A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Meme Stock Prediction

Anonymous ACL submission

Abstract

More and more investors and machine learning models rely on social media (e.g., Twitter and Reddit) to gather information and predict certain stocks' prices (meme stock). However, text-based models are known to be vulnerable to adversarial attacks, but whether stock prediction models have similar adversarial vulnerability is underexplored. In this paper, we experiment with a variety of adversarial attack configurations to fool three stock prediction victim models (*StockNet, FinGRU, FinLSTM*). We address the task of adversarial generation by solving combinatorial optimization problems with semantics and budget constraints. Our results show that the proposed attack method can **achieve consistent success rates**, with capabilities of causing **thousands of dollars** loss (with *Long-Only Buy-Hold-Sell* investing strategy) by simply concatenating a perturbed but semantically similar tweet.

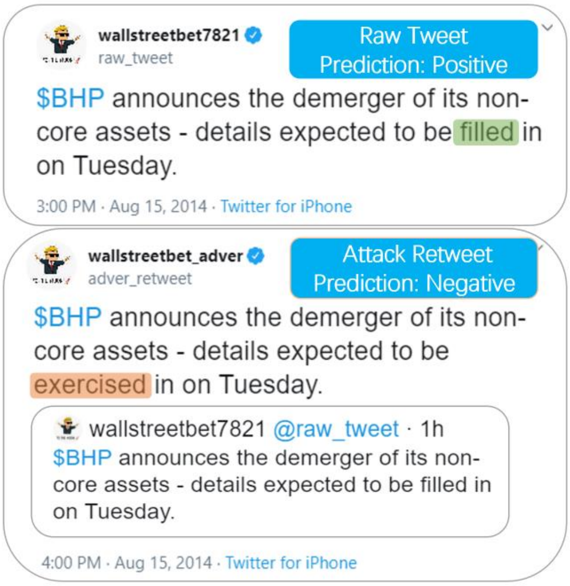

Figure 1: An adversarial sample with *concatenation attack* and *replacement-perturbation* on *Stocknet* as victim model. (Top) benign tweet leads to Stocknet predicting stock going up; (Bottom) adversarial retweet leads to Stocknet predicting stock going down.

1 Introduction

The advance of deep learning based language models are playing a more and more important role in the financial context, including convolutional neutral network (CNN) (Ding et al., 2015), recurrent neutral network (RNN) (Minh et al., 2018), long short-term memory network (LSTM) (Hiew et al., 2019; Sawhney et al., 2021), graph neutral network (GNN) (Sawhney et al., 2020a,b), transformer (Yang et al., 2020), autoencoder (Xu and Cohen, 2018), etc. For example, Antweiler and Frank (2004) find that comments on Yahoo Finance can predict stock market volatility after controlling the effect of news. Cookson and Niessner (2020) also show that sentiment disagreement on Stocktwits is highly related to certain market activities. Readers can refer to these survey papers for more details (Dang et al., 2020; Zhang et al., 2018; Xing et al., 2018). It is now known that text-based deep learning models may be vulnerable to adversarial attacks (Szegedy et al., 2013; Goodfellow et al., 2014). The perturbation can be done at the sentence level (e.g., Iyyer et al., 2018; Ribeiro et al., 2018) or the word level (e.g., Zhang et al., 2019; Alzantot et al., 2018; Zang et al., 2020; Jin et al., 2020; Lei et al., 2018). We are interested in whether such adversarial attack vulnerability also exists in stock prediction models, as these models embrace more and more user-generated public data (e.g., Twitter, Reddit, or Stocktwit (Xu and Cohen, 2018; Sawhney et al., 2021)). The adversarial robustness may be a more critical topic in the context of stock prediction as any one can post perturbed tweets to influence predicting models. As one example, a fake news ("Two Explosions in the White House and Barack Obama is Injured") posted by a hacker using the AssociatedPress's Twitter account on 04/23/2013 erased $136 billion in stock market in just 60 seconds (Fisher, 2013).

