CryptoWho: Auditing YouTube Investment Recommendations using GPT-3

Brian Huang, Lily Yu

Abstract

Since 2021, an estimated half a billion dollars were lost to crypto scams originating from social media platforms. The main victims of these scams were often younger individuals, who were three times more susceptible to these scams. With the recent rise of 'Finfluencers'[1], individuals who promote investment advice on social media; young investors are flocking to these platforms to become financially literate. While investment scams have always existed, the untrackable nature of cryptocurrency and the influx of personal finance autodidacts creates a perfect environment for common scams to churn out cash. In this paper, we set out to detect the distribution of crypto videos and traditional investment YouTube videos recommended to two age groups. YouTube is chosen over other platforms due to the seamless data pipeline that can be built on top of its API, something that is harder to do with TikTok for example. The goal of this audit is not to identify scams on YouTube, but instead to quantify how individuals in varying age buckets are recommended investment advice.

We approach this by creating three watch histories based on our labels: traditional, crypto, and mixed. We hypothesize that across all age buckets, users should receive the recommendations most in line with their labeled watch history.

Introduction

According to the Federal Trade Commission, since the start of 2021, over a billion dollars was lost to crypto scams with nearly half of the losses originating from a social media recommendation or advertisement. Within these losses people, ages 20-49 were more than three times as likely to be scammed as opposed to their older counterparts, with individuals in their 30s being hit the hardest, attributing 35% of fraud losses to crypto since 2021.[2]

Accompanying these losses is a rise in the 'Finfluencer' content space. Throughout all social media platforms, content creators specialized in personal finance, investing, and cryptocurrency has become increasingly popular.

For this paper, the main platform we're focusing on is YouTube. The main reasons for choosing youtube are

1. The seamless data pipeline that can be built on top of its API
2. It was one of the first user-generated video platforms
3. Many videos from other platforms inevitably end up on YouTube and vice-versa

[1] Lowry, Erin. "Personal Finance: Should You Trust Tiktok, YouTube, Instagram Finfluencers?" *Bloomberg.com*, Bloomberg, 22 Feb. 2022, https://www.bloomberg.com/opinion/articles/2022-02-22/personal-finance-should-you-trust-tiktok-youtube-instagram-finfluencers.

[2] Staff, the Premerger Notification Office, et al. "Reports Show Scammers Cashing in on Crypto Craze." *Federal Trade Commission*, 11 Aug. 2022, https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/06/reports-show-scammers-cashing-crypto-craze.

YouTube has become one of the most important information sources for news and other topics[3], however, it has long struggled with dealing with misinformation, something that was highlighted during the COVID-19 pandemic.[4] In tandem with the untrackable nature of cryptocurrency, the inability to moderate misinformation has made crypto scams appear a dime a dozen. As many influencers rely on sponsorships to run their channels, they may unwittingly promote scams. Oftentimes, influencers may not be financially literate themselves and often rely on social media hype to validate their sponsors. This can be seen in cases such as Sam Bankman Fried's FTX[5] scandal and Logan Paul's NFT 'CryptoZoo'[6]. This audit doesn't provide opinions on whether the individuals tied to these projects/companies are culpable but does use these examples to highlight how YouTube, social media, and its influencers have played a role in promoting these scandals.

While there is no way to moderate influencers, due to the sheer scale of YouTube and other media platforms, we can observe if YouTube's recommender system is playing a role or not. Investing is a fundamental part of building a secure lifestyle, and oftentimes a singular scam can deter an individual from investing further, including ones that have stood the test of time such as the S&P 500.[7] Motivated to protect young investors, we set out to characterize and quantify how YouTube's recommendations impact investors exposure to assets. In particular, we set out to answer the following questions:

> **RQ1.** Can we effectively identify if a video is discussing crypto or traditional investments?
>
> **RQ2.** What is the proportion of videos that are being recommended to different age buckets for each of the labels we identified above on the homepage and video recommendations section? How does the age bucket our user falls into affect this and does it differ from what we expect given a user's predetermined watch history?

Methodology

There are two major parts to our methodology to consider. To perform our audit, we use the browser automation library Selenium[8] to automatically watch a set of seed videos, recording data on recommendations along the way. Then, to process our collected data at scale, we utilize

[3] N. Newman, R. Fletcher, A. Schulz, S. Andi, and R. K. Nielsen. Reuters Digital News Report. http://bit.ly/rtrs-report, 2020.

