Earnings Calls and Stock Price

11

Quan Zhou 805872079
Olivar Pan

Abstract

Our project focuses on how the language used in earnings calls, specifically the CEO's pre-written scripts, can predict or affect the short-term stock price movements. Instead of only focusing on some traditional information like revenue or earnings per share, we investigated the impact of linguistic patterns and tone on the stock's price. Our dataset includes 56 earnings call transcripts from 7 companies, collected over the past two years, with each company contributing 8 quarterly calls. For each report, we recorded whether the stock price rose or fell within 24 hours. We applied three text vectorization methods: TF-IDF, word embeddings (GloVe), and BERTopic for building a confusion matrix and predicting the stock price. The result shows that TF-IDF is the best method in this case and gives us 63.6% accuracy. Then, We conduct the word clouds for both dfm matrix and the TF-IDF matrix to find the underlying linguistic patterns.Finally, we used a logistic regression model to examine how the timing of the earnings call affects the likelihood of a stock price increase or decrease.The result shows thatb linguistic features such as using confident, quantitative language ("we've," "revenue," "really") correlated with positive price movement, while vague language ("adjusted," "approximately") correlated with declines. By analyzing the word cloud, we discovered that quarter timing plays a major role in stock movement. We used a logistic regression model to examine how the timing of the earnings call affects the likelihood of a stock price increase or decrease. Stock prices were more likely to rise after Q4 reports (released in January–February) and more likely to fall after Q1 reports (April).

Introduction

Stock Trading has become one of the most accessible ways for individuals to participate in the financial market and build up personal wealth. Earnings calls have become important moments in

today's financial market, serving as high-stakes presentations where company leaders relay not just raw numbers but also their vision for the future. A single turn of phrase, even a little unexpected financial surprises, can start rapid and sometimes dramatic stock price swings. This scenario shows us that markets are not driven solely by quantitative data, but also by the nuances of human communication: tone, emphasis, and the framing of future guidance.

However, even though the CEO's script portion of earnings calls is significantly important, most of the financial analysis for earnings call remains surprisingly under examined regarding this portion. The traditional analysis and models lean heavily on quantitative data such as revenue growth or earnings per share,often overlooking the rich semantic cues embedded in the CEO's language. At a time when retail investors increasingly rely on real time sentiment analysis and hedge funds deploy NLP-driven algorithms, understanding how carefully chosen words influence immediate market reactions is both socially and technically important.

In this paper, we focus on the CEO's script portion of the earnings call. Our analysis is guided by four research questions:

- RQ1. Can textual features from the CEO's script predict stock movement (up or down)?
- RQ2. Which modeling method—TF-IDF, word embeddings, or BERTopic—offers the most accurate prediction?
- RQ3. How is language used differently between earnings calls followed by stock increases versus those followed by declines?

To answer these questions, we collected 56 earnings call transcripts(CEO's script portion) from the last two years using Seeking Alpha. Seeking Alpha is a well-known platform of financial news and analysis. It provides investors with access to earnings call transcripts, stock research, market insight,s and opinion articles. We then processing the data and end our paper by discussing the key findings, addressing the limitations of our study, and outlining directions for future research.

Related Works

Early studies in text analysis of earnings conference calls established that the linguistics of these calls carries significant information for investors. Price et al. (2012) used computer-aided content analysis on quarterly calls and found that the sentiment embedded in call wording is a significant predictor of abnormal returns and trading volume, even after controlling for financial surprises [1]. Loughran and McDonald (2011) developed a finance-specific sentiment dictionary, showing that domain tailored lexicons better capture the true polarity of financial text than general dictionaries [5]. Complementing this, Chambers and Anne E(1984) found that the timeliness of earnings announcements significantly influences stock price reaction, with earlier disclosures leading to stronger market responses.[6]

With the rise of machine learning and deep learning, researchers have moved beyond lexicon counts to better feature sets and models. Chin and Fan apply TF-IDF, readability measures, and BERT embeddings to earnings call transcripts, extracting features from both presentation and Q&A sections to

predict future stock returns (alliancebernstein.com). Ma et al. (2020) propose an attention based deep learning framework that encodes entire transcripts into vectors for discriminative classifiers, showing good performance to traditional ML baselines [3]. More recently, Medya et al. (2022) employ graph neural networks on a decade's worth of calls (about 100,000 transcripts) and showed that semantic features of calls often outperform numerical hard data (e.g., sales, EPS) in predicting post-call stock movements [4].

