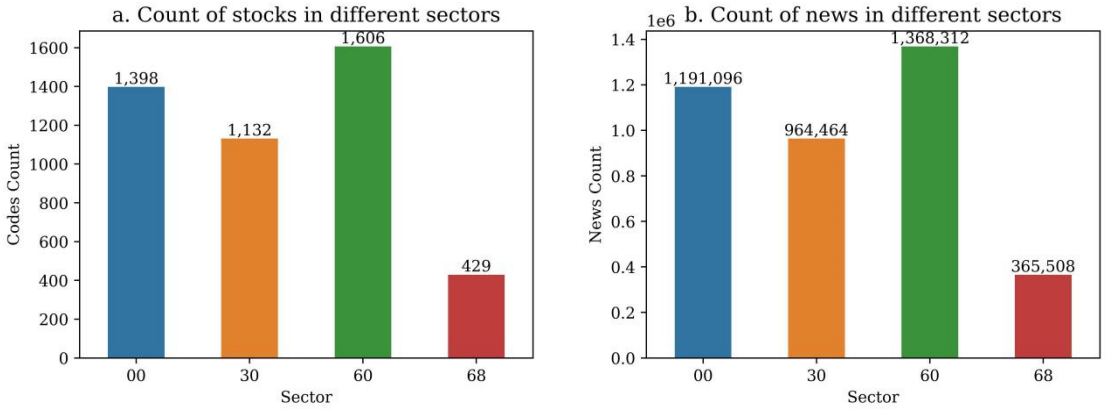
Supplementary Information for

A deep learning approach with extensive sentiment analysis for quantitative investment

Wang Li, Chaozhu Hu and Youxi Luo *

School of Science, Hubei University of Technology, Wuhan 430068, China

*: To whom correspondence should be addressed: 20051038@hbut.edu.cn



Supplementary Figure S1. Statistics of the collected stock and news data. a) Number of stocks in the three SSE boards; b) Number of news articles collected for each board.

Supplementary Table S1. Example of the collected news.

Stock	000001 平安银行 Ping An Bank
Date	2022-05-29
Title	平安银行本周被深股通减持 4.7 亿元，周内减持市值两市排名第 11 Ping An Bank was divested by 470 million yuan through the Shenzhen-Hong Kong Stock Connect this week, ranking 11th in terms of the total value of divestments across both markets
Content	平安银行本周(2022/05/22-2022/05/29)累计跌幅达 5.59%，截至当前最新股价报收 14.18 元..... Ping An Bank fell 5.59% cumulatively this week (2022/05/22-2022/05/29). As of the latest closing price, it was quoted at 14.18 yuan...

Supplementary Table S2. Parameters settings for deep learning models.

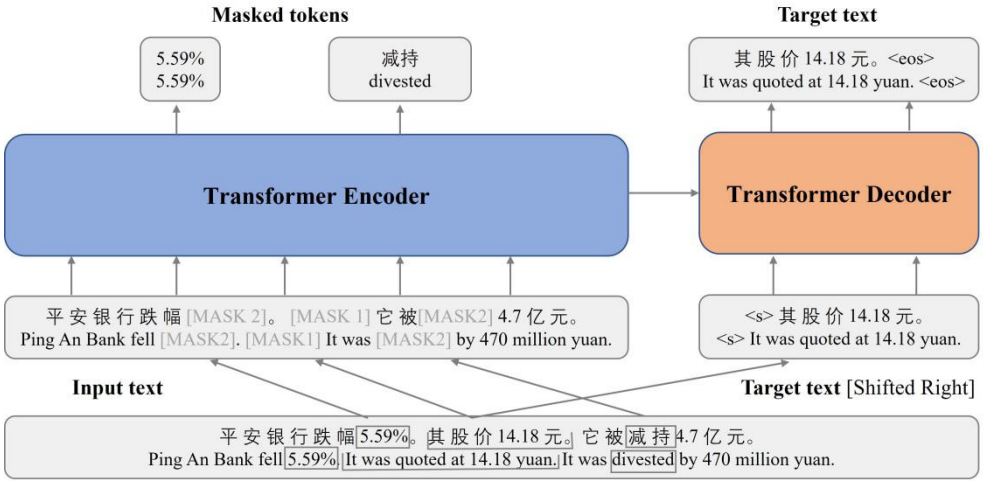
Parameter	LSTM	Transformer
Learning rate	0.001	0.001
Sliding window	20	64
Epochs	150	150
Batch size	16	16
Optimizer	Adam	Adam
Loss function	Cross-entropy	Cross-entropy
Num layers	2	6

