Predict S&P 500 Motion by Analyzing Reddit Posts from Relevant Subreddit Pages on the Previous Day

Yuval Arbel
yuval.arbel1@mail.huji.ac.il

Asaf Shul
asaf.shul@mail.huji.ac.il

Tamar Czaczkes
tamar.czaczkes@mail.huji.ac.il

0 Github repository

https://github.com/AsafShul/Predict_Market_Change_From_Reddit

1 Introduction

Predicting stock market trends is of great interest to both individuals and companies. However, due to the high volatility of stock markets, accurately predicting their behavior is a challenging task. Given the increasing significance of LLMs in various aspects of our lives, we sought to leverage their capabilities for stock market forcasting, using publicly available data. During our research, we came across several papers attempting to utilize Twitter tweets (Kolasani (2020) and Pagolu et al. (2016)), or financial articles (Sethia et al., 2022) for predicting market trends. In our work we explored the potential of Reddit data, which we believe could be highly indicative in this context, especially due to Reddit's huge impact on the Gamestop (GME) short squeeze[1] in 2021. Specifically, the *r/wallstreetbets* subreddit was described in Bradley et al. (2021) paper as "the most influential social finance site by many metrics". Our aspiration was to use state-of-the-art LLM models to develop a tool for this task or, at the very least, generate a valuable dataset in the process.

2 Data

We Collected Reddit posts from *r/WallStreetBets*[2] and *r/Stocks*[3] (Range: 2008-01-01 to 2022-12-31). These two subreddits contain approximately 200 posts per day, which translate to roughly 100GB of posts and comments in json format. Reddit's API is not available since last April, so we used data dumps[4] instead. After examining the data, we decided to exclude the comments and use only the posts, due to context length limitations.

2.1 Preprocessing

We filtered out irrelevant data fields and posts, such as polls, memes, and daily discussion posts. We removed posts submitted on days when the stock market was closed on the following day. HTML addresses were converted to 'http' tags to conform with the input requirements of fine-tuned model. To create a unified input for the model, we concatenated the title and content of each post using the separator '$$$'. This process resulted in a pandas DataFrame containing the essential information from the relevant Reddit data, with the post submission time serving as the index. Initially, we considered incorporating weights into our predictions based on the number of up-votes (similar to "likes" on Facebook) of the Reddit posts. However, due to the challenge of determining if these up-votes occurred on the same day as the post, we decided against it to prevent potential data leakage. Instead, we calculated the number of comments for each post, only counting comments from the same day to avoid data leakage and ensure data integrity. To reduce noise we kept only posts with $5 < |ValidComments| < 1000$. A data snippet example can be found in Fig. 1. Overall, our dataset contains ~215K samples, split to ~150K (train), ~20K (validation), ~50K (test) samples each.

2.2 Labelling

For the purpose of this study, we used financial data procured from '*Yahoo! Finance*'[5] API. This is a known and reliable public source of stock data. Each Reddit post is assigned one of three distinct labels: "Increased" ('2'), "Decreased" ('0') or "Neutral" ('1'). The labels are assigned in accordance with whether the value of the S&P 500 index (SPY) exhibits an increase or decrease in the day following the post. An increase is identified when the closing price exceeds the opening price by a margin of more than a certain epsilon. To establish

[1] https://en.wikipedia.org/wiki/GameStop_short_squeeze

[2] https://www.reddit.com/r/wallstreetbets/

[3] https://www.reddit.com/r/stocks/

[4] https://the-eye.eu/redarcs/

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

these labels, we used the following formulas, calculating the change in percentage as follows:

$$change = (1 - \frac{nextDayClose}{nextDayOpen}) \cdot 100$$

Upon obtaining the percentage change, the labels are assigned using the following criterion:

$$label = (|change| > \epsilon) \cdot sign(change) + 1$$

The label calculation takes into account the predefined threshold ϵ, where the absolute value of the change must exceed ϵ in order to be considered significant, similar methods are used in other market classification papers such as (Steinbacher, 2023). After examining different epsilon values, we chose $\epsilon = 0.3$, which results in a relatively balanced dataset (see Table 2).

3 Methods

In our study, we initially considered utilizing a pre-trained RoBerta (Liu et al., 2019) based sentiment analysis model (He et al., 2023) that had been fine-tuned on financial news articles. However, upon analyzing our Reddit post data, we observed that the nature of the posts was more like tweets rather than news articles.