In this work, we take the attack's physical implementation feasibility into the design consideration

—we aim to maximize the attack success rate while also preserving semantic meaning for the newly generated tweets so that potential human readers and models can not detect our adversarial tweets. To achieve that, we consider the adversarial tweet generation task as a combinatorial optimization problem. Also, as we believe it is not feasible to inject the adversarial data into the training dataset, we mimic a re-tweet or comment function on social media to feed the adversarial samples into the prediction dataset, inspired by concatenation attack design (Jia and Liang, 2017). As shown in Fig. 1, we locate a tweet, identify the token, perturb it, and inject this new tweet back to the prediction data by posting it as a comment or retweet with the same stock ticker (BHP is the ticker of BHP Group).

We then examine our attack method on three stock prediction victim models: **Stocknet** (Xu and Cohen, 2018), **FinGRU** (Cho et al., 2014), **FinLSTM** (Hochreiter and Schmidhuber, 1997) with both *attack success rate* and potential *profit and loss* as two evaluation metrics. Results show that our attack method design can consistently achieve good success rate on the three victim models. More astonishingly, the attack can cause an additional loss of $2,300 to $3,200 dollars, if the investor trades on model predictions with initial $10,000 on day 1 (Fig. 3). We conclude the paper with an analysis of the result.

2 Adversarial Attack on Stock Prediction Models with Tweet Data

Stock prediction with tweet data. Massive amountd of texs data are generated by billions of users on Twitter every day. And investors often use the Twitter *cashtag* function (a $ symbol followed by a ticker) to organize their particular thoughts around one single stock, e.g., $AAPL. Financial organizations and institutional investors often ingest the massive text data in real time and incorporate them or their latent representation into their stock prediction models.

Attack model: Adversarial tweets. In the case of Twitter, adversaries can post malicious tweets which are crafted to manipulate downstream models that take them as input. We propose to attack by posting these malicious tweets as re-tweets or comments on Twitter and other social media platforms, so that these newly generated text could be identified as relevant and being absorbed by the model only in the post-training prediction period. For example, as shown in Fig 1, the original authentic tweet posted by the user *wallstreetbet7821* was "*$BHP announces the demerger of its non-core assets - details expected to be filled in on Tuesday.*" and the model predicts the price goes up; But an adversarial sentence could be "*$BHP announces the demerger of its non-core assets - details expected to be exercised in on Tuesday.*". With this message added to the prediction data, the model predicts the price goes down.

The proposed attack method takes the practical implementation into its current design consideration, thus has many advantages. First, the adversarial tweets are crafted based on carefully-selected relevant tweets, so they are more likely to pass the model's data processing filter and enter the inference data corpus. Secondly, adversarial tweets are optimized to be semantically similar to original tweets so that they are not counterfactual and may very likely fool human sanity checks as well as the Twitter's content moderator mechanisms.

Attack generation: Hierarchical perturbation. The challenge of our attack method centers around how to select the optimal tweets and the token perturbations with semantic similarity constraints. In this paper, we formulate the task as an *hierarchical perturbation* consisting of three steps: *tweet selection*, *word selection* and *word perturbation*. In the first step, a set of optimal tweets is first selected as target tweets to be perturbed and retweeted. The number of tweets are determined by the retweeting budget. Traditional attack modifies benign text directly (**manipulation attack**) and used them as model input; However, in our case, adversarial retweets enter the model along with benign tweets (**concatenation attack**). It is more realistic as malicious Twitter users can not modify others' existing tweets, but rather to re-tweet it with a comment. Consequently, the selected tweets could be different between the two attack modes.