Introduction for Technical Indicators

Among the 16 factors used, both MA and EMA are variants of moving averages, commonly used tools for analyzing time series data that help identify trends in stock prices. MACD can help identify changes in trends. RSI analyzes the strength of buying and selling pressure based on the principle of balance between supply and demand in the stock market, by comparing the magnitude of price changes within a certain period for an individual stock or the rise and fall of the market index, thereby assessing the future market trend. WILLR is an indicator to measure stock price changes, aiding in determining if a stock is overbought or oversold. MOM and CMO help users assess the speed of stock price changes. ULTOSC and ADOSC help capture market oscillation signals. OBV assists in determining the trend in changes in trading volume.

Architecture of Pegasus

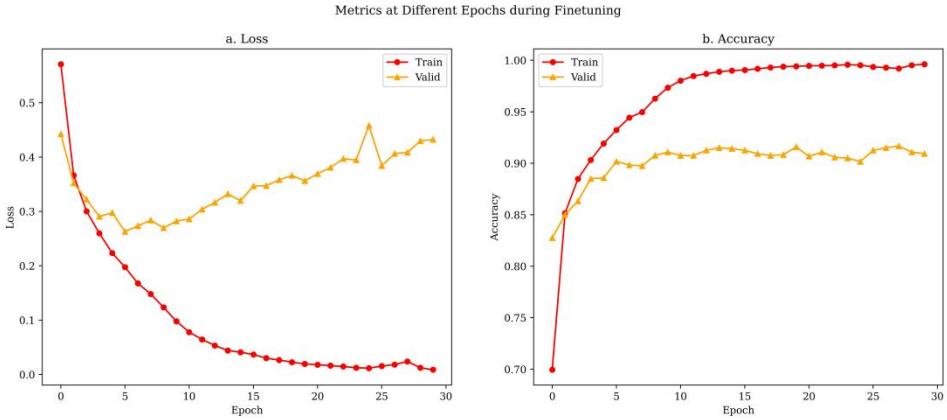
As shown in Supplementary Figure S2, Pegasus consists of the standard Transformer encoder and decoder. The figure presents two pretraining tasks - Gap Sentences Generation (GSG) and Masked Language Modeling (MLM). These two tasks apply different masking (MASK) to the input text, where the encoder is responsible for recovering the masked language model (MLM), and the decoder recovers the Gap Sentences Generation (GSG) task.



Supplementary Figure S2. The architecture of Pegasus.