[4] C. G. Weissman. Despite Recent Crackdown, YouTube Still Promotes Plenty of Conspiracies. http://bit.ly/youtube-consp, 2019.

[5] Reiff, Nathan. "The Collapse of FTX: What Went Wrong with the Crypto Exchange?" *Investopedia*, Investopedia, 27 Feb. 2023, https://www.investopedia.com/what-went-wrong-with-ftx-6828447.

[6] Silberling, Amanda. "YouTuber Logan Paul's CryptoZoo NFT Project Is a Total Mess." *TechCrunch*, 6 Jan. 2023, https://techcrunch.com/2023/01/06/youtuber-logan-pauls-cryptozoo-nft-project-is-a-total-mess/.

[7] Maverick, J.B. "S&P 500 Average Return." *Investopedia*, Investopedia, 15 Feb. 2023, https://www.investopedia.com/ask/answers/042415/what-average-annual-return-sp-500.asp.

[8] *Selenium*, https://www.selenium.dev/

OpenAI's GPT-3.5[9] as a language model to process textual data associated with videos such as titles, transcripts, and tags.

Audit Methodology

We focus on four labels in this audit: Blockchain, Traditional, Mixed, and None. A video is considered blockchain if it discusses or recommends any blockchain asset (cryptocurrency, NFTs, etc) while a video is considered traditional if it discusses any traditional investing topic (stocks, bonds, commodities, real estate, IRAs, etc). The mixed label accounts for any videos discussing both while the none labeled videos account for any other videos (such as Never Gonna Give You Up - Rick Astley).

According to a previous study, the minimum amount of videos to establish a watch history and receive personalized recommendations is 22.[10] With this in mind, we collect 40 videos per label and label these videos within our group and with our mentor. Due to the scope of this project, it is difficult to collect larger amounts of 'seed videos' (videos that are used to build watch history). In a follow-up experiment, we would like to collect more.

Since we're interested in the effects of age, we keep watch history and all other factors consistent across labels. The two age bucket users we're creating are ages 18-23 and 55-60. In order to create these users, we also collect an additional 20 videos that fall into the age bucket for these users. These videos are once again hand labeled and are videos that a stereotypical individual in the age bucket would watch. Examples of these videos include: "Asking Reddit what I should know when I turn 18", "My first day at College", "Getting ready for your first Colonoscopy", and "Things to do to stay active at 55". The videos are selected with the intent of not identifying a gender, however, due to the timeline and lack of large-scale labeling, it's unknown how well we were able to mitigate these effects.

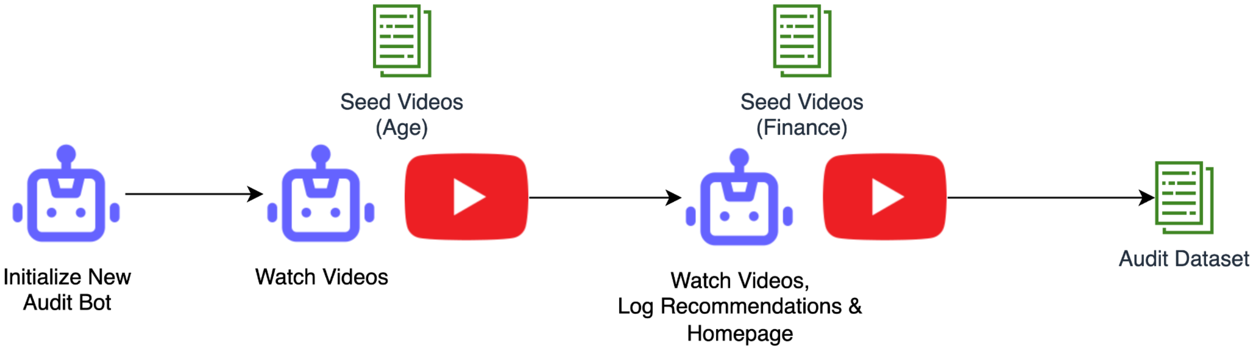


Figure 1: Browser automation flow for audits

[9] *OpenAI API*, https://platform.openai.com/docs/models
[10] *It's just a flu,* https://doi.org/10.48550/arXiv.2010.11638

The rationale for creating a watch history for each age is rooted in the idea that YouTube does not consider the age parameter in your account settings. Oftentimes, ages on YouTube can be faked, and we assume that YouTube's recommender system will use the videos watched to generate an age bucket rather than relying on a user's inputted age. Following this logic, we do not create any accounts.

