Analysis of Topics Related To Smart Contracts on Social Media

Giacomo **Ibba**[1], Marco **Ortu**[2] and Roberto **Tonelli**[3]

[1]*Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, Cagliari, Italy*

[2]*Department of Business School, University of Cagliari, V. Sant'Ignazio da Laconi 74, Cagliari, Italy*

[3]*Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, Cagliari, Italy*

Abstract

Blockchain technology popularity, particularly that of the Ethereum platform and smart contracts, keeps increasing over time and is one of the most exciting trends in computer science, economy, and finance. The importance of this technology is also witnessed by the fact that not only developers and researchers are interested in understanding smart contracts, but also generic users are turning their attention to blockchain. Notably, the focus of general users is on Non-Fungible-Tokens (NFT), whose popularity spread over the last three years and has become one of the most popular trends regarding blockchain technology. The users' interest started to apply also on social media (social networks, forums, blogs etc.), arising potentially interesting discussions about NFT and, in general, SCs programming. This work proposes an analysis of smart contract topics on Reddit and Twitter through a topic modeling approach to spot relevant arguments in users' discussions. Starting from a dataset containing subreddits and tweets (which were analysed separately), we built a transformer model to perform our topic modeling. Our results show that Reddit has several exciting topics related to smart contracts programming, such as games, Initial Coin Offering (ICO), Crowdsales, and complex Decentralized applications building. Twitter has mainly posts related to NFT giveaways and, generally, on NFTs promotion.

Keywords

Blockchain, Smart Contract, NFT, Topic Modeling, Natural Language Processing, Token, ICO, Smart Contracts Trends

1. Introduction

The Ethereum platform was officially launched in 2015, introducing a significant improvement and exciting innovation: Turing complete smart contracts (SC) programming. Over time, the Solidity programming language improved remarkably, allowing the development of increasingly sophisticated programs [1] [2]. Indeed, in the early years of the Ethereum lifecycle, SCs' purpose was to build gambling games, crowdfunding platforms [3], token implementation [4], and related programs [5]. Instead, Solidity evolution allows the creation of even kinds of Role Playing Games where it is possible to spend tokens to buy particular digital assets. Moreover, in 2017 Cryptokitties [1] was launched, contributing to the spread of Non-Fungible-Tokens (NFT) [6], which certifies the property of a specific digital asset. Particularly, NFTs caught the attention of Ethereum developers, researchers, and generic users; several games based on NFTs came

---

[1]https://www.cryptokitties.co/

out starting from right after the launch of Cryptokitties. In terms of interest, researchers and developers are more focused on the practical aspect of SCs; indeed, several researchers carried out works illustrating the potentialities and possibilities of SCs [7] [8]. Meanwhile, generic users are more interested in the economic intrinsic value of these programs and, generally, in the chance of gaining easy earnings by taking advantage of these contracts. The ever-growing interest in these programs is highlighted by the increasing discussion on social media forums about what smart contracts are and which possibilities are offered. Therefore, this work proposes an analysis of users' discussions about smart contracts, trying to spot the most exciting and relevant topics about SCs programming and categories. Remarkably, our aim is also to check possible matches with our findings from previous work [9] where we categorized a significant number of SCs, spotting classes such as:

- Token: programs that implement operations to create and deal with tokens.
- Certification and NFT (CNFT): programs implementing certification operations and Non Fungible Tokens.
- Bank: programs that implement functions to keep ETH safe (acting as a bank).
- Ether Lock Time Constraints (ELTC): programs that work as banks, but these particular contracts can keep ETH safe for a limited time.
- Bid: programs that implement Initial Coin Offering (ICO) and Crowdsales.
- Game: programs that implement role-playing games. With these particular SCs, buying items with spending tokens is possible.
- Gambling: programs that implement dice games, roulettes, blackjack, and other gambling games.
- Wallet: programs that implement wallet-like operations.
- Chain Management: contracts that implement functions to help users to interact easily with the blockchain.
- Money Investment: contracts that allow users to invest their money to gain interest tax.

Our findings confirm that the most popular categories are Token and CNFT programs in deployed SCs of the Ethereum main net [10]. In this work, we aim to check if the users' discussions focus more on tokens and Non-Fungible-Tokens as well. To perform this analysis, we extracted a sample of Reddit posts and tweets related to Ethereum smart contracts and performed a topic modeling analysis [11], taking advantage of the BERT transformer model.

