"Not another Ponzi!": Comparing adverts for cryptocurrency investment scams across platforms

DECISION: accept

AUTHORS: Gilberto Atondo Siu and Alice Hutchings

Summary of Reviews

- Review 1: 2 (4)
- Review 2: -1 (4)
- Review 3: -1 (4)

Reviews

Review 1

TOTAL SCORE: 2 (accept)
Overall evaluation: 2 (accept)
Reviewer's confidence: 4 (high)

The paper studies the advertisement of cryptocurrency investment scams on BitcoinTalk and some investment subreddits. The authors have previously performed a longitudinal analysis of these scams on BitcoinTalk, and here are focused on identifying improved textual classifiers, extending their analysis to Reddit, and searching for evidence of targeting at pensioners.

There are three classification tasks to be solved: thread type, post actor type, and lure type. The authors use sensible baselines and standards of evaluation throughout these tasks, and a reasonable selection of alternative methods are tested in each case. An active learning approach was trialled to improve performance of the best-performing approach, but there appears to be negligible (in some cases negative) impact on classifier performance. Overall, this evaluation is conducted to a high standard.

The evidence regarding scams targeted at pensioners is (understandably) limited. The authors find that there is not a higher prevalence of potential cryptocurrency scam advertisements in r/retirement than in r/investing, presenting evidence that at least one expected form of this targeting is not occurring. The authors present this result with due caution and discuss that an alternative explanation could be that r/retirement is well-moderated (i.e.,scammers are targeting this population, but not effectively in this venue). Another point I would suggest for consideration in future work is that targeting may occur through the advertisement content rather then venue (at the crudest level, explicitly mentioning pension pots in the advertisement). Establishing this may be difficult, especially as the sample of cryptocurrency scams identified from Reddit is currently small, but an initial exploration could involve examining whether the lures used in r/retirement differ from those used elsewhere.

While the selection of XGBoost relative to the other models trialled is well-defended, especially given the multi-class nature of the problem, the overall performance of the model (.82 F1 at scam prediction, .71 F1 at actor labelling) does leave room for improvement. The authors could perhaps make suggestions in the paper regarding what they think would be necessary to improve upon current performance. An error analysis may be helpful here. The confusion matrices in Figure 5 suggest that the classifier performance for true class label 1 (whichever this is,

see comment below) is close to random (recall 0.3, precision 0.46 for XGBoost-AL), suggesting that perhaps functional gains could be made by focusing on this class or combining it with another (perhaps class 2 if a joint definition of these is coherent).

Minor notes:
- Names such as 'section 2' should be capitalised.
- Typo, pg. 2 "discuss them on" -> "discuss them in"
- Typo, pg. 5 "not related with" -> "not related to"
- The numeric labels in Figure 5 are not clearly explained. 0 is presumably "Not investment scam", but the reader is left to guess which thread types the other labels refer to.
- The bar labels on Figures 1-3 are difficult to read.

Review 2

TOTAL SCORE: -1 (weak reject)
Overall evaluation: -1 (weak reject)
Reviewer's confidence: 4 (high)

This paper seeks to identify cryptocurrency scams posted on Reddit and Bitcointalk using data collected, in part by the author and in part, by somebody else. In addition, they seek to identify the actors posting the scams. They find that 5% of all posts are fraudulent. The authors then subject the data to a battery of machine learning models and highlight the one that fits best. Unsurprisingly, the winner is xgboost.

The authors also attempt to identify whether pensioners are targeted more, on average, relative to other victim types. However, they don't explain why they focus on this group, as opposed to any other vulnerable group. What was special about this group? Other people face the risk of losing a good deal of money, too.

Perhaps I missed it, but where did the ground truth about whether a post was an actual scam (or not) come from? i.e. how are the authors identifying the sample of fraudulent posts?

Overall, this is a fine descriptive paper, but it lacks strong motivation and conviction. What is the overall point of the work? What have we really learned from this? What issues does this resolve?

Review 3

TOTAL SCORE: -1 (weak reject)
Overall evaluation: -1 (weak reject)
Reviewer's confidence: 4 (high)

This study classifies advertisements for cryptocurrency investment scams from Reddit and Bitcointalk using various machine learning techniques. The authors identify the most efficient machine learning method for detecting investment scams, concluding that scam owners are the most common actors posting such scams, and the most prevalent luring type is the financial principle (~99%) compared to other principles.

While the study is interesting, the paper's focus is too methodological (i.e., determining which machine learning technique is better) for the WACCO venue. Additionally, the scope of the results is limited, preventing me from recommending the paper for publication.

Here are my main concerns:

- The authors seem to assume that all scam investments are Ponzi schemes. However, many scam investments are not Ponzi schemes; they are simply scams (with no returns whatsoever, regardless of the recruitment principles). A better conceptual definition of the subject matter is needed.

- I'm not sure why comparing different ML models is relevant, considering this venue, which focuses on understanding cyberattackers. Some of the content related to different ML models could be condensed into annotations in the text and the paper should could (should?) on what is relevant: the content of the posts, the strategies used, and how victims can be better protected based on the results.

Nevertheless, the scope of the results is limited. The lack of variance in the categories, once the machine learning models have predicted the data, limits the scope of the results: 97% are scam owners, and 99% use the financial principle. Isn't the finding that 97% of those who post scam investments are scam owners just a result of flagging scam investments?

Alsom the definition of the financial principle (99%) is: "The scammer takes advantage of users' 'need and greed' to promise enticing options and convince users to make an investment." I am uncertain whether this category does not encompass all investment scams just because it is a generic definition of investment scams.

The authors' third objective relates to "anecdotal evidence" suggesting that pensioners are targeted for investment scams. The authors should cite the source of this anecdotal evidence, as it is unclear where this hypothesis comes from (or develop further on the topic). Also, why would subreddits like /investment or /investing attract pensioners? The link between the hypothesis and the data is unclear, leaving the reader with the idea that specific subreddits gather pensioners. However, is this truly the case? As the text is written, I remain unconvinced.

Please, define HYIP before using the acronym.

The five percent mentioned in the abstract is not found in the text (the authors mention in the abstract that investment scams represent 5% of posts on both their datasets are potential scams).