Multimodal Multi-Task Financial Risk Forecasting

Ramit Sawhney
ramits.co@nsit.net.in
Netaji Subhas Institute of Technology

Puneet Mathur
puneetm@cs.umd.edu
University of Maryland, College Park

Ayush Mangal
amangal@cs.iitr.ac.in
IIT Roorkee

Piyush Khanna
piyushkhanna_bt2k17@dtu.ac.in
Delhi Technological University

Rajiv Ratn Shah
rajivratn@iiitd.ac.in
MIDAS, IIIT-Delhi

Roger Zimmermann
rogerz@comp.nus.edu.sg
National University of Singapore

ABSTRACT

Stock price movement and volatility prediction aim to predict stocks' future trends to help investors make sound investment decisions and model financial risk. Companies' earnings calls are a rich, underexplored source of multimodal information for financial forecasting. However, existing fintech solutions are not optimized towards harnessing the interplay between the multimodal verbal and vocal cues in earnings calls. In this work, we present a multi-task solution that utilizes domain specialized textual features and audio attentive alignment for predictive financial risk and price modeling. Our method advances existing solutions in two aspects: 1) tailoring a deep multimodal text-audio attention model, 2) optimizing volatility, and price movement prediction in a multi-task ensemble formulation. Through quantitative and qualitative analyses, we show the effectiveness of our deep multimodal approach.

CCS CONCEPTS

• **Social and professional topics** → *Economic impact*; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

multi-task learning; finance; speech processing; stock prediction

ACM Reference Format:
Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal Multi-Task Financial Risk Forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413752

1 INTRODUCTION

Context & Scope: Financial risk modeling is of great interest to capital market participants for making sound investment decisions. Stock volatility and price trends are vital indicators of a company's risk profile, and accurately predicting them has been an extensive area of research in finance [6, 52]. With unparalleled advances in multimodal learning, a massive amount of unstructured data is accessible to investors [42] for financial forecasting. One such rich source of information is the earnings call - a periodic conference held by the executives of publicly listed companies to compile company performance and answer questions raised by analysts [33, 75]. The call is a combination of a verbal presentation followed by a question-answer session aimed at analyzing the claims made by the executives [12, 45]. We analyze earnings calls for two reasons specifically, 1) transcripts and audio recordings are publicly available[1], and 2) they are often associated with high volatility primarily due to the market reaction to the earnings announcement [31]. Post Earnings Announcement Drift (PEAD), a documented phenomenon in financial research, shows significant stock price movements linger towards an earnings surprise several days after the call [4, 41]. Earnings calls often bring additional important information about company sales, investment, amongst other less tangible information. Thus, despite being held quarterly, earnings calls are related to abnormally high returns, even days after the calls [13].

Motivation: There is anecdotal evidence that Chief Executive Officer's (CEO) vocal cues, such as emotions, and voice tones, can be indicative and correlated with a firm's performance. Although existing research has used text for volatility prediction, only a few very recent studies exploit multimodality, especially vocal cues [77]. Multimodal approaches can extract complementary information from multiple modalities to improve financial modeling [53, 77]. Financial tasks, such as predicting a stock's price movement and volatility, are often strongly correlated, thus making multi-task learning a promising modeling choice for financial forecasting. Though homogeneous multi-task approaches [67, 94] have been studied for volatility and price movement prediction, there has been no research on heterogeneous Multi-Task Learning (MTL). Here strongly related tasks on the same feature space differ based on being classification and regression problems [95, 98]. Such cross-task multimodal learning builds on complex price movement prediction through volatility prediction. Despite their applications in finance, solutions for volatility prediction do not only benefit this domain. Rather, they hold merit across all settings in which the effect of newly disclosed language data on public perceptions of risk needs to be quantified. Financial forecasting has proven useful for tasks as manifold as forecasting presidential approval [35] and weather [88], and neuro-muscular activation modeling [27].

Contributions: We introduce a multi-task ensemble architecture **(Sec. 4)** that leverages audio and text for volatility and price movement prediction. We employ domain-specific textual and attention-based alignment for multimodal fusion through a text-audio aligned

[1]Transcripts are publicly available at SeekingAlpha and audio at EarningsCast

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413752

encoder. Through a set of comparative, qualitative, and simulation-based experiments **(Sec. 5)** on real-world S&P 500 index data, we show our model's generalizability across both tasks for both short and long term trading periods in **Sec. 6**. Lastly, we use a trading strategy that generates alpha *(signals to buy or sell)* [7, 19] from vocal, verbal, and financial cues, and show its utility through a market trading simulation, and then discuss future work **(Sec. 7)**.

Ethical Considerations: Examining a CEO's speech and tone in earnings calls is a well studied phenomenon in financial literature [20, 62]. Our work focuses only on calls for which transcripts and audio recordings are publicly released by companies for analysis. The data used in our study corresponds to earnings calls of S&P 500 companies. We acknowledge the presence of gender bias in our study, given the imbalance in the gender ratio of CEOs of S&P 500 companies. We also acknowledge the demographic bias in our study, as the S&P 500 companies are organizations listed in the US.

2 RELATED WORK

2.1 Multimodality in Financial Forecasting

Conventional Approaches: Forecasting stock volatility and price movement are paramount pillars across multiple domains [42]; 1) **theoretical**: quantitative financial models like *Modern Portfolio Theory* [28, 44], *Black-Scholes model* [8], fundamental analysis [22] etc. and 2) **practical**: investment strategies [9], portfolio management [36], and beyond finance [29, 72, 78]. Financial models have previously relied only on numerical features [56, 68] such as macroeconomic indicators [37]. This includes discrete (GARCH [11], rolling regression [71]), continuous (Markov chain [40] & stochastic volatility [2]), and neural approaches [48, 55, 58, 68].

Contemporary approaches: Newer work categorized under Fundamental Analysis [1] based on the Efficient Market Hypothesis [60] highlight the success of multimodal data in finance [53], as they capture a wider set of affecting knowledge and their interdependencies. Recent models used textual data such as social media posts, news reports, web searches, etc. [38, 59, 64, 92, 93]. Methods using images to analyze trends and graphs to use inter stock relations also showed the potency of multimodality for these problems. These approaches did not focus on highly volatile and macro events such as earning calls, where the market microstructure is highly uncertain [39, 79, 89]. Thus, making prediction tasks tough and risk-oriented [2]. Newer studies [77, 94] illustrated the gains obtained by using vocal cues from the CEO's earnings conference calls for volatility prediction. Leveraging deep multimodal neural networks, they better extract the interplay between text and audio features leading to improved performance. These models focus on volatility prediction and set the premise for further study of speech in financial applications. Building on this premise, we extend their work by employing financial features and alignment based fusion over price movement prediction for enhanced cross-task learning.

2.2 Quantitative Trading & Multimodality

Trading strategy design is a high-level task that uses signals (called alpha [16]) from lower-level tasks such as volatility and price movement prediction for making profit [19, 34]. These strategies and their underlying alpha generation tasks have heavily relied on numeric data [15, 21]. Recently, neural networks, have been proposed for long-term stock price prediction for trading [99], including deep q-networks [51]. Through our multimodal approach, we broaden information capture [14] for identifying profit creation opportunities in the market, for strategy design in a multi-task fashion. To the best of our knowledge, our work is the first to use text and audio modalities for alpha generation for strategies.

2.3 Multi-Task Learning in Multimedia

MTL aims to solve multiple related learning tasks simultaneously by using the information generated by the training signals of similar tasks [14]. This learning approach aims to provide better performance than a model trained on only a single task, allowing the learner to gain from inter-task associations and reduce the risks of overfitting. Recent advances in deep learning have led to a notable improvement in using multiple modalities for solving tasks such as emotion recognition [17], and audio-visual speech recognition [66]. Until recently, multimodal MTL has not been explored in the finance except for [94], which concentrates on the homogeneous tasks of stock volatility regression across various durations. We build upon their homogeneous approach to a heterogeneous multi-task [95] ensemble approach with the introduction of stock price classification, owing to the relatedness of financial tasks [67].

