Tackling the Myriads of Collusion Scams on YouTube Comments of Cryptocurrency Videos

Sadat Shahriar
University of Houston,
Texas, USA
sshahriar@uh.edu

Arjun Mukherjee
University of Houston,
Texas, USA
arjun@cs.uh.edu

Abstract

Despite repeated measures, YouTube's comment section has been a fertile ground for scammers. With the growth of the cryptocurrency market and obscurity around it, a new form of scam, namely "Collusion Scam" has emerged as a dominant force within YouTube's comment space. Unlike typical scams and spams, collusion scams employ a cunning persuasion strategy, using the facade of genuine social interactions within comment threads to create an aura of trust and success to entrap innocent users. In this research, we collect 1,174 such collusion scam threads and perform a detailed analysis, which is tailored towards the successful detection of these scams. We find that utilization of the collusion dynamics can provide an accuracy of 96.67% and an F1-score of 93.04%. Furthermore, we demonstrate the robust predictive power of metadata associated with these threads and user channels, which act as compelling indicators of collusion scams. Finally, we show that modern LLM, like *chatGPT*, can effectively detect collusion scams without the need for any training.

1 Introduction

The most popular online video-sharing platform YouTube has seen a surge of scams and spam comments since its creation in 2005. Although measures have been taken, financial frauds, especially related to cryptocurrency investment have not been slowed down (Dig, Accessed: 2023-05-14). Scammers have adapted their tactics to circumvent the scam-detection algorithm by adopting the disguise of genuine users, engaging in seemingly ordinary conversations, and perpetrating a previously undocumented form of deceit known as the "Collusion Scam". Due to their facades, such scams frequently go unnoticed by automated detection systems, posing a significant threat to users who may unwittingly fall victim to such schemes. Consequently, it has become imperative to employ rigorous linguistic, psycholinguistic, and metadata analyses to effectively detect and combat these collusion scams.

The "Collusion Scam" can be defined as a fake conversation where the participants pretend to be beneficiaries of a person or an entity to entrap users for their monetary gain. Typically, a scammer or a group of scammers will share their success and gratitude in working with a person or entity. Often another group joins the conversation by pretending to be curious or newbies, and on later turns they also express to be a beneficiary. In this method, the scammers share the entity's handles or contact information to get around YouTube's rules. Figure 1 shows an example of the collusion scam where some scammers engage in a conversation by pretending to be a beneficiary of a cryptocurrency investment through a claim expert.

The rise of cryptocurrencies has not only attracted genuine enthusiasts and investors but has also unfortunately attracted a surge in fraudulent activities. The absence of comprehensive regulations, limited awareness among users, and the inherent obscurity of cryptocurrency transactions have created fertile ground for scammers to exploit unsuspecting individuals. One prominent avenue for scams in the cryptocurrency space is YouTube, where misleading and collusive comments on cryptocurrency videos can deceive and manipulate unsuspecting viewers.

YouTube's own machine learning algorithms deleted over 950 million comments in Q4, 2021 (9to, Accessed: 2023-1-21), however, the measures were not adequate because of the evolving nature of these scams. Due to YouTube's policy on spam comments, it often deletes comments that strictly violate the policy. For example, the comments that trick others into leaving the site for another one, offer monetary incentives, repetitive, links to coun-

Proceedings of Recent Advances in Natural Language Processing, pages 1066–1075
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_114

terfeits, etc (You, Accessed: 2023-1-21). However, collusion scam is a fairly new approach of scamming where multiple scamming strategies are used to deceive the user, and current spam filters are not able to detect these contents. Hence, there is an urgent need to address the pervasive issue of collusion scams to establish trust, combat the distortion in information exchange, and ensure a safer online environment for the cryptocurrency community and beyond.

