

Sentiment classifier disagreement and stock returns: Evidence from CEO earnings calls

Luke Ridewood-Thomas^a, Yizhi Wang^{a,*}

^a*Cardiff Business School, Cardiff University, United Kingdom*

Abstract

This study examines whether disagreement between dictionary and contextual sentiment classifiers in CEO earnings call language is associated with subsequent returns. A disclosure ambiguity proxy is constructed from Loughran–McDonald and FinBERT classifications for 1,010 firm-quarter observations from 23 S&P 500 financial firms during 2014–2024. No significant linear association with 90-day abnormal returns is found. The highest disagreement decile has returns 2.04 percentage points below the middle deciles, but the estimate becomes small and insignificant after removing low-denominator observations and redefining the top decile. This exploratory tail association is sensitive to measurement and threshold choices.

Keywords: Earnings calls, Sentiment analysis, FinBERT, Model disagreement, Abnormal returns

1. Introduction

Financial disclosures can be public yet costly to interpret. Bloomfield’s (2002) incomplete revelation hypothesis links information processing costs to incomplete incorporation of disclosures into prices. Earnings calls provide a setting in which managerial language may complicate interpretation, particularly when executives respond to questions. Sentiment measures capture the polarity of this language, but disagreement between measures may also contain information about the text.

Dictionary and machine-learning approaches produce different measures of financial sentiment (Frankel et al., 2022; García et al., 2023). Angelo et al. (2025) examine tone differences between managers within earnings calls. This study instead applies two classifiers to the same sentences spoken by the CEO. Holding the sentence and speaker constant within each comparison separates cross-model disagreement from differences in what individual managers say. It does not hold CEO identity constant across calls.

The disclosure ambiguity index (DAI) measures the proportion of sentences labelled negative by a Loughran–McDonald (LM) rule-based classifier that FinBERT labels neutral or positive. If negative content is initially underweighted, disagreement could be negatively associated with subsequent returns. It could also reflect classification errors or differences between the models’ training domains. DAI is therefore an unvalidated proxy for ambiguity, not a direct observation of investor disagreement.

Using 23 S&P 500 financial firms, the analysis examines a linear association and an exploratory tail specification. There is no significant linear association with subsequent

*Corresponding author: Yizhi Wang.

Email address: WangY510@cardiff.ac.uk (Yizhi Wang)

90-day abnormal returns. The highest DAI decile has lower returns, but the result is sensitive to redefining the tail after excluding observations with small denominators. The contribution is an exploratory application of classifier disagreement to disclosure analysis, with explicit evidence on the limits of its predictive interpretation.

2. Data and methodology

Bloomberg earnings call transcripts cover 23 firms over 44 quarters during 2014–2024. Firms were selected from current S&P 500 financial-sector constituents with complete transcript and price histories, introducing survivorship and look-ahead selection. The initial panel contains 1,012 calls; two have no LM-negative sentences, leaving 1,010 observations with defined DAI. Pre-merger BB&T transcripts and prices are spliced into the Truist series.

Text comprises CEO contributions, including prepared remarks and answers; in a small number of calls without CEO speech, the most senior executive is used. Whitespace is normalised, text is divided at sentence-terminal punctuation followed by a capitalised token, and fragments shorter than 15 characters are excluded. Text is lowercased without stemming. The resulting corpus contains 168,573 sentences.

The rule-based classifier uses the LM financial word lists (Loughran and McDonald, 2011) with explicit hedging rules. A sentence containing negative words and hedging is negative; otherwise, more negative than positive words gives a negative label. More positive than negative words gives a positive label without hedging and a neutral label with hedging. Remaining sentences are neutral. These are study-specific classification rules, rather than validated sentence-level labels. Supplementary Section S1 provides further implementation details.

The contextual classifier is the ProsusAI/finbert checkpoint associated with Araci (2019). Sentences are tokenised using its tokenizer and truncated at 128 tokens. The highest-probability class supplies the label, so marginal and high-confidence classifications receive equal weight. For call c , let n_c denote the number of LM-negative sentences and d_c the number of these labelled neutral or positive by FinBERT. The index is

$$\text{DAI}_c = \frac{d_c}{n_c}, \quad n_c > 0. \quad (1)$$

A value of one means all LM-negative sentences are relabelled. It can arise from a single sentence and does not establish maximum economic ambiguity.

The buy-and-hold abnormal return (BHAR) is the firm's compounded price return minus the corresponding S&P 500 Financials index return (Bloomberg S5FINL). The window starts on the trading day after the call and extends to 90 calendar days after the call, using the next trading day when the endpoint is not a trading day. Bloomberg PX_LAST prices are used for firms and the benchmark; dividends are excluded. This sector-adjusted holding-period return is not fully risk-adjusted and can overlap the next earnings announcement.

DAI and BHAR are winsorised at the full-sample first and ninety-ninth percentiles. The linear model relates BHAR to DAI with firm and year-quarter fixed effects. The tail model replaces DAI with high- and low-tail indicators, using the middle eight deciles as the reference group. Full-sample thresholds are $\text{DAI} \geq 0.960$ and $\text{DAI} \leq 0.500$. The decile threshold is exploratory, not imposed by theory. Standard errors are clustered by firm; a wild-cluster bootstrap with Rademacher weights and 4,999 iterations addresses the limited number of clusters (Cameron and Miller, 2015). Neither model includes time-varying

firm fundamentals or event-level controls, so coefficients describe conditional associations. Additional specifications appear in the supplement.