Using our three labeled watch histories and our two age-bucketed watch histories, we create six types of users. We have three young users interested in each of the labels and three old users interested in each of the labels. We conduct the audit by using a web driver to collect home page recommendations after the watch history every 10 videos after the age seed videos have been watched (i.e starting from video 20 we return to the homepage, then video 30, etc.). We collect the top ten recommendations off of the sidebar for each video that we watch.

Video Classification

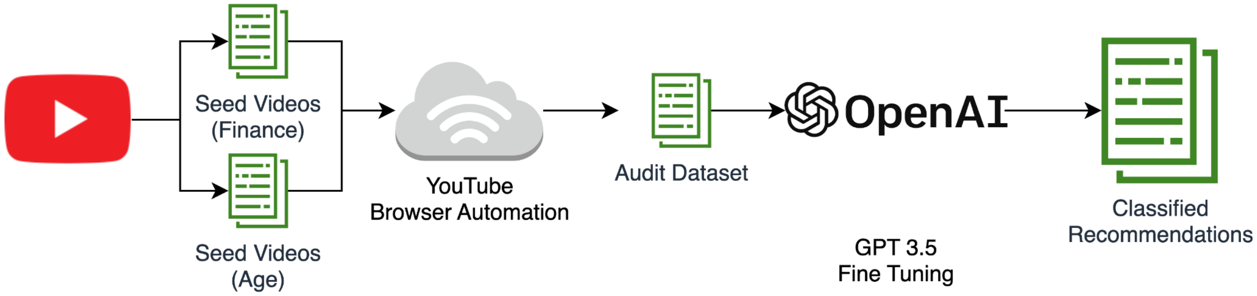


Figure 2: Video classification pipeline

In order to classify the recommended videos, we use OpenAI's GPT-3.5[11]. We decided to use GPT-3.5 for reasons including but not limited to: the time constraints of the project, interest in prompt engineering, better performance achievable by utilizing an existing model, and interpretability via prompting. We assessed the accuracy of varying prompts and temperature by using our seed videos which we labeled and seeing how consistently GPT-3.5 performed. We originally started with GPT-3 DaVinci as our model engine of choice, however starting March 1st, 2023, OpenAI published GPT3.5 Turbo, which had significant cost reductions and accuracy improvements.

Designing prompts was critical to reaching high accuracies in our classifier. Unlike traditional classifiers, our seed video set functioned both as the set used to create watch history and the test set for our classifier. This is due to the fact that GPT-3.5 does not require a labeled training set (however it does help to have more data) to fine-tune/train. In the future, given more time, we would like to fine-tune the model using a training set and testing set. This is something future researchers can delve into when using GPT to audit. Given the relatively short lifetime of GPT 3.5, there are many different ways to approach using it as a classifier that we did not cover.

[11] *OpenAI Chat Completion Model*, https://platform.openai.com/docs/guides/chat

Starting from GPT 3.5 prompts are strung together to create a message. The message is sent to the Chat Completions API where we collect our final prediction. Messages are formatted as follows:

System: {Prompt to bias our chat completion towards our task}
User: {Messages from our ‘user’. Can be used to provide further instructions, data, etc.}
Assistant: {Replies from the API. You can leverage pre-written assistant messages to make your outputs more consistent. Examples are shown below.}

In all of our prompts, we pass in a YouTube video snippet, which is the title concatenated with a summarized transcript (using TextRank) and the top video tags determined using TF-IDF. It outputs a predicted label with the probabilities and a short explanation.

We had multiple iterations of different prompts. We follow a message format starting with a system message to bias our model towards our task, a user message providing details on our labels and how-to return formats, and an assistant message to reiterate instructions so our classifier is more consistent with output. We follow up with one last user message when we make our API call containing our video snippet. The API returns a completion object with the label and rationale behind its prediction.

