Introduction

It is the very nature of the stock market that even the most modest of advantages (e.g. speed of trade) can be parlayed into significant financial rewards, and thus traders have long been attracted to the idea of using historical data to predict future stock market trends. However, the stochastic nature of the stock market has proved to be very challenging when it comes to provide accurate future forecasts, especially when it comes to optimize a real-world trading strategy, is the predicted stock return. While this has been explored by using financial news data (Ding et al. 2014, 2015; Chang et al. 2016; Duan et al. 2018; Yang et al. 2019; Ye, Qin, and Xu 2020), to predict the subsequent stock price fluctuation over a specified period (e.g., three days or seven days) after the earnings call (Bollerslev, Patton, and Quaedvlieg 2016; Rekabsaz et al. 2017). An even more challenging, yet potentially more valuable signal, especially when it comes to optimizing a real-world trading strategy, is the predicted stock return. While this has been explored by using financial news data (Ding et al. 2014, 2015; Chang et al. 2016; Duan et al. 2018; Yang et al. 2019; Du and Tanaka-Ishii 2020) and analyst reports (Kogan et al. 2009; Loughran and McDonald 2011; Chen, Huang, and Chen 2021b), the use of earnings call data remains largely unexplored. One notable exception is Sawhney et al. (2020), which considers stock movement prediction as an auxiliary task for enhancing the financial risk predictions. In particular, they show that multi-task learning is useful for improving volatility prediction at the expense of lower price movement prediction accuracy.

The main objective of this work is to explore the use of earnings calls data for stock movement prediction. The starting point for this work is the multi-task learning approach described by (Yang et al. 2020), which leverages textual and audio earnings call data with additional vocal features extracted by Praat (Boersma and Van Heuven 2001). We quantify the effectiveness of textual and vocal information from earnings calls to predict stock movement using this baseline reports (Duan et al. 2018; Kogan et al. 2009) to real-time social media (Xu and Cohen 2018; Fung et al. 2018) — which may lead to be effective forecasting predictions. Of particular relevance to this paper is the earnings call data (Kimbrough 2005; Wang and Hua 2014; Qin and Yang 2019) that typically accompany the (quarterly) earnings reports of publicly traded companies. The multi-modal data associated with these reports include the textual data of the report itself plus the audio of the so-called earnings call where the report is presented to relevant parties, including a question-and-answer session with relevant company executives. The intuition is that the content of such a report and the nature of the presentation and Q&A may encode valuable information to determine how a company may perform in the coming quarter and, more immediately relevant, how the market will respond to the earnings report.

Previous work on using earnings conference calls has mostly considered the volatility prediction (Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020; Ye, Qin, and Xu 2020), to predict the subsequent stock price fluctuation over a specified period (e.g., three days or seven days) after the earnings call (Bollerslev, Patton, and Quaedvlieg 2016; Rekabsaz et al. 2017). An even more challenging, yet potentially more valuable signal, especially when it comes to optimizing a real-world trading strategy, is the predicted stock return. While this has been explored by using financial news data (Ding et al. 2014, 2015; Chang et al. 2016; Duan et al. 2018; Yang et al. 2019; Du and Tanaka-Ishii 2020) and analyst reports (Kogan et al. 2009; Loughran and McDonald 2011; Chen, Huang, and Chen 2021b), the use of earnings call data remains largely unexplored. One notable exception is Sawhney et al. (2020), which considers stock movement prediction as an auxiliary task for enhancing the financial risk predictions. In particular, they show that multi-task learning is useful for improving volatility prediction at the expense of lower price movement prediction accuracy.