Baseline: We used a pre-trained Roberta-based sentiment analysis model (Loureiro et al., 2022) as baseline, acquired via the HuggingFace Hub[6]. We observed that the baseline consistently predicted "neutral" sentiment as the most frequent label for each day. This outcome is not surprising, given that the number of "neutral" posts significantly outweighs the number of positive or negative ones (due to large amount of noise in the dataset).

Comment-weighted naive model: In order to improve the model's performance, we counted the comments received by each post within the same day (as detailed in subsection 2.1). To predict a unified daily sentiment from the sentiments of each post, we computed a weighted average using softmax over the comment counts. The rationale behind this approach is that irrelevant posts would likely receive fewer comments, while meaningful posts tend to generate higher engagement.

Comment-weighted fine-tuned model: We further improved our model by fine-tuning it on the dataset we created. The fine-tuning was done post-by-post. Testing of the model was done using the comments-weighted-average method described above. The training was done with Huggingface trainer and AdamW (Loshchilov and Hutter, 2017) optimizer with hyperparameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $lr = 5/e^5$, for 12 epochs using Cross-Entropy loss.

Unified-signal model: Finally, aiming to derive more "semantic" signals during the fine-tuning process, we sought to utilize gradients from all available posts per day (instead of processing one post at a time as done in the previous models). To achieve this, we proposed a new model architecture (see Fig. 2). This architecture predicts the label for the top K posts (padding if necessary), and passes them to a linear head, alongside each post's normalized number of comments. The linear head consists of 2 fully-connected layers $(size : [K \cdot 4] \rightarrow [K \cdot 4 \cdot \Phi] \rightarrow [3])$, to predict a single label. The order of the top-K posts (and their respective comments) was shuffled in each epoch. The model was trained for both 50 and 150 epochs with the same hyper-parameters as before. With our available GPU resources, $K = 10$ with a batch size of 2 was the largest we could run.

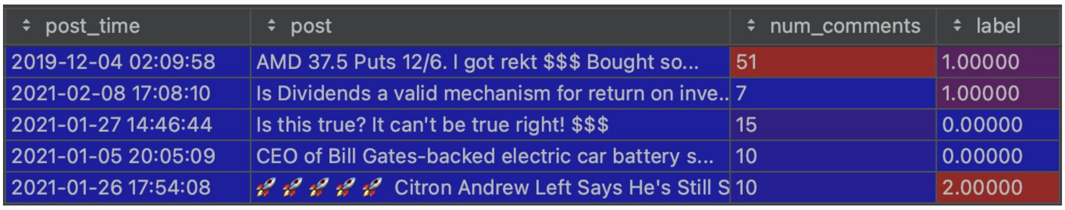
4 Results

Compared to the *baseline* model results, the *comment-weighted naive model* yielded a modest improvement in the F1-score. The *comment-weighted fine-tuned model* achieved a slight improvement on the baseline but performed worst than the *naive* model. We believe this happened because the labels (the day's market sentiment) passed to the optimization did not accurately portray the post's sentiment itself. For this reason we turned to the *unified-signal model* which led to further performance improvements (see Table 1).

5 Conclusions

Our work demonstrates the potential of using social media data and advanced language models for stock market trend analysis and prediction. The results showed that the final "semantic" approach of the *unified-signal model* achieved the best performance in predicting stock market trends. However, as these results do not surpass random-level choice, there is plenty of room for improvements with exhaustive hyper-parameter search and more compute power. The innovative dataset we curated, involving the collection, preprocessing, and labeling of a substantial amount of data, can be invaluable to future researchers in training more advanced models.

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest



post_time	post	num_comments	label
2019-12-04 02:09:58	AMD 37.5 Puts 12/6. I got rekt $$$ Bought so...	51	1.00000
2021-02-08 17:08:10	Is Dividends a valid mechanism for return on inve..	7	1.00000
2021-01-27 14:46:44	Is this true? It can't be true right! $$$	15	0.00000
2021-01-05 20:05:09	CEO of Bill Gates-backed electric car battery s...	10	0.00000
2021-01-26 17:54:08	🚀 🚀 🚀 🚀 🚀 Citron Andrew Left Says He's Still S	10	2.00000

Figure 1: A random sample from our created dataset. Index: post creation time; Data: post (title and content, pre-tokenization), number of comments, and the generated label.

