A Novel Earnings Call Dataset for Stock Return Prediction

Dilan SriDaran, Maxime Wolf, and Nuobei Zhang

April 28, 2024

Abstract

We introduce a multi-modal, text-audio earnings-call dataset, initially based on 4 companies over a 5-year period. We present the hybrid analytical approach used to process the text and audio data through a scalable pipeline, which we intend to extend to a larger number of transcripts drawn from more diverse companies and over a longer time horizon. The pipeline includes extracting structured features through sentiment analysis, topic modeling, similarity over time, audio measures, and other quantitative metadata extraction tools, which provides a comprehensive basis to understand the key themes within transcripts. When combined with traditional financial indicators, this fills a significant gap in the current modeling landscape and sets the stage for future research that could enhance models for stock price predictions.

1 Introduction

Over the past few decades, there has been substantial research into modeling stock market movements using both statistical and machine learning models [14]. Traditional financial models, employing techniques like AutoRegressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and more recently, Long Short-Term Memory (LSTM) networks, have focused predominantly on historical data and market indicators to predict stock price movements [12, 16, 17]. This study seeks to extend the boundaries of this research by leveraging deep learning advances to harness insights from multimedia data from company earnings call transcripts [13].

Earnings calls, which are structured conference calls involving management of public companies, external analysts, investors, and media, are held following the disclosure of financial results for a specific period—usually quarterly or annually. These calls act as a vital platform for senior management to both provide context for past performance and to discuss future strategic and operational plans. The discussions typically cover a range of topics, spanning financial results, growth projections, risk factors, investments and acquisitions, liabilities, legal matters, share buybacks, dividend policies, and executive leadership changes. Earnings calls consist of two main sections: a presentation (herein "Presentation") of recent financial performance by senior executives ("Company"), which is usually scripted and moderated by an "Operator", and a question and answer ("Q&A") session, which allows for unscripted interactions between the company management and market participants ("External Analysts"). This Q&A session, in particular, provides a more authentic gauge of a company, revealing insights that the scripted portion may not.

Stock markets often show increased volatility and activity prior to earnings announcements due to heightened uncertainty regarding company performance [7]. Accurately predicting the movements of stock prices based on earnings releases can therefore be highly profitable. Given the depth of insight that earnings calls offer into a company's operations

and outlook, we believe that they play a significant role in shaping investor confidence and consequently influencing stock prices. However, although textual analysis has been increasingly integrated into financial models through sentiment analysis of news headlines, financial reports, and social media posts [6], earnings calls have not yet been as systematically exploited. This project aims to fill this gap by developing a predictive model that not only analyzes the textual and audio content of earnings calls but also integrates this information with other structured quantitative data in a multi-modal framework. This hybrid approach aims to offer a more comprehensive understanding of potential stock price movements, thus making a novel contribution to the field of financial predictions.

The remainder of this paper is structured as follows. We discuss the data used in Section 2, the methods employed in Section 3, and preliminary results in Section 4. Noting that this work remains in progress, Section 5 outlines key considerations for future work and development.

2 Data

Our research utilizes data from two primary sources: Yahoo Finance [4], which provides daily stock price data, and the London Stock Exchange Group (LSEG) [3], which supplies the text transcripts (in txt format) and audio recordings (in mp3 format) of earnings calls. We have compiled a dataset consisting of 80 quarterly earnings call transcripts from major technology companies: Apple (AAPL), Google (GOOG), Microsoft (MSFT), and Nvidia (NVDA), with a temporal span of January 2018 to April 2024. On average, each company has 20 transcripts for this period. We segment each text transcript into unique utterances, which we define as uninterrupted segments of speech by the same person within the same section of the call.

We recognize that the dataset of 80 transcripts is not sufficient for a robust statistical analysis. However, the primary objective of this project is to investigate deep learning architectures capable of extracting insights from unstructured data and to integrate these capabilities into a scalable analytical pipeline. The intention for future work involves expanding the application of this pipeline to a larger dataset of earnings calls.

3 Methods

The goal of our pipeline is to create a transcript-level dataset that incorporates both metadata, such as the stock ticker symbol and the date of the earnings call, and features derived from the text and audio data. Certain features, such as the number of words, sentences, questions, and mentioned financial figures, are directly calculated at the transcript level. Other features, including sentiment, similarity metrics, and topics, are initially analyzed at the sentence or utterance level and then aggregated to provide an overall transcript-level perspective.