2. State of Art

The Ethereum platform, particularly smart contracts, has caught the interest of researchers since the spread of this technology in 2017. Several studies have been conducted on SCs analysis [12] and classification. Apart from the first-ever taxonomy conceived [13], which nowadays is quite limited for reasons bound to the limits of the Ethereum platform in its early years, several studies analyse SCs' design patterns [14, 15] and other aspects related to classification. For instance, several works try to detect the design pattern of a specific contract by exploring its transactions network [16], and others perform analyses based on the source code and the

bytecode [17] [18]. Moreover, other works focus their aim on spotting vulnerabilities inside SCs [19] [20]. Still, in the current literature, there is not much previous work about Ethereum programs topics on social media, except for research discussing SCs issues [21]. Therefore, we decided to extend our research from the programs deployed on the Ethereum blockchain to users' discussions about SCs on social media. This kind of analysis could help to enrich the understanding of the opinion and the interest of two disjointed groups of users: one includes researchers and developers, the other is composed by generic users more interested in these programs' economic aspects and on the possibility of easy earnings. Moreover, this work could provide also hints from a statistical point of view, highlighting exciting topics about smart contracts and mostly confirming our findings of the massive SCs categorization, looking for possible matches on most popular programs categories and most popular topics discussed by users.

3. Dataset

The data collection process is a crucial step of this work and must be done carefully. Notably, we are studying the topics related to SCs discussions on social networks and forums. Therefore, the provenance of our corpus should come from those two types of platforms. We mainly decided to retrieve our data from Reddit and Twitter [2], respectively, one of the most popular forums and social networks. Potentially, Reddit [3] could be a precious source of data since it is a platform used by most different users; indeed, it is possible to find both SCs repositories [22] and subreddits dedicated to SCs programming. On the one hand, from Reddit, we expect to highlight exciting topics related to SCs, such as how to program specific types of contracts, build decentralized applications, information about Non-Fungible Tokens and related games, and other popular arguments related to Ethereum (differences between standards, ICO, Crodwdsale, and others). On the other hand, on Twitter, we expect to find mainly NFTs giveaways, or at most, we hope to spot sponsored crowdfunding platforms. Indeed, it is highly improbable to find any smart contracts programming topic on Twitter. To build our dataset, we took advantage of Phantom Buster [4]. This cloud-based platform allows developers to extract social media, forums, and e-mail data thanks to 'phantoms,' which are programs focused on a specific text-scraping task. Other use cases of Phantom Buster include data and e-mail enrichment, lead extraction, and social media automation, providing solutions for a wide range of social network analysis tasks. Specifically, we used the Reddit subreddit post extractor and the Twitter hashtag search export phantoms. The first one expects as input the subreddits in which we are interested (in our case, smart contracts), and as output, the phantom returns titles, links, the number of comments, the creation date, and other info for each extracted post. The second one expects an array of hashtags; then, the phantom will pull the tweets containing those hashtags, returning the tweet, the hashtags related to that tweet, the creation date, and other info. However, the Phantom settings must be done carefully since the research and the data extraction will be performed very specifically. Indeed, any specific Phantom Buster

---

[2]https://twitter.com/
[3]https://www.reddit.com/
[4]https://phantombuster.com/

solution includes instructions on what the tool needs as input and what returns as output, including a tutorial. Remarkably, in our case, the Twitter hashtag search export phantom needs to be run with different parameters to perform exhaustive research because it will look for tweets with the same array of hashtags given as input. Therefore, if we insert as hashtags '#smartcontracts#ico#crowdsale#nft#nfts#smartcontractgames#eth#token', the phantom will look exclusively for tweets containing this list of hashtags, which could be pretty limiting. Indeed, our analysis on Twitter was performed by looking for tweets containing a single hashtag and then running the algorithm cyclically combining several hashtags to ensure variety in our data. Concerning the time interval of the subreddits and tweets, we did not look only for a specific one since we thought that any particular year of the Ethereum lifecycle could be interesting for design patterns development and smart contracts evolution. Instead, we generally looked for every post from the oldest to the newest. At the end of the procedure, we retrieved a sample of subreddits from 2016 to 2022 and tweets posted from 2017 to 2022.

The idea is to perform two disjointed analyses: the first consists of training a transformer model, giving as input the corpus of subreddits; during the second one, instead, the input text will be composed of tweets. The reason behind this analysis lies in the dissimilarities of the two selected platforms; indeed, we expect to find various topics between the two platforms and, specifically for Twitter, only a small subset of the arguments found on Reddit. The next section illustrates the research methodology we followed to deliver our analysis.