For each target tweets in the target set, the word selection problem is then solved to find one or more best sites to apply perturbation, depending on word budget. Word budget quantifies the strength of perturbation within each tweet. How should we perturb the target words? We consider word replacement and deletion as two different approaches for word perturbation. In the case of replacing perturbation, the final step is to find the optimal candidate for the replacement. Synonym as replacement is widely adopted in the word-level attack since it is a natural choice to preserve semantics (Zang et al., 2020; Dong et al., 2021; Zhang et al., 2019; Jin et al., 2020). Therefore, we replace target words by their synonyms chosen from synonym sets which contains semantically closest words measured by similarity of the GLOVE embedding (Jin et al., 2020). The proposed hierarchical perturbation can then be cast as a combinatorial problem for tweet selection, word selection and replacement selection. To solve the resulting combinatorial optimization problem, we follow the convex relaxation approach developed in (Srikant et al., 2021). Specifically, the Boolean variables (for tweet and word selection) would be relaxed into the continuous space so that they can be optimized by gradient-based methods over a convex hull. There exist two main implementations of the optimization-based attack generation method: *joint optimization* (JO) solver and *alternating greedy optimization* (AGO) solver. JO calls projected gradient descent method to optimize the tweet and word selection variables and word replacement variables simultaneously. AGO uses an alternative optimization procedure to sequentially update the discrete selection variables and the replacement selection variables.

3 Experiments

Dataset & Task. We evaluate our adversarial attack using an stock prediction dataset (Xu and Cohen, 2018). The dataset contains both tweets and historical prices (e.g., open, close, high, etc) for 88 stocks of 9 industries. The data sampling period spans from 01/01/2014 to 01/01/2016. We follow the same data processing procedure and task formulation: the stock prediction task is considered as binary classification; a stock going up more than 0.55% in a day is labeled as positive, and going down more than -0.5% is labeled as negative, and the minor moves in between are filtered out.

In the experiments, we name our attack mechanism as *concatenation attack* whereas the traditional attack mechanism as *manipulation attack*. It is worth to separate the two attack formulations and compare their performance since they differ on the philosophy of searching adversarial tweets. For example, suppose that the tweet in Figure 1 posted by *wallstreetbet7821* is the most important predictor for the victim models, manipulation attack can directly amend the original tweet to mitigate its influence. However, concatenation attack has to create a new retweet to offset its impact. Such difference leads to different adversarial generation and attack performances.

Evaluation metrics. As aforementioned, we evaluate the attack performance on three victim models (**Stocknet** (Xu and Cohen, 2018), **FinGRU** (Cho et al., 2014), **FinLSTM** (Hochreiter and Schmidhuber, 1997)) on a binary classification task. Attack performance is evaluated on correctly classified instances by two metrics: *Attack Success Rate* (ASR) and victim model's *F1* drop after attack. ASR is defined as the percentage of the attack efforts that make the victim model misclassify the instances that are originally correctly classified. F1 indicates the prediction performance of the victim model, and the pre-attack F1 is 1. The drop of the F1 score of a model demonstrates the success of the attack method. More successful attack leads to higher ASR and lower post-attack F1.

Last but not least, we also use ***Profit and Loss*** as an additional metric. This widely-used financial indicator measures the profitability of a trading strategy. There are many trading strategies can be used together with a binary classification model, and in our paper, we use the simple ***Long-Only Buy-Hold-Sell*** strategy (Sawhney et al., 2021; Feng et al., 2019). This trading strategy *buy* stock(s) on Day T if the model predicts these stocks go up on Day $T + 1$, *hold* for one day, and *sell* these stocks the next day no matter what prices will be, and repeat it. It does not *short* a stock even when the model predict a negative move in the second day. Assume an investor's initial assets are $10,000 dollars, and accumulate profits and losses for each trade action, we can then calculate the final *profit and lost* for a model.

4 Results

Effect of attack budget. First, we report the effect of different attack budgets on the attach performance in Fig. 2. We observe that the more budgets allowed (perturbing more tweets and words), the better attack performance, but the increase is not significant. Moreover, the attack performance becomes saturated if we keep increasing the attack budget, thus in the following analysis we only show the the case that budgets are equal to 1.