Despite these advances, there are also disadvantages. Most work analyzes entire call transcripts including spontaneous analyst Q&A or focuses on forecasting earnings surprises over multiple month periods rather than immediate price reactions. Few studies isolate CEO pre-scripted script, examine simple structural features such as script length, or apply unsupervised topic modeling (e.g., BERTopic) to unearth latent thematic differences. Moreover, much of the literature relies on large, cross sectional corpora, leaving small sample, short term outcome contexts under explored [2]. Because the CEO's prewritten script is more able to reflect the company's view for this quarter. Which is more stable and easier to observe the difference between each quarter.

Our project builds on this foundation by concentrating exclusively on CEO prepared script and correlating linguistic features with stock movement within 24 hours of the call. We integrate multiple NLP methods TF-IDF classification via random forests, chi-square keyness tests, word embeddings, and BERTopic to capture both discriminative and thematic signals. Additionally, we analyze script length and seasonal (quarter-specific) vocabulary patterns, achieving a 63.64% classification accuracy with TF-IDF and revealing novel insights into how confident versus cautious language maps to immediate stock performance. This targeted, short window approach and combination of supervised and unsupervised techniques differentiate our work and situate it as a complementary extension to existing research.

Data and Methods

To analyze the hidden sentimental and linguistic clues, we collected a total of 56 earnings call transcripts from Seeking Alpha. We extracted the prepared script by the CEOs and excluding the Q&A section to ensure consistency. These transcripts come from seven different companies, each company contributing eight calls across the past two years. For each call, we recorded the company name, the date, and a binary label for whether the stock price went up or down within 24 hours of the report. To perform the text vectorization techniques, we first process the transcripts. We cleaned our data by converting all text to lowercase, removing punctuation, numbers, symbols, and stopwords.

For building the **confusion matrix**, we began by vectorizing the CEO transcripts using two different vectorization methods. Each vectorized dataset was then split into two parts: 80% of the data was used for training the model, 20% was reserved for testing. For both TF-IDF and GloVe vectorization, we trained separate Random Forest classifiers using the caret package in R. We applied 5-fold cross-validation and tuned both the number of trees and the maximum depth of the model to find the best performing combination for each method. After the optimal model is identified, we evaluate each

vmodel using the confusionMatrix function to generate key classification metrics, including accuracy, precision, recall, and F1-score. Comparing the confusion matrices from the two methods allowed us to assess which vectorization technique provided better predictive performance.

To test which vectorization has the best predictive outcome of the groups, we also used the **BERTopic** in Python to vectorize and classify our data. We began by vectorizing each script using SentenceTransformer to vectorize the data. Then, we applied UMAP to reduce the dimensionality of the vectorization. Next, HDBSCAN was used to cluster the reduced vectors into topic groups. Finally, we used the pandas library to organize the output and visualize the topic distributions and frequencies. And also visualized and compared the most relevant words within the topics.

To explore the underlying linguistic patterns, we created **word clouds** to visually represent the most relevant words in each group. We used the texplot_wordcloud function from the quanteda.textplots package to build the word clouds. We first built word clouds using the document-feature matrices(dfm), which is based on raw word frequencies. Then we convert the DFM into TF-IDF to emphasize distinctive terms rather than simply the most common ones.

Beyond the frequency-based visualization, we conducted a **chi-square test** to find the statistically significant terms for each group. Using the text_keyness function from the quanteda.textstats package in R, we computed the keyness scores to determine how strongly each word was associated with each group. Either the up or down outcome. The results of the keyness scores were visualized using ggplot2. Based on the result of the word clouds and chi-square test, we did a further analysis about how does the quarterly timing affects the stock movement by conducting a **logistic regression model**.

Results

Our first findings came out when we were organizing the data. We first split our data into up and down groups based on the stock performance afterward. Across our sample of 56 CEO-only calls, we observed that the up group has a significantly longer script than the down group. As shown in Figure 1, the mean length of the up group is around 2800 tokens, while the mean of the down group is around 2400. This 400 tokens gap (18%) suggests that the earning calls that have a longer script for the CEO tend to be in the up group. This indicates that the presenters are more confident, providing richer performance detail tends to generate more favor for the investor. What's more, after we excluded all the extreme outliers(top and lowest 5%), this pattern remains. The mean difference held at around 300 tokens($p < 0.05$ by two-sample t test). This indicates that verboisity in the earnings call is a reliable marker of investor optimism.