4. Research Methodology

We already discussed the research methodology's first step; the data collection process, was explained in the previous section, but further background information on the phantom settings to scrape data are helpful. The phantom's options for Reddit data extraction are trivial since we only have to give as input the subreddit or subreddits in which we are interested. Indeed, we are interested only in subreddits discussing SCs. The settings for Twitter are pretty different since, in this case, the phantom needs a list of hashtags to perform the scraping task. So, since it could be rather constraining to use only '#smartcontracts' as a hashtag, we also set the phantom with other terms related to SCs such as #ico, #crowdsale, #nft, #solidity, #solidityprogramming, and others. The idea is that we probably have more chances to find topic variety by adding these key terms to our search. Once we have our data collected, the next step involves choosing the topic modeling technique to manage and analyze our data. One of the first topic modeling choices is usually the Latent Dirichlet Allocation Model [23](LDA). Nonetheless, an LDA analysis depends on the hyperparameters setting and does not consider the text's contextual nature. So, we opted for BERT [24], a method that takes advantage of transformers [25] and a class based TF-IDF (c-TF-IDF), which is used to calculate words interesting to each topic such to produce easily interpretable topics. Moreover, BERT leverages contextual embeddings and can capture the text's contextual nature. The BERTopic model will return the number of topics (associated with documents) and include keywords for each class, allowing us to deliver an accurate analysis of the arguments discussed by users on social. First, the corpus should be converted to numerical data, creating embeddings. Notably, we took advantage of 'sentence-transformers,' which include a set of models optimized to build vector representations considering semantic similarity. The

next step consists of dimensionality reduction since, generally, most clustering algorithms do not perform well when dealing with high dimensionality. We delivered dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) technique [26], which approximates data samples to a lower dimension, assuming they are evenly distributed over a topological space. We decided to cluster documents to group documents with similar topics and spot them within these clusters taking advantage of the hierarchical density-based clustering algorithm (HDBSCAN) [27], which extends the DBSCAN [28] algorithm but, as the name suggests, it converts the algorithm itself into a hierarchical one. After dimensionality reduction (if needed) and document clustering, we can build our topic representation using BERT and eventually reduce the topics if the number returned by the model is enormous. Once the model produces the topics, they must be interpreted, which should not be difficult since BERT creates easily interpretable topics. The whole process should be applied both for the Reddit corpus and the Twitter one since we want to keep the topics related to each platform disjointed. Once the topics' probabilities are associated with our documents, we can deliver relevant analysis, such as:

- The most and the less relevant topics
- The differences between the discussions of Reddit and Twitter users
- Discover if there are any matches between SCs categories in the Ethereum chain and topics discussed by social network users