The main objective of this work is to explore the use of earnings calls data for stock movement prediction. The starting point for this work is the multi-task learning approach described by (Yang et al. 2020), which leverages textual and audio earnings call data with additional vocal features extracted by Praat (Boersma and Van Heuven 2001). We quantify the effectiveness of textual and vocal information from earnings calls to predict stock movement using this baseline reports (Duan et al. 2018; Kogan et al. 2009) to real-time social media (Xu and Cohen 2018; Fung et al. 2018) — which may lead to be effective forecasting predictions. Of particular relevance to this paper is the earnings call data (Kimbrough 2005; Wang and Hua 2014; Qin and Yang 2019) that typically accompany the (quarterly) earnings reports of publicly traded companies. The multi-modal data associated with these reports include the textual data of the report itself plus the audio of the so-called earnings call where the report is presented to relevant parties, including a question-and-answer session with relevant company executives. The intuition is that the content of such a report and the nature of the presentation and Q&A may encode valuable information to determine how a company may perform in the coming quarter and, more immediately relevant, how the market will respond to the earnings report.
model and then go on to extend this model (Yang et al. 2020) by describing three auxiliary loss functions to investigate the utility of a more sophisticated representation of numerical data during prediction. Thus the central advance in this work is a novel, numeric-oriented hierarchical transformer model (*NumHTML*) for the prediction of stock returns, using multi-modal aligned earnings calls data by taking advantage of the auxiliary task (volatility prediction), different categories of numerical data (monetary, temporal, percentages etc.), and their magnitude, motivated by the fact that volatility is a relevant factor to future stock trends and the assumption that better numerical understanding can benefit forecasting. These components are integrated through a novel structured adaptive pre-training strategy and Pareto Multi-task Learning.

We present the results of a comprehensive evaluation on a real-world earnings call dataset to show how the model can be more effective than the baseline system, facilitating more accurate stock returns predictions without compromising volatility prediction (Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020). Also, the results of a realistic trading simulation show how our approach can generate a significant arbitrage profit using a real-world trading dataset. All code and datasets will be released on GitHub.

**Related Work**

This paper brings together several areas of related work – Stock Movement Predictions, Multi-modal Aligned Earnings Call, and Representing Numbers in Language Models – and in what follows, we briefly summarise the relevant state-of-the-art in each of these areas as it relates to our approach. We are the first to examine whether pre-trained models with better numerical understanding can improve performance on financial forecasting tasks based on the multi-modal data.

**Stock Movement Prediction**

While there has been a long-standing effort when it comes to applying machine learning techniques to financial prediction (Du, Engelberg, and Gao 2015; Xing, Cambria, and Welsh 2018; Xing, Cambria, and Zhang 2019), typically by using time-series pricing data, reliable and robust predictions have proven to be challenging due to the stochastic nature of stock markets. However, recent work has shown some promise when it comes to predicting stock price movements using deep neural networks with rich textual information from financial news and social media (primarily Twitter) (Liu and Tse 2013; Ding et al. 2014, 2015; Xu and Cohen 2018; Duan et al. 2018; Yang et al. 2018; Feng et al. 2018). By taking advantage of much richer sources of relevant data (news reports, export commentaries etc.), deep learning techniques have been able to generate more robust and accurate predictions even in the face of market volatility.

Elsewhere, researchers have considered the role of opinions in financial prediction. For example, one recent study (Chen, Huang, and Chen 2021a) has shown that the opinions from company executives and managers or financial analysts can be more effective than the opinions of amateur investors when it comes to predicting the stock price. However, previous works using earnings calls data typically focus on the financial risk (volatility) prediction, while whether volatility prediction can help predict movement has been less well covered.

**Representing Numbers in Language Models**

Current language models treat numbers within the text input as plain words without understanding the basic numeric concepts. Given the ubiquity of numbers, their importance in financial datasets, and their fundamental difference with words, developing richer representations of numbers could improve the model’s performance in downstream financial applications (Chen et al. 2019b; Sawhney et al. 2020). Progress towards deeper numerical representations has been limited but promising. For example, previous work, represented by the DROP (Dua et al. 2019), presents a variety of numerical reasoning problems. Different from the existing works that pay attention to explore the capability of pretrained language models for general common-sense reasoning (Zhang et al. 2020), and math word problem-solving (Wu et al. 2021), we focus on improving the numerical understanding ability of language models for financial forecasting, motivated by the fact that financial documents often contain massive amounts of numbers. In particular, we consider two tasks – Numeral Category Classification (Chen, Wei, and Huang 2018; Chen et al. 2019b; Chen, Huang, and Chen 2021a) and Magnitude Comparison (Wallace et al. 2019; Naik et al. 2019) – using a structured adaptive pre-training strategy to improve the capability of pretrained language models for multi-task financial forecasting.

