NUS
FINTECH
SOCIETY

Project Report

Integrating Crowd-based Trade Signals from Social Media for Algo Trading

AY 25/26 (Sem 1)

Tech Lead:

Roydon Tay

Tech Analysts:

Wayne Tan, Mya Thet Chai, Park Seunghwi, Quek Jin Kai

Abstract:

Cryptocurrencies and decentralized finance (DeFi) applications present unique analytical challenges, characterized by high volatility, a lack of conventional fundamental valuation methods, and a pronounced dependence on social media for informational cues. This study aims to integrate signals derived from Reddit posts with established time series analysis techniques to predict price trends for Ethereum (ETH). We study the effectiveness of predicting price trends using time series price data and social media sentiment based features, and propose a CNN-GRU architecture to perform time series prediction with integration of aggregated text-based sentiments as tabular features. The proposed architecture outperforms baseline LSTM for time series modelling, and integration of sentiment features have also shown to improve model predictions and algorithmic trading performance.

Related Work:

The research presented in *On the enrichment of time series with textual data for forecasting agricultural commodity prices* demonstrates the efficacy of integrating tabular data features with time series models. Specifically, it shows that incorporating low-dimension text-based features to provide domain knowledge as extra context can improve forecasting accuracy in certain scenarios (Filho et al., 2022). This idea of augmenting time series data with textual information was the main motivation of our study.

Previous studies have used various forms of aggregation metrics for measuring financial sentiment on social media platforms (Liu & Son, 2024). These metrics provide ways to aggregate sentiment over large volumes of posts (per timestep in the case of time series analysis). Other studies found greater success by extracting explicit trading advice from social media posts, such as explicit mentions of buying or selling (Haase et al., 2025). These methods give us multiple directions of extracting trade signals from social media posts, and motivated us to conduct experiments to find out which method was most suitable for our case.

Finally, when selecting models to use, we referred closely to a study where researchers utilised a CNN model on time series data to make trend predictions to identify optimal entry and exit points (Zaar et al., 2023). The study achieved promising results using a CNN based architecture to predict price trends, and hence the model we eventually propose follows their architecture closely, but also includes an additional GRU layer at the end.

Research Questions:

In this study, we aim to answer the following questions:

1. Can the CNN-GRU architecture we propose beat a baseline time series model (LSTM)?
2. Can sentiments / trade signals on Cryptocurrencies from Reddit posts be useful as features for time series price trend prediction?

To answer the first question, we conducted an experiment of the time series price trend prediction task between our proposed model and a baseline LSTM model. Both models were trained on daily closing price data from 2017-2024, and had to make trades based on their trend predictions (invest all capital in long positions when predicting uptrend, short position if downtrend). The metric of comparison is the profitability by the end of the backtesting period (2024-2025). We found that our proposed method performed better than the LSTM model.

To answer the second question, we experimented with different methods of extracting and aggregating sentiments from Reddit posts, and experimented time series prediction tasks with and without text-based sentiment features on top of price data, using our proposed model to see if including them would improve price trend predictions.

i. Prefixed text for sentiment classification:

We tested if adding a prefix to textual posts to provide additional context could provide more meaningful sentiments. For example, adding “ETH price outlook: “ to the front of text from posts. For the rest of the report, we will refer to text processed in this way as Prefixed text. The results showed that doing so created sentiment features that perform better for time series trend prediction.

ii. Extracting trade signals vs Sentiments of Cryptocurrency

We also attempted extracting buy / sell intentions from the posts and comments directly. We combined pretrained language model zero-shot classification of posts and comments for trading intents (buy / sell / hold / neutral) and keyword detection to extract trading intent of the Reddit users. Aggregation methods tested include intent ratios (e.g. buy/sell ratio). In our experiments, our methods identified mostly neutral intentions in many timesteps, and we concluded that this method is less effective in identifying possible signals that influence crypto prices.

iii. Sentiment aggregation metrics:

We also experimented with three different sentiment aggregation metrics, B-A4, B-B1 and B-B2, shown in figure 1. In all the formulas, N refers to the number of posts / comment within the batch (hourly timestep) of the sentiment class (Positive / Neutral / Negative). In B-B1, ‘score’ refers the softmax probability of the ‘Positive’ class, while W (weight) refers to the number of upvotes the post / comment had (defaults to 1 when there is none). Based on experimental results, we found that B-B1 was the best metric, suggesting that using information from upvotes made the feature more useful.

$$IS_t^{B-A4} = \frac{N_t^{positive}}{N_t^{positive} + N_t^{negative} + N_t^{neutral}} = \frac{N_t^{positive}}{N_t};$$

$$IS_t^{B-B1} = \frac{\sum_{i=0}^{N_t} score_{t,i} * W_{t,i}}{\sum_{i=0}^{N_t} W_{t,i}}$$

$$IS_t^{B-B2} = \frac{N_t^{positive} - N_t^{negative}}{N_t} = \frac{1}{N_t} \sum_{i=0}^{N_t} score_{t,i}$$

Figure 1: Sentiment aggregation functions used

Methodology:

1. Data Collection:

A data pipeline was built using Github Actions and the Praw package. From 29/10/2025 to 22/11/2025, the pipeline ran hourly, collecting the 100 most recent posts each time. For hourly closing price data, we used the yfinance package to get data from Yahoo Finance.

