NumHTML: Numeric-Oriented Hierarchical Transformer Model for Multi-Task Financial Forecasting

Linyi Yang,[1,2] **Jiazheng Li,**[4] **Ruihai Dong,**[3] **Yue Zhang,**[1,2] **Barry Smyth**[3]
[1] Westlake Institute for Advanced Study, Westlake University
[2] School of Engineering, Westlake University
[3] School of Computer Science, University College Dublin
[4] Department of Computer Science, University of Warwick
linyi.yang, yue.zhang@westlake.edu.cn, ruihai.dong, barry.smyth@insight-centre.org, jiazheng.li@warwick.ac.uk

Abstract

Financial forecasting has been an important and active area of machine learning research because of the challenges it presents and the potential rewards that even minor improvements in prediction accuracy or forecasting may entail. Traditionally, financial forecasting has heavily relied on quantitative indicators and metrics derived from structured financial statements. Earnings conference call data, including text and audio, is an important source of unstructured data that has been used for various prediction tasks using deep earning and related approaches. However, current deep learning-based methods are limited in the way that they deal with numeric data; numbers are typically treated as plain-text tokens without taking advantage of their underlying numeric structure. This paper describes a numeric-oriented hierarchical transformer model (*NumHTML*) to predict stock returns, and financial risk using multi-modal aligned earnings calls data by taking advantage of the different categories of numbers (monetary, temporal, percentages etc.) and their magnitude. We present the results of a comprehensive evaluation of NumHTML against several state-of-the-art baselines using a real-world publicly available dataset. The results indicate that NumHTML significantly outperforms the current state-of-the-art across a variety of evaluation metrics and that it has the potential to offer significant financial gains in a practical trading context.

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 relying on pricing data alone (Moskowitz, Ooi, and Pedersen 2012; Kristjanpoller, Fadic, and Minutolo 2014; Manela and Moreira 2017; Zheng et al. 2019; Pitkäjärvi, Suominen, and Vaittinen 2020). However, recent advances in natural language processing (NLP) and deep learning (DL) introduce novels sources of data — textual data in the form of financial news (Ding et al. 2014; Yang et al. 2018; Hu et al. 2018; Chen et al. 2019a; Du and Tanaka-Ishii 2020) and financial reports (Duan et al. 2018; Kogan et al. 2009) to real-time social media (Xu and Cohen 2018; Feng 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 earning 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

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 shows 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 (Da, Engelberg, and Gao 2015; Xing, Cambria, and Welsch 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 numeral 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 pretraining 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

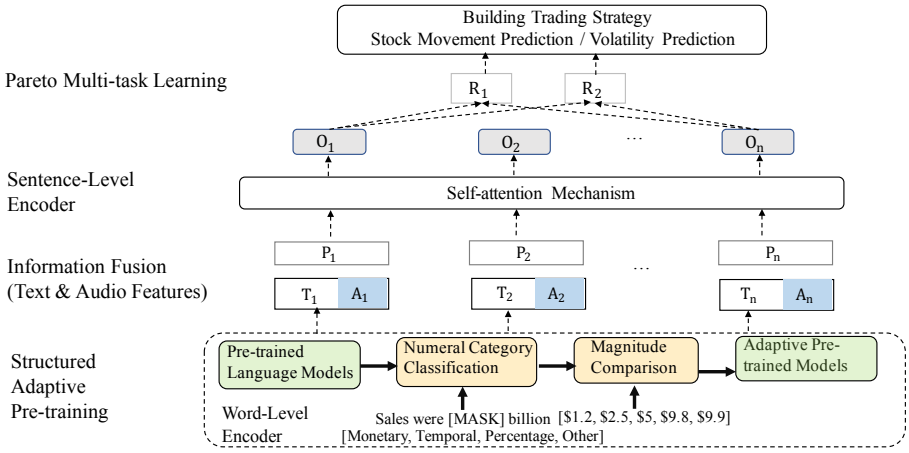

Figure 1: Overall architecture of NumHTML. We use the surrounding tokens around [MASK] to classify the numeral categories.