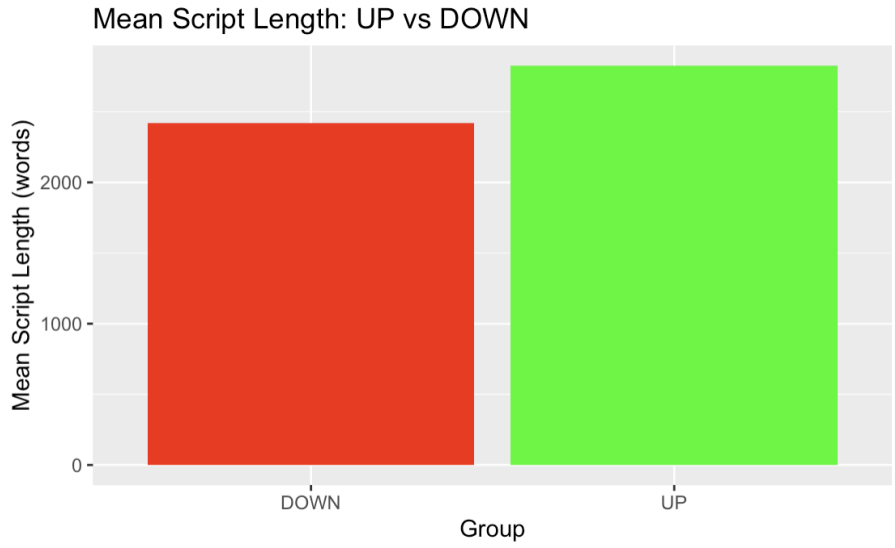

Figure 1: Distribution of CEO script length by outcome.

To predict the stock movement, we conducted two models using two vectorization methods(TF-IDF and GloVe) and trained them using Random Forest classifiers. Figure 2 shows the result of the TF-IDF model's confusion matrix. It achieved 63.64% accuracy on the 20% hold-out test set, substantially exceeding the 50% baseline of random guessing. Precision and recall score are really even and high for the down group, and the recall is really low in the up group, which means that there are fewer true positives for the up group. These results underscore the model's strength in flagging negative sentiment patterns.

Class <chr>	Precision <dbl>	Recall <dbl>	F1_Score <dbl>	Accuracy <dbl>
Down	0.625	0.833	0.714	0.636
Up	0.667	0.400	0.500	0.636

Figure 2: Confusion matrix of the Random Forest classifier trained on TF-IDF features

By contrast, the trained model using the word embedding vectorization method only achieved 0.455 accuracy, which is even less than the 50% baseline of random guessing. As we can see from the precision and recall scores shown in Figure 3, both groups have only half of the data are predicted correctly. Indicating the powerlessness of this model. This performance gap highlights that simple embedding may average the critical lexical features, but TF-IDF preserves them through term weighting.

Class <chr>	Precision <dbl>	Recall <dbl>	F1_Score <dbl>	Accuracy <dbl>
Down	0.5	0.5	0.5	0.455
Up	0.4	0.4	0.4	0.455

Figure 3: Confusion matrix of the Random Forest classifier trained on averaged GloVe embeddings

We also experimented with BERTopic to classify the data using contextual sentence embeddings. Aiming to see if a more advanced embedding vectorization method can have a better performance. The classification result is shown in Figure 4. It only gave us a topic classification with an accuracy of 57.14%, which is less than the TF-IDF method. Therefore, we conclude that for these three types of vectorization methods (sparse, word embedding, sentence embedding), TF-IDF offers the best performance in the case of predicting stock movement from earnings call script.