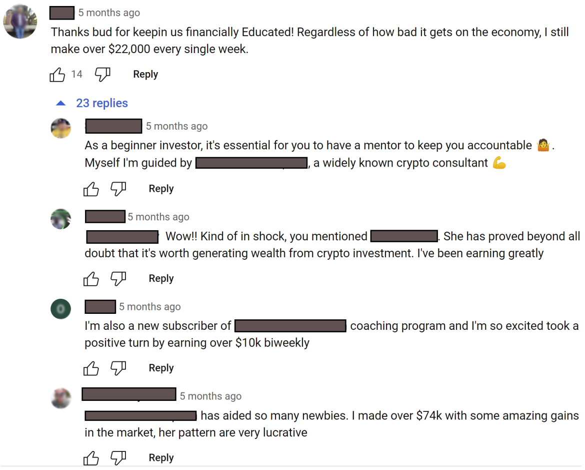
In this research, we collect 7,335 conversation threads (comment-replies) from 112 cryptocurrency-related YouTube videos. We manually label them for the presence or absence of collusive scams. Next, we delve into a comprehensive analysis of the linguistic patterns, as well as an exploration of the persuasive strategies employed within these conversations. We also analyze the collusion dynamics within a conversation by using a BERT-LSTM architecture. Furthermore, we explore how the collusion scam detection performance improves with the progression of the thread, and find that we can obtain 96.67% accuracy and 93.04% F1-score when utilizing the initial comment, and all subsequent replies in a conversation. Additionally, we explore how different metadata, like the timespan between the comments and replies, the number of *like* counts, age of the users' channels can provide strong cues for collusion scams. Finally, we examine the performance of *chatGPT* in the realm of collusion scam detection. The main contributions of our research can be summarized as follows:

- To the best of our knowledge, we build the first dataset for collusion scam detection in cryptocurrency-related YouTube videos
- We show how deep learning techniques can be useful in understanding the collusion scam dynamics
- We demonstrate the efficacy of leveraging metadata in collusion scam detection

The data is publicly available at https://github.com/sadat1971/YouTube_Collusion_Scam.

2 Related Works

Researchers explored several aspects of YouTube comments, such as, analyzing the user interactions, sentiment analysis, hate speech, and bias and misinformation (Thelwall et al., 2012; Bhuiyan et al., 2017; Döring and Mohseni, 2020; Jiang et al., 2019). A number of studies worked on spam detection in YouTube videos. Alberto et al. (2015) proposed a machine learning-based automated spam comment filtering system. Similar work has been conducted by Abdullah et al. (2018), Aiyar and Shetty (2018), and Das et al. (2020), highlighting the ongoing research efforts in this domain. Using network analysis, O'Callaghan et al. (2012) explored how spammers use multiple spam bots to post similar comments on multiple popular YouTube videos. However, while these studies have made notable contributions to combating spam, scams constitute a more sinister category. Due to their deceptive nature and nefarious objectives, it is imperative to undertake meticulous research specifically geared toward detecting scam comments.


[redacted] 5 months ago
Thanks bud for keepin us financially Educated! Regardless of how bad it gets on the economy, I still make over $22,000 every single week.

[redacted] 5 months ago
As a beginner investor, it's essential for you to have a mentor to keep you accountable 🤷.
Myself I'm guided by [redacted], a widely known crypto consultant 💪

[redacted] 5 months ago
[redacted] Wow!! Kind of in shock, you mentioned [redacted]. She has proved beyond all doubt that it's worth generating wealth from crypto investment. I've been earning greatly

[redacted] 5 months ago
I'm also a new subscriber of [redacted] coaching program and I'm so excited took a positive turn by earning over $10k biweekly

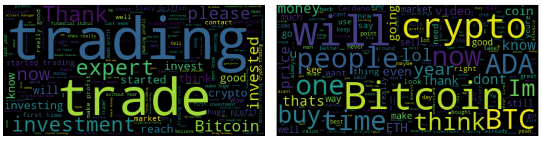
[redacted] 5 months ago
[redacted] has aided so many newbies. I made over $74k with some amazing gains in the market, her pattern are very lucrative

Figure 1: An example of Collusion Scam in YouTube .

There are some research initiatives around scams on YouTube. Tripathi et al. (2022) performed a comparative analysis of machine learning algorithms to detect monetary scam videos. Bouma-Sims and Reaves (2021) explored the metadata aspect of scam videos on YouTube. They found that scammers' accounts have lesser activity and scam videos have less longevity than non-scam videos. However, these works do not address cryptocurrency-related scam comments or collusion scams. Notably, researchers have explored bitcoin-related scam comments and relevant keywords on platforms like *Bitcointalk* (Atondo Siu et al., 2022). Other studies have also investigated cryptocurrency scams, albeit with a primary focus on Ponzi schemes and pump-and-dump schemes,

Category	# of threads	# of replies
Collusion Scam	1,174	20,341
Spam	332	1,428
Non-Scam	1,272	8,409
Unlabeled	4,557	5,933
Total	7,335	36,111

Table 1: Data collection in different categories for YouTube comment-replies threads.