Attack performance under single perturbation. The experiment results for the concatenation attack with word replacement perturbation mechanism is shown in Table 1 (with tweet and word budgets

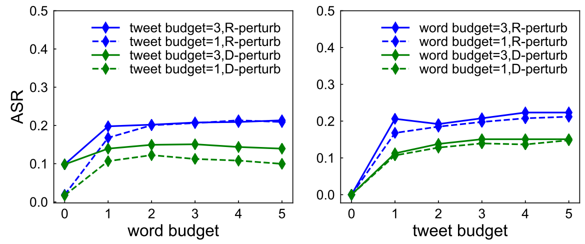

Figure 2: Effect of attack budgets on ASR with Stocknet as victim model and with JO solver. r-perturb: word replacement; d-perturb: word deletion.

both as 1). As we can see, for both JO and AGO optimization methods, ASR increase by roughly 10% and F1 drop by 0.1 on average in comparison to RA. Such performance drop is considered significant in the context of stock prediction given that the state-of-the-art prediction accuracy of interday return is only about 60%.

Model	ASR(%)				F1			
	NA	RA	JO	AGO	NA	RA	JO	AGO
Stocknet	0	4.5	**16.8**	11.8	1	0.96	**0.84**	0.88
FinGRU	0	5.1	**16.4**	14.1	1	0.95	**0.85**	0.87
FinLSTM	0	11.9	16.5	**19.7**	1	0.89	0.85	**0.78**

Table 1: Performance of the various adversarial attacks. NA: no attack; RA: random attack; JO: joint optimization; and AGO: alternating greedy optimization.

Effect on profit and loss. The ultimate measure of a stock prediction model's performance is profitability. Figure 3 plots the profit and loss of the trades with and without an attack. The attacks are optimized by JO solver on stocknet, and the results on the other two victim models are listed in Appendix. Net values of three scenarios are set as $10,000 at the beginning. Even a single word replacement on one tweet can cause a $3.2K additional loss in this benchmark dataset. Our result alerts investors who use text-based stock prediction models.

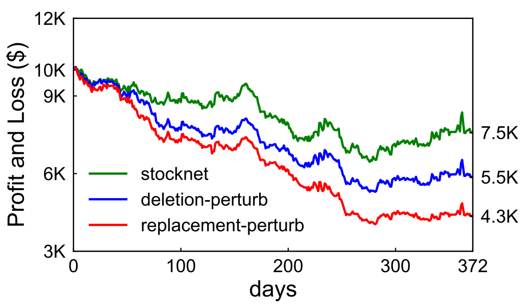

Figure 3: Effect on *Profit and Loss* with stocknet as victim model using a *Long-Only Buy-Hold-Sell* strategy. Green line: trade using stocknet prediction without attack; Blue line: deletion perturbation with concatenation attack; Red line: replacement perturbation.

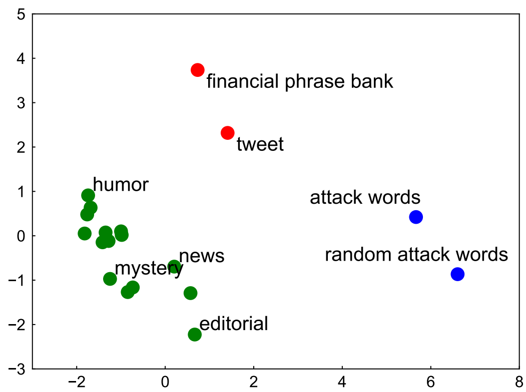

Figure 4: Corpora clusters. 18 corpora are grouped into 3 clusters based on features from LIWC. Principal component analysis is applied to the features to find the first 2 principal components, which are then used as x-axis and y-axis to generate this figure.