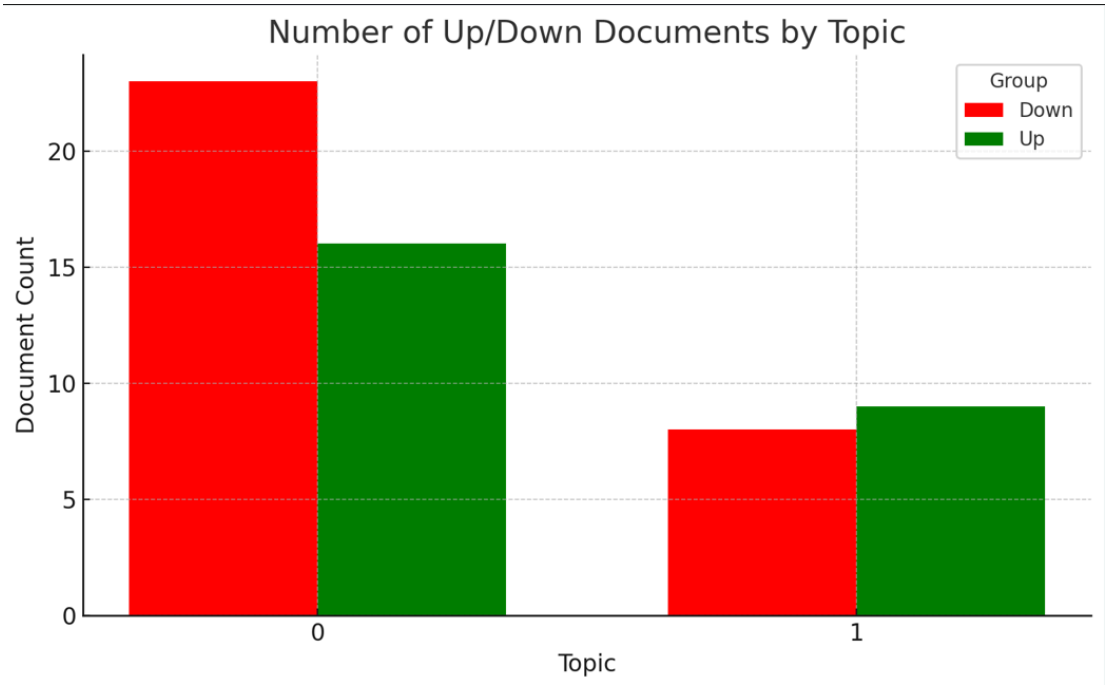

Figure 4: UMAP projection of topic clusters from BERTopic.

To qualitatively illustrate these lexical patterns, we generated word clouds under two schemes. The raw-frequency clouds in Figure 4 remain dominated by high-frequency function words (“the,” “and,” “we”), offering limited discriminative insight. However, after applying TF-IDF weighting, the clouds in Figure 5 reveal stark contrasts: the “Up” group emphasizes action-oriented and positive-framing terms such as “revenue,” “million,” “guidance,” and “confident,” whereas the “Down” group surfaces more cautious or technical terms like “approximately,” “adjusted,” “liabilities, ”and “loans.” These visualizations corroborate our quantitative finding that investor optimism is linguistically marked by forward-looking, performance-driven language, while cautionary or hedged phrasing portends negative

reactions.