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 *Numeral 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 $W_i = \left(w_i^1, w_i^2, ..., w_i^{|W_i|}\right)$ be a sentence, where $|W_i|$ is the length and $w_i^{|W_i|}$ is an artificial `EOS` (end of sentence) token. The word embedding matrix associated with W_i is initialized as

$$\begin{aligned} \mathbf{E}_i &= \left(\mathbf{e}_i^1, \mathbf{e}_i^2, \ldots, \mathbf{e}_i^{|t_i|}\right) \\ \text{where } \mathbf{e}_i^j &= e\left(w_i^j\right) + \mathbf{p}_j. \end{aligned} \tag{1}$$

$e(\cdot)$ maps each token to a d dimensional vector using WWM-BERT, and p_j is the position embedding of w_i^j with the same dimension d. Consequently, $e_i^j \in \mathbb{R}^d$ 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:

$$\begin{aligned} \mathcal{D}^{(k)} &= \left(s_1^{(k)}, s_2^{(k)}, \ldots, s_M^{(k)}\right) \\ \text{where } \mathbf{s}_i^{(k)} &= \left((T_i^{(k)}, A_i^{(k)}) + P_i\right). \end{aligned} \tag{2}$$

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

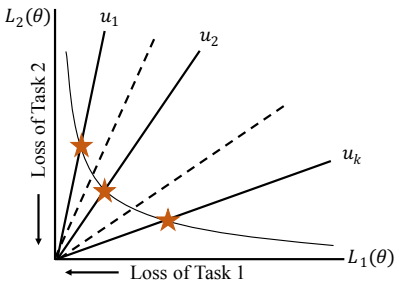

Figure 2: Pareto MTL aims to find a set of Pareto solutions in different restricted preference regions (Lin et al. 2019).

$P_i \in \mathbb{R}^{M \times d_s}$ denotes the trainable sentence-level position embedding. M is the maximum number of sentences.

Pareto Multi-task Learning

We adopt the Pareto Multi-task Learning algorithm (Pareto MTL) proposed by Lin et al. (2019) to integrate stock movement prediction and volatility prediction by finding a set of well-distributed Pareto solutions that can represent different trade-offs between both tasks. Pareto MTL decomposes a Multi-Task Learning (MTL) problem into multi-objective subproblems with multiple constraints. An average pooling operation is first applied to the output of the sentence-level Transformer encoder. Then, we find a set of well-distributed unit preference vectors $\{u_1, u_2, \ldots, u_K\}$ in R_+^2; $K = 10$ in this work. The multi-objective sub-problem corresponding to the preference vector u_K is defined as:

$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), \cdots, \mathcal{L}_m(\theta))^{\mathrm{T}}, \text{s.t.} \mathcal{L}(\theta) \in \Omega_k \quad (3)$$

The idea of Pareto MTL is shown in Figure 2, where $\mathcal{L}_m(\theta)$ is the loss of task m and Ω_k(k=1,..., K) is a sub-region in a objective space and with $u_j^T v$ as the inner product between the preference vector u_j and a given vector v:

$$\Omega_k = \left\{ \boldsymbol{v} \in R_+^2 \mid \boldsymbol{u}_j^T \boldsymbol{v} \leq \boldsymbol{u}_k^T \boldsymbol{v}, \forall j = 1, \ldots, K \right\} \quad (4)$$

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 multi-modal representations.

Initial Solution: To find an initial feasible solution θ_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 WWM-BERT model. The update rule used is $\theta_{t+1} = \theta_t + \eta d_t$ where η is the step size and d_t is the search 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 (α_{patero_1} and α_{patero_1}) 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.

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

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 earning 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 [1] 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

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

n-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-day* volatility (that is, the average volatility of the following n days); and (2) cumulative *n-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^s_{d-n} 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): there are no transaction fees, we can only purchase a single share (but for multiple companies) for each time period, and intra-day trading is not considered [2].

