The main original contribution is a completely auditable pipeline from raw text to score to statistical test. Others can replicate, expand or challenge it. In addition, we can go further:

(i) tape of all verbatim transcripts and larger panel spanning several years may facilitate adding controls for earnings surprise and industry, (ii) execute the above scoring exactly as defined, and employ a formal test for the predicted relationship of the lexicon score with the FinBERT score, and (iii) incorporate the question-and-answer part of the interview, and score speech parameters if available (Mayew and Venkatachalam, 2012).

REFERENCES

[1] Angela K. Davis, J. P. (2012). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*. *Contemporary Accounting Research*.
[2] Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arxiv.org*.
[3] Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *CONTEMPORARY ACCOUNTING RESEARCH*.
[4] Kearney, C. &. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33, 171–185*.
[5] Larcker, D. F. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research, 50(2), 495–540*.
[6] Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature, 29, 143–165*.
[7] Loughran, T. &. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66(1), 35–65*.
[8] Loughran, T. &. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research, 54(4), 1187–1230*.
[9] Matsumoto, D. P. (2011). What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review, 86(4), 1383–1414*.
[10] Mayew, W. J. (2012). The power of voice: Managerial affective states and future firm performance. *The Journal of Finance, 67(1), 1–43*.
[11] Price, S. M. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance, 36(4), 992–1011*.
[12] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62(3), 1139–1168*.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)