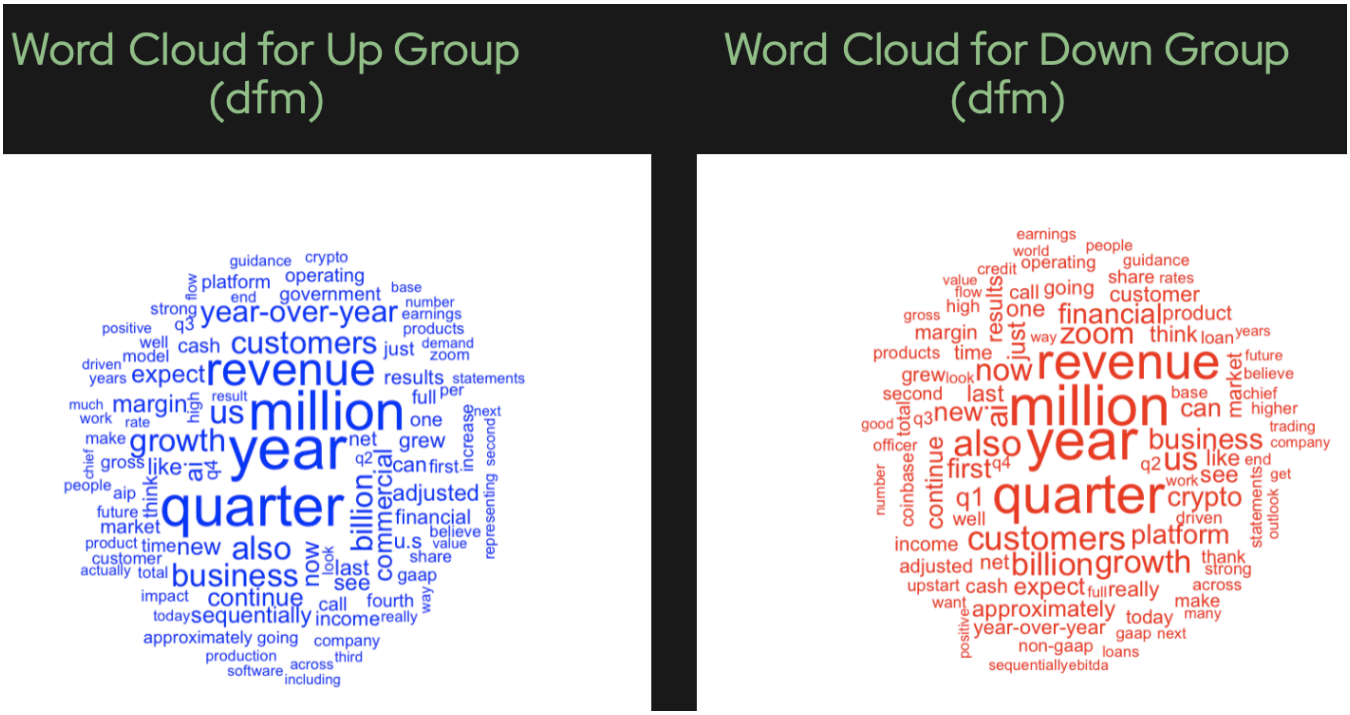

Figure 5: Word cloud (frequency‑based) for Up group (left) and Down group (right).

