Modeling Financial Uncertainty with
Multivariate Temporal Entropy-based Curriculums: Supplementary Material

Ramit Sawhney*[1] **Arnav Wadhwa***[1] **Ayush Mangal**[2] **Vivek Mittal**[3] **Shivam Agarwal**[4] **Rajiv Ratn Shah**[1]

[1]IIIT Delhi
[2]Indian Institute of Technology Roorkee
[3]Indian Institute of Technology Mandi
[4]Manipal Institute of Technology

The text throughout this supplementary material assumes familiarity with the main article for this work.

A FINCLASS DETAILS

A.1 DIFFICULTY SCORE COMPUTATION

On computing the Stock-complexity S_d: First, we compute the bearish, neutral, and bullish intent of each news item or tweet in the lookback period using class probabilities obtained via fine-tuned Financial BERT (FinBERT) for English tweets [Araci, 2019] and Chinese financial text [Rao et al., 2021]. For the US S&P 500 dataset, we use the pre-trained implementation of FinBERT provided here.[1] We use the logits from the last layer to get the sentiment scores for bearish, bullish and neutral intent for each English tweet posted in a day in the lookback window. For the China & Hong Kong dataset we use a BERT model pretrained on Chinese data provided here.[2] We finetune the model on a dataset for Chinese financial short text sentiment analysis, for which we adopt the corpus as provided here.[3] We add a dense head on the top, and finetune the model for financial sentiment analysis. We then use the logits extracted from the fine-tuned model to get the bearish, bullish, and neutral intents for the Chinese news headline data.

Permutation Entropy: Permutation Entropy (PE) is a robust time series tool which provides a quantification measure of the complexity of a dynamic system by capturing the order relations between values of a time series and extracting a probability distribution of the ordinal patterns. PE has been shown to be more sensitive to minor changes, and it can better reflect trends of complexities in stock market related environments and systems. We observe that it captures richer information about the differences between the US and Chinese stock indexes. We use the implementation provided here.[4]

Multivariate Permutation Entropy: For computing the MPE, we use the implementation as provided here.[5]

Dynamic Time Warping: To analyze how the temporal evolution of stock-affecting signals varies across days, we adopt Dynamic Time Warping (DTW) distance [Müller, 2007], which measures the similarities between time-series of the three intents across consecutive days in a *time-aware fashion*. In time series analysis, DTW is one of the algorithms for measuring similarity between two temporal sequences, which may vary in timescale or speed. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Since the number of tweets in a day might not be consistent across days, and their release times also differ, we use DTW between the time-series of the

*Indicates equal contribution to this work.
[1]Pretrained FinBERT model (for English text): `https://huggingface.co/ProsusAI/finbert`
[2]Pretrained Chinese Bert models: `https://huggingface.co/bert-base-chinese`
[3]Financial short-text sentiment analysis corpus: `https://github.com/fip-lab/Sentiment_Analysis`
[4]Permutation Entropy Implementation: `https://github.com/nikdon/pyEntropy`
[5]MPE implementation: `https://github.com/alberto-ara/Multivariate-multiscale-permutation-entropy`

three intents across consecutive days in a time-aware fashion as a measure of inter-day complexity and entropy. DTW is particularly useful for classifying stock market multivariate time series data related to daily returns, volatility, daily stocks returns, commodity prices, volume trading, index, enhanced index tracking portfolio, and much more, being robust to outliers. We use the implementation for DTW as provided here.[6]

We fuse the DTW values obtained from each consecutive pair of days in the lookback window by calculating the PE of the values obtained for each pair to get an intra-day entropy/difficulty measure.

B EXPERIMENTS AND SETUP

B.1 MORE ABOUT PUBLIC DATA

US S&P 500 [Xu and Cohen, 2018] This dataset comprises $109,915$ *tweets* in the English language, related to 88 high-trade-volume-stocks from the *NASDAQ* Stock Exchange forming the *S&P 500* index. NASDAQ is a fairly volatile US exchange. Xu and Cohen [2018] categorize the stocks in this data into 9 industries:[7] Basic Materials, Consumer Goods, Healthcare, Services, Utilities, Conglomerates, Financial, Industrial Goods and Technology.