(a) (b)

Figure 2: Word-cloud representation of the YouTube threads: a)collusion-scam b)non-scam

which differ from the intricacies of collusion scams. (Li et al., 2022; Nghiem et al., 2021; Mirtaheri et al., 2021). Ponzi schemes involve promising high returns to investors using funds from new participants, while pump-and-dump schemes manipulate asset prices through coordinated buying and selling. In contrast, collusion scams employ social and psychological strategies, such as mimicking regular conversations and leveraging social proof, to deceive users. Hence, the existing research lacks in effectively detecting and addressing the nuances of collusion scams.

3 Methodology

Our work involves a meticulous data collection process, labeling, and employing machine learning techniques to detect collusion scam.

3.1 Data Collection

The data collection process begins with YouTube searches, utilizing specific keywords like *Crypto Investment Suggestions*, *Bitcoin Suggestions*, *CNN Crypto News*, and *Fox Crypto News* to locate cryptocurrency-related videos. From each search results page, we retrieve the top ten videos that have accumulated at least 10,000 views. Furthermore, we identify popular YouTube channels offering cryptocurrency suggestions through a Google search, selecting the most informative ones, and gathering recent uploads with a minimum of 10,000 views. All view counts were recorded from their uploads up to January 10, 2023. In total, our dataset comprises 112 YouTube videos focused on cryptocurrency.

To collect the data, we leverage the *YouTube Data API v3*, utilizing various API calls such as *channels*, *comments*, and *commentThreads*. Due to the API limitations, allowing only 10,000 queries per day, the data collection process spanned multiple weeks. In total, we collect 7,335 threads with comments and 36,111 replies. Among the metadata, we collect the number of *likes* on comments, and replies, timestamps of postings, and the video published time. Additionally, we collect channel information for all users involved in the threads, encompassing details such as channel join dates, view counts, and subscriber counts.

3.2 Labeling

We manually annotate the dataset to indicate the presence or absence of a collusion scam within each thread, employing two raters for the labeling process. However, we only label threads that surpass the threshold of three replies. This selection criterion is based on our observation that threads below this threshold often remain in a developmental stage, lacking clear indications of being a scam or non-scam threads. We find a total of 1,174 collusion scam threads, 1,272 non-scam threads, and 4,557 threads were unlabeled. Additionally, we identify 332 spam threads that evade YouTube's spam filtering algorithm, representing instances where one individual comment on a financial coach and shares their WhatsApp number across multiple replies, exemplifying a typical form of such spam threads. Table 1 summarizes the data distribution for our research. The wordcloud visualization (Figure 2) highlights the frequent use of words like "trading" and "expert" in collusion scam threads.

To further validate our manual labeling process, we collect the annotation from two other annotators for 5% of the collusion scam and non-scam threads. To help with the annotation process, we provide them with a short PowerPoint presentation, and 10 examples of collusion scams. We find the Cohen Kappa inter-annotator agreement as 0.91 and 0.96 respectively (Cohen, 1960). The high inter-annotator agreement scores provide strong evidence of the reliability and consistency of our manual labeling process. The data is available at https://github.com/sadat1971/YouTube_Collusion_Scam.

3.3 Detection Models

We use two modes of detection strategy for collusion scams. In the static mode, we use a 2-layer

Fully-Connected (FC) neural network architecture, followed by a softmax layer to classify threads for being a scam or non-scam. This mode is utilized during training with a single comment or training with metadata only. To leverage the collusion dynamics present within the comment threads, we use the dynamic mode of learning. In this mode, we utilize a Bi-directional Long Short-Term Memory (Bi-LSTM) model with an attention mechanism, followed by an FC layer and softmax layer (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2014).

To extract textual features, we utilize 768-dimensional pretrained BERT embeddings obtained from the output of the $[cls]$ token, due to their capability of capturing contextualized and semantic representations of text (Devlin et al., 2018). To obtain better explainability, we also incorporate tf-idf-based features and train a logistic regression model to detect the collusion scams solely based on the comments.

3.4 Experimental Setup

For all the experiments, we use 70% of the data to train, and 30% to test. To ensure robustness, we repeat the experiments using five random splits of the data. Within the training set, 20% of the data is set aside to determine the optimal hyperparameters, including batch size, hidden layer size, learning rates, and epochs. For performance evaluation, we report accuracy and F1-score.

4 Collusion Scam Detection

4.1 It All Starts with the Comment

In a comment-reply thread, the comment gives the first cue for detecting the collusion scam. Often the comment entices the readers into reading the full conversation that sets up the trap. Typical scammer opening lines include some tangential reference to the video's subject matter, followed by boasts about their accomplishment, while working with a person or entity, e.g., *The contents of this channel is so lovely!! Despite the economy situation , I'm so blessed to make withdrawal of my $124k profits out of my crypto trading investment.*

Results and Discussion Using solely the comments, the tf-idf-based logistic regression model gives an accuracy of 90.33%, and an F1-score of 88.66%. Figure 3 shows the most important words in the comments that separate scams from non-scam conversations. Non-scam comments involve specific cryptocurrency-based discussions, like "ada", "cardano", while scam comments tend to describe generalized opinions and focus more on their luring strategy.