**Multi-modal Aligned Earnings Call Data**

Earnings conference call is typically presented by leading executives of publicly listed companies and provide an opportunity for the company to present an explanation of its quarterly results, guidance for the upcoming quarter, and an opportunity for some in-depth Q&A between the audience and company management (Keith and Stent 2019). An early study (Larcker and Zakolyukina 2012) mentioned that text-based models could reveal misleading information during earnings calls and cause stock price swings in financial markets and most unstructured data resources are still text data. So far, the multi-modal aligned earnings call data mainly refers to the sentence-level text-audio paired data resource, represented by Qin and Yang (2019) and Li et al. (2020). While previous works (Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020) mainly explore the benefits of audio features for volatility predictions, we propose a different research question that whether adaptive pre-training and volatility prediction loss can benefit the performance of stock prediction by using multi-modal aligned earnings call data.

**Approach**

The NumHTML model proposed in this paper is shown in Figure 1 and is made up of four components: (1) word-level encoder; (2) multimedia information fusion; (3) sentence-level encoder; and (4) pareto multi-task learning. Briefly, a key innovation in this work is the use of a novel structured
adaptive pre-training approach to improve how numeric information is treated during the word-level encoding of earnings calls transcripts. Then, sentence-level text features are aligned with 27 classical audio features extracted from earnings calls audio (Boersma and Van Heuven 2001) based on the approach described by Qin and Yang (2019). Next, the information fusion layer is responsible for combining the resulting text and audio features into a single representation for use by the sentence-level encoder to generate a multi-modal input that is suitable for training and prediction.

Structured Adaptive Pre-training

The pre-training process consists of two main tasks, Numeral Category Classification and Magnitude Comparison, in order to improve the representation of numerical data. During Numerical Category Classification, sentences that contain numeric data are categorised as belong to one or more of four main classes: monetary, temporal, percentage, and other. This categorization process uses a set of trigger tokens and rules so that, for example, in the sentence "During 2020 profits increased by 13% to $205m" presumably this is tagged as monetary (because of the $205m), temporal (2020) and percentage (%). We freeze the penultimate layer of fine-tuned whole-word-masked BERT (WWM-BERT) model before fine-tuning for the next numeral understanding task, Magnitude Comparison.

Following Wallace et al. (2019), Magnitude Comparison is probed in an argmax setting. Given the embeddings for five numbers, the task is to predict the index of the maximum number. Each list consists of values of similar magnitude within the same numeral type in order to conduct a fine-grained comparison. For example, for a given list containing five monetary numbers [$1.2, $2.5, $5, $9.8, $9.9], the training goal is to find the largest position value within this five values. In this given example, the golden label should be [0, 0, 0, 0, 1]. Softmax is used to assign a probability to each index using the hidden state trained by the negative log-likelihood loss. In practice, we shape the training/test set by uniformly sampling raw earnings call transcript data without placing back. A BiLSTM network will be fed with the list of token embeddings – varying from the pre-trained embeddings to the fine-tuned embeddings – connected with a weight matrix to compare its performance. The token-level encoder tuned by the structured adaptive pre-training is used for shaping the sentence-level textual embedding.

Sentence-level Transformer Encoder

We adopt a sentence-level Transformer encoder fed with sentence-level representations of long-form multi-modal aligned earnings call data (usually contains more than 512 tokens) for multi-task financial forecasting. Let $W_i = \langle w_1^i, w_2^i, \ldots, w_{|W_i|}^i \rangle$ be a sentence, where $|W_i|$ is the length and $w_{|W_i|}^i$ is an artificial EOS (end of sentence) token. The word embedding matrix associated with $W_i$ is initialized as

$$E_i = (e_1^i, e_2^i, \ldots, e_{|W_i|}^i)$$

$$e(\cdot)$$ maps each token to a $d$ dimensional vector using WWM-BERT, and $p_j$ is the position embedding of $w_{|W_i|}^i$ with the same dimension $d$. Consequently, $e_1^i \in \mathbb{R}^d$ for all $j$.