Evaluation Metrics

For stock movement prediction we report the F1 score and Mathew's Correlation Coefficient (MCC) for stock price prediction. MCC performs more precisely when the data is skewed by accounting for the true negatives. For a given confusion matrix:

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \quad (6)$$

Then the predicted average $n-day$ volatility is compared with the actual volatility (Eq. 7) to compute the mean squared error for each hold period: $n \in \{3, 7, 15, 30\}$.

$$v_{[0,n]} = \ln\left(\sqrt{\frac{\sum_{i=1}^{n}(r_i - \bar{r})^2}{n}}\right) \quad (7)$$

r_i is the stock return on day i and $\bar{r}$ is the average stock return (using adjusted closing price) in a window of n days.

$$MSE = \frac{\sum_i (\hat{y}_i - y_i)^2}{n} \quad (8)$$

For the stock trading simulation we use the cumulative profit and Sharpe Ratio metrics. The cumulative profit generated by a simple trading strategy is defined as:

$$\text{Profit} = \sum_{s \in S} \left(p^s_d - p^s_{d-\tau}\right) * (-1)^{\text{Action } {}^{d-\tau}_{s}} \quad (9)$$

where (p^s_d) indicates the stock price of stock s on the day d, and Action $^{d-\tau}_{s}$ is a binary value depended on the stock movement prediction result; it equals to 0 if the model predicts a rise in price for stock s on day d, otherwise it is 1. Sharpe Ratio evaluates the performance of investments using their average return rate r_x, risk-free return rate R_f and the standard deviation σ across the investment x:

$$\text{Sharpe Ratio} = \frac{r_x - R_f}{\sigma(r_x)} \quad (10)$$

[2]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.

Model	LRAP	ROC_AUC
Glove	0.870	0.858
WWM-BERT	0.920	0.904
WWM-BERT+NCC	0.973	0.977

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

Baselines

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.

Numeral Category Classification The results of multi-label numeral category classification (NCC) on the validation set are shown in Table 1. The aim is to show how this task significantly enhances the token-level embeddings. Both Label ranking average precision (LRAP) and

Model	Monetary	Temporal	Percentage	All
GloVe	0.84	0.78	0.89	0.82
WWM-BERT	0.90	0.71	0.95	0.88
WWM-BERT+NCC	0.89	0.72	0.95	0.88
WWM-BERT+NCC+MC	0.93	0.85	0.99	0.94

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

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-model 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-

Model	Price Movement Predictions							
	MCC_3	MCC_7	MCC_{15}	MCC_{30}	$F1_3$	$F1_7$	$F1_{15}$	$F1_{30}$
Price-based LSTM	0.069	0	0.097	0	0.271	0.694	0.200	0.765
Price-based BiLSTM-ATT	0	0	0	0	0.149	0.342	0.200	0.721
SVM	-0.069	0.015	-0.048	-0.003	0.524	0.683	0.645	0.734
HAN (Glove)	0.090	-0.005	0.266	-0.042	0.591	0.621	0.598	0.703
Text-only Methods								
MDRM	0.117	-0.107	0.032	-0.085	0.675	0.500	0.571	0.601
HTML	0.195	0.007	0.119	0.022	0.623	0.688	0.648	0.700
Ensemble	0.204	0.008	**0.132**	0.024	0.675	0.690	0.636	0.703
NumHTML	**0.229****	**0.009**	0.122	**0.031**	**0.689***	**0.691**	**0.644****	**0.727***
Multi-modal Methods								
MDRM (Multi-modal)	0.095	0.056	0.159	-0.065	0.628	0.690	0.452	0.590
HTML (Multi-modal)	0.280	0.125	0.196	0.131	0.696	0.695	0.703	0.748
Ensemble (Multi-modal)	0.321	0.128	0.191	0.128	0.702	0.698	0.702	0.761
NumHTML (w/o Pareto MTL)	0.293	**0.129**	0.198	0.133	0.701	**0.700**	0.711	0.759
NumHTML (w/o Adaptive Pre-training)	0.282	0.121	0.199	0.130	0.697	0.668	0.705	0.746
NumHTML (Multi-modal)	**0.325****	0.126	**0.206****	**0.136***	**0.722***	0.697	**0.716***	**0.770****

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.