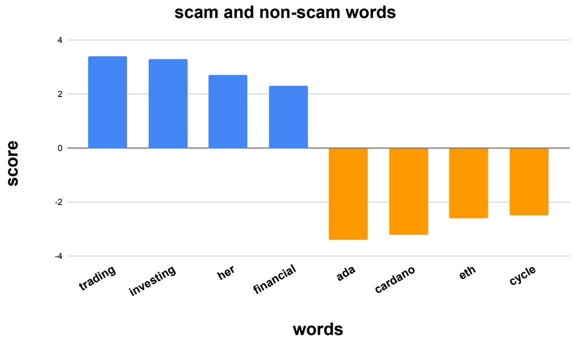


Figure 3: The most important words in detecting collusion scams only from the comments. Words with positive scores indicate a higher contribution to detecting scams.

To obtain more insights, we perform the training with specific Parts-of-Speech (POS) tagged words, and find the top three performances (F1-score) come from Nouns (84.21%), Verbs (79.88%), and Adjectives (70.73%). Hence, collusion scams can be recognized from *what* is being said in the comments. Finally, the BERT-FC model provides a better performance than the tf-idf model, by achieving 92.26% in accuracy, and 91.42% in F1-score, due to the richer textual representation obtained from the BERT embeddings.

4.2 Collusion Scam Conversation Dynamics

We explore the dynamics of collusion scam conversations and gain insights into the strategies used by scammers to deceive readers.

Persuasion Strategy In a collusion scam thread, the scammer(s) lure the readers into believing a fictitious scenario by depicting a fake conversation. The goal of the conversation is to persuade the readers by using several persuasion techniques (Cialdini and Cialdini, 2007; Gragg, 2003; Stajano and Wilson, 2011). Utilization of such techniques are observed in fake review detection, and phishing email detection (Munzel, 2016; Shahriar et al., 2022). Table 2 shows some examples of strategies used in a collusion scam.

Most collusion scam starts with a generic advisory and "call-for-urgency" message. Such texts can encode the *Authority* technique of persuasion, where a scammer pretends to be an experienced veteran and advise the general users. The scammers may use this technique to avoid being flagged as spam by users. In and of itself, the message is often harmless, suggesting inexperienced users pursue a financial coach and explaining its benefits. However, such messages create a facade of collusion, which comes as the next step for the scam.

The scammers often pretend to be a novice who needs help with investment, with the goal of gaining the victim's trust and providing them a feeling of sharing the same predicament. By pretending to be a newbie, the scammer uses social engineering to create a false sense of familiarity and establish a relationship of trust with the victim, which they will later exploit for their own benefit.

Various techniques are utilized to emphasize the contact information and credentials of the target individual or organization. Scammers often split the contact information, such as phone numbers, WhatsApp, or Telegram, into multiple responses to avoid detection by YouTube's algorithm for scams. In the Name-dropping technique, scammers frequently post responses from multiple accounts with slight variations in language, claiming to have benefited from a particular individual and expressing gratitude. These responses can project commitment, integrity, and consistency, thus enhancing the trust level among users.

Scammers use the scarcity principle to persuade readers to invest their money in fraudulent schemes. They create a sense of urgency by suggesting that it is the best time to invest, and that if the reader does not act quickly, they will miss out on a lucrative opportunity. Scammers may use various tactics to entice people into investing, such as promising huge profits, using fear-mongering techniques, or creating a sense of panic around a particular investment opportunity.

Results and Discussion To examine the dynamics of collusion, we feed the BERT embeddings of comments and replies to the BiLSTM-Attention-FC network. Our results indicate that the performance of the model improves with an increase in the number of replies, as illustrated in Figure 4. For instance, with one reply, the model achieves an average accuracy of 79.28% and an F1-score of 66.74% across all five folds. By adding one more reply, we observed a 5.49% increase in accuracy and a 9.03% increase in F1-score. When using the maximum number of replies, the model achieved the highest performance, with an accuracy of 96.67% and an F1-score of 93.04%. Thus, our model learns more about collusion as the conversation progresses.