The enhanced WWM-BERT model after structured adaptive pre-training is adopted as the token-level Transformer encoder. The sentence representation $T_i \in \mathbb{R}^{d_t}$ of the sentence $W_i$ is calculated through the average pooling operating over the second last layer of the network. $d_t$ represents the default dimensions of word embeddings. The sentence representations are aligned with sentence-level audio features in the information fusion layer later. Finally, the multi-modal representation of a single earnings call is represented as:

$$D^{(k)} = (s_1^{(k)}, s_2^{(k)}, \ldots, s_M^{(k)})$$

$$T_i^{(k)}, A_i^{(k)} \in \mathbb{R}^{d_t \times d_s}$$

$$T_i^{(k)}$$ and $A_i^{(k)}$ represent the textual and audio features of sentence $i$ of document $D^{(k)} \in \mathbb{R}^{M \times d_s}$, respectively, and

Figure 1: Overall architecture of NumHTML. We use the surrounding tokens around [MASK] to classify the numeral categories.
The idea of Pareto MTL is shown in Figure 2, where $L$ is the loss of task objective space and with the preference vector $u$. The multi-objective sub-problem corresponding to the preference vector $u_k$ is defined as:

$$\min_{\theta} L(\theta) = (L_1(\theta), L_2(\theta), \ldots, L_m(\theta))^T, \text{s.t. } L(\theta) \in \Omega_k$$ \hspace{1cm} (3)

The set of possible solutions in different sub-regions is $\Omega_k = \{ v \in R^d | u_k^T v \leq u_j^T v, \forall j = 1, \ldots, K \}$. (4)

Hence, the set of possible solutions in different sub-regions represent different trade-offs among these two tasks. A two-step, gradient-based method is used to solve these multi-objective sub-problems based on the sentence-level multimodal representations.

**Initial Solution:** To find an initial feasible solution $\theta_0$ for a high-dimension, constrained optimization problem, we use a sequential gradient-based method since the straightforward projection approach is too expensive to calculate directly for the 345-million parameter WMM-BERT model. The update rule used is $\theta_{t+1} = \theta_t + \eta d_t$ where $\eta$ is the step size and $d_t$ is the direction, which can be obtained from the rule of Pareto critical (Zitzler and Thiele 1999). Iteration terminates once a feasible solution is found or the max number of iterations is met.

**Achieving Pareto Efficiency:** The next step is to solve the constrained subproblems in order to find a set of distributed solutions that can achieve the Pareto efficiency. Following Lin et al. (2019), we obtain a restricted Pareto critical solution for each training goal by using constrained multi-objective optimization, which generalizes the steepest descent method for unconstrained multi-objective optimization problems. Due to the high-dimensionality of the problem, we change the decision space from the parameter space to a more tractable objective and constraint space; see Lin et al. (2019) for a proof of this and the algorithm used. The result is a reduction in the dimension of the optimization problem from 345 millions to seven (two objective functions plus five activated constraints), which allows the Pareto MTL to be scaled and optimized for our task as shown in Equation 5, where $\hat{y}_i$ and $\hat{y}_j$ are the predicted values for the main and auxiliary tasks, respectively, and $y_i$ and $y_j$ denote the corresponding true values. The output of Pareto MTL is a set of weight allocation strategies ($\alpha_{patero_1}$ and $\alpha_{patero_0}$) for both tasks. We use Adam (Kingma and Ba 2014) as the optimizer and adopt the trick of learning-rate decay with increasing steps to train the model until it converges.

$$F = \alpha_{patero_1} \sum_i (\hat{y}_i - y_i)^2 + \alpha_{patero_2} \sum_j (\hat{y}_j - y_j)^2$$ \hspace{1cm} (5)

**Evaluation**

We make a comprehensive comparison of NumHTML with several state-of-the-art baselines using a publicly available dataset, by first focusing on stock movement prediction and then by testing various stock prediction techniques in a realistic long-term trading simulation. The hyper-parameters for our method and baselines are all selected by a grid search on the validation set. In each case, we demonstrate the significant advantage of NumHTML compared to baselines. Prior to these studies, we describe the intermediate results for the adaptive training used by NumHTML to demonstrate how it significantly outperforms the conventional pre-trained model.