3.1 Call Attributes

Attributes from earnings calls, such as the duration of the Presentation and the number of questions during Q&A sessions, provide crucial insights into analyst engagement and speaker transparency. We derive seven key call attributes: word count, sentence count, and number-to-word ratio for both the Presentation and Q&A Sections, and a question count for the Q&A section. These are calculated directly at the transcript level.

Word, sentence, and question counts are calculated using basic counting rules of words, periods, and question marks respectively. The number-to-word ratio calculation leverages **spaCy** [10] to identify and tally financial entities.

3.2 Readability

Readability scores reflect how easy text is to read and understand. For each utterance, we determine overall readability scores by taking the simple average of seven readability metrics: Automated Readability Index, Coleman-Liau Index, Dale-Chall Score, Flesch-Kincaid Grade Level Ease, Flesch Ease, Gunning Fog Index, and SMOG Index. These indices focus on two key areas: the use of difficult words and the length of sentences, where lower values indicate that the text is clearer, more concise, and easier to comprehend [5].

From the utterance-level readability scores, four transcript-level features are created, reflecting the average readability of the Operator and Company during the Presentation, and the Company and External Analysts during the Q&A.

3.3 Topic Modeling

We examine two methodologies for extracting topics from earnings calls. The initial method is centered on specific keywords, while the second utilizes **BERTopic** [9].

3.3.1 Keywords

For the keyword-based approach, we select key financial terms including "margin," "cost," "revenue," "earnings," "growth," "debt," "dividend," and "cash flow," drawing inspiration from existing literature [5]. In this method, each sentence is assigned to one or more of these keywords based on their presence in the sentence.

3.3.2 BERTopic

Searching for keywords may ignore general themes or synonyms. To address this limitation, we use a two-pronged topic modeling approach. We first redact entities, such as names, locations, and organizations, using **spaCy**. This ensures that the subsequent topic modeling learns to extract general themes from transcripts, rather than company-specific topics.

We then use **BERTopic**, a technique that leverages transformer-based models to dynamically identify themes based on the transcripts' inherent semantic structures. By applying **BERTopic** to the redacted texts at a sentence level, the analysis can reveal prevalent financial and strategic themes discussed during the earnings calls, providing insights into corporate discourse patterns.

3.4 Sentiment Analysis

We develop a sentiment analysis pipeline using **FinBERT** [11], a transformer model fine-tuned on large financial datasets. This model, sourced from HuggingFace, assesses the sentiment of earnings call transcripts by outputting three distinct sentiment scores: positive, neutral, and negative. Additionally, we calculate a fourth metric, polarity, reflecting the difference between positive and negative scores relative to the total sentiment score.

Our analysis involves processing each sentence through **FinBERT** to capture these sentiment scores. We then performed a multi-level aggregation of these scores to derive a comprehensive sentiment profile for each transcript. This includes calculating the mean, median, standard deviation, minimum, and maximum sentiments across various segments: by speaker (e.g., Company and External Analysts) and by section (e.g., Presentation and Q&A). Sentiment aggregation was also conducted based on specific financial keywords and the 26 topics identified via **BERTopic**. In cases where particular topics were absent in a transcript, a default sentiment score was assigned. This framework generates a robust set of 202 distinct sentiment features, providing a nuanced understanding of the sentiment landscape within each transcript.

3.5 Similarity

We explore the similarity between current earnings call transcripts and their predecessors, focusing exclusively on the Presentation sections delivered by the Company. This decision was based on the premise that consistent messaging across calls could indicate business stability or management alignment. To manage this, we adopted a three-step process, considering that a single transcript might include presentations from up to four different speakers and might exceed the maximum token limit of our chosen model.

First, we summarize each speaker's presentation using **Google Gemma** (**gemma-2b-it** [2]), a model designed for compact yet comprehensive content representation. We concatenate these individual summaries to form a single, cohesive transcript-level summary for each earnings call. Finally, we employ a sentence transformer model (**all-MiniLM-L6-v2** [1]) to calculate similarity scores between the current transcript and up to three of its predecessors. This approach allows us to quantitatively assess the continuity and thematic consistency of corporate communications over time.

3.6 Audio Features

We extend our feature extraction beyond text data, incorporating analysis of audio, which offers additional insights through features like speech tempo, tone, and vocal sentiment.

Considering the resource-intensive nature of audio data processing, for practicality, we segment each earnings call into 15 equal parts and use the **librosa** [15] library for audio processing. This facilitates the transformation of audio files into spectrograms from which we derive multiple acoustic features such as RMS coefficients, spectral centroids, and tempo, extracting a total of 8 distinct features per segment, culminating in 120 features overall.