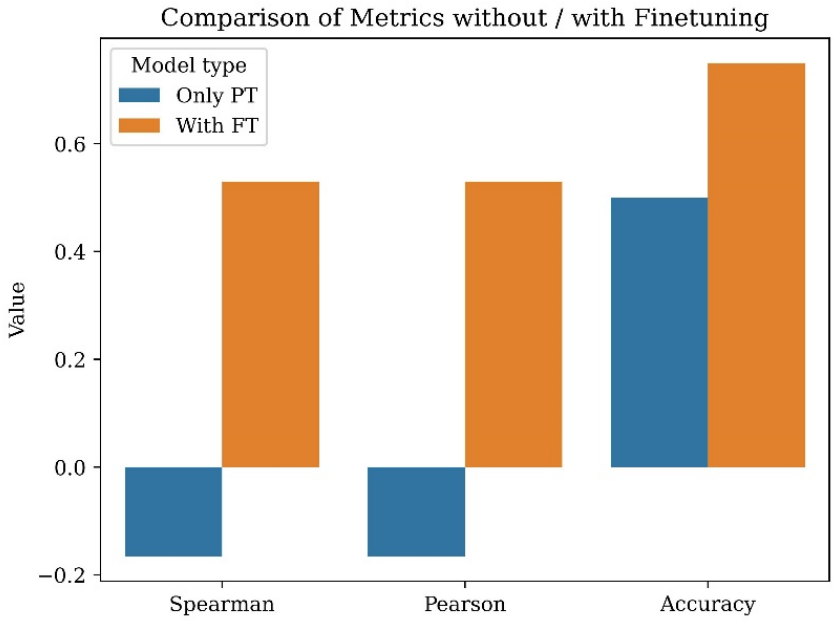
Details of the finetuning for ALBERT

During fine-tuning of the sentiment classification prediction model, the Adam optimizer is used to minimize the cross-entropy loss function, with $learning\_rate = 2 \times 10^{-5}, \lambda = 1 \times 10^{-4}$. Mean squared error (MSE) and accuracy are used as evaluation metrics. The sentiment prediction model fine-tuning curves are shown in Supplementary Figure S3. It can be observed that in the first few rounds, the validation metrics are better than the training set, reflecting the effect of the model's pre-training knowledge. As training iterates, the model starts performing better on the training set from round 3, and mostly converges after round 10, achieving the best validation result of 0.9167 accuracy at round 28. At the best round, model generalization is tested by computing the metrics on the test set, obtaining 0.4135 loss and 0.9125 accuracy, indicating strong generalization capability.

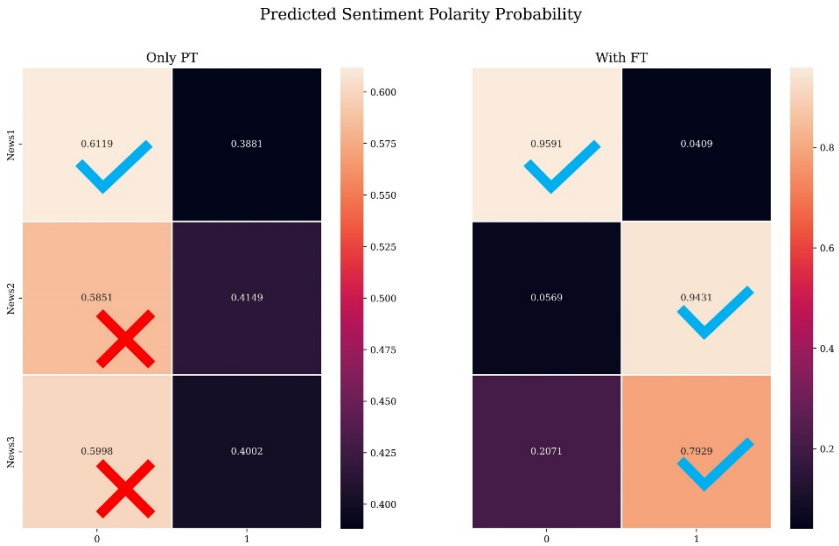


Supplementary Figure S3. The curves of finetuning.

To verify the improvement in model sentiment prediction accuracy from fine-tuning, we compare the performance of models with and without fine-tuning on the ChnSentiCorp test set, as shown in Supplementary Figure S4. It can be seen that on all three evaluation metrics of Spearman, Pearson and Accuracy, the fine-tuned model has significantly improved precision in predicting news sentiment, with higher consistency between predicted and true sentiment labels. Supplementary Figure S5 provides a comparison of the predicted sentiment polarity probabilities on three stock news examples by the two models. While the merely pre-trained model correctly predicts the first news (negative) albeit with low confidence, it wrongly predicts the second and third news (positive). In contrast, the fine-tuned model correctly predicts both positive and negative news with high confidence.



Supplementary Figure S4. A comparison of sentiment prediction. These are the results for 100 randomly selected samples from the ChnSentiCorp test set. Only PT represents the model after pre-training only, while With FT denotes the model that underwent additional fine-tuning after pre-training.



Supplementary Figure S5. A comparison of model prediction performance on stock market news before and after fine-tuning. Only PT represents the model after pre-training only, while With FT denotes the model that underwent additional fine-tuning after pre-training.