5. Results

5.1. Reddit Analysis

As a result of the Reddit corpus analysis, we highlighted 14 different topics inside the users' discussions, though many documents were classified as outliers and not associated with any specific category. The colored bar chart on the y-axis represents the number of clusters and the color associated with each group. To compute the similarity between documents, we used the cosine measure. Therefore, we converted the similarity matrix to a distance measure to plot the distance between the clusters represented by the x and y axis of Figure 1. As highlighted in 1, Reddit's most discussed smart contracts topics concern Non-Fungible-Tokens and tokens. This result confirms the NFT category as the most popular SC trend and our previous work findings. It is exciting to observe the variety of topics found inside the corpus. Indeed, except for classic arguments and discussions such as the need for help, advice, and general questions about SCs programming and development, there are pretty exciting topics. Remarkably, games, smart contracts hacking, and audits are the most relevant. The 'Game' design pattern has been widespread, especially in the early years of the Ethereum lifecycle, when the development of gambling games could be a possible way of gaining ETH. Over time, developers start to build Role-Playing-Games where it is possible to buy particular digital assets by spending tokens. The topics spotted inside the Reddit corpus witness the further evolution of game programs; indeed, the users' focus is more aimed at those games where it is possible to collect NFTs, which is another proof of their popularity. The audit topic is exciting, too, because, as a term of economics
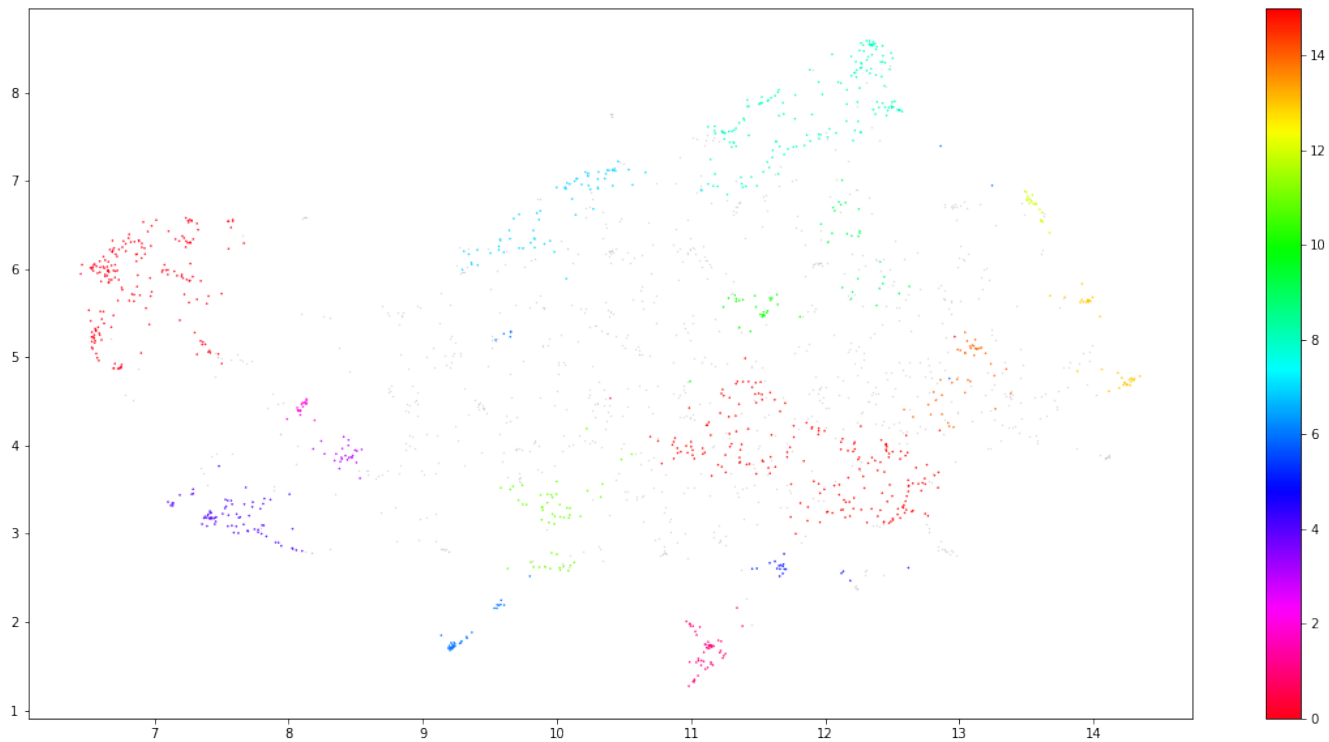
Figure 1: Result of document clustering

Topic	Samples	Description	Keywords
NFT	113	Discussions on Non-Fungible-Tokens	nft, nfts, erc721, collect, marketplace
Tokens	94	Discussions on regular Tokens	token, tokens, erc20, tokenomics, standard
Transactions	71	Discussions on transactions functioning	transactions, transaction, uniswap, pending, failed
Developers Search	69	Search for SCs developers	developer, hire, smart, contracts, company
SCs Developing	54	Tips for SCs development	development, tips, concurrent, smart, contracts
Audit on SCs	52	Discussions on SCs audit	audit, report, smart, contracts, security
SCs programming	48	Help and advice for SCs programming	smart, contract, advice, help, need
Games	47	Discussions on games where is possible to collect NFTs	cryptosoccr, game, nfts, collect, play
Solidity	40	General questions on Solidity programming	smart, contracts, solidity, learn, interview
Web3	36	General questions on Web3	web3, develop, dapp, js, solidity
NFTs minting	35	Discussions on NFTs minting	nft, nfts, mint, minting, how
Token standards	31	Discussions on ERC standards	erc20, erc721, erc1155, erc, standard
Hacking	31	Questions about SCs hacking	hack, prevent, dos, smart, contracts
Faucets	28	Discussions on testnets token faucet	ropsten, eth, rinkeby, goerli, faucet

Table 1

Table resuming the topics highlighted in the Twitter corpus

science, an audit should verify the fairness of the financial statement data and the correctness of business procedures. In terms of smart contracts, it is unclear what an audit should provide since, manually checking the subreddits, the only information available is relative to programs delivering an 'audit'. According to the available tools, a SC audit should provide quality code checking and look for any bugs with gas consumption. Therefore, a tool performing an audit