2. Data Processing and Feature engineering

Text from the Reddit posts comes in an unstructured prose format, which necessitates additional labeling to categorize the content and facilitate the aggregation of results across all posts. We utilised a text transformer for classification to determine the sentiment expressed in each post (Positive / Negative / Neutral). For labelling of sentiments, we used a DistilRoBERTa model finetuned for classifying financial news sentiments from Huggingface, which outputs a softmax probability distribution for each textual input; and assigned a sentiment class based on threshold (softmax score of class > 0.8).

3. Trend labelling

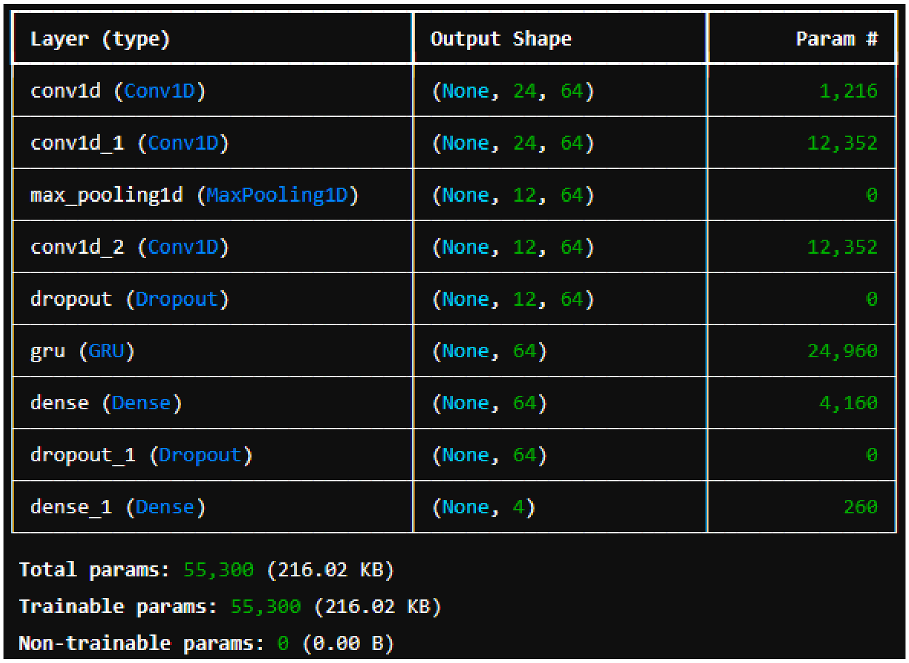
Trend labelling was performed using a rule-based, forward-looking approach to identify meaningful price trends while filtering short-term noise. We computed the Hull Moving Average (HMA) of ETH closing prices and its first-order difference to estimate local trend direction, and detected candidate reversals when the gradient changed sign. A reversal was confirmed only if future price movement over a 15-timestep look-ahead exceeded a volatility-adjusted threshold of 2.25× the 14-days Average True Range (ATR), ensuring that labels corresponded to economically significant trend changes. Following each confirmed reversal, trend continuation labels (uptrend or downtrend) were propagated forward until the next reversal, resulting in a temporally consistent trend label for every timestep used in supervised model training.

Limitation:
The labelling scheme relies on fixed parameters, including the volatility (ATR) window, look-ahead horizon (15 days), and threshold multiplier (2.25×), which were selected heuristically

and not optimized. Consequently, the definition of a “trend” is tied to a specific daily time scale and may not generalize optimally across different market regimes or assets. Exploring more data-driven or regime-adaptive labelling methods is left for future work.

4. Model Architecture and training parameters

In the experiments conducted, we did not conduct hyperparameter tuning and kept train parameters constant, to observe the effects of engineered features. All models had the same architecture (except for input layer, depending on dimensions of input). All models had a lookback of 24 timesteps, and were trained on 40 epochs.



```
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ conv1d (Conv1D)                 │ (None, 24, 64)         │         1,216 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ conv1d_1 (Conv1D)               │ (None, 24, 64)         │        12,352 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ max_pooling1d (MaxPooling1D)    │ (None, 12, 64)         │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ conv1d_2 (Conv1D)               │ (None, 12, 64)         │        12,352 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dropout (Dropout)               │ (None, 12, 64)         │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ gru (GRU)                       │ (None, 64)             │        24,960 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                   │ (None, 64)             │         4,160 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dropout_1 (Dropout)             │ (None, 64)             │             0 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 4)              │           260 │
└─────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 55,300 (216.02 KB)
 Trainable params: 55,300 (216.02 KB)
 Non-trainable params: 0 (0.00 B)
```