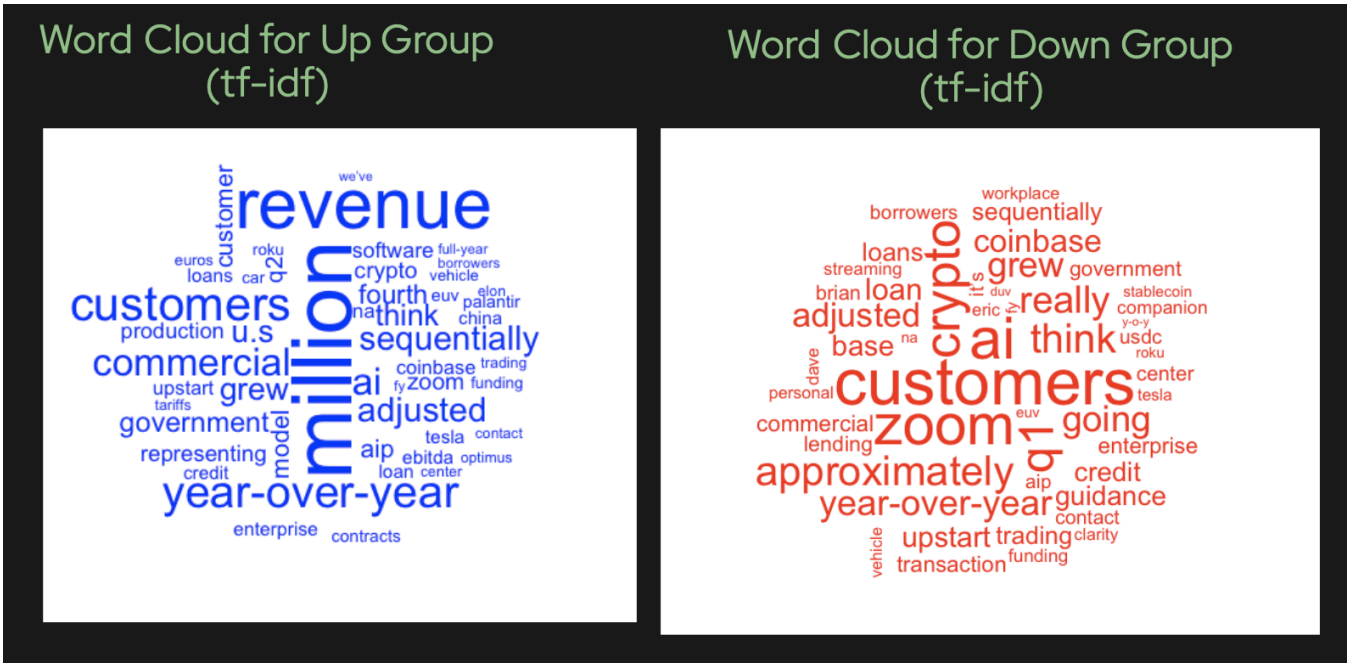

Figure 6: Word cloud (TF-IDF–weighted) for Up group (left) and Down group (right).

To formalize which words are most relevant to each group, we conducted a chi-square keyness analysis on the document-feature matrix. Figure 6 ranks the top ten terms by their keyness scores ($p < 0.01$). Intriguingly, quarter-designators emerge as the most discriminative tokens: “Fourth” appears with a strong positive association to “Up” calls, whereas “Q1” is significantly tied to “Down” calls. This temporal signal prompted us to fit a logistic-regression model using the quarter (Q1–Q4) as a categorical predictor of stock movement. The regression coefficients as shown in Figure 7confirm that Q4 calls (January–February earnings) increase the probability of a positive 24-hour change by a factor of approximately 1.9 ($p = 0.03$), while Q1 calls (April earnings) decrease those odds (odds ratio -1.3, $p = 0.04$). We hypothesize that end-of-year reporting fosters investor enthusiasm (year-end bonus cycles, fresh budgeting outlooks), whereas first-quarter disclosures often trigger caution as companies reset forecasts.

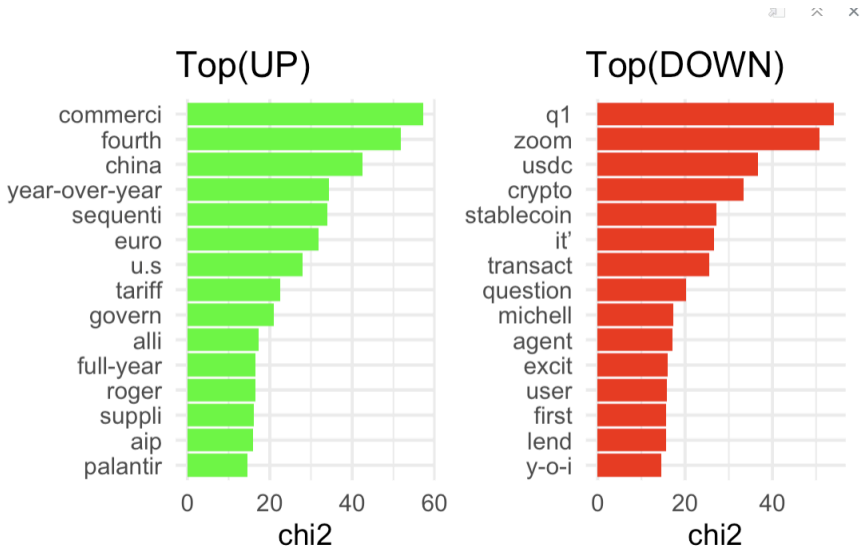

Figure 6: Top keywords by chi-square keyness, showing words most over-represented in each group.

```
Call:
glm(formula = group ~ quarter, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2993     0.6513  -1.995   0.0461 *
quarterQ2     0.7115     0.8575   0.830   0.4067
quarterQ3     1.5870     0.8461   1.876   0.0607 .
quarterQ4     1.8871     0.8575   2.201   0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 76.988  on 55  degrees of freedom
Residual deviance: 70.168  on 52  degrees of freedom
AIC: 78.168

Number of Fisher Scoring iterations: 4
```

Figure 7: logistic regression model based on quarterly timing and the stock price movement

Taken together, our multi analysis demonstrates that (1) extended, confidence-signaling CEO script correlate with upward price movements; (2) TF-IDF–weighted lexical features reliably forecast short-term stock direction, outperforming both word-embedding averages and latent-topic clusters; and

(3) temporal markers—specifically quarter references—carry meaningful predictive weight. These findings highlight practical implications for corporate communicators aiming to shape investor perceptions: emphasizing detail, framing commentary with positive performance metrics, and understanding seasonal investor psychology can materially influence market response.