Advanced models, **Wave2Vec** [18] and **CLAP** [8], were considered for their capabilities in generating embedding vectors and extracting complex features such as sentiment or speaker characteristics. However, generating these features was not computationally feasible given our available resources, and therefore not scalable to a larger number of transcripts. We therefore restrict our current implementation to **librosa**.

4 Results

In this section, we present preliminary findings for topic modeling and sentiment analysis. This includes an evaluation of the accuracy of **FinBERT** for sentiment analysis. The other features are not extensively validated, either because they are derived from deterministic formulae (e.g., call attributes and readability) or have no ground truth against which to compare (e.g., topic modeling, similarity measures, and audio features).

4.1 Topic Modeling

BERTopic identifies a number of commonly themes within the data, some of which are summarized in Table 1. The topics show coverage across the 4 companies considered, demonstrating the effectiveness of **BERTopic** in identifying commonalities across companies. Note, the topics identified are data driven, and therefore future expansion of the pipeline to consider more transcripts will likely change the number and distribution of topics.

4.2 Sentiment Analysis

We manually labeled 500 sentences and assessed the performance of our sentiment analysis using the **FinBERT** model. By comparing the model's predictions against our labeled data, we achieve an accuracy of approximately 87.8%. The results are summarized in Table 2.

Table 1: Notable topics identified by BERTopic.

Theme	**Count**	**Example**
Cloud services	2,263	Azure continues to take share
Gaming	361	Gaming are areas where we've seen some softness
Investments	311	We're well on our way towards meeting the investment projections
Subscriptions	218	Paid subscriptions continued to show very strong growth
Dividends	216	Returned nearly $27 billion to shareholders
Generative AI	208	Generative AI, we have – obviously, we have work going on.
Vehicles	184	We've been investing in self-driving cars.
Cybersecurity	183	As cybersecurity elevates and concern across companies
Supply chain	172	Freight is a huge challenge

Table 2: Confusion matrix for manually labeled sentences.

Actual \ Predicted	**Positive**	**Neutral**	**Negative**
Positive	173	31	8
Neutral	5	140	10
Negative	4	3	126