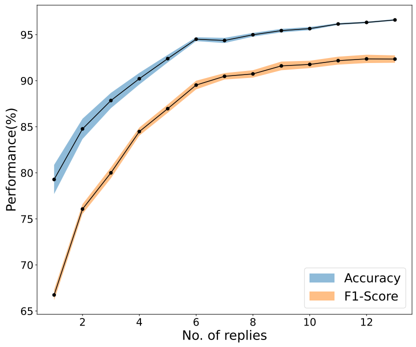


Figure 4: Collusion scam detection performance with an increase in the number of replies in a comment-reply thread.

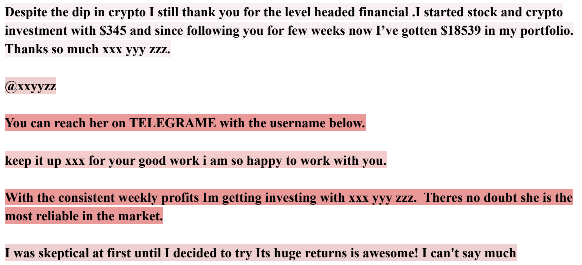
We further explore the attention weights used by our model to identify the conversation threads containing collusion scams. Our investigation indicates that the model primarily focuses on replies that mention individuals. Figure 5 displays the areas of high attention during a scam conversation. It further demonstrates that the model's attention mechanism is particularly drawn to replies that contain Name-dropping and expressions of admiration or appreciation. Hence, such characteristics can provide a significant indication of collusion scams.

We conducted an error analysis to investigate mislabeling patterns in our model. Our findings indicate that genuine conversations discussing common topics associated with collusion scams can be erroneously classified as such. For example, collusion scam threads employ the persuasion strategy of "scarcity" by stating how risky it is for inexperienced people to invest without a financial coach. When non-scam threads involve users discussing various cryptocurrencies and sharing personal investment mistakes without any intention to deceive others, our model may mistakenly identify them as collusion scams.

We observe another common error where conversations include a mix of legitimate comments and collusion scam comments. We find that in 16.03% of the cases, collusion scam threads have non-scam comments or completely unrelated comments. In cases where the non-scam comments outnumber the scam-related ones, our model misclassifies the threads as non-scams. This highlights the need to consider alternative approaches, such as multiclass

Description	Example	Persuasion-technique
Urgency and Advisory	*If you are not conversant with the markets Id advise you to get some kind of advise or assistance from a financial investing coach. It might sound basic or generic but getting in touch with an investment broker was how I was able to outperform the market*	Authority
Social Engineering	*Please how can I reach her Im a newbie and know nothing about crypto investment*	Social Proof/Compliance
Name-dropping	*Wow you really know expert XYZ? Im a living testimony of her good expertise she has been trading for me for months now*	Commitment, Integrity, and Consistency
Panic and Possibilities	*Most coins are going to 10x this Year. The recent bitcoin correction down from its all-time high has had the market in a panic in the past week. However, not everyone has seen it as a bad omen*	Scarcity, Need and Greed

Table 2: Example of collusive conversation text, and the persuasion strategies used to convince the readers to invest



Despite the dip in crypto I still thank you for the level headed financial .I started stock and crypto investment with $345 and since following you for few weeks now I've gotten $18539 in my portfolio. Thanks so much xxx yyy zzz.

@xxyyzz

You can reach her on TELEGRAME with the username below.

keep it up xxx for your good work i am so happy to work with you.

With the consistent weekly profits Im getting investing with xxx yyy zzz. Theres no doubt she is the most reliable in the market.

I was skeptical at first until I decided to try Its huge returns is awesome! I can't say much

Figure 5: Attention weights visualization in a collusion scam conversation. The regions with darker shed indicate higher attention.

classification or formulating the problem as a regression task, to measure the "degree" of collusion scam presence in a conversation. Addressing these challenges will be the focus of our future work.

4.3 The Cues from Metadata

In this section, we will investigate the collusion scam patterns from the metadata available on YouTube video pages.