Strategy	**Profit**	**Sharpe Ratio**
Simple Baselines		
Buy-all	$36.59	0.76
Short-sell-all	-$36.59	-0.77
Random	-$25.78	-0.58
Multi-modal Methods		
MRDM	$38.75	0.81
HTML	$72.47	1.52
NumHTML (w/o Pareto)	$73.90	1.53
Ensemble	$75.73	1.59
NumHTML	$**77.81**	**1.62**

Table 4: Cumulative profit across different trading strategies.

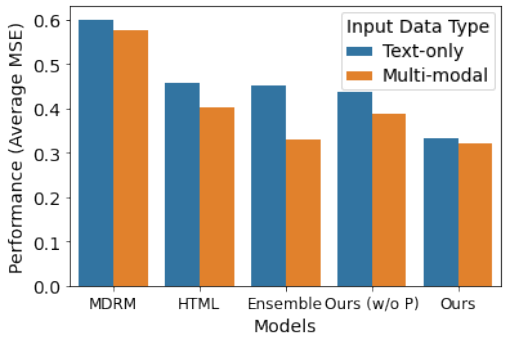

Figure 3: The results of volatility prediction. 'Ours (w/o P)' indicates NumHTML without Pareto MTL. Text-only and multi-modal methods are presented by different colors.

out benefiting the Sharpe Ratio. Furthermore, the single Pareto MTL component provides significant performance benefits in terms of profit (73.90 vs 77.81) and Sharpe Ratio (1.53 vs 1.62), which suggests that our method benefits considerably from the trade-off considerations.

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).

Our 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-

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 2022SDXHDX0003. Yue Zhang and Barry Smyth are co-corresponding authors.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Boersma, P.; and Van Heuven, V. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10): 341–347.

Bollerslev, T.; Patton, A. J.; and Quaedvlieg, R. 2016. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1): 1–18.

Chang, C.-Y.; Zhang, Y.; Teng, Z.; Bozanic, Z.; and Ke, B. 2016. Measuring the Information Content of Financial News. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3216–3225. Osaka, Japan.

Chen, C.; Zhao, L.; Bian, J.; Xing, C.; and Liu, T.-Y. 2019a. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2376–2384. New York, NY, USA: Association for Computing Machinery.

Chen, C.-C.; Huang, H.-H.; and Chen, H.-H. 2021a. Evaluating the Rationales of Amateur Investors. In *The World Wide Web Conference*.

Chen, C.-C.; Huang, H.-H.; and Chen, H.-H. 2021b. From Opinion Mining to Financial Argument Mining. *Springer Briefs in Computer Science*, 1–95.

Chen, C.-C.; Huang, H.-H.; Takamura, H.; and Chen, H.-H. 2019b. Numeracy-600k: learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, 6307–6313.

Chen, Y.; Wei, Z.; and Huang, X. 2018. Incorporating Corporation Relationship via Graph Convolutional Neural Networks for Stock Price Prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, 1655–1658. ISBN 9781450360142.

Da, Z.; Engelberg, J.; and Gao, P. 2015. The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, 28(1): 1–32.

Ding, X.; Zhang, Y.; Liu, T.; and Duan, J. 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)*, 1415–1425.

Ding, X.; Zhang, Y.; Liu, T.; and Duan, J. 2015. Deep Learning for Event-Driven Stock Prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, 2327–2333. Buenos Aires, Argentina.

Du, X.; and Tanaka-Ishii, K. 2020. Stock embeddings acquired from news articles and price history, and an application to portfolio optimization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3353–3363.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Duan, J.; Zhang, Y.; Ding, X.; Chang, C. Y.; and Liu, T. 2018. Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING-18)*, 2823–2833.

Feng, F.; Chen, H.; He, X.; Ding, J.; Sun, M.; and Chua, T.-S. 2018. Enhancing stock movement prediction with adversarial training. *arXiv preprint arXiv:1810.09936*.

Hu, Z.; Liu, W.; Bian, J.; Liu, X.; and Liu, T.-Y. 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*, WSDM '18, 261–269. New York, NY, USA: Association for Computing Machinery.

Keith, K.; and Stent, A. 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*, ACL '19, 493–503. Florence, Italy.