3 PROBLEM FORMULATION

We first present the tasks of volatility and price movement prediction and enlist the notations used across the paper. Let $s \in S$ denote a stock, $c \in C$ be an earnings call for s. For each stock s, there exist multiple earnings calls c that are held periodically. Each call c can be segmented into a set of $a_c^i \in A_c$ audio clips, and corresponding $t_c^i \in T_c$ text sentences for $i \in [1, N]$, where N is the maximum number of audio clips in a call. For each stock s, there exists a closing price p_d^s that represents the price of s at the end of the day d.

Formalizing stock volatility: Following [49, 77, 94], we define stock volatility as the natural log of the standard deviation of return prices r in a window of τ days. This form of volatility, known as realized volatility [2, 63] for the period $d-\tau$ to d, is formalized as:

$$v_{[d-\tau,d]} = ln\left(\sqrt{\frac{\sum_{i=0}^{\tau}(r_{d-i} - \bar{r})^2}{\tau}}\right) \tag{1}$$

where $r_i = \frac{p_i - p_{i-1}}{p_{i-1}}$ is the return price on day i for a given stock, and $\bar{r}$ is the average return price over a period of τ days.

Formalizing price movement prediction: Following [93], we define price movement $y_{d-\tau,d}$ over a period of τ days as a binary movement. Here, we use the close price [93] of a given stock that can either rise or fall on day d compared to a previous day $d-\tau$, as:

$$y_{[d-\tau,d]} = \begin{cases} 1, p_d > p_{d-\tau}, \\ 0, p_d \leq p_{d-\tau} \end{cases} \tag{2}$$

Multi-Task Learning objective: Our main objective is to simultaneously predict stock volatility $v_{[d-\tau,d]}$ and price movement $y_{[d-\tau,d]}$ using earnings call data (A_c, T_c) on day $d-\tau$.

Trading strategy for profit generation: To assess the performance of our approach in a real-world scenario; we define a trading strategy [54]. In general, creating profitable strategies is the primary goal of financial forecasting [69]. The designed strategy uses movements from stock movement prediction models to make decisions

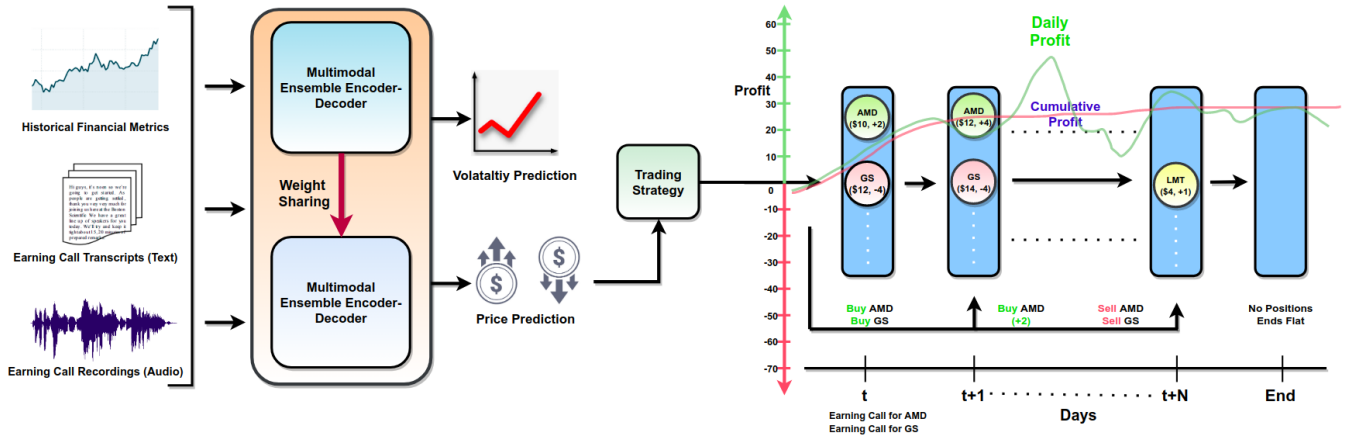


Figure 1: Schematic diagram of our approach. The extracted text, audio and financial features are passed through a multi-task weight sharing ensemble for volatility and price prediction, signals from which are used in a strategy for trading simulation.

on whether to buy or sell stock s. The buy/sell trading decision is based on the price movement over a period of τ days. If the strategy predicts a rise in price $p^s_{d-\tau}$ from day $d-\tau$ to d for stock s, the strategy decides to buy the stock on day $d-\tau$, and then sell it on day d. Otherwise, the strategy speculates a fall in price and performs a short sell.[2] The short sell is a transaction in which the strategy sells a borrowed stock on day $d-\tau$ in anticipation of a price fall; the seller is then required to return the stock on the day d [32].

The profit generated by this trading strategy is then defined as:

$$\text{Profit} = \sum_{s \in S} (p^s_d - p^s_{d-\tau}) * (-1)^{\text{Action}^{d-\tau}_s} \tag{3}$$

where Action^d_s is a binary value; 0 if the strategy buys the stock *i.e., predicts a rise in price* for stock s on day d, otherwise it is 1.

4 METHODOLOGY

4.1 Domain Specialized Textual Pragmatics

Earnings calls are complex sources of naturally-occurring discourse. These are different from other sources like news articles and social media, which have previously been modeled using specialized shallow linguistic features and correlation analyses [70, 86]. To integrate specialized linguistic features for earnings calls modeling, we retrofit word representations based on a vocabulary focusing on the pragmatics of these calls. Such an embedding retrofitting better captures financial context within earnings calls and has shown to be correlated with investment forecasts [46].

Embedding Retrofitting: As shown in the bottom left portion of Fig. 2 we leverage Mittens [24] to enhance pre-trained word embeddings with financial vocabulary. Thereby, updating representations for domain-specific words for improved semantic and structural quality based on pragmatics. Formally, we retrofit GloVe by borrowing financial tokens from lexical corpora R in literature for finance [57, 76, 87]. We develop a joint vocabulary by taking the union of tokens in earning calls and R. Then, for $\hat{w}_i$ in the existing pre-trained embeddings and $r_i \in R$, we optimize $J_{mittens}$ as:

$$J_{mittens} = J + \nu \sum_{i \in R} ||\hat{w}_i - r_i|| \tag{4}$$

where J is the GloVe objective function, and ν is a hyper-parameter controlling the influence of the specialized pragmatic vocabulary.

4.2 Harnessing Vocal Cues

Audio-based features provide prosodic cues related to the affective state of speakers [65]. Capturing the emotional valence of the CEO can alter the understanding of the underlying linguistic utterances in an earnings call [81]. Acoustic features such as pitch, energy, speaking rate have been known to help estimate the degree of deception and sincerity in speech [82]. Features like jitter and shimmer are a good correspondence to variability in the frequency and amplitude of vocal-fold vibration [3]. Patterns of pitch, intensity, and temporal acoustic features perceive confidence levels [73]. Confident expressions are associated with higher mean amplitude and range, whereas unconfident ones are slower in speaking rate with more frequent pauses [43]. Thus, vocal cues can be correlated to one's trustworthiness or persuasiveness [10, 50].

Feature design: We use 26 distinct speech features using the Parselmouth library[3] for extracting features corresponding to pitch, intensity, voice, and harmonicity as shown in the bottom right of Fig. 2. We add more affective state-related features such as APQ (Amplitude Perturbation Quotient) Shimmer, Jitter, Spectral Density Energy, Voiced frames, Voiced to unvoiced ratio to the feature set used by works [77, 94]. The resultant input vector of $\max(N_k) \times f_l$ where $\max(N_k)$ is the maximum length of conference clips and f_l is the feature length ($f_l = 26$, $\max(N_k) = 520$ audio clips, conference calls with less than maximum number of clips were padded).

