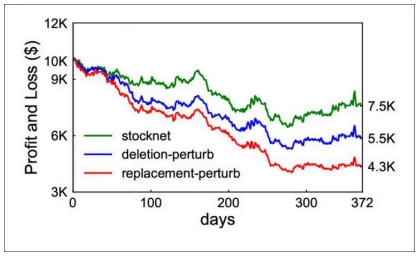
Devaluing Stocks With Adversarially Crafted Retweets

Originally published at https://www.unite.ai/devaluing-stocks-with-adversarially-crafted-retweets/

A joint research collaboration between US universities and IBM has formulated a proof-of-concept adversarial attack that's theoretically capable of causing stock market losses, simply by changing one word in a retweet of a Twitter post.



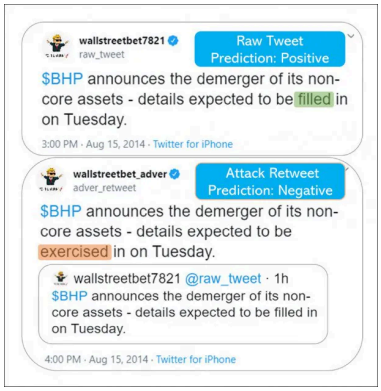
In one experiment, the researchers were able to hobble the Stocknet prediction model with two methods: a manipulation attack and a concatenation attack. Source: https://arxiv.org/pdf/2205.01094.pdf

The attack surface for an adversarial attack on automated and machine learning stock prediction systems is that a growing number of them are relying on organic social media as predictors of performance; and that manipulating this 'in-the-wild' data is a process that can, potentially, be reliably formulated.

Besides Twitter, systems of this nature ingest data from Reddit, StockTwits, and Yahoo News, among others. The difference between Twitter and the other sources is that retweets are editable, even if the original tweets are not. On the other hand, it's only possible to make additional (i.e. commentary or related) posts on Reddit, or to comment and rate – actions which are rightly treated as partisan and self-serving by the data sanitation routines and practices of ML-based stock prediction systems.

In one experiment, on the Stocknet prediction model, the researchers were able to cause notable drops in stock value prediction by two methods, the most effective of which, manipulation attack (i.e. edited retweets), was able to cause the most severe drops.

This was effected, according to the researchers, by simulating a single substitution in a retweet from a 'respected' financial Twitter source:



Words matter. Here, the difference between 'filled' and 'exercised' (not an overtly malicious or misleading word, but just about categorized as a synonym) has theoretically cost an investor thousands in stock devaluation.

The paper states:

> *'Our results show that the proposed attack method can achieve consistent success rates and cause significant monetary loss in trading simulation by simply concatenating a perturbed but semantically similar tweet.'*

The researchers conclude:

> *'This work demonstrates that our adversarial attack method consistently fools various financial forecast models even with physical constraints that the raw tweet can not be modified. Adding a retweet with only one word replaced, the attack can cause 32% additional loss to our simulated investment portfolio.*
>
> *'Via studying financial model's vulnerability, our goal is to raise financial community's awareness of the AI model's risks, so that in the future we can develop more robust human-in-the-loop AI architecture.'*

The paper is titled *A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction*, and comes from six researchers, hailing variously from the University of Illinois Urbana-Champaign, the State University of New York at Buffalo, and Michigan State University, with three of the researchers affiliated to IBM.

Unfortunate Words

The paper examines whether the well-studied field of adversarial attacks on text-based deep learning models are applicable to stock market prediction models, whose forecasting prowess depends on some very 'human' factors which can only be roughly inferred from social media sources.

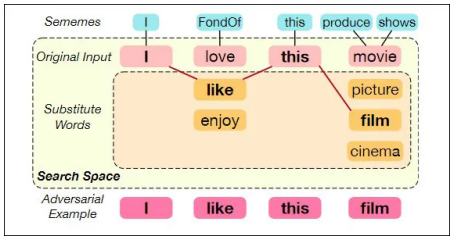
As the researchers note, the potential of social media manipulation to affect stock prices has been well-demonstrated, though not yet by the methods proposed in the work; in 2013 a malicious Syrian-claimed tweet on the hacked Twitter account of the Associated Press wiped $136 billion USD of equity market value in around three minutes.

The method proposed in the new work implements a concatenation attack, which leaves the original tweet untouched, whilst misquoting it:

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.

From the supplementary material for the paper, examples of re-tweets containing substituted synonyms that change the intent and significance of the original message, without actually distorting it in such a way that humans or simple filters might catch – but which can exploit the algorithms in stock market prediction systems.

The researchers have approached the creation of adversarial retweets as combinatorial optimization problem – the crafting of adversarial examples capable of fooling a victim model, even with a very limited vocabulary.



Word substitution using sememes – the 'minimum semantic unit of human languages'.
Source: https://aclanthology.org/2020.acl-main.540.pdf

The paper observes:

> *'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 semantically similar adversarial tweets as retweets on Twitter, so that they could be identified as relevant information and collected as model input.'*

For each tweet in a specially selected pool, the researchers solved the word selection problem under the constraints of word and tweet budgets, which place severe restrictions in terms of semantic divergence from the original word, and the substitution of a 'malicious/benign' word.

The adversarial tweets are formulated based on pertinent tweets that are likely to be allowed into downstream stock prediction systems. The tweet must also pass unhindered through Twitter's content moderation system, and must not appear to be counterfactual to the casual human observer.

Following prior work (from Michigan State University, together with CSAIL, MIT and the MIT-IBM Watson AI Lab), selected words in the target tweet are replaced with synonyms from a limited pool of synonym possibilities, all of which must be semantically very near to the original word, whilst maintaining its 'corrupting influence', based on inferred behavior of stock market prediction systems.

The algorithms used in the subsequent experiments were the Joint Optimization (JO) solver and the Alternating Greedy Optimization (AGO) solver.

Datasets and Experiments

This approach was tried out on a stock prediction dataset comprising 10,824 examples of pertinent tweets and market performance information across 88 stocks between 2014-2016.

Three 'victim' models were chosen: Stocknet; FinGRU (a derivative of GRU); and FinLSTM (a derivative of LSTM).

Evaluation metrics consisted of Attack Success Rate (ASR), and a drop in the victim model's F1 score after the adversarial attack. The researchers simulated a Long-Only Buy-Hold-Sell strategy for the tests. Profit and Loss (PnL) was also calculated in the simulations.

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**

Results of the experiments. Also see first graph at the top of this article.

Under JO and AGO, ASR rises by 10%, and the F1 score of the model drops by 0.1 on average, compared to a random attack. The researchers note:

> *'Such [a] 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%.*'

In the Profit-and-Loss tranche of the (virtual) attack on Stocknet, the results of adversarial retweets were also noteworthy:

> *'For each simulation, the investor has $10K (100%) to invest; the results show that the proposed attack method with a retweet with only a single word replacement can cause the investor an additional $3.2K (75%-43%) loss to their portfolio after about 2 years.'*

First published 4th May 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More