Attack word analysis. To qualitatively understand what kinds of words and tweets are being selected in the perturbation and retweet, we compare our tweet corpus and the selected word replacements with 15 corpora of different genres in Brown corpus via Linguistic Inquiry and Word Count program (LIWC) (Tausczik and Pennebaker, 2010). As Brown corpus does not have a financial genre, we also use Financial Phrase Bank (Malo et al., 2014). We then run K-means clustering these 18 corpus based on the feature matrix from LIWC. As shown in Figure 4, financial corpora (red), Brown general word corpus (green), and attack words (blue) are grouped into three clusters, indicating the inherent difference of those text genres. Moreover, we observe that target words identified by our solvers (red "tweet" and blue "attack words" dots) are closer to financial corpora than "random attack words".

5 Conclusion

In summary, we hypothesize the text-based stock prediction models are also vulnerable to adversarial attack, and we prove it by formulating a new adversarial attack task on a financial tweet dataset and three victim models. The experiment results demonstrate that our adversarial attack mechanism is consistent in attacking various prediction models. With one single word replacement on one tweet, the attack can cause a $3,200 additional loss to a $10,000 investment portfolio. Through studying stock prediction models' vulnerability, our goal is **to raise awareness for the community, and to develop more robust empirical models in the financial industry.**

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Werner Antweiler and Murray Z Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

J Anthony Cookson and Marina Niessner. 2020. Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.

Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural language word substitutions. *arXiv preprint arXiv:2107.13541*.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. 2019. Temporal relational ranking for stock prediction. *ACM Trans. Inf. Syst.*, 37(2).

Max Fisher. 2013. Syrian hackers claim AP hack that tipped stock market by $136 billion. Is it terrorism? *Washington Post*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Joshua Zoen Git Hiew, Xin Huang, Hao Mou, Duan Li, Qi Wu, and Yabo Xu. 2019. Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G Dimakis, Inderjit S Dhillon, and Michael Witbrock. 2018. Discrete adversarial attacks and submodular optimization with applications to text classification. *arXiv preprint arXiv:1812.00151*.

Pekka Malo, Ankur Sinha, Pekka J. Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Dang Lien Minh, Abolghasem Sadeghi-Niaraki, Huynh Duc Huy, Kyungbok Min, and Hyeonjoon Moon. 2018. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *Ieee Access*, 6:55392–55404.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020a. Deep attentive learning for stock movement prediction from social media text and company correlations.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020b. Voltage: volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Shah. 2021. Fast: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2164–2175.

Shashank Srikant, Sijia Liu, Tamara Mitrovska, Shiyu Chang, Quanfu Fan, Gaoyuan Zhang, and Una-May O'Reilly. 2021. Generating adversarial computer programs using optimized obfuscations. *arXiv preprint arXiv:2103.11882*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Yumo Xu and Shay B Cohen. 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)*, pages 1970–1979.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of The Web Conference 2020*, pages 441–451.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

A Effect of Iteration Number

We experiment with the optimizer to perform gradient descent or greedy search for up to 10 rounds before yielding the final solution. To visualize the effect of iteration, we plot the loss trajectory and ASR along with the optimization iterations in Figure 5. We also collect the average model loss of attack instances at each iteration, and then normalize the loss to set the initial loss as 1. Therefore, the loss trajectory visualization reveals the percentage loss drop during the optimization. We consider two different perturbations (replacement and deletion) under concatenation attacks. The attack is optimized with the JO solver.

The three charts on the first row of Figure 5 show that optimizations on all three victim models quickly converge after 4 iterations in our experiment. Accordingly, ASRs rise gradually during the first 4 iterations, but then flattens or even slides afterward. Such results suggest that our optimizer solvers can find the convergence in just a few iterations. Therefore, it makes our attack computationally effective, and insensitive to hyperparameter of iteration number.

B Supplemental Experiment Results

We report results for concatenation attack with only the *replacement perturbation* result in the main text in Table 1. Here we also report results for the *deletion perturbation* in Table 2. Attacks conducted via deletion perturbation in general performs worse than the replacement perturbation results. We observe ASRs via JO and AGO fall by 5.1% and 4.1% respectively compared with the replacement perturbation. Accordingly, F1 slightly increases as attack performance worsens. There is no significant difference between the two optimizers (JO and AGO) in the case of deletion perturbation, but JO is preferable in terms of optimization efficiency.