4.3.1 Response Time of Comments and Replies

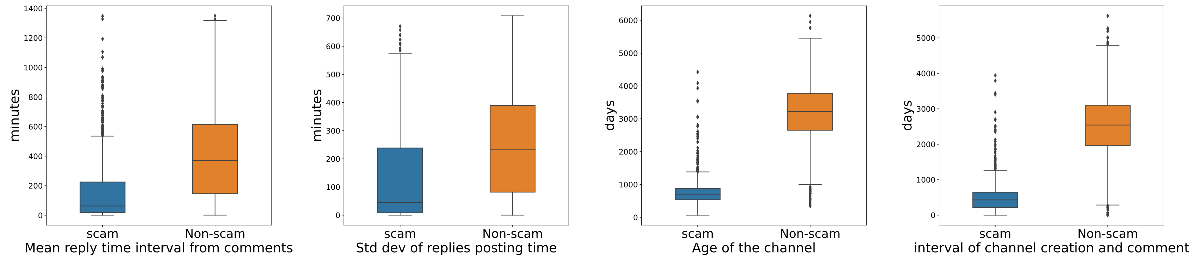
First, we explore the response time of replies posted under the comments in the conversation thread. Our investigation reveals that the replies within collusion scam comments exhibit a significantly shorter response time than those in non-scam comments (p-value < 0.05). As depicted in Figure 6a, the average time interval between the posting time of comments and replies is 161.01 minutes in the case of scams, and 404.93 minutes for non-scam conversations. Furthermore, we investigate the time intervals patterns within the replies, as illustrated in Figure 6b. In scam comments, the average standard deviation within the replies is 135.41 minutes, compared to 244.36 minutes in non-scam conversations. This suggests that scammers adopt a more aggressive approach to engaging users and luring them into their fraudulent schemes.

Scammers expose two crucial trends by creating conversations and replying promptly to comments: i) Unlike regular non-scam conversations, in a collusion scam, the scammers do not engage in a genuine conversation that may require time to respond. With the intention of generating more engagement, they frequently post identical or slightly altered answers praising a person or an entity, ii) scammers respond quickly to a conversation to create an illusion of legitimacy and trustworthiness by making the collusion conversation more voluminous of replies than the non-scam conversation. This tactic is evident in the average length of collusion scam replies, which stands at 21.78, while non-scam replies average at 6.91. Hence, the findings imply that analyzing the response time and time interval patterns within comments and replies can be an effective technique to identify collusion scam patterns.

4.3.2 Number of *Likes*

The number of *Likes* can act as a form of social validation, and scammers can exploit that metric to engage viewers. Comments usually serve as conversation starters and, consequently, receive more attention (and thus, more *likes*). Replies, on the other hand, are merely discussions on the comment, and hence, receive less attention. We found that scam comments receive an average of 72.58 *likes*, while non-scam comments receive an average of 52.29 *likes*. However, the scenario flips in the case of replies. While a collusion-scam conversation receives an average of 0.23 likes per reply, a non-scam conversation receives an average of 1.25 likes per reply.

Scam comments are designed to be more



(a)

(b)

(c)

(d)

Figure 6: Visualizing the metadata from the comment threads: a) Mean time interval between comment and replies posted, b) the standard deviation of the posting time among the replies in a thread, c) Age of the user channels who posted comments or replies in the threads, d) how long it took to post after the video is published

attention-grabbing or emotional than non-scam comments. This can make them more likely to elicit a strong response from viewers, including *likes*. However, once viewers skim through the conversation, they may start to become suspicious and less likely to continue engaging with or rewarding them with likes. On the other hand, non-scam conversations may be more genuine and focused on the topic at hand, making them more enjoyable or informative to read, and thus, more likely to receive *likes* on replies. However, it should be noted that the number of *likes* may not always be a definitive indicator of collusion scams. This metric may be influenced by various other factors such as the content of the video, the number of subscribers, and the number of viewers. Consequently, it would be erroneous to rely solely on the number of *likes* as a standalone indicator of a collusion scam.

4.3.3 Age of Scammer Account

Scammers frequently use the approach of constantly creating new accounts as their prior ones are reported or deleted. This is due to the fact that their fraudulent operations are frequently detected and reported by attentive users or platform administrators. Scammers want to avoid detection and prolong their fraudulent activities by regularly cycling through different accounts. Figure 6c shows the distribution of channel age for the users recorded during our data collection process. We find that scammers possess accounts with an average age of 797.74 days, significantly lower than the average age of 3172.74 days observed for the non-scammers' accounts.

This disparity in account age reflects the ephemeral nature of scammers' online presence. Their accounts, which have very brief lifespans, are a direct result of their deceptive actions and the repercussions they face. Genuine users, on the other hand, have accounts that have been active for considerably longer lengths of time, indicating their real and long-term participation in the online community.

In addition to creating new accounts frequently, scammers tend to comment on these fraudulent schemes shortly after their account creation. Figure 6d illustrates the distribution of the time it takes for users to comment on a video after creating their channel. On average, scammers begin commenting on collusion scams approximately 468.61 days after creating their accounts. In contrast, genuine users, with authentic intentions, take an average of 2509.46 days for commenting on cryptocurrency-related posts. Furthermore, our analysis demonstrates that 11.27% of scammers begin commenting on collusion scams within just a month of creating their accounts. This rapid initiation into fraudulent activities highlights their aggressive approach, aiming to exploit vulnerable individuals as quickly as possible. On the contrary, genuine users exhibit a significantly lower rate of early engagement, with only 2.38%.