**Dataset & Methodology**

**Dataset:** In line with previous work (Yang et al. 2020; Sawhney et al. 2020) for multi-task financial forecasting, in this evaluation, we use a publicly available Earning Conference Calls dataset constructed by Qin and Yang (2019). This dataset contains 576 earnings calls recordings, correspond to 88,829 text-audio aligned sentences, for S&P 500 companies in U.S. stock exchanges. The dataset also includes the corresponding dividend-adjusted closing prices from Yahoo Finance\footnote{https://finance.yahoo.com/} for calculating volatility and stock returns. To facilitate a direct comparison with the current state-of-the-art (Sawhney et al. 2020), we split the dataset into mutually exclusive training/validation/testing sets in the ratio of 7:1:2 (refers to instances) in chronological order, since future data cannot be used for prediction.

**The Stock Prediction Task:** We evaluate the stock prediction task as a classification problem — that is, the task is to predict whether a stock moves up (positive) or down (negative) due to the earnings call — in order to ensure a fair comparison with (Sawhney et al. 2020). The prediction of
\[ n \text{-day stock movement will be a rise if the regression results of the stock return is a positive value and vice versa.} \]

The Trading Simulation Task: To perform a trading simulation based on the multi-modal multi-task learning architecture, we aim to optimize for (1) average \( n \text{-day} \) volatility (that is, the average volatility of the following \( n \) days); and (2) cumulative \( n \text{-day} \) stock return (that is, the cumulative profit \( n \) days after an earnings call). In the trading simulation, stock movement predictions are used to decide whether to buy or sell a stock after \( n \) days. To ensure a fair comparison, we use the trading strategy implemented by (Sawhney et al. 2020), which relies on the results of stock movement prediction when \( n = 3 \). Thus, if the prediction is a rise in price \( p_{d-n}^s \) from day \( d-n \) to \( d \) for stock \( s \), the strategy buys the stock \( s \) on day \( d-n \), and then sells it on day \( d \). In addition, we perform a short sell if the prediction is a fall in price.

We maintain the same trading environment with Sawhney et al. (2020), which relies on the results of stock movement prediction when \( n = 3 \). Thus, if the prediction is a rise in price \( p_{d-n}^s \) from day \( d-n \) to \( d \) for stock \( s \), the strategy buys the stock \( s \) on day \( d-n \), and then sells it on day \( d \). In addition, we perform a short sell if the prediction is a fall in price.

We consider several different baselines (Wang et al. 2016; Yang et al. 2016; Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020), which, to the best of our knowledge, offer the best available stock prediction methods at the time of writing. These baselines can be grouped according to whether they use historical (numeric) pricing data, textual earnings call data, or multi-modal earnings call data.

1. LSTM+ATT (Wang et al. 2016): The best performing price-based model (LSTM with attention) in which the \( n \)-day volatility in the training data is predicted using pricing data only.

2. HAN (Glove) (Yang et al. 2016): Uses textual data in which each word in a sentence is converted to a word embedding using the pre-trained Glove 300-dimensional embeddings and trained by a hierarchical Bi-GRU models (Bahdanau, Cho, and Bengio 2014).

3. MDRM (Qin and Yang 2019): This recent work was the first to consider volatility prediction a multi-modal deep regression problem based on a newly proposed multi-modal aligned earnings calls dataset.

4. HTML (Yang et al. 2020): This recent hierarchical transformer-based, multi-task learning framework is designed specifically for volatility prediction using multi-modal aligned earnings call data.

5. Multi-Modal Ensemble Method (Sawhney et al. 2020): This multi-modal, multi-task learning approach represents the current state-of-the-art in the task of stock movement predictions using a combination of textual and audio earnings calls data.

Evaluating Structured Adaptive Training

To begin with, we present the results on the validation set for the adaptive training used by NumHTML.

Table 1: The four-class numeral category classification results varying from different embeddings.

<table>
<thead>
<tr>
<th>Model</th>
<th>LRAP</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>0.870</td>
<td>0.858</td>
</tr>
<tr>
<td>WWM-BERT</td>
<td>0.920</td>
<td>0.904</td>
</tr>
<tr>
<td>WWM-BERT+NCC</td>
<td>0.973</td>
<td>0.977</td>
</tr>
</tbody>
</table>

We consider several different baselines (Wang et al. 2016; Yang et al. 2016; Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020), which, to the best of our knowledge, offer the best available stock prediction methods at the time of writing. These baselines can be grouped according to whether they use historical (numeric) pricing data, textual earnings call data, or multi-modal earnings call data.

\[ \text{Sharpe Ratio} = \frac{r_n - R_f}{\sigma(r_x)} \]

Evaluating Structured Adaptive Training

To begin with, we present the results on the validation set for the adaptive training used by NumHTML.