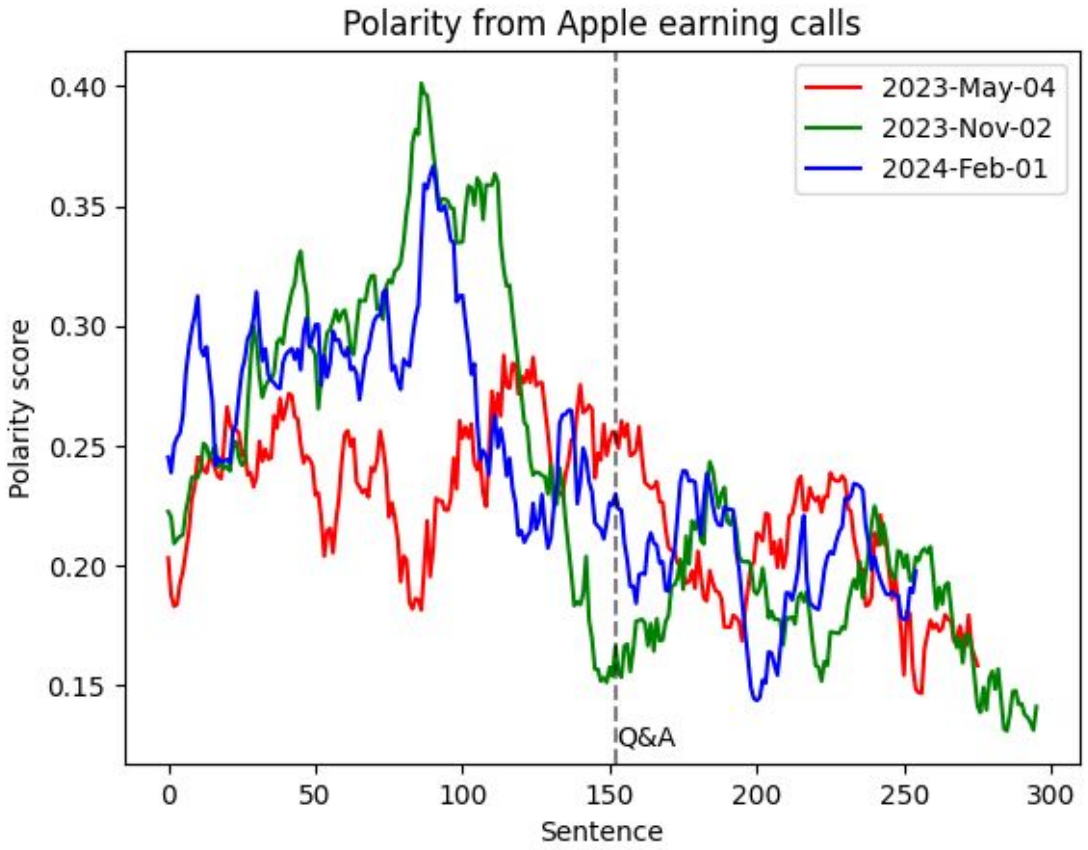

Figure 1: Polarity evolution by time for three Apple calls.

While most classes are predicted with a high degree of accuracy, positive sentences are somewhat frequently misclassified as neutral. However, upon examining the positive and negative scores for these cases, it is evident that the positive scores significantly outweigh the negative scores. As a result, even though these samples are technically "misclassified," they still fall on the correct side of the sentiment spectrum (Figure 3) indicating less of a concern from a modeling standpoint.

Figure 1 illustrates a clear pattern of decreasing polarity from the onset of the Q&A section in three selected Apple calls. This behavior is expected as Company Presentations are usually scripted to overly highlight positive elements of performance, whereas Q&A allows External Analysts to ask more balanced questions related to strengths and weaknesses of the company. The alignment of our results to this known characteristic of calls provides further validation of our approach.

5 Discussion

The work conducted establishes a solid foundation for future research. We have developed a robust and scalable analysis pipeline which we plan to extend to a larger set of transcripts. Initially, our focus will be on expanding our dataset to include a wider range of companies within the technology sector, covering a longer time period than that of this initial study.

With this expanded dataset, we aim to assess the predictive capabilities of the transcripts concerning stock price movements across various time frames (1, 7, and 30 days) after earnings calls. Our models may analyze text and audio data either separately or in conjunction with traditional data types such as historical stock prices and other financial metrics. Figure 2 depicts the intended final output of our analysis, which currently utilizes data from 80 transcripts. Despite the limitations of a small sample size, this demonstrates our primary goal: to predict stock prices accurately and in a way that is understandable. An alternative approach could involve generating embeddings directly from the transcripts using a pre-trained large language model (LLM), and concatenating these with the features derived from our analysis. This final model could potentially offer improved performance by incorporating a broader context from the transcripts, though it may reduce interpretability.

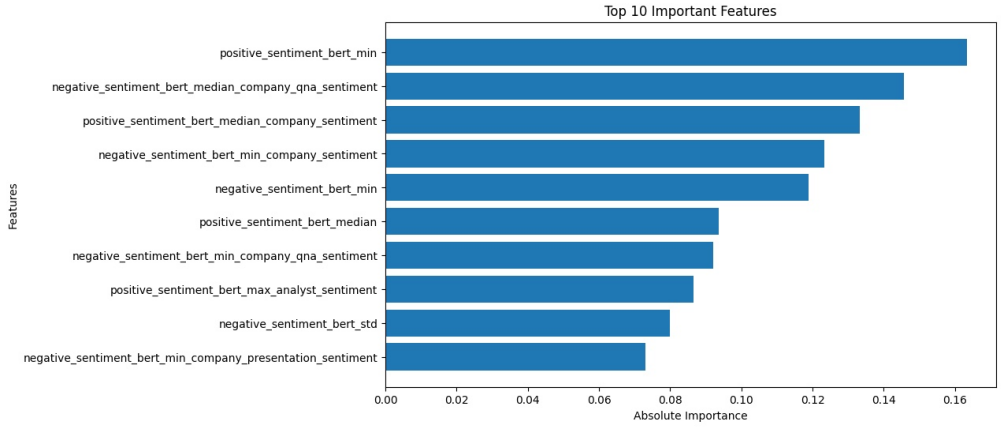

Figure 2: Illustrative example of feature importance plot.

A significant challenge in financial forecasting involves the potential for look-ahead bias, which might occur if the pre-trained language models have been exposed to future data during training. This risk can be mitigated by restricting our analysis to a test set that was collected after the training period of the language models used in our pipeline.

References

[1] all-MiniLM-L6-v2. Hugging Face, 2024. Accessed: 2024-04-01.

[2] Gemma Open Models. Google AI for Developers, 2024. Accessed: 2024-04-01.

[3] LSEG Workspace. London Stock Exchange Group, 2024. Accessed: 2024-04-01.

[4] Yahoo Finance. Yahoo, 2024. Accessed: 2024-04-01.