Discussion

Our study demonstrates that the language used in CEO earnings call has a strong predictive power on stock movement. For RQ1 and RQ2, we conclude that the TF-IDF vectorization method has the beset performance with a 63.3% accuracy.This result significantly shows that the textual features from the CEO script are able to predict stock price movement. By comparing the vectorization method, we are able to see that the importance of words' raw frequency exceeds the semantics of the call. This finding supports our initial hypothesis that textual features alone can signal market reactions. It also highlights the importance of word choice and linguistic framing in earnings calls. The relatively poor performance of semantic embedding may be due to the fact that most transcripts cover similar content. Making it difficult for semantic embedding to distinguish between different stock outcomes. In contrast, the tone and emphasis of the language can highlight more directly, and appear to play a more critical role in influencing investor sentiment. This finding leads us to RQ3, the result of word clouds suggests that a confident and action-oriented language would lead to the rise of the stock price, and a vague or hedged language would lead to a drop in the stock price.

What's unexpected to us is that our logistic regression shows that quarterly timing alone is a strong predictor. Q4 earnings calls (January–February) have a 71% probability of leading a stock increase, while Q1 calls (April) have a 78% probability of leading a stock decrease. The accuracies are even higher than our prediction model. This suggests that investor psychology or seasonal expectations may also play an important role, regardless of the actual language content.

Together, these findings show that both linguistic and timing influence short-term stock behavior. While our data is limited to 56 calls, the patterns we observed may generalize a broader financial communication setting. Especially those where sentiment and framing carry the same weight as the financial metrics.

Limitations and Future Work

The biggest limitation of our project is obvious- small sample size. We only collected 56 calls from the last two years, which is too limited to make a strong prediction about the whole financial kingdom. In addition, we are not proficient in using these methods so there could be many hidden insights that we missed or technical errors that affected our results. To be more specific, we are not able to perform the

feature selection in our TF-IDF method due to the time limit, which may weaken its performance. There could also be some hidden mistake in our code that we are not able to fully detect.

Looking ahead, future work should be sampling a more diverse range and a larger sample size. We only collected 7 tech-based companies so our sample likely suffers from selection bias, which could limit the validity of our conclusions. Furthermore, given our findings about the influence of quarterly timing on stock movement, we believe there is strong potential to build a more powerful predictive model by combining quarterly timing with advanced text vectorization techniques.

To conclude with this paper, this is our first experience conducting a research project using machine learning techniques. Throughout the process, we discovered how exciting and powerful this type of work can be. It sparked our curiosity and gave us a deeper application for data analysis. This project has been an unforgettable learning experience, and we will become better researchers in the future!

References

[1]S. McKay Price a, a, b, AbstractQuarterly earnings conference calls are becoming a more pervasive tool for corporate disclosure. However, E. Amir, V. Bernard, A. Brav, et al. “Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone.” Journal of Banking & Finance, October 19, 2011. https://www.sciencedirect.com/science/article/pii/S0378426611002901?via%3Dihub.

[2] China, Andrew, and Yuyu Fan. “Leveraging Text Mining to Extract Insights ...” www.joim.com. Accessed June 13, 2025. https://www.alliancebernstein.com/content/dam/global/insights/insights-whitepapers/leveragingtextmining toextractinsights-chinfan.pdf.

[3] Ma, Zhiqiang, Grace Bang, Chong Wang, and Xiaomo Liu. “Towards Earnings Call and Stock Price Movement.” arXiv.org, August 23, 2020. https://arxiv.org/abs/2009.01317.

[4] Medya, Sourav, Mohammad Rasoolinejad, Yang Yang, and Brian Uzzi. “An Exploratory Study of Stock Price Movements from Earnings Calls.” arXiv.org, January 31, 2022. https://arxiv.org/abs/2203.12460.

[5] When is a liability not a liability? textual analysis, dictionaries, and 10‐KS - loughran - 2011 - the journal of finance - wiley online library. Accessed June 13, 2025. https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2010.01625.x.

[6]Chambers, Anne E., and Stephen H. Penman. “Timeliness of Reporting and the Stock Price Reaction to Earnings Announcements.” Journal of Accounting Research 22, no. 1 (1984): 21–47. https://doi.org/10.2307/2490706.