Results and Discussion By examining the metadata associated with conversations, we have discovered that they serve as important indicators for identifying collusion scams. Leveraging this insight, we build a collusion scam detector relying solely on metadata analysis. We find that using the above-discussed metadata results in an average of 87.08% accuracy, and 88.42% F1-Score. The metadata-based collusion scam detector, excluding textual content, offers a streamlined and effective approach for the early identification of fraudulent activities. Its focus on metadata analysis enables efficient detection without the need for complex text processing systems.

5 *ChatGPT* and Collusion Scam

Among the family of Large Language Models (LLM), *chatGPT* has shown enormous promise, due to its language generation and comprehension abilities (ChatGPT). First, we use the *chatGPT* prompt to provide the following instruction: *The Collusion Scam can be defined as a fake conversation where the participants pretend to be beneficiaries of a person or an entity to entrap the users for their monetary gain. I will provide some examples, can you tell me if they are creating a collusion scam or not?* Subsequently, we provide it with a set of threads involving both collusion scams and non-scams. These prompts are presented within a single chat session. We manually extract the output from the response.

We find that *chatGPT* as collusion scam detector yields an accuracy of 89.40%, with an F1-score of 88.54%. In 8.53% of the cases, it does not provide any direct answer, and we use the prompt to ask further questions to have a clear response. We also find that after an average of 8.33 responses, *chatGPT* seems to forget the task, and we provide the task description again. Since *chatGPT* provides a linguistic response, it first summarizes the conversation, and then the verdict with its reasoning. Although its performance falls short of our BERT-LSTM model, its explanations accompanying the responses can enhance collusion scam detection reliability for users. However, given the large number of collusion scams on YouTube and the lack of a fine-tunable architecture, further research is necessary to incorporate *chatGPT* into collusion scam detection.

Nevertheless, it is crucial to acknowledge that while *chatGPT* demonstrates responsible behavior by refraining from offering harmful or improper responses, it remains susceptible to manipulation by scammers (Hacker et al., 2023). For example, when prompted with instructions for writing a comment in a YouTube video about being financial beneficiaries of a person, *chatGPT* answers with a legitimate-sounded response with a specific amount of "profit" and "investment". Thus, collusion scam detection in the post-AI era may require more careful work and sophistication with a responsible AI research.

Data	Model	Accuracy	F1-score
Comments only	tf-idf BERT-FC	90.33 92.26	88.66 91.42
full thread	BERT-LSTM	**96.67**	**93.04**
Metadata	FC	87.08	88.42
No Training	*chatGPT*	89.40	88.54

Table 3: Summary of the collusion scam detection approaches.

6 Conclusion and Future Work

In this research, we address the issue of collusion scams within YouTube's comment section, particularly in the cryptocurrency market. We have demonstrated scammers' deceptive tactics, luring unsuspecting users through social interactions. We also explore different collusion scam detection strategies, where the comments may have an important initial signal, and the thread dynamics can further bolster the detection performance. Additionally, our study of YouTube metadata shows promising discriminators between collusion scams and genuine discussions, including *likes*, reply patterns, and user channel age. Table 3 provides a comprehensive summary of our approaches and the corresponding detection performances. Future research directions of this work include:

- The collusion scam threads may contain replies from genuine users, ranging from the inquisitive ones seeking information to experienced individuals who raise suspicions about the scam. In a few cases, the scammers also engage in conversations refuting the accusations. Future research on exploring these exchanges can help gain deeper insights and a better understanding of the dynamics surrounding collusion scams.

- Investigating the scalability and generalizability of our proposed detection strategies for other online platforms, like, Reddit, Twitter, and Facebook would be an interesting direction of work.

- Whether the modern text generative LLMs like GPT-4, chatGPT, BARD are more susceptible to generating effective collusion scams, making it harder for the AI to combat them, can be a valuable research direction.

7 Ethics and Broader Impact Statement

Throughout this research, we have prioritized fairness and adhered to ethical practices in our data collection strategy, strictly abiding by YouTube's terms of service and community guidelines. Additionally, we ensured compliance with YouTube's API "Terms of Service", aligning with the laws and regulations of the country where this research took place. We also respected and adhered to the API's quota limit, ensuring responsible data usage.