Kimbrough, M. D. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80(1): 189–219.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kogan, S.; Levin, D.; Routledge, B. R.; Sagi, J. S.; and Smith, N. A. 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*, NAACL '09, 272–280.

Kristjanpoller, W.; Fadic, A.; and Minutolo, M. C. 2014. Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, 41(5): 2437–2442.

Larcker, D. F.; and Zakolyukina, A. A. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2): 495–540.

Li, J.; Yang, L.; Smyth, B.; and Dong, R. 2020. MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3063–3070.

Lin, X.; Zhen, H.-L.; Li, Z.; Zhang, Q.-F.; and Kwong, S. 2019. Pareto multi-task learning. In *Advances in Neural Information Processing Systems*, 12060–12070.

Liu, S.; and Tse, Y. K. 2013. Estimation of monthly volatility: An empirical comparison of realized volatility, GARCH and ACD-ICV methods. *Finance Research Letters*.

Loughran, T.; and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1): 35–65.

Manela, A.; and Moreira, A. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1): 137–162.

Moskowitz, T. J.; Ooi, Y. H.; and Pedersen, L. H. 2012. Time series momentum. *Journal of financial economics*, 104(2): 228–250.

Naik, A.; Ravichander, A.; Rose, C.; and Hovy, E. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3374–3380.

Pitkäjärvi, A.; Suominen, M.; and Vaittinen, L. 2020. Cross-asset signals and time series momentum. *Journal of Financial Economics*, 136(1): 63–85.

Qin, Y.; and Yang, Y. 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*, 390–401. Florence, Italy: Association for Computational Linguistics.

Rekabsaz, N.; Lupu, M.; Baklanov, A.; Dür, A.; Andersson, L.; and Hanbury, A. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1712–1721.

Sawhney, R.; Mathur, P.; Mangal, A.; Khanna, P.; Shah, R. R.; and Zimmermann, R. 2020. Multimodal Multi-Task Financial Risk Forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 456–465. Association for Computing Machinery.

Wallace, E.; Wang, Y.; Li, S.; Singh, S.; and Gardner, M. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5310–5318.

Wang, W. Y.; and Hua, Z. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1155–1165.

Wang, Y.; Huang, M.; Zhu, X.; and Zhao, L. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 606–615.

Wu, Q.; Zhang, Q.; Wei, Z.; and Huang, X. 2021. Math Word Problem Solving with Explicit Numerical Values. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5859–5869. Online: Association for Computational Linguistics.

Xing, F. Z.; Cambria, E.; and Welsch, R. E. 2018. Intelligent asset allocation via market sentiment views. *ieee ComputatioNal iNtelligeNCe magaziNe*, 13(4): 25–34.

Xing, F. Z.; Cambria, E.; and Zhang, Y. 2019. Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, 176: 68–76.

Xu, Y.; and Cohen, S. B. 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)*, 1970–1979.

Yang, L.; Dong, R.; Ng, T. L. J.; and Xu, Y. 2019. Leveraging BERT to Improve the FEARS Index for Stock Forecasting. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, 54–60. Macao, China.

Yang, L.; Ng, T. L. J.; Smyth, B.; and Dong, R. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, 441–451.

Yang, L.; Zhang, Z.; Xiong, S.; Wei, L.; Ng, J.; Xu, L.; and Dong, R. 2018. Explainable text-driven neural network for stock prediction. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 441–445. IEEE.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 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*, 1480–1489.

Ye, Z.; Qin, Y.; and Xu, W. 2020. Financial Risk Prediction with Multi-Round Q&A Attention Network. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence*, 4576–4582. International Joint Conferences on Artificial Intelligence Organization.

Zhang, X.; Ramachandran, D.; Tenney, I.; Elazar, Y.; and Roth, D. 2020. Do Language Embeddings Capture Scales? *arXiv preprint arXiv:2010.05345*.

Zheng, J.; Xia, A.; Shao, L.; Wan, T.; and Qin, Z. 2019. Stock volatility prediction based on self-attention networks with social information. In *2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 1–7. IEEE.

Zitzler, E.; and Thiele, L. 1999. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4): 257–271.