Figure 2: Model Architecture of proposed CNN-GRU model

Results & Conclusions:

Model trained with B-B1 Aggregation	Model trained with B-B1 Aggregation (prefixed text)
$10040.78	**$10622.86**

Price only	B-B1 (prefixed text)	B-B2 (prefixed text)	B-A4 (prefixed text)
$10104.19	**$10622.86**	$10455.51	$10419.19

Figure 3: Tables comparing final equity obtained by models after trading over backtesting period with starting capital of $10000

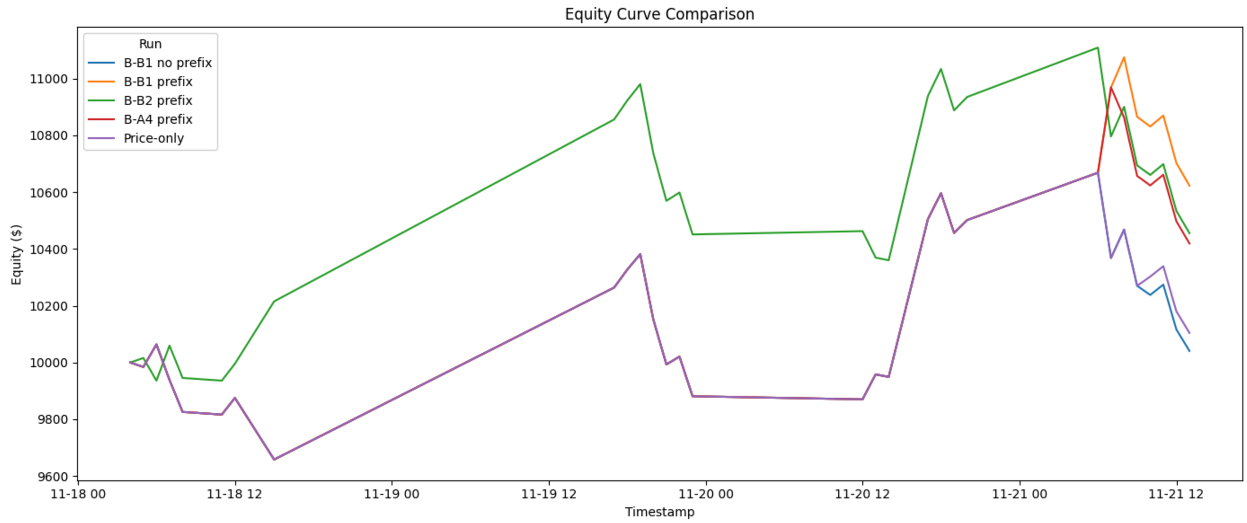


Figure 4: Plots of backtesting equity plots between different models trained

From the results in figure 3, we observed that using prefixed text in sentiment labelling improved the model's profitability, which suggests prefixing texts in Reddit posts makes the sentiment feature obtained more meaningful, and creates a better feature for the price trend prediction task. Hence, subsequent experiments compared models trained with prefixed text. These experiments concluded that B-B1 was the best metric for aggregating sentiments, out of those we tested. The results also shown that models with any sentiment features performed better than model predicting with just price time series alone.

Limitations:
For simplicity, we conducted backtesting assuming absence of bid-ask spread, and with the simple strategy of investing all capital in either long or short positions only each time.

In conclusion, we observed that aggregated cryptocurrency sentiments from Reddit can be useful for time series price trend prediction. The methods discussed here could potentially be used with other social media platforms as well.

References:

Liu, Q., Son, H. Methods for aggregating investor sentiment from social media. Humanit Soc Sci Commun 11, 925 (2024). https://doi.org/10.1057/s41599-024-03434-2

Haase, F., Celig, T., Rath, O., & Schoder, D. (2025). Wisdom of the crowd signals: Predictive power of social media trading signals for cryptocurrencies. Electronic Markets, 35(1). https://doi.org/10.1007/s12525-025-00815-6

Filho, I. J. R., Marcacini, R. M., & Rezende, S. O. (2022). On the enrichment of time series with textual data for forecasting agricultural commodity prices. MethodsX, 9, 101758. https://doi.org/10.1016/j.mex.2022.101758

Zaar, A. E., Benaya, N., Moubtahij, H. E., Bakir, T., Mansouri, A., & Allati, A. E. (2023). Ethereum Cryptocurrency Entry Point and Trend Prediction using Bitcoin Correlation and Multiple Data Combination. International Journal of Advanced Computer Science and Applications, 14(5). https://doi.org/10.14569/ijacsa.2023.0140506

Appendix:

GitHub repository: https://github.com/RoydonTay/crowd_based_signals_for_crypto_trading