To further preserve fairness and mitigate any potential biases in our models, we implemented a masking technique to anonymize user names in the conversation threads, where applicable. By masking user names, we aim to prevent any unintended profiling or bias that may arise based on specific individuals or their characteristics. This approach serves to enhance the fairness and integrity of our research outcomes.

Our work contributes to fostering a safer online environment where users can engage, free from the pervasive threat of scams and fraudulent activities. This research can also help improve YouTube's platform responsibility in battling collusion scams. By raising awareness, improving detection mechanisms, and promoting collaborative efforts, we strive to create a positive and trustworthy digital ecosystem for all users.

Acknowledgments

This research was supported in part by grant ARO W911NF-20-1-0254. The views and conclusions contained in this document are those of the authors and not of the sponsors.

We also express our sincere gratitude to Navid Ayoobi, Shanto Roy, and Tomin Chazhikat for their invaluable assistance in labeling the data used in this study.

References

Accessed: 2023-05-14. *Crypto Scammers Are Prowling YouTube Comment Sections to Target Users*. DigitNews.

Accessed: 2023-1-21. *Spam, deceptive practices, scams policies*. YouTubePolicy.

Accessed: 2023-1-21. *YouTube is finally doing something about comment spam that impersonates creators*. 9to5Google.

Abdullah O Abdullah, Mashhood A Ali, Murat Karabatak, and Abdulkadir Sengur. 2018. A comparative analysis of common youtube comment spam filtering techniques. In *2018 6th international symposium on digital forensic and security (ISDFS)*, pages 1–5. IEEE.

Shreyas Aiyar and Nisha P Shetty. 2018. N-gram assisted youtube spam comment detection. *Procedia computer science*, 132:174–182.

Túlio C Alberto, Johannes V Lochter, and Tiago A Almeida. 2015. Tubespam: Comment spam filtering on youtube. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pages 138–143. IEEE.

Gilberto Atondo Siu, Alice Hutchings, Marie Vasek, and Tyler Moore. 2022. "invest in crypto!": An analysis of investment scam advertisements found in bitcointalk. APEG.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Hanif Bhuiyan, Jinat Ara, Rajon Bardhan, and Md Rashedul Islam. 2017. Retrieving youtube video by sentiment analysis on user comment. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 474–478. IEEE.

Elijah Bouma-Sims and Brad Reaves. 2021. A first look at scams on youtube. *arXiv preprint arXiv:2104.06515*.

ChatGPT. Chatgpt, may 3 version.

Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rama Krushna Das, Sweta Shree Dash, Kaberi Das, and Manisha Panda. 2020. Detection of spam in youtube comments using different classifiers. In *Advanced Computing and Intelligent Engineering*, pages 201–214. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nicola Döring and M Rohangis Mohseni. 2020. Gendered hate speech in youtube and younow comments: Results of two content analyses. *SCM Studies in Communication and Media*, 9(1):62–88.

David Gragg. 2003. A multi-level defense against social engineering. *SANS Reading Room*, 13:1–21.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. *arXiv preprint arXiv:2302.02337*.

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

Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 278–289.

Sijia Li, Gaopeng Gou, Chang Liu, Chengshang Hou, Zhenzhen Li, and Gang Xiong. 2022. Ttagn: Temporal transaction aggregation graph network for ethereum phishing scams detection. In *Proceedings of the ACM Web Conference 2022*, pages 661–669.

Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2021. Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, 8(3):607–617.

Andreas Munzel. 2016. Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. *Journal of Retailing and Consumer Services*, 32:96–108.

Huy Nghiem, Goran Muric, Fred Morstatter, and Emilio Ferrara. 2021. Detecting cryptocurrency pump-and-dump frauds using market and social signals. *Expert Systems with Applications*, 182:115284.

Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. 2012. Network analysis of recurring youtube spam campaigns. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 531–534.

Sadat Shahriar, Arjun Mukherjee, and Omprakash Gnawali. 2022. Improving phishing detection via psychological trait scoring. In *Proceedings of the IADIS International Conference Web Based Communities 2022 (part of MCCSIS 2022)*, pages 131–139.

Frank Stajano and Paul Wilson. 2011. Understanding scam victims: seven principles for systems security. *Communications of the ACM*, 54(3):70–75.

Mike Thelwall, Pardeep Sud, and Farida Vis. 2012. Commenting on youtube videos: From guatemalan rock to el big bang. *Journal of the American society for information science and technology*, 63(3):616–629.

Ashutosh Tripathi, Mohona Ghosh, and Kusum Bharti. 2022. Analyzing the uncharted territory of monetizing scam videos on YouTube. *Social Network Analysis and Mining*, 12(1).