Model	F1 score
Baseline	0.1486
Comment-weighted naive model	0.2657
Comment-weighted fine-tuned model	0.1966
Unified-signal model (50 epochs)	0.2843
Unified-signal model (150 epochs)	0.3153

Table 1: The F1-scores of the different suggested models.

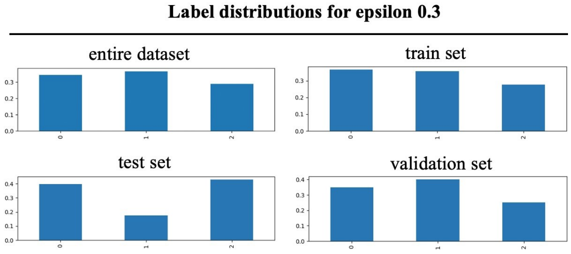


Table 2: Label distribution of each data split for $\epsilon = 0.3$.

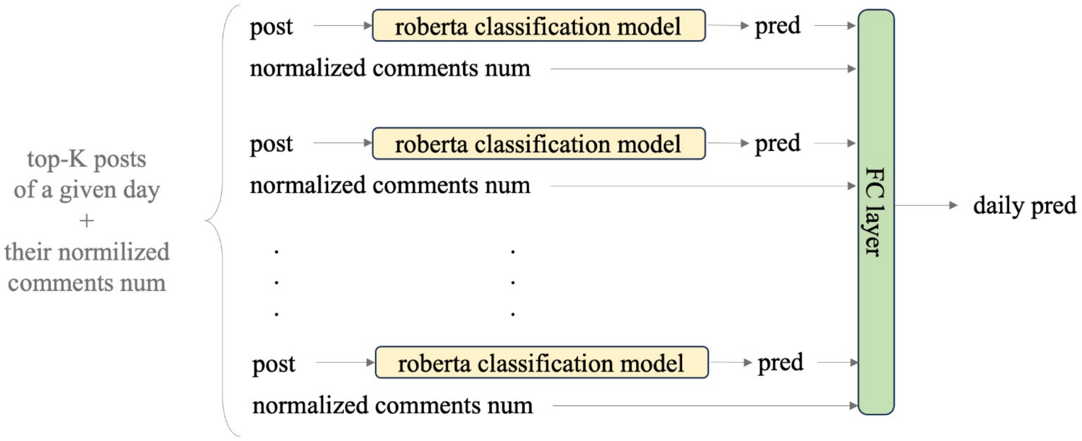


Figure 2: A block diagram of our new *Unified-signal model* architecture.

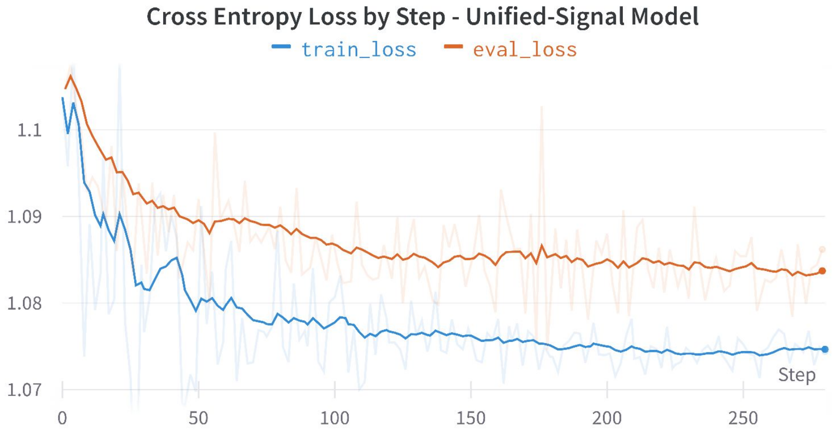


Figure 3: Loss over the train and validation sets of our *Unified-signal model* during the training process.

References

Daniel Bradley, Jan Hanousek Jr, Russell Jame, and Zicheng Xiao. 2021. Place your bets? the market consequences of investment research on reddit's wallstreetbets.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Sai Vikram Kolasani. 2020. Predicting stock movement using sentiment analysis of twitter feed with neural networks. volume 8, pages 309–319. Scientific Research Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system*, pages 1345–1350. IEEE.

Divyasikha Sethia et al. 2022. Stock price prediction using news sentiment analysis. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6.

Matej Steinbacher. 2023. Predicting stock price movement as an image classification problem.