4.3 Text-Audio Aligned Encoder - Architecture

Often each word-level utterance may not be uniformly informative for the prediction tasks. It is vital to align the acoustic and linguistic utterances to extract fine-grained temporal relationships between text and speech. Text-based audio alignment uses soft attention

[2] Short sell:https://en.wikipedia.org/wiki/Short_(finance)

[3] https://pypi.org/project/praat-parselmouth/

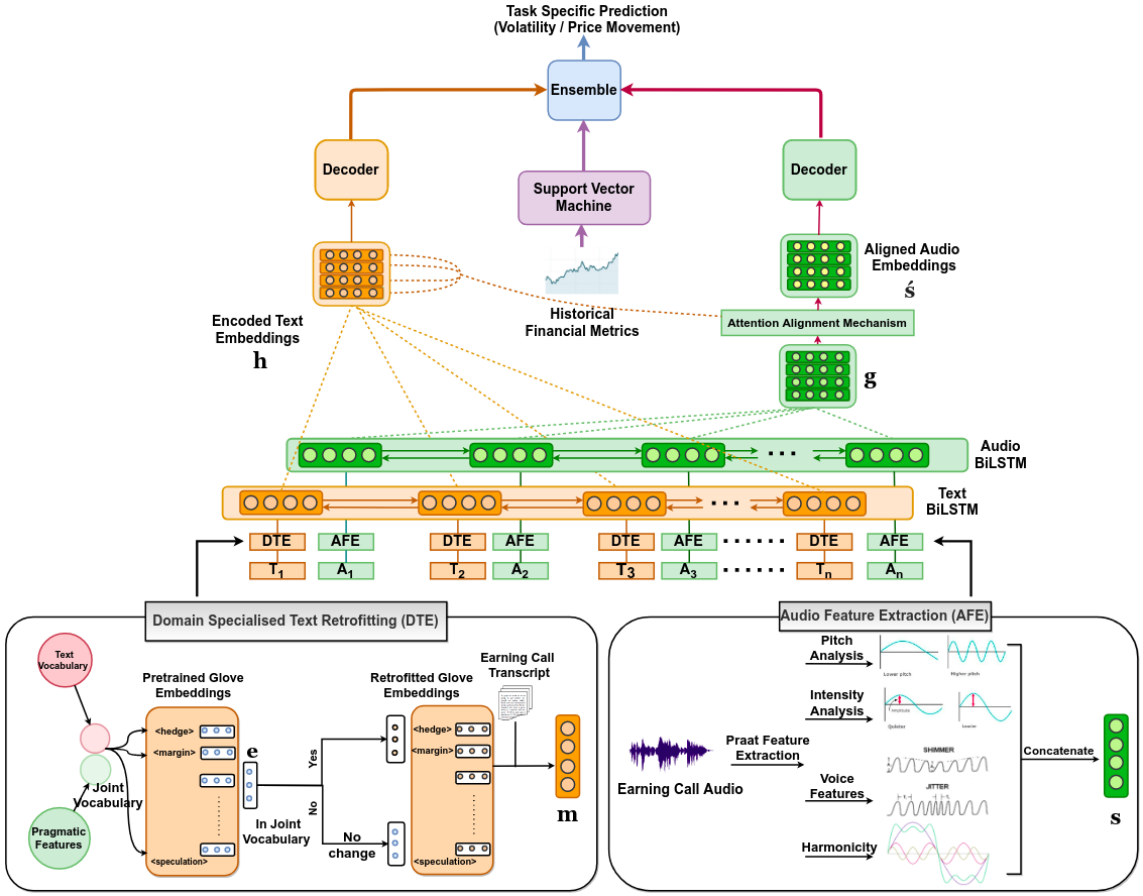


Figure 2: Architecture of the text-audio aligned encoder. We show the individual building blocks a) text retrofitting b) audio feature extraction c) their subsequent alignment, d) financial features, and their combination and final ensemble output.

mechanisms at the sentence level. We then combine the final output of text encoder and aligned audio encoder with financial features as described in the following subsections and as shown in Fig. 2.

4.3.1 Speech and Text Encoders: We represent the sequence of audio representations of the call utterances as $[s_1, \ldots, s_{N_k}]$ and $[e_1, \ldots, e_{N_k}]$ to be the sequence of sentences (represented by average of constituent word embeddings) in the transcript of call c_k. We embed each e_i vector into a vector space illustrated by $m_i = Me_i$, where M is Mittens embedding obtained from Section 4.1. As shown in Fig. 2 we employ BiLSTMs to generate the audio and textual representations individually. Let g_i and h_i represent the generated audio and text encoding of the i^{th} clip of conference call c_k from the BiLSTM encoder as represented by Equations 7 and 8, respectively.

$$\overrightarrow{g_i^{(f)}} = \text{BiLSTM}^{(f)}(m_i, g_{i-1}^{(f)}) \tag{5}$$

$$\overleftarrow{g_i^{(b)}} = \text{BiLSTM}^{(b)}(m_i, g_{i+1}^{(b)}) \tag{6}$$

$$g_i = \left[\overrightarrow{g_i^{(f)}}, \overleftarrow{g_{N_k-i}^{(b)}}\right] \tag{7}$$

$$\text{Similarly, } h_i = \left[\overrightarrow{h_i^{(f)}}, \overleftarrow{h_{N_k-i}^{(b)}}\right] \tag{8}$$

4.3.2 Attention Alignment Mechanism. We propose an attention mechanism that learns the alignment weights between audio features and text sentence sequences simultaneously, as shown in Fig. 2. Consider the i^{th} audio clip feature sequence and j^{th} sentence sequence. The attention weight is calculated between the hidden state g_i of the audio BiLSTM and hidden state h_j of the text BiLSTM at utterance i through Equation 9. Here, u, v and c are trainable parameters. The attention weights thus obtained are normalized using softmax, as shown in Equation 10, which gives the soft alignment between speech and text. Finally, a weighted summation of the aligned speech feature vector is taken with the hidden states of the speech encoding and passed through another BiLSTM layer to give the aligned audio embeddings $\tilde{s}_j$ as shown in Equation 11.

$$\alpha_{i,j} = tanh(u^T g_i + v^T h_j + c) \tag{9}$$

$$a_{i,j} = \text{softmax}(\alpha_{i,j}) \tag{10}$$

$$\tilde{s}_j = \text{BiLSTM}\left(\sum_i a_{i,j} * s_i\right) \tag{11}$$

Attention Decoding: The encoded text h_j and aligned audio $\tilde{s}_j$ are then decoded using task specific decoders as shown in Fig. 2.

For the regression task, we decode the embeddings using dense layers with linear activation ϕ and learnable weights W_1 and W_2 as:

$$\hat{H} = \phi(W_1^T h_j) \quad \text{and} \quad \hat{S} = \phi(W_2^T \tilde{s}_j) \tag{12}$$

where, $\hat{H}$ and $\hat{S}$ are the decoded text and aligned audio outputs. Similarly, we decode text h_j and aligned audio $\tilde{s}_j$ using dense layers with softmax activation for price movement classification.

4.3.3 Ensembling Historical Stock Prices: Literature suggests that ensembles of historical numeric data improve the performance of base learners in financial prediction given that multiple data sources may have varying frequencies [18]. Thus, we employ an ensemble approach, given the different frequencies of data (*price: daily*, and earnings calls: quarterly). Our approach uses the historical prices for the past 30 days preceding the earning call to exploit market data. We use Support Vector Regression (SVR) to predict volatility from historical price, and ensembled it with text $\hat{H}$ and aligned audio $\hat{S}$ outputs of encoders trained for regression. Similarly, we pass price data through a Support Vector Classifier (SVC) to obtain price movement probabilities, and then ensemble them with outputs from the encoders trained for classification, initialized with weights from the encoders trained for regression, as explained next.