Table 1: The four-class numeral category classification results varying from different embeddings.

<table>
<thead>
<tr>
<th>Model</th>
<th>LRAP</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove</td>
<td>0.870</td>
<td>0.858</td>
</tr>
<tr>
<td>WWM-BERT</td>
<td>0.920</td>
<td>0.904</td>
</tr>
<tr>
<td>WWM-BERT+NCC</td>
<td>0.973</td>
<td>0.977</td>
</tr>
</tbody>
</table>

Obviously, this represents a simplified trading strategy, given that it is limited to single share purchases. It was adopted here to align with previous work (Sawhney et al. 2020) but is an obvious avenue for future work to implement more sophisticated trading strategies.
ROC-AUC scores of the financial numeral category classification have been increased with the benefit of adaptive pre-training, which suggests that our approach (BERT+NCC) can classify numeral categories better than the raw pre-trained embeddings, including BERT and Glove. In particular, the performance of the adaptive pre-trained model is improved around 5.3% in LRAP and 7.3% in ROC-AUC.

**Magnitude Comparison** The accuracy of the Magnitude Comparison (listed maximum 5-numbers) based on different methods are shown in Table 2. We find that the ‘NCC’ task cannot guarantee the accuracy benefits for the maximum list task. However, the adaptive pre-training directly on the magnitude comparison task can significantly improve performance (94% vs 88% on average). We also notice that the ‘percentage’ classification can achieve the highest accuracy among four types, while the secular values are the hardest to predict (85% vs 99%). We speculate that numbers representing percentages are between 0 and 99, making it easier to predict the largest number among them. On the other hand, the numbers representing years usually contain four digits and are similar, posing a challenge for the magnitude comparison.

### Evaluating Stock Movement Prediction

The stock movement predictions results are presented in Table 3, using each of the baselines and several variations of the NumHTML model for 3, 7, 15, and 30-day prediction periods. The results indicate that NumHTML using multi-modal data generally outperforms all alternative methods, including the current state-of-the-art, multi-modal Ensemble method. The NumHTML variants generate predictions with the highest MCC and F1 scores, compared with the similar multi-modal, multi-task approach of the Ensemble alternative (Sawhney et al. 2020). This means that our single-model approach achieves statistically significant performance improvements over the Ensemble method for all cases when using multi-modal versions. In addition, using text-only data, our approach also achieves some meaningful improvements in almost all settings, excluding $n=15$.

To further understand the benefits of NumHTML, in what follows, we also consider several ablation studies to determine the efficacy of different NumHTML components.

#### On the Utility of Structured Adaptive Pre-Training:

In Table 3, we see NumHTML prediction performance exceeds that of NumHTML without the structured adaptive pre-training, for all $n$. In other words, by better modeling the numerical aspects of earnings call data, it is possible to significantly improve subsequent prediction performance.

#### On the Utility of Volatility Prediction as an Auxiliary Prediction Task:

Overall, NumHTML also significantly outperforms baseline methods in the volatility prediction task. Moreover, by comparing our approach with and without Pareto MTL, in Table 3, we see that the volatility prediction task consistently improves the results as an auxiliary task when predicting the stock movement; the single exception is for $n=7$. Moreover, Figure 3 shows that NumHTML can even achieve the best average performance (least MSE error over four sub-tasks) for the auxiliary task, which is ignored in previous works (Yang et al. 2020). Thus, by comparing the volatility prediction results of our approach with and without Pareto MTL, we find that the trade-off considerations between two tasks can significantly improve the performance of the auxiliary task.

#### On the Utility of Audio Features:

While existing work (Qin and Yang 2019; Yang et al. 2020; Sawhney et al. 2020) only explores the utility of audio features for volatility prediction, Table 3 shows how multi-modal learning consistently outperforms methods, which are purely based on the textual features, for stock movement prediction also. In particular, we observe consistent improvements by using the vocal cues compared with text-only versions among the four multi-modal methods. Improvements of this scale, relative to the corresponding text-only versions, are likely to translate into substantial practical benefits and suggest significant value in the use of audio features for a range of financial forecasting problems.