Moreover, we also simulate the trading profit and loss based on FinGRU and FinLSTM. For the sake of consistency, the two models are under concatenation attack with replacement perturbation. The results are illustrated in Figure 6. Same as our main results, the attack is optimized by JO solver. The simulation results are reported in Figure 6, which provide further evidence for the potential monetary loss caused by our adversarial attack. Replacement perturbation again outperforms deletion perturbation in the case of FinGRU and FinLSTM.

C Regularization on Attack Loss.

The experiment results reported in the main text have a sparsity regularization. We also run ablation experiments that remove sparsity regularization. The results are consistent with our conclusion. Furthermore, inspired by (Srikant et al., 2021), we try smoothing attack loss to stabilize the optimization. We add Gaussian noise to optimization variables and evaluate the attack 10 times. The loss average is then used as the final loss for back-propagation. The results show that loss smoothing does not contribute to attack performance in our experiment as it does in (Srikant et al., 2021).

D Example of Adversarial Retweet

Table 3 reports 10 adversarial retweets generated in concatenation attack mode with JO and AGO solver and replacement perturbation. For all the examples, the victim model predicts positive outcomes originally, and but predicts negative outcomes after adding the adversarial retweet.

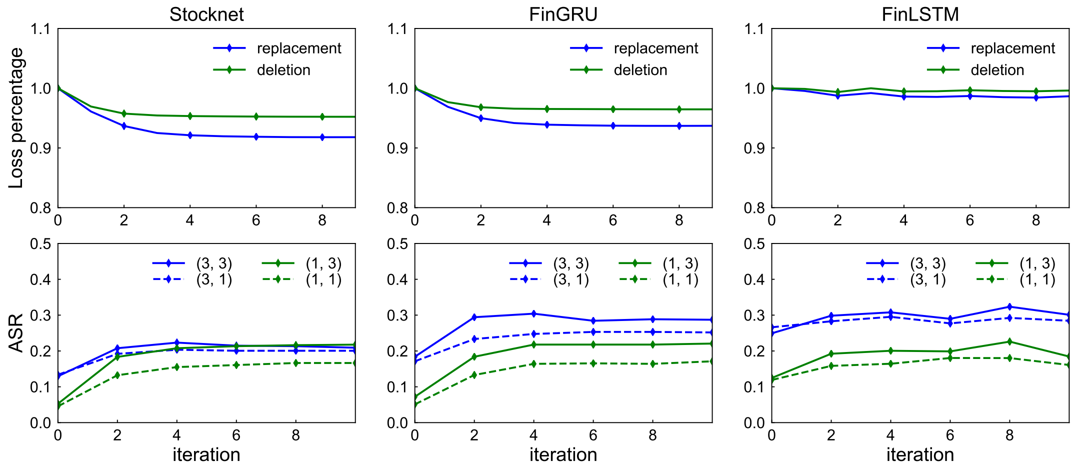

Figure 5: Iteration Number Effect on Prediction Loss and Attack Success Rate. The three plots on the first row show the loss trajectory during optimization for the three victim models, and the bottom row reports the ASRs trajectory. The legends for the bottom-row charts read as (tweet budget, word budget).

Model	**ASR(%)**				**F1**			
	NA	RA	JO	AGO	NA	RA	JO	AGO
Stocknet	0	3.6	12.1	11.0	1	0.97	0.89	0.89
FinGRU	0	4.0	10.2	10.6	1	0.96	0.85	0.91
FinLSTM	0	11.9	12.1	11.6	1	0.89	0.89	0.89

Table 2: Results for concatenation attack with deletion perturbation and budgets 1. NA and RA stand for no attack and random attack respectively, serving as benchmarks.