4.4 Model Optimization

Given a set of learning tasks where a subset of them are related, Multi-task Learning aims to improve learning over all tasks through either generalization, parameter sharing, or knowledge transfer between all or some of the tasks [14, 26, 90]. MTL for the volatility and price prediction tasks is a case of heterogeneous MTL with different label sets (regression and classification). Such context-sensitive parameter sharing allows multiple tasks to share the common feature extractors augmented by backpropagation of error training. For the regression task, we train the text, aligned audio encoders and the SVR by optimizing the mean squared loss between the actual realized volatility v_k and the predicted volatility $\hat{y}_k$ as:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_k \|v_k - \hat{y}_k\|^2 \tag{13}$$

For the more complex, price movement classification task, we first initialize the weights of the text and aligned audio encoder using the weights of the encoders optimized for the regression task. These text and aligned audio encoders, along with the SVC are trained by optimizing the binary cross entropy loss between the actual price movement x_k and predicted price movement probabilities of the encoders and SVC $\hat{x}_k$ as:

$$\mathcal{L}_{clf} = \sum_k (x_k log(\hat{x}_k) + (1 - x_k)\, log\,(1 - \hat{x}_k)) \tag{14}$$

5 EXPERIMENTS

5.1 Dataset

We used the S&P 500 2017 Earnings Conference Calls dataset [77] for all the experiments. Each conference call is made up of a sequence of sentences partially aligned with their corresponding audio clips as spoken by a company executive during a live broadcast. The dataset comprises 274 unique constituent companies in the S&P-500 2017 index, summing up to a total of 562 earnings calls with several companies holding more than one such call in a financial year. Following prior work, we temporally divide the dataset into train, validation, and a test set in the ratio of 70:10:20 respectively to ensure that future data is not used for training. We use Yahoo Finance[4] for the period 1 January 2017 to 31 December 2017 for historical price data.[5] For classification, the average class distribution across all periods is *BUY:* 52.97%, *SELL:* 47.03%.

5.2 Baselines

We contrast against many modern and traditional baselines across varied domains and modalities across both tasks as follows:

5.2.1 Price-based: These methods use historical price exclusively.

- V_{past} **[25]:** Past volatility is often a strong indicator of future volatility [77]. Hence, we use the volatility of the previous T days, termed as V_{past}, to predict it T days after the call.
- **Price LSTM [47, 97]:** Historical price data often encapsulates patterns that can be exploited through LSTMs.
- **BiLSTM + ATT [85]:** Incorporating attention over a BiLSTM helps to capture a larger context along with focus on a specific period of historical data.

5.2.2 Text-based: The following approaches utilize text data from the call transcripts alone for both tasks.

- **TF-IDF + SVM [23, 91]:** We use Term-Frequency Inverse Document Frequency (TF-IDF) features from call transcripts with Support Vector Regression (SVR) and Classifiers (SVC) for volatility and price movement prediction respectively.
- **Hierarchical Attention Network (HAN) [96]:** HAN employs dual attention layers at the word and sentence level for stratified contextual learning. The Bi-GRU version is used to encode each call transcript as a single document, which is then used for both the prediction tasks.

5.2.3 Comparative baselines: Here, we present contemporary multimodal methods. Speech-based methods are limited in existing literature with MDRM [77], and HTML [94] being the most recent.

- **MDRM [77] [Audio + Text]:** The Multimodal Deep Regression Model (MDRM) utilizes BiLSTM layer to extract context-dependent unimodal features, and then fuses unimodal features together using another layer of BiLSTM to extract multimodal inter-dependencies for the regression task. We experiment with three (text, audio, and multimodal) variants of MDRM with task specific loss functions.
- **Hierarchical Transformer-based Multi-task Learning (HTML) [94] [Audio + Text]:** HTML is the most recent, and a transformer based multi-task architecture using the text and audio data from earnings conference calls. The homogeneous multi-task framework is applied to predict future volatility. We directly apply this for both tasks.
- **bc-LSTM [Audio + Text] [74]:** Contextual feature extractor. Following [77, 94] we use the speech and text models.
- **Multi-Fusion CNN [83] [Audio + Text]:** 1-D CNN-based emotion classification model is adopted for our tasks which exploits multimodal fusion of speech and text features.

[4] https://finance.yahoo.com/

[5] We were unable to map price data for 11 data points, which were then dropped.

5.3 Training Setup and Experiment Settings

Training Setup: Hyper-parameters for our model were tuned on the validation set to find the best configurations. We summarize the range of our model's hyper parameters such as: number of hidden layers $(1, 2, 3)$, size of hidden layers (LSTM, BiLSTM, Dense), embedding size $d \in \{100, 200, 300\}$, dropout $\delta \in \{0.2, 0.3, 0.4, 0.5.0.6\}$, learning rate $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, weight decay $\omega \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, optimizer $\{Adam, Adadelta\}$, batch size $b \in \{16, 32, 64\}$ and epochs (<100). We used grid search with regularization parameter $C \in \{0.001, 0.01, 0.1, 1, 10\}$, kernel coefficient $\gamma \in \{0.001, 0.01, 0.1, 1\}$ and kernel $\in \{rbf, linear, polynomial\}$ to tune the SVMs. To find the best configuration for ensembling, we experimented by varying the ensemble weight assigned to each component with a step-size of 0.01.

Experiment Settings: We experiment with trading periods $\tau \in \{3, 7, 15, 30\}$ days allowing experimentation across both short and medium-term periods. For trading simulations, we assume the traded stock quantity to be 1, and transaction cost to be \$0. We also do not consider intraday trading. We acknowledge that these simplifications hinder the immediate deployability of our model to real-world trading scenarios. Our broad focus through this work is to analyze the correlations between earnings calls and stock prices and to design a neural architecture for financial forecasting.

5.4 Evaluation Metrics

Regression: Following [77, 89, 94], the predicted volatility is compared with the actual volatility to compute the mean squared error for each hold period; $n \in \{3, 7, 15, 30\}$, and $\overline{MSE}$. We also report the coefficient of determination $R^2 = 1 - \frac{MSE_{model}}{MSE_{v_{past}}}$ in Fig. 3.

Classification: We report the F1 score and Mathew's Correlation Coefficient (MCC) for the classification task [61]. We use MCC because unlike F1 score, MCC avoids bias due to data skew as it does not depend on the choice of the positive class and also accounts for the True Negatives. For a given confusion matrix $\begin{pmatrix} tp & fn \\ fp & tn \end{pmatrix}$:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{15}$$

Returns: For real-world application comparisons of price movement prediction, we contrast strategies across profit, and the Sharpe Ratio [84]. The Sharpe Ratio evaluates the performance of investments using their average return rate r_x, risk-free return rate R_f[6] and the standard deviation σ across the investment x, defined as:

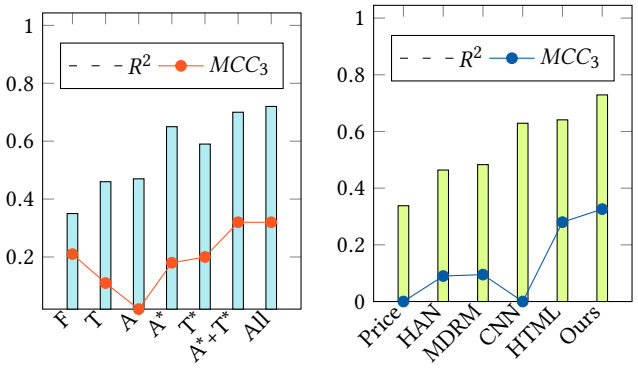
$$\text{Sharpe Ratio} = \frac{r_x - R_f}{\sigma(r_x)} \tag{16}$$

6 RESULTS AND DISCUSSION

We now present an evaluation of our approach through quantitative comparisons, trading simulations, and qualitative case studies.