China & Hong Kong Huang et al. [2018] This dataset comprises $90,361$ financial *news headlines* in the *Chinese* language, originally aggregated by *Wind*[8] from major financial website like Sina[9] and Hexun.[10] Huang et al. [2018] extract news from these websites, which cover corporate news across the regions of Mainland China and Hong Kong.

Pre-processing We pre-process English tweets using the NLTK (Twitter mode), for treatment of URLs, identifiers (@) and hashtags (#). We adopt the Bert-Tokenizer for tokenization of tweets. For the English tweets, we use the pre-trained BERT-base-cased[11] model. For the Chinese news, we adopt the Chinese-BERT-base model, having 12 layers and 110M parameters. We use character-based tokenization for the Chinese headlines.

B.2 MORE ABOUT TRAINING SETUP

We explore the lookback period for THA-Net $T \in$ range$[2, 10]$ days. Lower stock prediction performance indicates the inability of shorter lookbacks to capture stock affecting market information, likely as public information requires time to absorb into price movements. As we increase T, we find that larger lookbacks allow the inclusion of stale information from older days having relatively lower influence on prices, thus deteriorating the stock prediction and volatility regression performance. We observe the best performance for mid-sized (5-day) lookback periods.

B.3 EVALUATION

Classification Following prior research for stock prediction [Xu and Cohen, 2018, Sawhney et al., 2020], we use accuracy and MCC (Implementations from *sklearn*[12]) for classification performance.

Matthew's Correlation Coefficient (MCC) MCC produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

Regression To evaluate the volatility regression performance, we adopt the Mean Squared Error (MSE) to compute the error between actual and the predicted volatility values. The MSE is defined as:

$$mse = (\hat{v_\tau} - v_\tau)^2 \tag{1}$$

where $\hat{v_\tau}$ is predicted volatility by the model.

[6]Dynamic Time Warping implementation: `https://pypi.org/project/fastdtw/`
[7]Industry classification source: `https://finance.yahoo.com/industries`
[8]Wind: `https://www.wind.com.cn/en/wft.html`
[9]Sina: `http://finance.sina.com.cn/`
[10]Hexun: `http://www.hexun.com/`
[11]Pretrained BERT-base-cased models: `www.github.com/google-research/bert`
[12]sklearn: `https://scikit-learn.org`

Profit *Sharpe ratio* A portfolio with a higher Sharpe ratio is considered superior relative to its peers. Sharpe ratio is widely used for robust performance hypothesis testing in financial domain. If two funds offer similar returns, the one with higher standard deviation will have a lower Sharpe ratio. In order to compensate for the higher standard deviation, the fund needs to generate a higher return to maintain a higher Sharpe ratio. Sharpe shows how much additional return an investor earns by taking additional risk. We use the implementation of the annualized Sharpe ratio provided in the following footnote.[13]

Maximum Drawdown (MDD) MDD is an indicator of downside risk over a specified time period. MDD is used to assess the relative riskiness of one stock screening strategy versus another, as it focuses on capital preservation, which is a key concern for most investors. MDD is widely used for measuring the sensitivity of a portfolio optimisation strategy. A low maximum drawdown is preferred as this indicates that losses from investment were small. If an investment never lost any money, the MDD would be zero. The worst possible maximum drawdown would be -100%, where the investment is completely worthless. We use the implementation of MDD provided in the following footnote.[14]

B.4 MORE ABOUT BASELINES

Here, we provide detailed descriptions and implementation details (if available) regarding some of the baseline methods we adopt to contrast the performance of THA-Net + FinCLASS.