Preliminary exploration of integrating social media data

Incorporating other forms of data might provide additional perspectives on market sentiment. In this section, we investigate the contribution of social media data to quantitative investment and compare its effectiveness to that of news. We collected social media data related to stocks on EastMoney (www.eastmoney.com). Example of news and social media sentiment of three stocks from three different boards are shown in Supplementary Table S3. It can be observed that news content is more concise and objective. Further correlation analysis between sentiment changes in news media and social media and the stock price movements is presented in Supplementary Table S4. The results indicate that sentiment changes in news media exhibit a stronger correlation with stock price movements, with Spearman correlation coefficients consistently exceeding 0.2, while those for social media are close to 0. This suggests that social media sentiment is less representative and cannot comprehensively reflect the attitudes of investors. Consequently, it is reasonable to consider sentiment analysis from a news media perspective as more effective and conducive to improving model performance.

Supplementary Table S3. Comparison of news and social media.

Stock	News	Social Media
000524 (2022-08-23)	岭南控股股东户数下降 1.59%，户均持股 21.91 万元。岭南控股股东户数高于行业平均水平。根据 Choice 数据，…… Lingnan Holdings saw a 1.59% decrease in the number of shareholders, with an average holding of 219,100 RMB per shareholder. The number of shareholders in Lingnan Holdings is higher than the industry average. According to Choice data, ...	岭南垃圾，拿了接近两个月没一次涨超过 5%的垃圾。 Garbage Lingnan, which hasn't seen a single increase of over 5% in nearly two months, is truly disappointing.
300718 (2022-07-18)	长盛轴承盘中最高 24.63 元，股价连续两日创近一年新高。截至收盘，长盛轴承最新价为 22.7 元，…… Changsheng Bearings reached an intraday high of 24.63 RMB, with its stock price hitting a near one-year high for two consecutive days. As of the closing, Changsheng Bearings' latest price was 22.7 RMB, ...	开盘诱多，千万别追，坐等跌停。 Opening with a trap for long positions, it's better not to chase. Just wait for the limit down.
603398 (2022-05-12)	沐邦高科 05 月 12 日被沪股通减持 11.99 万股。最新持股量为 39.92 万股，占公司 A 股总股本的 0.12%。 On May 12th, Mubang High-tech had 119,900 shares reduced by the Shanghai Stock Connect. The latest holding is 399,200 shares, accounting for 0.12% of the company's A-share total share capital.	这货亏那么多钱，还在涨真好。 It's surprising that this stock is still rising despite such heavy losses.

Supplementary Table S4. Relational analysis between stock price movements and sentiment from different forms of data.

Data	000524		300718		603398	
	Pearson's R	**Spearmanr's R**	**Pearson's R**	**Spearmanr's R**	**Pearson's R**	**Spearmanr's R**
News	0.2558	0.2709	0.2304	0.2538	0.3166	0.3155
Social Media	0.0845	0.0469	-0.0019	-0.0305	-0.0493	-0.0470

Baseline models performance

XGBoost and RNN, which belong to machine learning and deep learning respectively, are utilized as the additional baseline models. Both RNN and LSTM are recurrent neural networks, while XGBoost is a well-performing machine learning model in many research domains. Following the same experimental approach as with LSTM and Transformer, the backtesting results are shown in Supplementary Table S5. The results on RNN and XGBoost are consistent with those on LSTM and Transformer, with the model that incorporates both news titles and content summaries performing the best. Specifically, RNN achieved the best ARR and MDR of 29.11% and 11.68%, respectively, while XGBoost performed less favorably with ARR and MDR of 12.92% and 5.92%, respectively. XGBoost performed significantly worse than the other models. The lower returns and higher risks of RNN compared to LSTM are expected since LSTM is an improved model over RNN. In contrast to previous studies that solely considered sentiment analysis in either title or content, our approach comprehensively integrates both title and content information, fully utilizing the rich information embedded in news. The results of the baseline models along with LSTM and Transformer collectively confirm the effectiveness of our approach.

Supplementary Table S5. Backtesting performance of baselines models.

Group	RNN		XGBoost	
	ARR	**MDR**	**ARR**	**MDR**
Vanilla	15.03%	25.29%	8.67%	6.29%
Title	20.75%	6.31%	11.94%	9.13%
Title + Content	29.11%	11.68%	12.92%	5.92%