### Cumulative Profit in a Trading Simulation

Next, we consider the value of the various approaches to stock movement prediction in the context of a more realistic trading simulation. Table 4 presents the results (cumulative profit achieved and Sharpe Ratio) for various approaches. We use three standard trading strategies as baseline strategies: *Buy-all, Short-sell-all, Random* that are commonly used as benchmarks. We also compare our method with three strong multi-modal baselines, namely MRDM (Qin and Yang 2019), HTML (Yang et al. 2020), and Ensemble method (Sawhney et al. 2020). Once again, the NumHTML approach outperforms all of the alternatives in terms of both profit achieved and the Sharpe Ratio (higher is better). The profit achieved by NumHTML significantly exceeds that of the S&P 500 over the same period.

We have also interested in the individual effect of *structured adaptive pre-training and pareto multi-task learning*, respectively. Comparing the NumHTML to HTML with adaptive pre-training only (shown as NumHTML w/o Pareto), Table 4 shows that HTML with adaptive pre-training can improve the arbitrage profit somewhat, but with-

<table>
<thead>
<tr>
<th>Model</th>
<th>Monetary</th>
<th>Temporal</th>
<th>Percentage</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe</td>
<td>0.84</td>
<td>0.78</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>WWM-BERT</td>
<td>0.90</td>
<td>0.71</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>WWM-BERT+NCC</td>
<td>0.89</td>
<td>0.72</td>
<td>0.95</td>
<td>0.88</td>
</tr>
<tr>
<td>WWM-BERT+NCC+MC</td>
<td>0.93</td>
<td>0.85</td>
<td>0.99</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 2: The Magnitude Comparison Results (List Maximum from 5-numbers).
Price Movement Predictions

<table>
<thead>
<tr>
<th>Model</th>
<th>$MCC_3$</th>
<th>$MCC_7$</th>
<th>$MCC_{15}$</th>
<th>$MCC_{30}$</th>
<th>$F_{13}$</th>
<th>$F_{17}$</th>
<th>$F_{15}$</th>
<th>$F_{130}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price-based LSTM</td>
<td>0.069</td>
<td>0.097</td>
<td>0</td>
<td>0.271</td>
<td>0.694</td>
<td>0.200</td>
<td>0.765</td>
<td></td>
</tr>
<tr>
<td>Price-based BiLSTM-ATT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.149</td>
<td>0.342</td>
<td>0.200</td>
<td>0.721</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>-0.069</td>
<td>0.015</td>
<td>-0.048</td>
<td>-0.003</td>
<td>0.524</td>
<td>0.683</td>
<td>0.645</td>
<td>0.734</td>
</tr>
<tr>
<td>HAN (Glove)</td>
<td>0.090</td>
<td>-0.005</td>
<td>0.266</td>
<td>-0.042</td>
<td>0.591</td>
<td>0.621</td>
<td>0.598</td>
<td>0.703</td>
</tr>
</tbody>
</table>

### Text-only Methods

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_{13}$</th>
<th>$F_{17}$</th>
<th>$F_{15}$</th>
<th>$F_{130}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDRM</td>
<td>0.117</td>
<td>-0.107</td>
<td>0.032</td>
<td>-0.085</td>
</tr>
<tr>
<td>HTML</td>
<td>0.195</td>
<td>0.007</td>
<td>0.119</td>
<td>0.022</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.204</td>
<td>0.008</td>
<td>0.132</td>
<td>0.024</td>
</tr>
<tr>
<td>NumHTML</td>
<td>0.229**</td>
<td>0.009</td>
<td>0.122</td>
<td>0.031</td>
</tr>
</tbody>
</table>