[5] Andrew Chin and Yuyu Fan. Leveraging text mining to extract insights from earnings call transcripts. JOIM, 2022.

[6] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 2327–2333. AAAI Press, 2015.

[7] W.M. Donders, Monique, Roy Kouwenberg, and C. F. Vorst, Ton. Options and earnings announcements: an empirical study of volatility, trading volume, open interest and liquidity. *European Financial Management*, 6(2):149–171, 2000.

[8] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning Audio Concepts From Natural Language Supervision, 2022.

[9] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[10] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in Python. 2020.

[11] Allen H. Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 2022.

[12] Xingdan Huang, Panlu You, Xiaolian Gao, and Dapeng Cheng. Stock Price Prediction Based on ARIMA-GARCH and LSTM. In *Proceedings of the 2nd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2023)*, pages 438–448. Atlantis Press, 2023.

[13] Jiazheng Li, Barry Smyth, Linyi Yang, and Ruihai Dong. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. page 3063–3070. CIKM, 2020.

[14] Zhiqiang Ma, Grace Bang, Chong Wang, and Xiaomo Liu. Towards earnings call and stock price movement, 2020.

[15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, volume 8, 2015.

[16] Adil Moghar and Mhamed Hamiche. Stock Market Prediction Using LSTM Recurrent Neural Network. *Procedia Computer Science*, 170:1168–1173, 2020.

[17] Ruochen Xiao, Yingying Feng, Lei Yan, and Yihan Ma. Predict stock prices with ARIMA and LSTM, 2022.

[18] Ye Yuan, Guangxu Xun, Qiuling Suo, Ke bin Jia, and Aidong Zhang. Wave2Vec: Learning Deep Representations for Biosignals. *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1159–1164, 2017.

6 Appendix

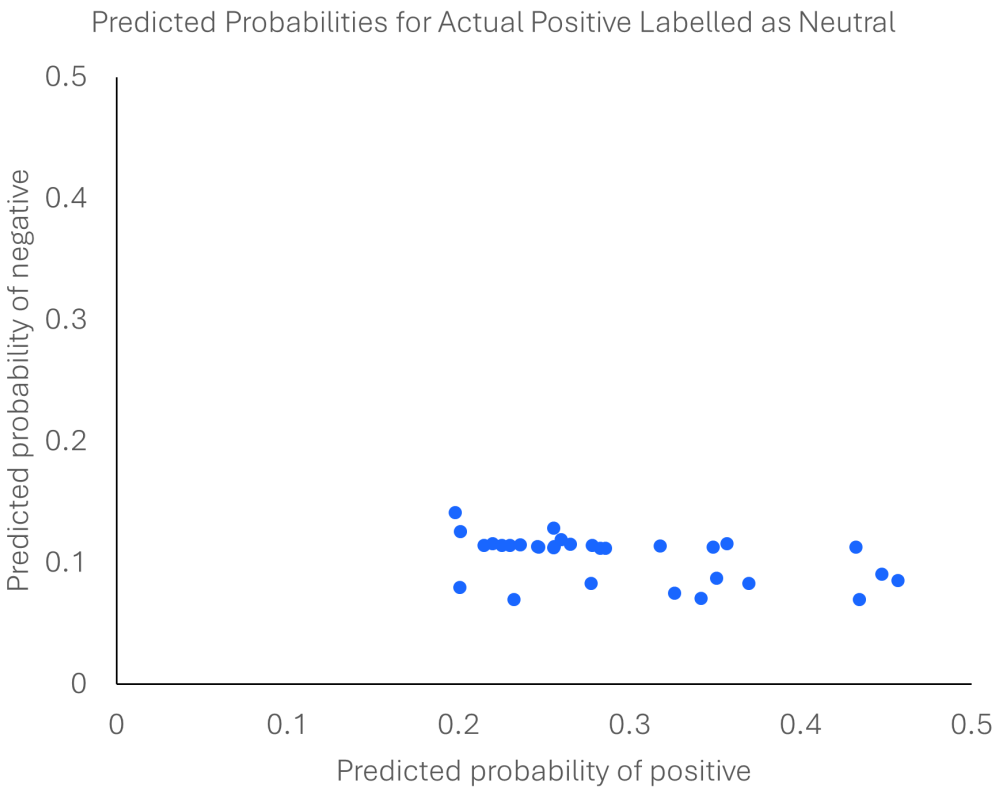

Figure 3: Underlying predicted probabilities for sentences "incorrectly" labelled as "Neutral" rather than "Positive."