6.1 Ablation: Impact of Multimodality

The ablation results in Table 1 and Fig. 3a validate the potency of multimodal features over unimodal counterparts, for both tasks, as observed across the financial domain. We observe significant (for n=3, 7, 15: $p < 0.001$ and for n=30: $p < 0.05$ based on Wilcoxon's test) improvements across both text *(T: GloVe → T*: Domain specialized text embeddings)* and audio modalities *(A: Unaligned audio features → A*: Aligned audio features)* individually. This increase shows the effectiveness of synthesized; pragmatic domain-specialized text features over-generalized GloVe embeddings. We attribute gains in the audio modality to the learned alignment between earning call transcripts and audio through attention mechanisms in the temporal domain. This alignment sets the premise for using both text and audio, even in the absence of text transcripts, as text can be inferred through Automatic Speech Recognition systems, and subsequently aligned. The blend of both text(T*) and audio(A*) T^*+A^*, ensembled with financial features lead to improvements across both tasks A^*+T^* *(MCC* : +57.3%, R^2 : +18.12%*)*, *All (MCC* : +59.8%, R^2 : +22.31%*)*, in contrast to the best unimodal model (T*), empirically validating evidence in support of multimodal fusion.



(a) Ablation study over components (b) Comparison study with baselines

Figure 3: Model names on the X-axis are as shown in Table 1. We report the coefficient of determination R^2 for volatility regression, and *MCC* for price movement classification.

6.2 Comparative Analysis

As shown in Fig. 3b and Table 1, our approach achieves the highest performance across both tasks in the short and long term. Specifically for text, other than HTML (WWM-BERT), existing works rely on generalized embeddings, where our domain-specialized embeddings better capture word-level context. Similar to our approach of retrofitting expert based pragmatic financial vocabulary, HTML leverages BERT to develop context, thus achieving similar results along with the text modality. We observe substantial improvements across the comparative state-of-the-art *HTML* models using both audio and text modalities. This gain is likely due to HTML's treatment of audio features hindering alignment-based fusion with text features. We show in the following subsections, the addition of additional audio features, namely, *DDA shimmer*, and *voiced to unvoiced ratio*, also add towards these performance gains. Combining financial features that add the context of market events and historical price trends as opposed to MDRM and HTML, also shows improvements in performance.

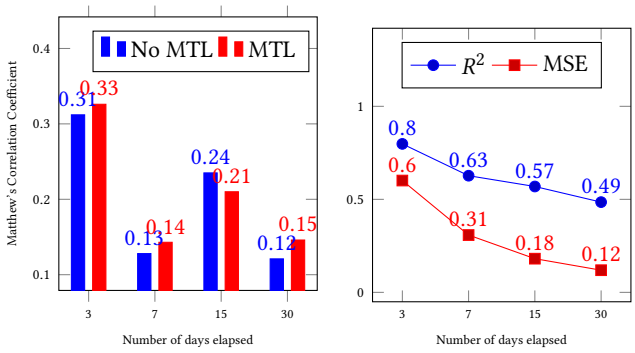
[6]Treasury bills, values taken from https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=billrates

	Model	Volatility Prediction					Price Prediction							
		$\overline{MSE}$	MSE_3	MSE_7	MSE_{15}	MSE_{30}	$F1_3$	$F1_7$	$F1_{15}$	$F1_{30}$	MCC_3	MCC_7	MCC_{15}	MCC_{30}
Baselines	V_{past} [25]	1.116	2.986	0.826	0.420	0.231	-	-	-	-	-	-	-	-
	Price LSTM [47, 97]	0.746	1.970	0.459	0.320	0.235	0.271	0.694	0.200	0.765	0.069	0	0.097	0
	BiLSTM + ATT [85]	0.739	1.983	0.435	0.304	0.233	0.149	0.342	0.200	0.721	0	0	0	0
Comparative	SVM(TF-IDF) [23, 91]	0.696	1.695	0.498	0.342	0.249	0.524	0.683	0.645	0.734	-0.069	0.015	-0.048	-0.003
	HAN(Glove) [96]	0.598	1.426	0.461	0.308	0.198	0.591	0.621	0.598	0.703	0.090	-0.005	0.266	-0.042
	MDRM - Text only [77]	0.600	1.431	0.439	0.309	0.219	0.675	0.500	0.571	0.601	0.117	-0.107	0.032	-0.085
	MDRM - Audio only [77]	0.598	1.412	0.440	0.315	0.224	0.333	0.675	0.000	0.675	0.028	0.050	0.000	-0.001
	bc-LSTM [74]	0.594	1.418	0.436	0.304	0.219	0.538	0.632	0.642	0.708	0.044	0.004	0.119	0.037
	MDRM - Multimodal [77]	0.577	1.371	0.420	0.300	0.217	0.628	0.690	0.452	0.590	0.095	0.056	0.159	-0.065
	Multi-Fusion CNN [83]	0.414	0.732	0.353	0.293	0.276	0.598	0.694	0.214	0.214	0.000	0.000	0.000	0.018
	HTML - Text [94]	0.458	1.175	0.372	***0.153***	0.133	0.623	0.688	0.648	0.700	0.195	0.009	0.119	0.022
	HTML - Multimodal [94]	*0.401*	*0.845*	*0.349*	0.251	*0.158*	*0.696*	*0.695*	***0.703***	*0.748*	*0.280*	*0.126*	*0.196*	*0.131*
Ablation	SVM Financial (F)	0.718	1.880	0.445	0.319	0.229	0.283	0.664	0.325	0.755	0.209	0.000	0.000	0.000
	Audio only (A)	0.592	1.409	0.420	0.317	0.221	0.333	0.675	0.000	0.675	0.028	0.050	0.000	-0.001
	Text only (GloVe) (T)	0.600	1.431	0.439	0.309	0.219	0.602	0.500	0.571	0.601	0.117	-0.107	0.032	-0.085
	Domain specialized text (T*)	0.450	1.156	0.338	0.169	0.137	0.675	0.690	0.636	0.703	0.204	0.008	0.132	0.024
	Attention aligned audio (A*)	0.580	1.417	0.428	0.277	0.197	0.635	0.690	0.534	0.593	0.184	0.056	0.049	0.103
	(A* + T*)	0.331	0.608	0.337	0.190	0.188	0.712*	0.698**	0.712**	0.761*	0.321*	0.128	0.191	0.128
	Ours (A* + T* + F)	**0.302**	**0.601**	**0.308**	0.181	**0.119**	**0.725***	**0.702***	0.726**	**0.771****	**0.326***	**0.133**	**0.210***	**0.146****

Table 1: Results for both tasks. Bold and *italics* represent the best and current state-of-the-art respectively. * and ** indicate statistically significant improvements over HTML - Multimodal with p<0.001, p<0.05 respectively, under Wilcoxon's test.

6.3 Diminishing Performance Gains over Time

Similar to prior work [77], Fig. 4b highlights that short term volatility prediction is complex, given the erratic price fluctuations around earnings calls. These price fluctuations settle over long periods after the earnings calls, as per PEAD [4, 5, 80]. We observe, that PEAD holds to price movement prediction and that over longer durations, the improvements R^2 over simpler baselines *(V_{past})* diminish. We attribute this diminishing performance to the dilution of the cues extracted from earning calls as we "drift" away from them. This finding motivated us to ensemble the financial modality, to balance the diminishing gains over time. Over longer periods, we note the benefit of financial data in Table 1: 36.7% drop in MSE_{30} vs. 1.1% in MSE_3 over A^*+T^* → A^*+T^*+F.