Topic	Samples	Description	Keywords
Ethereum	144	Tweets about Ethereum information and news	ethereum, blockchain, nft, fabric, ethereumnews
NFT	134	NFT giveaway and NFT sale	nft, giveaway, sale, nftcommunity, tag

Table 2
Table resuming the topics highlighted in the Twitter corpus

on SCs could be matched with a general smart contract checker and vulnerabilities detector. The last relevant topic concerns smart contract hacking and, in particular, information about how to avoid specific attacks on SCs, explanations of how famous hacks were delivered against precise contracts, and how to prevent the hijack of ETH transfer calls by malicious users. So, notably, the interest is focused on how to avoid SCs hacking and not the contrary.

5.2. Twitter Analysis

In terms of smart contracts, Twitter does not explicitly provide posts related to SCs but provides tweets that discuss the Ethereum blockchain in general. Others promote NTF giveaways and platforms that purposefully allow users to collect tokens, but most documents are separate from any specific topic. The main difference concerning Reddit is that the Twitter analysis returned only two clusters and, therefore, much less variety of topics 2. The lack of topics on Twitter concerning Reddit is not a surprise since the last one is more suitable for discussions, while the first is ideal for disclosing information. The reasons behind the few topics lie in the different categories of users using the two platforms; indeed, the Twitter platform is operated by a wide range of users with general interests, and apart from a few accounts, such as the Ethereum one, is pretty unusual to find a user promoting or managing information relative to smart contracts. Reddit, instead, has a subreddit dedicated to smart contracts and is more suitable for retrieving the information we are interested in since developers and researchers take advantage of this platform.

6. Conclusions and Future Work

The most exciting platform for a number and variety of topics between Reddit and Twitter is the first one. Notably, we saw how Twitter highlights arguments of interest as smart contracts auditing, games purposefully aimed to collect Non Fungible Tokens, and also critical topics such SCs vulnerabilities and hacking. Nevertheless, Twitter confirms the great importance and widespread of NFTs since many of the documents' arguments concern token and NFTs giveaways. Moreover, we found matches with our previous work [10], noting that NFT and Token are the most deployed categories of SC on Ethereum's main net and are also the most discussed topics by Reddit users. Given the exciting results, we aim to collect more data from Reddit and to spread our research to other social media and forums in future work, such as Facebook, Ethereum StackExchange, and Github, which could be potentially interesting, both for variety and for the number of topics. It could not be easy to look for potentially exciting data on Facebook because of the nature of this social network since it is more used for social interaction, videos, media, daily news, and picture sharing, considering that users' data could not be public and therefore not available. Nonetheless, it could be exciting to look for Facebook

groups discussing smart contracts and generally smart contracts topics, which would be the primary data source for potential analysis. GitHub and Ethereum StackExchange could be exciting data sources because of the information on the users' profiles, and the topics and issues shared on the platforms, where (compared to other social networks) it could be easier to find specific discussions about SCs. We do not intend to analyze social media as Instagram since users almost exclusively use it for media sharing, such as pictures and videos, and it is not suitable for topics and issues discussions. Moreover, as an extension of the work, we aim to perform sentiment analysis on data to know users' opinions about the smart contract technology and, more specifically, about the SCs' topics and the trend of the specific arguments spotted inside our dataset.

References

[1] S. N. Khan, F. Loukil, C. Ghedira-Guegan, E. Benkhelifa, A. Bani-Hani, Blockchain smart contracts: Applications, challenges, and future trends, Peer-to-peer Networking and Applications (2021) 1–25.

[2] L. Marchesi, M. Marchesi, R. Tonelli, Abcde—agile block chain dapp engineering, Blockchain: Research and Applications 1 (2020) 100002.

[3] V. Bracamonte, H. Okada, An exploratory study on the influence of guidelines on crowdfunding projects in the ethereum blockchain platform, in: G. L. Ciampaglia, A. Mashhadi, T. Yasseri (Eds.), Social Informatics, Springer International Publishing, Cham, 2017, pp. 347–354.

[4] M. di Angelo, G. Salzer, Tokens, types, and standards: Identification and utilization in ethereum, in: 2020 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS), 2020, pp. 1–10. doi:10.1109/DAPPS49028.2020.00001.

