DATASHEET FOR “Unveiling the Risks of NFT Promotion Scams”

Authors: Sayak Saha Roy, Dipanjan Das, Priyanka Bose, Christopher Kruegel, Giovanni Vigna, Shirin Nilizadeh

The dataset documentation was guided by the methodologies and questions proposed in:
Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.

1. For what purpose was the dataset created?

The dataset has been created (and was subsequently used) for training a machine learning classifier (PhishNFT) that utilizes various social media, and Web3 features to identify fraudulent NFT projects that are promoted on Twitter (Now X).

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by the research paper authors: Sayak Saha Roy, Dipanjan, Priyanka Bose, Christopher Kruegel, Giovanni Vigna and Shirin Nilizadeh.

3. Who funded the creation of the dataset?

See the ‘Acknowledgement’ section of the paper.

4. Any other comments?

None.

5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The dataset (groundtruth.csv) contains instances of fraudulent and benign NFT projects and their accompanying features from Twitter (such as post engagement metrics, and profile information), Cryptocurrency transaction information, and website information (of the project's homepage). These features can be utilized to train an ML-based classifier to differentiate between legitimate and malicious NFT projects.

6. How many instances are there in total (of each type, if appropriate)?

There are 494 malicious NFT projects and 649 benign NFT projects in the dataset.

7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset contains complete information of the features covered for each NFT project.

8. What data does each instance consist of?

The features provided by each instance of the data has been tabulated below:

Feature	Description
URL_ID	Random ID for the NFT Project's official URL
matches_official_slug_url	Check if URL matches that of a verified NFT project.
If_official_contract_address	Check if contract address of the project matches with a verified project's contract address.
No_of_ether_addresses	No. of contract addresses found from the website source-code
Twitter_link	Check if NFT project website has a vaid link.
Twitter_active	Check if Twitter account for the project is active.
Twitter_match	Profile matches with that of a verified NFT project.
Followers	No. of followers that the NFT Project's Twitter page has
Age	Age of the NFT Project's Twitter page.
Opensea_match	Check if NFT project has been published on Opensea
eth_track	Check if atleast one contract address has been flagged as suspicious by Etherscan/Solscan.

9. Is there a label or target associated with each instance?

Yes, the 'Label' column for each project denotes Malicious (1) or Benign (0)

10. Is any information missing from individual instances?
No.

11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

No.

12. Are there recommended data splits (e.g., training, development/validation, testing)?

Our model was trained using a 70:30 training test data split, but users can utilize any ratio of data split.

13. Are there any errors, sources of noise, or redundancies in the dataset?

No.

14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained and sufficient to train the model.

15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

No.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

17. Does the dataset identify any subpopulations (e.g., by age, gender)?

No.

18. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

No.

19. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? I

No.

20. How was the data associated with each instance acquired?

As explained in Section "Methodology" of our paper, we used the Twitter API V2 to collect a sample of tweets that promoted NFT projects, from which several data points were extracted. To collect transaction data, we queried EtherScan, BSCScan, and Solscan.

21. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

The data was collected using Twarc (Twitter API V2).

22. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

This dataset is not a part of a larger dataset.

23. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Data was collected by the paper's authors.

24. Over what timeframe was the data collected?

June 15th to August 20th,2022

25. Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

26. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

No, we used Twitter API V2.

27. Did the individuals in question consent to the collection and use of their data?

Given the public nature of the tweets sourced from Twitter, explicit consent from individuals for the collection and use of their tweets was not directly obtained. However, the collection and analysis were conducted in accordance with the platform's terms of service and use policy (https://twitter.com/en/tos), which states that publicly posted tweets may be accessed, analyzed, and shared by other users, researchers, and third parties. Additionally, we have anonymized any identifiable characteristics from the dataset.

28. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

No.

29. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The ground truth was manually evaluated by the authors.

30. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

No.

31. Is the software that was used to preprocess/clean/label the data available?

No.

32. Has the dataset been used for any tasks already?

The dataset was utilized to train our ML based model to detect NFT Scams on Twitter.

33. Is there a repository that links to any or all papers or systems that use the dataset?

Here is the pre-print of our paper: https://arxiv.org/pdf/2301.09806.pdf

34. What (other) tasks could the dataset be used for?

The dataset can be utilized to improve prevalent detection models against NFT-based threats.

35. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No.

36. Are there tasks for which the dataset should not be used?

No.

37. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

No.

38. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset has been shared on Zenodo and GitHub.
Link: https://zenodo.org/records/10884589

39. When will the dataset be distributed?

See above.

40. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

41. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

42. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

43. Who will be supporting/hosting/maintaining the dataset?

The authors of the paper.

44. How can the owner/curator/manager of the dataset be contacted?

The first author (Sayak Saha Roy) can be contacted at sayak.saharoy@mavs.uta.edu, contact information for authors can be found in our paper.

45. Is there an erratum?

No.

46. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

No.

47. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?

No.

48. Will older versions of the dataset continue to be supported/hosted/maintained?

Currently, this is the only version of the dataset.

49.If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset can be trained using any ML-based framework that relies on distinct features. We also provide scripts for training our Random Forest Classifier.