(a) Impact of MTL on MCC **(b) Volatility prediction trend**

Figure 4: Analysis of long and short-term periods after calls

6.4 On Heterogeneous Multi-task Learning

Fig. 4a shows the improvements (for n=3, 7: $p < 0.005$, and for n=30: $p < 0.05$ in MCC) for the classification task, when the text and audio aligned encoders are initialized with the weights of the encoders trained for realized volatility regression. These findings are in line with and expand on the claim supporting the generalizability across heterogeneous related tasks in the financial domain [67]. Such weight sharing based ensemble learning boosts overall performance on classification in comparison to individual isolated learning.

6.5 Trading Simulation: Profitability Test

For a thorough analysis of our approach, we perform a real-world trading simulation for the period of 24^{th} Oct 2017 to 20^{th} Dec 2017 *(test split of the dataset)*. To ensure a fair comparison, we use the strategy described in Section 3 with a hold period $\tau = 3$, which solely relies on how well the model performs in price movement prediction. We use three standard baseline strategies *Buy-all, Short-sell-all, Random* that are commonly used as benchmarks in trading simulations. Table 5 shows the profit earned by comparable strategies in USD ($) and the Sharpe Ratio. We observe that our approach

Strategy	Profit *(USD)*	Sharpe Ratio
Buy-all	$36.59	0.76
Short-sell-all	-$36.59	-0.77
Random	-$50.94	-1.08
Text only (T*)	$60.09	1.25
MRDM - Multimodal	$38.75	0.81
HTML - Multimodal	$72.47	1.52
Ours	**$75.73**	**1.59**

Figure 5: Profit metrics across trading strategies

outperforms all comparative strategies, with HTML being a close second, which is in line with the performance of the underlying models across both tasks. We look further into this simulation by presenting three case studies in the next section.

6.6 Case Studies: A Real-World Outlook

We analyze the decisions taken by different models across three **high risk** situations, marked as ⬠ in Fig. 6. The earnings calls

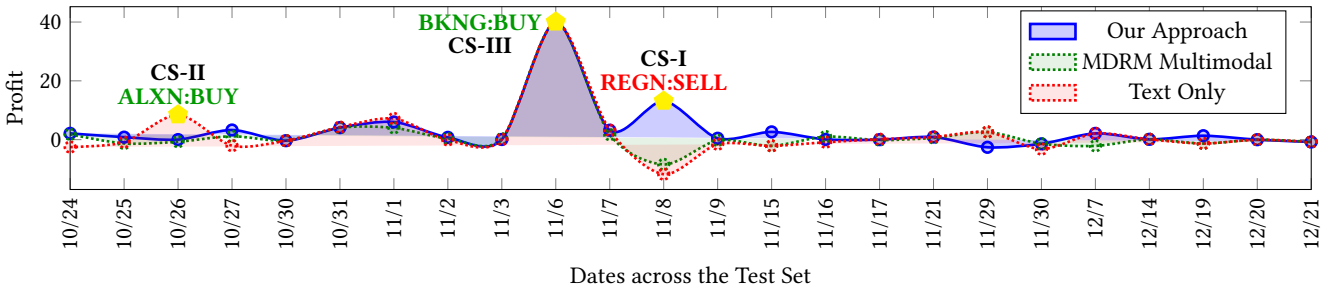
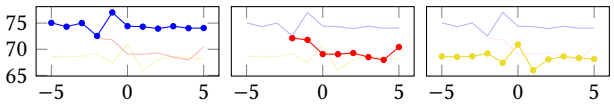
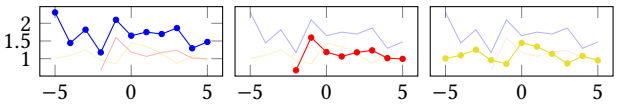


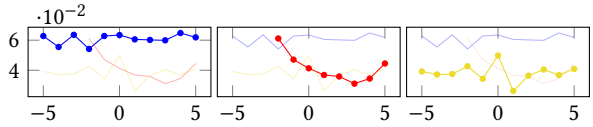
Figure 6: Profit analysis over our multimodal, text-only models and MDRM over the test period



(a) Mean Intensity around target sentence



(b) V/UV ratio around target sentence



(c) SD energy around target sentence

Stock	Sentences
BKNG	*CS-III:* Gross profit for our U.S operations amounted to $372 million, which grew about 11% compared to the prior year.
ALXN	*CS-II:* As we had forecasted and previously communicated, growth this quarter was lower than the previous two quarters due to a few factors.
REGN	*CS-I:* There were no new safety concerns observed with EYLEA in this study.

(d) Target Sentences of stocks under Case Study

Figure 7: Qualitative study of the variations in audio features around sentences of CEO's earnings conference calls

excerpts shown in Fig. 6 (d) are based on our observations.

Case Study I: REGN Earning Call (*Q*3, 2017**)** We study the earnings call for REGN (Regeneron), a pharmaceutical company. REGN's investors are likely to be disappointed on 8^{th} November 2017 following a price drop of 2.5% over the next three days. We look into the audio recording along with the text transcript of the CEO Leonard Schleifer for the third quarter of 2017. Fig. 7 shows vocal cues around the sentence (CS-I) *"There were no new safety concerns observed with EYLEA in this study"* spoken by the CEO. The language is neutral to positive; however, appears inconsistent with the CEO's voice features. Specifically, around this sentence, we see a jump in mean intensity of the CEO's voice by 9.2%, above his average mean intensity. After the earning call, it is revealed that their mid-stage eye disease studies (EYLEA) didn't perform as expected and potentially has long-lasting effects in patients. Existing work in acoustics and psycholinguistics [30, 43], shows this might be indicative of hesitation. We observe similar inconsistencies throughout the call, across other vocal features as well, such as SD Energy and voiced to unvoiced ratio in Fig. 7c and 7b respectively. Based on Fig. 6, we see that both MDRM and text-only models incorrectly predict a price increase. Our model likely captures such correlations between stock prices and these subtle cues through audio features.

Case Study II: ALXN Earning Call (*Q*3, 2017**)** We examine Alexion Pharmaceuticals's Q3 call as error analysis, where both MDRM and our approach make a wrong prediction. Here the text-only model gains a profit by correctly predicting a decrease in the price of ALXN by 6.84%. We analyze the earning call across audio features to identify likely deviations between the CEO's vocal and verbal cues. A specific instance being the fluctuations in audio features, as seen in Fig. 7, when the CEO is heard, saying sentence *CS-II* in Fig. 7d. We observe the language indicates a subpar performance, as correctly predicted by the text-only model *see CS-II; Fig. 6.* However, models that employ audio features, MDRM and our multimodal model incorrectly predict the stock's trend. We attribute this to potential overfitting, or noise in terms of audio features, and this case presents an interesting case to analyze in future work.

Case Study III: BKNG Earning Call (*Q*3, 2017**)** The third quarter *July - Sept 2017* of 2017 was a stellar period for Booking Holdings Inc. The great performance led to an increase of $42.03 in its price. The earning call is indicative of positive company performance. We dive deeper into the vocal cues of the CEO, by analyzing the variation shown in Fig. 7, specifically *CS-III* shown in fig 7d. We observe that the vocal cues (Fig. 7(a, d) are *relatively* stable, likely correlating with confidence, and agreement between the CEO's vocal and verbal cues. Here, we observe that all three strategies correctly predict a price increase, as shown in Fig. 6.

7 CONCLUSION AND FUTURE WORK

In this paper, we formulated a heterogeneous multi-task approach for stock prediction comprising two related financial tasks: price movement classification, and volatility regression. Following modern economic theories and contemporary work, our approach highlighted the potential of leveraging the CEO's vocal and verbal cues in company earnings calls for financial tasks. The core of the method is a neural attentive alignment model, focusing on interdependencies across vocal and verbal modalities. Experimental results on the publicly available earnings calls of S&P 500 companies showed the effectiveness of our proposed multimodal multi-task solution. Our future work includes incorporating additional constraints for trading, such as transaction costs, and intraday transactions.