[5] G. Fenu, L. Marchesi, M. Marchesi, R. Tonelli, The ico phenomenon and its relationships with ethereum smart contract environment, in: 2018 International Workshop on Blockchain Oriented Software Engineering (IWBOSE), 2018, pp. 26–32. doi:10.1109/IWBOSE.2018.8327568.

[6] Q. Wang, R. Li, Q. Wang, S. Chen, Non-fungible token (nft): Overview, evaluation, opportunities and challenges, arXiv preprint arXiv:2105.07447 (2021).

[7] W. Zou, D. Lo, P. S. Kochhar, X.-B. D. Le, X. Xia, Y. Feng, Z. Chen, B. Xu, Smart contract development: Challenges and opportunities, IEEE Transactions on Software Engineering 47 (2019) 2084–2106.

[8] Z. Zheng, S. Xie, H.-N. Dai, W. Chen, X. Chen, J. Weng, M. Imran, An overview on smart contracts: Challenges, advances and platforms, Future Generation Computer Systems 105 (2020) 475–491.

[9] G. Ibba, M. Ortu, Analysis of the relationship between smart contracts' categories and vulnerabilities, in: 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), IEEE, 2022, pp. 1212–1218.

[10] G. Ibba, A smart contracts repository for top trending contracts, in: 2022 IEEE/ACM 5th International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB), IEEE, 2022, pp. 17–20.

[11] H. M. Wallach, Topic modeling: Beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 977–984. URL: https://doi.org/10.1145/1143844.1143967. doi:10.1145/1143844.1143967.

[12] G. A. Pierro, R. Tonelli, M. Marchesi, Smart-corpus: an organized repository of ethereum smart contracts source code and metrics, arXiv preprint arXiv:2011.01723 (2020).

[13] M. Bartoletti, L. Pompianu, An empirical analysis of smart contracts: platforms, applications, and design patterns, in: International conference on financial cryptography and data security, Springer, 2017, pp. 494–509.

[14] M. Wöhrer, U. Zdun, Design patterns for smart contracts in the ethereum ecosystem, in: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2018, pp. 1513–1520. doi:10.1109/Cybermatics.2018.00255.

[15] Y. Liu, Q. Lu, X. Xu, L. Zhu, H. Yao, Applying design patterns in smart contracts, in: S. Chen, H. Wang, L.-J. Zhang (Eds.), Blockchain – ICBC 2018, Springer International Publishing, Cham, 2018, pp. 92–106.

[16] T. Hu, X. Liu, T. Chen, X. Zhang, X. Huang, W. Niu, J. Lu, K. Zhou, Y. Liu, Transaction-based classification and detection approach for ethereum smart contract, Information Processing and Management 58 (2021) 102462. URL: https://www.sciencedirect.com/science/article/pii/S0306457320309547. doi:https://doi.org/10.1016/j.ipm.2020.102462.

[17] G. Tian, Q. Wang, Y. Zhao, L. Guo, Z. Sun, L. Lv, Smart contract classification with a bi-lstm based approach, IEEE Access 8 (2020) 43806–43816. doi:10.1109/ACCESS.2020.2977362.

[18] C. Shi, Y. Xiang, R. R. M. Doss, J. Yu, K. Sood, L. Gao, A bytecode-based approach for smart contract classification, arXiv preprint arXiv:2106.15497 (2021).

[19] W. Dingman, A. Cohen, N. Ferrara, A. Lynch, P. Jasinski, P. E. Black, L. Deng, Classification of smart contract bugs using the nist bugs framework, in: 2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), 2019, pp. 116–123. doi:10.1109/SERA.2019.8886793.

[20] R. Camino, C. F. Torres, R. State, A data science approach for honeypot detection in ethereum, CoRR abs/1910.01449 (2019). URL: http://arxiv.org/abs/1910.01449. arXiv:1910.01449.

[21] A. Ayman, S. Roy, A. Alipour, A. Laszka, Smart contract development from the perspective of developers: Topics and issues discussed on social media, in: International Conference on Financial Cryptography and Data Security, Springer, 2020, pp. 405–422.

[22] M. Ortner, S. Eskandari, Smart contract sanctuary (????). URL: https://github.com/tintinweb/smart-contract-sanctuary.

[23] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[24] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[25] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, arXiv preprint arXiv:2106.04554 (2021).

[26] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection

for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[27] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering., J. Open Source Softw. 2 (2017) 205.

[28] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: kdd, volume 96, 1996, pp. 226–231.