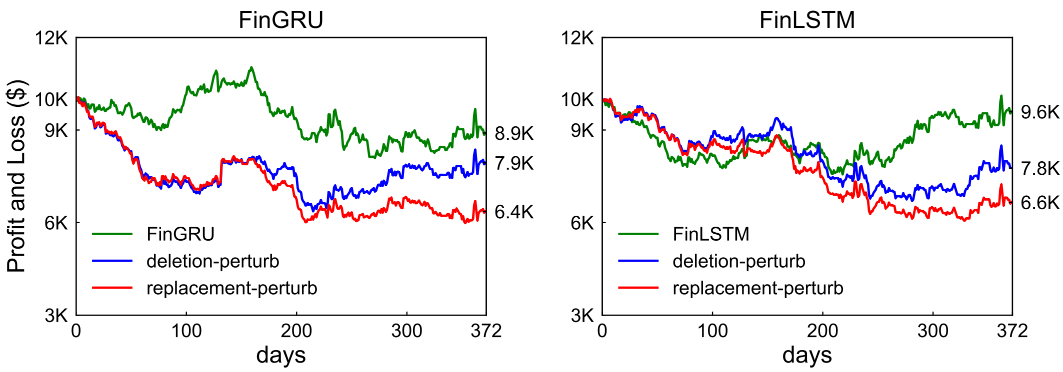

Figure 6: Effect on *Profit and Loss* of various perturbation methods on FinGRU and FinLSTM.

Adversarial Retweets Generated by AGO
Benign tweet: $BHP announces the demerger of its non-core assets - details expected to be filled in on Tuesday. **Adversarial retweet:** $BHP announces the demerger of its non-core assets - details expected to be exercised in on Tuesday
Benign tweet: Mover and Shakers... Losers- $KO $ABX $DD. Winners- $LAND $CHL $BHP. **Adversarial retweet:** Shoulder and Shakers... Losers- $KO $ABX $DD. Winners- $LAND $CHL $BHP.
Benign tweet: Latest information on #stocks like $TDS $DIS $CPWR $BLOX Give it a try. **Adversarial retweet:** Latest advance on #stocks like $TDS $DIS $CPWR $BLOX Give it a try.
Benign tweet: Top Buy Stocks for July 29: $AMX, $EXC, $NLY, $ORCL, $XRX. **Adversarial retweet:** Top unsettled Stocks for July 29: $AMX, $EXC, $NLY, $ORCL, $XRX.
Benign tweet: $JKHY #Stocks you might want to buy $HRS $SRCL $PCLN #singedinvestor. **Adversarial retweet:** $JKHY #Stocks you might confirm to buy $HRS $SRCL $PCLN #singedinvestor.
Adversarial Retweets Generated by JO
Benign tweet: #Stocks you migh want to Sell $CERN $CAT $PX $DO Try this http://t.co/95PUim108L. **Adversarial retweet:** #Stocks you migh announce to Sell $CERN $CAT $PX $DO Try this http://t.co/95PUim108L.
Benign tweet: The Most Accurate Stock Alert Service Made $240,000 Yesterday http://t.co/8WvnrVmkBS .. $SNY $SO $SOL. **Adversarial retweet:** The Most Accurate Stock preserving Service Made $240,000 Yesterday http://t.co/8WvnrVmkBS .. $SNY $SO $SOL.
Benign tweet: Is it breakout time on $HON $CF $LEG $IBM Give it a try. **Adversarial retweet:** Is it appealing time on $HON $CF $LEG $IBM Give it a try.
Benign tweet: Latest information on #stocks like $PETM $UTX $BRCD $CI #moneymanagement. **Adversarial retweet:** Latest discovery on #stocks like $PETM $UTX $BRCD $CI #moneymanagement.
Benign tweet: $BABA actually showing signs of life...would love a move back toward 90 although seems unlikely at moment. **Adversarial retweet:** $BABA actually showing signs of life...would love a move back toward 90 although seems unlikely at playday.

Table 3: Ten examples of adversarial retweets generated by concatenation attack