REFERENCES

[1] Jeffrey S Abarbanell and Brian J Bushee. 1997. Fundamental analysis, future earnings, and stock prices. *Journal of accounting research* 35, 1 (1997), 1–24.

[2] Leif Andersen. 2008. Simple and efficient simulation of the Heston stochastic volatility model. *Journal of Computational Finance* 11, 3 (2008), 1–43.

[3] Jo-Anne Bachorowski. 1999. Vocal expression and perception of emotion. *Current directions in psychological science* 8, 2 (1999), 53–57.

[4] Victor L Bernard and Jacob K Thomas. 1989. Post-earnings-announcement drift: delayed price response or risk premium? *Journal of Accounting research* 27 (1989), 1–36.

[5] Ravi Bhushan. 1994. An informational efficiency perspective on the post-earnings announcement drift. *Journal of Accounting and Economics* 18, 1 (1994), 45–65.

[6] Jędrzej Białkowski, Katrin Gottschalk, and Tomasz Piotr Wisniewski. 2008. Stock market volatility around national elections. *Journal of Banking & Finance* 32, 9 (2008), 1941–1953.

[7] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* 22, 7 (2000), 719–725.

[8] Fischer Black and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of political economy* 81, 3 (1973), 637–654.

[9] David C Blitz and Pim Van Vliet. 2007. The volatility effect. *The Journal of Portfolio Management* 34, 1 (2007), 102–113.

[10] Bibi Boehme. 2014. How trustworthy is your voice? The effects of voice manipulation on the perceived trustworthiness of novel speakers. (2014).

[11] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31, 3 (1986), 307–327.

[12] Robert M Bowen, Angela K Davis, and Dawn A Matsumoto. 2002. Do conference calls affect analysts' forecasts? *The Accounting Review* 77, 2 (2002), 285–316.

[13] Michael W Brandt, Runeet Kishore, Pedro Santa-Clara, and Mohan Venkatachalam. 2008. Earnings announcements are full of surprises. *SSRN eLibrary* (2008).

[14] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[15] Spyros K Chandrinos, Georgios Sakkas, and Nikos D Lagaros. 2018. AIRMS: A risk management tool using machine learning. *Expert Systems with Applications* 105 (2018), 34–48.

[16] Godfrey Charles-Cadogan. 2011. Alpha Representation For Active Portfolio Management and High Frequency Trading In Seemingly Efficient Markets. *JSM Proceedings, Business and Economic Statistics Section* (2011), 673–687.

[17] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-Task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, California, USA) *(AVEC '17)*. Association for Computing Machinery, New York, NY, USA, 19–26. https://doi.org/10.1145/3133944.3133949

[18] Cheng Cheng, Wei Xu, and Jiajia Wang. 2012. A Comparison of Ensemble Methods in Financial Market Prediction. In *Proceedings of the 2012 Fifth International Joint Conference on Computational Sciences and Optimization (CSO '12)*. IEEE Computer Society, USA, 755–759. https://doi.org/10.1109/CSO.2012.171

[19] C Cooper and Double-Digit Numerics. 2010. Alpha Generation and Risk Smoothing Using Volatility of Volatility. *Risk Professional* (2010).

[20] Belinda Crawford Camiciottoli. 2011. Ethics and ethos in financial reporting: Analyzing persuasive language in earnings calls. *Business Communication Quarterly* 74, 3 (2011), 298–312.

[21] Rajashree Dash and Pradipta Kishore Dash. 2016. A hybrid stock trading framework integrating technical analysis with machine learning techniques. *The Journal of Finance and Data Science* 2, 1 (2016), 42–57.

[22] Ilia D Dichev and Vicki Wei Tang. 2009. Earnings volatility and earnings predictability. *Journal of accounting and Economics* 47, 1-2 (2009), 160–181.

[23] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1415–1425. https://doi.org/10.3115/v1/D14-1148

[24] Nicholas Dingwall and Christopher Potts. 2018. Mittens: an Extension of GloVe for Learning Domain-Specialized Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 212–217. https://doi.org/10.18653/v1/N18-2034

[25] Ning Du and David V Budescu. 2007. Does past volatility affect investors' price forecasts and confidence judgements? *International Journal of Forecasting* 23, 3 (2007), 497–511.

[26] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, 845–850. https://doi.org/10.3115/v1/P15-2139

[27] Graham Elliott and Allan Timmermann. 2013. *Handbook of economic forecasting*. Elsevier.

[28] Edwin J Elton, Martin J Gruber, Stephen J Brown, and William N Goetzmann. 2009. *Modern portfolio theory and investment analysis*. John Wiley & Sons.

[29] Claude B Erb, Campbell R Harvey, and Tadas E Viskanta. 1994. Forecasting international equity correlations. *Financial analysts journal* 50, 6 (1994), 32–45.

[30] Karyn Fish, Kathrin Rothermich, and Marc D Pell. 2017. The sound of (in) sincerity. *Journal of Pragmatics* 121 (2017), 147–161.

[31] George Foster, Chris Olsen, and Terry Shevlin. 1984. Earnings releases, anomalies, and the behavior of security returns. *Accounting Review* (1984), 574–603.

[32] Merritt B Fox, Lawrence R Glosten, and Paul C Tetlock. 2009. Short selling and the news: a preliminary report on empirical study. *NYL Sch. L. Rev.* 54 (2009), 645.

[33] Richard M Frankel, Jared N Jennings, and Joshua A Lee. 2017. Using Natural Language Processing to Assess Text Usefulness to Readers: The Case of Conference Calls and Earnings Prediction. *Available at SSRN 3095754* (2017).

[34] Guido Giese. 2012. Optimal design of volatility-driven algo-alpha trading strategies. *The Journal of Risk* 1, 1 (2012), 34.

[35] Paul Gronke and John Brehm. 2002. History, heterogeneity, and presidential approval: a modified ARCH approach. *Electoral Studies* 21, 3 (2002), 425–452.

[36] Alexandre Hocquard, Sunny Ng, and Nicolas Papageorgiou. 2013. A constant-volatility framework for managing tail risk. *The Journal of Portfolio Management* 39, 2 (2013), 28–40.

[37] Ehsan Hoseinzade, Saman Haratizadeh, and Arash Khoeini. 2019. U-CNNpred: A Universal CNN-based Predictor for Stock Markets. arXiv:cs.LG/1911.12540

[38] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 261–269. https://doi.org/10.1145/3159652.3159690

[39] Dušan Isakov and Christophe Perignon. 2001. Evolution of market uncertainty around earnings announcements. *Journal of banking & finance* 25, 9 (2001), 1769–1788.

[40] Eric Jacquier, Nicholas G Polson, and Peter E Rossi. 2002. Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics* 20, 1 (2002), 69–87.

[41] Narasimhan Jegadeesh and Joshua Livnat. 2006. Post-earnings-announcement drift: The role of revenue surprises. *Financial Analysts Journal* 62, 2 (2006), 22–34.

[42] Weiwei Jiang. 2020. Applications of deep learning in stock market prediction: recent progress. *arXiv preprint arXiv:2003.01859* (2020).

[43] Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication* 88 (2017), 106–126.

[44] C Kenneth Jones. 2017. Modern Portfolio Theory, Digital Portfolio Theory and Intertemporal Portfolio Choice. *American Journal of Industrial and Business Management* 7 (2017), 833–854.

[45] Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33 (2014), 171–185.

[46] Katherine Keith and Amanda Stent. 2019. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 493–503. https://doi.org/10.18653/v1/P19-1047

[47] Ha Young Kim and Chang Hyun Won. 2018. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications* 103 (2018), 25–37.

[48] Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction. *arXiv preprint arXiv:1908.07999* (2019).