3. Results

DAI averages 0.734. The denominator has a median of 16 sentences. Of 101 calls in the highest decile, 90 have DAI equal to one; 76 of these 90 calls have fewer than ten LM-negative sentences. Extreme index values are therefore concentrated among observations for which the ratio has a small denominator. Supplementary Table S1 reports the cross-classification counts and denominator distribution.

Table 1 shows no significant linear association: the DAI coefficient is -0.0043 ($p = 0.777$). The result remains insignificant with the wild-cluster bootstrap ($p = 0.780$) and across the pooled and one-way fixed-effects specifications. The tail model yields a coefficient of -0.0204 for the highest decile ($p = 0.002$), corresponding to 2.04 percentage points lower BHAR than the middle deciles. The lowest-decile coefficient is -0.0047 ($p = 0.422$). The high-tail bootstrap p-value is 0.001. An exploratory quadratic specification has a negative squared term ($p = 0.057$), which becomes positive and insignificant after removing the highest decile (Supplementary Table S2).

Table 1: Main results

Model and term	Estimate	SE	p	WCB p	N
Linear model					
DAI	-0.0043	0.0152	0.777	0.780	1,010
Tail model					
High DAI	-0.0204	0.0067	0.002	0.001	1,010
Tail model					
Low DAI	-0.0047	0.0059	0.422	—	1,010

Notes: The dependent variable is BHAR(90). Both models include firm and year-quarter fixed effects. SE denotes the standard error clustered by firm. WCB denotes the wild-cluster bootstrap p-value. The tail model includes both indicators, with the middle eight deciles omitted. A dash indicates that no bootstrap p-value is reported for that term.

Table 2 distinguishes two ways of assessing denominator sensitivity. Keeping the original high-tail classification after excluding calls with fewer than ten LM-negative sentences leaves the high-tail coefficient at -0.0194 ($p = 0.023$). Weighting by the denominator also retains a negative association (-0.0199 ; $p = 0.015$). Thus, deleting low-denominator calls alone does not eliminate the original association.

However, excluding the same calls and redefining the highest decile within the remaining sample produces a coefficient of 0.0027 ($p = 0.706$). With the less restrictive five-sentence cutoff, recutting yields -0.0129 ($p = 0.079$). Fixed flags test the original high-DAI group after cleaning, whereas recutting tests relative extremes within the cleaned distribution. Their divergence shows that the tail finding depends on group definition. It neither confirms a general threshold effect nor establishes that the baseline result is entirely a small-denominator artefact.

Further checks qualify this interpretation. The baseline high-tail association remains negative at 30-, 60- and 120-day horizons, while the linear estimates remain insignificant. Alternative tail thresholds yield p-values of 0.337 at the eightieth percentile, 0.068 at the eighty-fifth and 0.014 at the ninety-fifth. The pre-call placebo coefficient is 0.0003 ($p =$

Table 2: Sensitivity to denominator restrictions

Specification	Estimate	SE	p	N
Full sample	−0.0204	0.0067	0.002	1,010
Denominator ≥ 5 ; original high-tail flag	−0.0246	0.0071	0.001	942
Denominator ≥ 5 ; redefined high tail	−0.0129	0.0074	0.079	942
Denominator ≥ 10 ; original high-tail flag	−0.0194	0.0085	0.023	746
Denominator ≥ 10 ; redefined high tail	+0.0027	0.0071	0.706	746
Weighted by denominator	−0.0199	0.0081	0.015	1,010

Notes: All rows report the high-tail coefficient from the two-way fixed-effects model, with firm-clustered standard errors. The denominator is the number of LM-negative sentences. Original flags retain the full-sample high-tail classification after exclusions. Redefined tails use the highest decile of the remaining sample. Weighting uses each call's denominator. The dependent variable is BHAR(90).

0.974). Excluding 2020 or individual firms generally preserves the negative association, but excluding the life and health insurance subsector reduces it to -0.0091 ($p = 0.180$). Supplementary Tables S3–S5 provide these results. These checks use overlapping samples and exploratory specifications; they are not independent confirmations and do not resolve the denominator sensitivity.

4. Conclusion

Cross-model disagreement in CEO earnings call sentiment has no significant linear association with subsequent abnormal returns in this sample. An association concentrated in the original highest decile survives several specification changes but does not survive redefining the tail within a denominator-cleaned sample. This limits its interpretation as evidence of delayed incorporation of ambiguous disclosures. The index's construct validity, the retrospectively selected 23-firm sample, price-only returns and omitted firm and event controls further constrain inference. Classifier disagreement remains a potential disclosure characteristic to investigate; the present evidence supports neither a validated measure of ambiguity nor an established return-prediction rule.

References

- Angelo, B., Johnston, M., Singh, A., Wan, Y.Q., 2025. Tone distance: Managerial tone divergence and market reaction to earnings announcements. *Financial Review* 60, 1415–1435. doi:10.1111/fire.70002.
- Araci, D., 2019. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. doi:10.48550/arXiv.1908.10063.
- Bloomfield, R.J., 2002. The “incomplete revelation hypothesis” and financial reporting. *Accounting Horizons* 16, 233–243. doi:10.2308/acch.2002.16.3.233.
- Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* 50, 317–372. doi:10.3368/jhr.50.2.317.
- Frankel, R., Jennings, J., Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science* 68, 5514–5532. doi:10.1287/mnsc.2021.4156.

García, D., Hu, X., Rohrer, M., 2023. The colour of finance words. *Journal of Financial Economics* 147, 525–549. doi:10.1016/j.jfineco.2022.11.006.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66, 35–65. doi:10.1111/j.1540-6261.2010.01625.x.