- **W-LSTM:**[15] In this method, the stock price time series is decomposed by a wavelet transformation to eliminate noise. Then, stacked autoencoders are applied to generate high-level features, which are fed into an LSTM for generating the predictions [Bao et al., 2017].
- **RandForest:**[16] Random Forest classifier trained over word2vec [Mikolov et al., 2013] embeddings for text.
- **TSLDA:** Topic-Sentiment Latent Dirichlet Allocation, is a generative model which simultaneously exploits topics (nouns) and sentiments (opinions, adjectives) from texts to predict stock movements [Nguyen and Shirai, 2015].
- **SN - HFA:**[17] StockNet - HedgeFundAnalyst, A deep generative variational autoencoder approach that applies attention on texts and days [Xu and Cohen, 2018]. The HFA variant exploits both text and price information. Other settings:
- **SN - DA:** StockNet - DiscriminativeAnalyst, A deep generative variational autoencoder approach that applies attention on texts and days [Xu and Cohen, 2018]. The discriminative StockNet directly optimizes the likelihood objective, and we take out the effects of the Kullback-Leibler term.
- **Chaotic:**[18] A hierarchical attention network that uses GRU encoders with temporal attention applied to texts and days in the lookback [Hu et al., 2018].
- **StockEmb:** This method generates stock embeddings using prices and a dual vector (word-level, context-level) representation of texts. The word level embeddings are acquired using the TFIDF-weighted word embeddings with the word2vec approach, and the context level vectors are obtained through BERT followed by a principal component analysis (PCA) to reduce the 1024-dimensional embeddings to a 256-dimensional space [Du and Tanaka-Ishii, 2020].

C ETHICAL CONSIDERATIONS

There is an ethical imperative implicit in this growing influence of automation in market behavior, and is worthy of serious study [Cooper et al., 2020]. While financial markets are transparent, and heavily regulated, we discuss the ethical considerations pertaining to our work. Following [Cooper et al., 2016], we emphasize on two ethical criteria for trading systems, and discuss FinCLASS's design with respect to these criteria.

Blocking Price Discovery A trading system should not block price discovery, and not interfere with the ability of other market participants to add to their own information [Angel and McCabe, 2013]. For example, placing an extremely large volume of orders to block competitor's messages *(Quote Stuffing)* or intentionally trading with itself to create the illusion of market activity *(Wash Trading)*. FinCLASS does not block price discovery in any form.

[13] Sharpe ratio implementation: `https://github.com/quantopian/empyrical`
[14] Maximum Drawdown implementation: `https://github.com/quantopian/empyrical`
[15] Implementation for WLSTM: `https://github.com/mlpanda/DeepLearning_Financial`
[16] Implementation for RandForest: `https://github.com/johnberroa/RandomForest-StockPredictor`
[17] Implementation for SN-HFA and SN-DA: `https://github.com/yumoxu/stocknet-code`
[18] Implementation for Chaotic: `https://github.com/gkeng/Listening-to-Chaotic-Whishpers--Code`

Circumventing Price Discovery A trading system should not hide information, such as by participating in dark pools or placing hidden orders Zhu [2014]. We evaluate the performance of FinCLASS only on public data (tweets and financial news items) pertaining to highly regulated stock markets.

We follow broad ethical guidelines to design and evaluate FinCLASS and THA-Net, and encourage readers to follow both regulatory and ethical considerations pertaining to the stock markets.

References

James J Angel and Douglas McCabe. Fairness in financial markets: The case of high frequency trading. *Journal of Business Ethics*, 112(4):585–595, 2013.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLOS ONE*, July 2017.

Ricky Cooper, Michael Davis, Ben Van Vliet, et al. The mysterious ethics of high-frequency trading. *Business Ethics Quarterly*, 26(1):1–22, 2016.

Ricky Cooper, Michael Davis, Andrew Kumiega, and Ben Van Vliet. Ethics for automated financial markets. *Handbook on Ethics in Finance*, pages 1–18, 2020.

Xin Du and Kumiko Tanaka-Ishii. 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*, pages 3353–3363, Online, July 2020. Association for Computational Linguistics.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 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, page 261–269, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810.

Jieyun Huang, Yunjia Zhang, Jialai Zhang, and Xi Zhang. A tensor-based sub-mode coordinate algorithm for stock prediction, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, Beijing, China, July 2015. Association for Computational Linguistics.

Dongning Rao, Sihong Huang, Zhihua Jiang, Ganesh Gopal Deverajan, and Rizwan Patan. A dual deep neural network with phrase structure and attention mechanism for sentiment analysis. *Neural Computing and Applications*, pages 1–12, 2021.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.676. URL `https://www.aclweb.org/anthology/2020.emnlp-main.676`.

Yumo Xu and Shay B. Cohen. 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)*, pages 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Haoxiang Zhu. Do dark pools harm price discovery? *The Review of Financial Studies*, 27(3):747–789, 2014.