### Multi-modal Methods

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_{13}$</th>
<th>$F_{17}$</th>
<th>$F_{15}$</th>
<th>$F_{130}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDRM (Multi-modal)</td>
<td>0.095</td>
<td>0.056</td>
<td>0.159</td>
<td>-0.065</td>
</tr>
<tr>
<td>HTML (Multi-modal)</td>
<td>0.280</td>
<td>0.125</td>
<td>0.196</td>
<td>0.131</td>
</tr>
<tr>
<td>Ensemble (Multi-modal)</td>
<td>0.321</td>
<td>0.128</td>
<td>0.191</td>
<td>0.128</td>
</tr>
<tr>
<td>NumHTML (w/o Pareto MTL)</td>
<td>0.293</td>
<td>0.129</td>
<td>0.198</td>
<td>0.133</td>
</tr>
<tr>
<td>NumHTML (w/o Adaptive Pre-training)</td>
<td>0.282</td>
<td>0.121</td>
<td>0.199</td>
<td>0.130</td>
</tr>
<tr>
<td>NumHTML (Multi-modal)</td>
<td>0.325**</td>
<td>0.126</td>
<td>0.206**</td>
<td>0.136*</td>
</tr>
</tbody>
</table>

Table 3: Results for the future n-day stock movement prediction (higher is better). * and ** indicate statistically significant improvements over the state-of-the-art ensemble method with $p < 0.05$, $p < 0.01$ respectively, under Wilcoxon’s test.

### Conclusion

This work contributes to multi-task financial forecasting, with a particular focus on the stock movement prediction, by using multi-modal earnings conference calls data. In particular, we propose a novel, numeric-oriented hierarchical transformer-based model (NumHTML) by using structured adaptive pre-training to improve how numeric data is represented and used in the pre-trained language model. A comprehensive comparative evaluation demonstrates significant performance benefits accruing to NumHTML, compared to a variety of state-of-the-art baselines and in the context of stock prediction and extended trading tasks. This evaluation also includes an ablation study to clarify the utility of different NumHTML components (adaptive pre-training, auxiliary volatility prediction, and the use of audio features).

Out work may be extended in several ways. More sophisticated numeric representations can be imagined in order to improve the representation of numeric data. Likewise, it may be feasible to develop similar representations for other categories of useful data in due course. In this work, we focused on stock movement prediction and trading, but the approaches described may be of value in a range of financial forecasting tasks such as portfolio management/design, hedging, financial fraud or accounting errors, etc. The current trading simulation, based on (Sawhney et al. 2020), imposes a significant single-stock purchasing limit per time period, as discussed. Going forward it will be necessary to consider more sophisticated trading policies.

### Acknowledgments

We acknowledge with thanks the discussion with Boyuan Zheng and Cunxiang Wang from Westlake University, as well as the many others who have. We would also like to thank anonymous reviewers for their insightful comments and suggestions to help improve the paper. This publication has emanated from research conducted with the finan-

Table 4: Cumulative profit across different trading strategies.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Profit</th>
<th>Sharpe Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Baselines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buy-all</td>
<td>$36.59</td>
<td>0.76</td>
</tr>
<tr>
<td>Short-sell-all</td>
<td>-$36.59</td>
<td>-0.77</td>
</tr>
<tr>
<td>Random</td>
<td>$-25.78</td>
<td>-0.58</td>
</tr>
<tr>
<td>Multi-modal Methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDRM</td>
<td>$38.75</td>
<td>0.81</td>
</tr>
<tr>
<td>HTML</td>
<td>$72.47</td>
<td>1.52</td>
</tr>
<tr>
<td>NumHTML (w/o Pareto)</td>
<td>$73.90</td>
<td>1.53</td>
</tr>
<tr>
<td>Ensemble</td>
<td>$75.73</td>
<td>1.59</td>
</tr>
<tr>
<td>NumHTML</td>
<td>$77.81</td>
<td>1.62</td>
</tr>
</tbody>
</table>

Figure 3: The results of volatility prediction. 'Ours (w/o P)' indicates NumHTML without Pareto MTL. Text-only and multi-modal methods are presented by different colors.

We thank the discussion with Boyuan Zheng and Cunxiang Wang from Westlake University, as well as the many others who have. We would also like to thank anonymous reviewers for their insightful comments and suggestions to help improve the paper. This publication has emanated from research conducted with the finan-
cial support of Postdoctoral Funding Sponsored by Zhejiang Province and Rong Hui Jin Xin Company under Grant Number 10313H041801 and the financial support of the Pioneer and "Leading Goose" R&D Program of Zhejiang under Grant Number 2022SDXHXDX0003. Yue Zhang and Barry Smyth are co-corresponding authors.

References