[49] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 272–280.

[50] David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50, 2 (2012), 495–540.

[51] Jinho Lee, Raehyun Kim, Yookyung Koh, and Jaewoo Kang. 2019. Global stock market prediction based on stock chart images using deep Q-network. *IEEE Access* 7 (2019), 167260–167277.

[52] Katharina Lewellen. 2006. Financing decisions when managers are risk averse. *Journal of Financial Economics* 82, 3 (2006), 551–589.

[53] Qing Li, Jinghua Tan, Jun Wang, and HsinChun Chen. 2020. A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[54] Yelin Li, Junjie Wu, and Hui Bu. 2016. When quantitative trading meets machine learning: A pilot survey. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE, 1–6.

[55] Yu-Fei Lin, Tzu-Ming Huang, Wei-Ho Chung, and Yeong-Luh Ueng. 2020. Forecasting Fluctuations in the Financial Index Using a Recurrent Neural Network Based on Price Features. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2020).

[56] Jiexi Liu and Songcan Chen. 2019. Non-stationary Multivariate Time Series Prediction with Selective Recurrent Neural Networks. In *PRICAI 2019: Trends in Artificial Intelligence*, Abhaya C. Nayak and Alok Sharma (Eds.). Springer International Publishing, Cham, 636–649.

[57] Tim Loughran and Bill Mcdonald. 2011. When Is a Liability NOT a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66 (02 2011), 35 – 65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

[58] Rui Luo, Weinan Zhang, Xiaojun Xu, and Jun Wang. 2018. A neural stochastic volatility model. In *Thirty-second AAAI conference on artificial intelligence*.

[59] Eduardo Jabbur Machado and Adriano César Machado Pereira. 2018. Proposal and Implementation of Machine Learning Models for Stock Markets Using Web Data. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (Salvador, BA, Brazil) *(WebMedia '18)*. Association for Computing Machinery, New York, NY, USA, 61–64. https://doi.org/10.1145/3243082.3264663

[60] Burton G Malkiel. 2003. The efficient market hypothesis and its critics. *Journal of economic perspectives* 17, 1 (2003), 59–82.

[61] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[62] William J Mayew and Mohan Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *The Journal of Finance* 67, 1 (2012), 1–43.

[63] Michael McAleer and Marcelo C Medeiros. 2008. Realized volatility: A review. *Econometric Reviews* 27, 1-3 (2008), 10–45.

[64] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C Anastasiu. 2019. Stock Price Prediction Using News Sentiment Analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 205–208.

[65] Claude Montacié and Marie-José Caraty. 2018. Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge.. In *Interspeech*. 541–545.

[66] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2130–2134.

[67] Thi-Thu Nguyen and Seokhoon Yoon. 2019. A Novel Approach to Short-Term Stock Price Movement Prediction using Transfer Learning. *Applied Sciences* 9, 22 (2019), 4745.

[68] Mahla Nikou, Gholamreza Mansourfar, and Jamshid Bagherzadeh. 2019. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management* 26 (12 2019). https://doi.org/10.1002/isaf.1459

[69] Giuseppe Nuti, Mahnoosh Mirghaemi, Philip Treleaven, and Chaiyakorn Yingsaeree. 2011. Algorithmic trading. *Computer* 44, 11 (2011), 61–69.

[70] Dan Palmon, Ke Xu, and Ari Yezegel. 2016. What Does' But'Really Mean?–Evidence from Managers' Answers to Analysts' Questions During Conference Calls. In *Evidence from Managers' Answers to Analysts' Questions During Conference Calls (August 29, 2016)*.

[71] Yaohao Peng, Pedro Henrique Melo Albuquerque, Jader Martins Camboim de Sá, Ana Julia Akaishi Padula, and Mariana Rosa Montenegro. 2018. The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression. *Expert Systems with Applications* 97 (2018), 177–192.

[72] Gabriel Perez-Quiros and Allan Timmermann. 2000. Firm size and cyclical variations in stock returns. *The Journal of Finance* 55, 3 (2000), 1229–1262.

[73] Heather Pon-Barry. 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In *Ninth Annual Conference of the International Speech Communication Association*.

[74] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 873–883. https://doi.org/10.18653/v1/P17-1081

[75] S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance* 36, 4 (2012), 992–1011.

[76] Anna Prokofieva and Julia Hirschberg. 2014. Hedging and Speaker Commitment.

[77] Yu Qin and Yi Yang. 2019. What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 390–401. https://doi.org/10.18653/v1/P19-1038

[78] Robert Rich and Joseph Tracy. 2004. Uncertainty and labor contract durations. *Review of Economics and Statistics* 86, 1 (2004), 270–287.

[79] Jonathan L Rogers, Douglas J Skinner, and Andrew Van Buskirk. 2009. Earnings guidance and market uncertainty. *Journal of Accounting and Economics* 48, 1 (2009), 90–109.

[80] Ronnie Sadka. 2006. Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk. *Journal of Financial Economics* 80, 2 (2006), 309–349.

[81] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan Gielen. 2001. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis, Vol. 1. 87–90.

[82] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, Keelan Evanini, et al. 2016. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION (INTERSPEECH 2016), VOLS 1-5*. 2001–2005.

[83] Jilt Sebastian and Piero Pierucci. 2019. Fusion techniques for utterance-level emotion recognition combining speech and transcripts. In *Proc. Interspeech*. 51–55.

[84] William F Sharpe. 1994. The sharpe ratio. *Journal of portfolio management* 21, 1 (1994), 49–58.

[85] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. A Comparative Analysis of Forecasting Financial Time Series Using ARIMA, LSTM, and BiLSTM. arXiv:cs.LG/1911.09512

[86] Kate Suslava. 2017. 'Stiff Business Headwinds and Unchartered Economic Waters': The Use of Euphemisms in Earnings Conference Calls. *Available at SSRN 2876819* (2017).

[87] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37 (06 2011), 267–307. https://doi.org/10.1162/COLI_a_00049

[88] James W Taylor and Roberto Buizza. 2002. Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power systems* 17, 3 (2002), 626–632.

[89] Kilian Theil, Samuel Broscheit, and Heiner Stuckenschmidt. 2019. PRoFET: Predicting the Risk of Firms from Event Transcripts. 5211–5217. https://doi.org/10.24963/ijcai.2019/724

[90] Sebastian Thrun and Lorien Pratt (Eds.). 1998. *Learning to Learn*. Kluwer Academic Publishers, USA.

[91] Ming-Feng Tsai and Chuan-Ju Wang. 2014. Financial Keyword Expansion via Continuous Word Vector Representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1453–1458. https://doi.org/10.3115/v1/D14-1152

[92] Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review* 50, 1 (2018), 49–73.

[93] Yumo Xu and Shay B. Cohen. 2018. Stock Movement Prediction from Tweets and Historical Prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1970–1979. https://doi.org/10.18653/v1/P18-1183

[94] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. HTML: Hierarchical Transformer-Based Multi-Task Learning for Volatility Prediction. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 441–451. https://doi.org/10.1145/3366423.3380128

[95] Xiaolin Yang, Seyoung Kim, and Eric P. Xing. 2009. Heterogeneous Multitask Learning with Joint Sparsity Constraints. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'09)*. Curran Associates Inc., Red Hook, NY, USA, 2151–2159.

[96] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 1480–1489. https://doi.org/10.18653/v1/N16-1174

[97] ShuiLing Yu and Zhe Li. 2018. Forecasting stock price index volatility with LSTM deep neural network. In *Recent Developments in Data Science and Business Analytics*. Springer, 265–272.

[98] Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* (2017).

[99] Feng Zhou, Hao-min Zhou, Zhihua Yang, and Lihua Yang. 2018. EMD2FNN: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. *Expert Systems with Applications* 115 (07 2018). https://doi.org/10.1016/j.eswa.2018.07.065