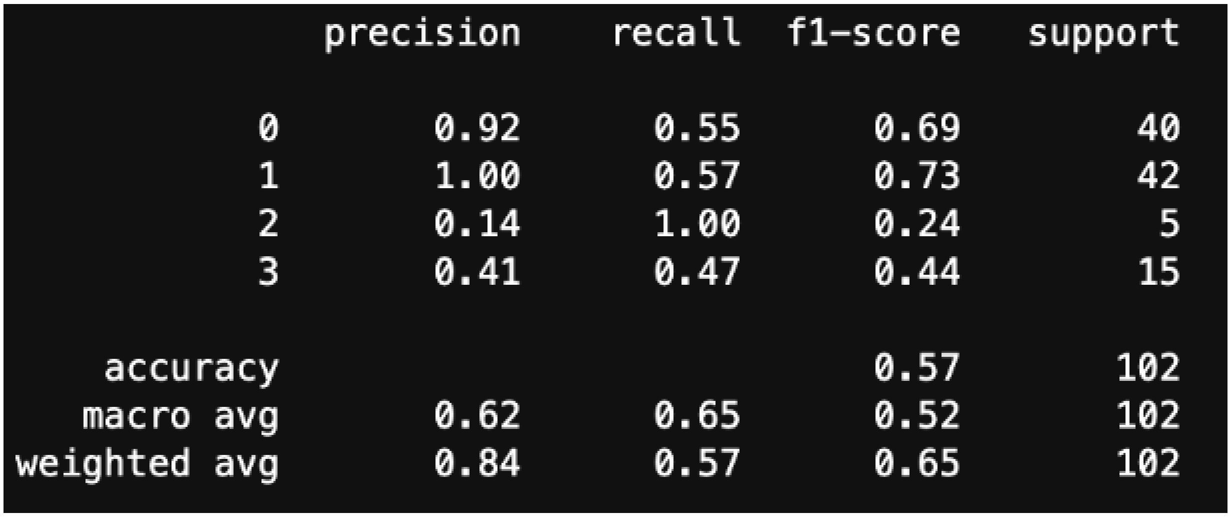
The final prompt we decided to use is shown below:

```
{"role": "system", "content" :

"You are a classifier that determines if a YouTube video snippet falls under a label. A snippet is a concatenation of the video title,
summarized transcript, and video tags. The labels and additional instructions will be included in the first user message."},

{"role": "user", "content" :

"""Labels:

Traditional: Videos that recommend or educate about stocks, bonds, real estate, commodities, retirement accounts, or other traditional
investments or keywords related to them.
Blockchain: Videos that recommend or educate about cryptocurrency (BTC, ETH, etc.), NFTs, or other Web3 investments or keywords related to
them.
Mixed: Videos that recommend or educate about both blockchain and traditional investments or keywords related to both.
Unrelated: Videos that do not recommend or educate about either blockchain or traditional investments or keywords related to them.

Instructions:
- The classifier should consider the context and meaning of the keywords used to determine whether the snippet is related to traditional
or blockchain investments.
- If talks about making money from jobs, side hustles, or other alternative assets (cars, watches, artificial intelligence, trading cards,
art, etc), they are Unrelated.
- A video that is only downplaying an investment or discussing it negatively should be classified as Unrelated.
- Please return predictions in the format" {Label} : {20 word or shorter rationale}"""},

{"role": "assistant", "content":

"""Understood. I will classify YouTube video snippets based on the provided labels and instructions. Here's how I will format the
predictions:

    {Label} : {20-word or shorter rationale}

Please provide me with the YouTube video snippet you would like me to classify."""}
```

Note that across all prompts we use, the system and assistant message stays the same. What we change is the first user message.

Although our final prompt performed extremely well, earlier iterations of prompts struggled. Our very first prompt had an accuracy of 51%. We used GPT-3 starting out and migrated to GPT-3.5 later. The first two prompts below were using GPT-3.

```
I am a YouTube video classifier that takes in video snippets (title + shortened transcript + tags) and
outputs one of the following labels if the video recommends or teaches about:

1. Blockchain: Cryptocurrency, NFTs, or anything related to the blockchain
2. Traditional: Stocks, Bonds, Real Estate, Commodities
3. Mixed: Both blockchain and traditional investments
4. None: Not related to the labels above.
```



```
              precision    recall  f1-score   support

           0       0.92      0.55      0.69        40
           1       1.00      0.57      0.73        42
           2       0.14      1.00      0.24         5
           3       0.41      0.47      0.44        15

    accuracy                           0.57       102
   macro avg       0.62      0.65      0.52       102
weighted avg       0.84      0.57      0.65       102
```

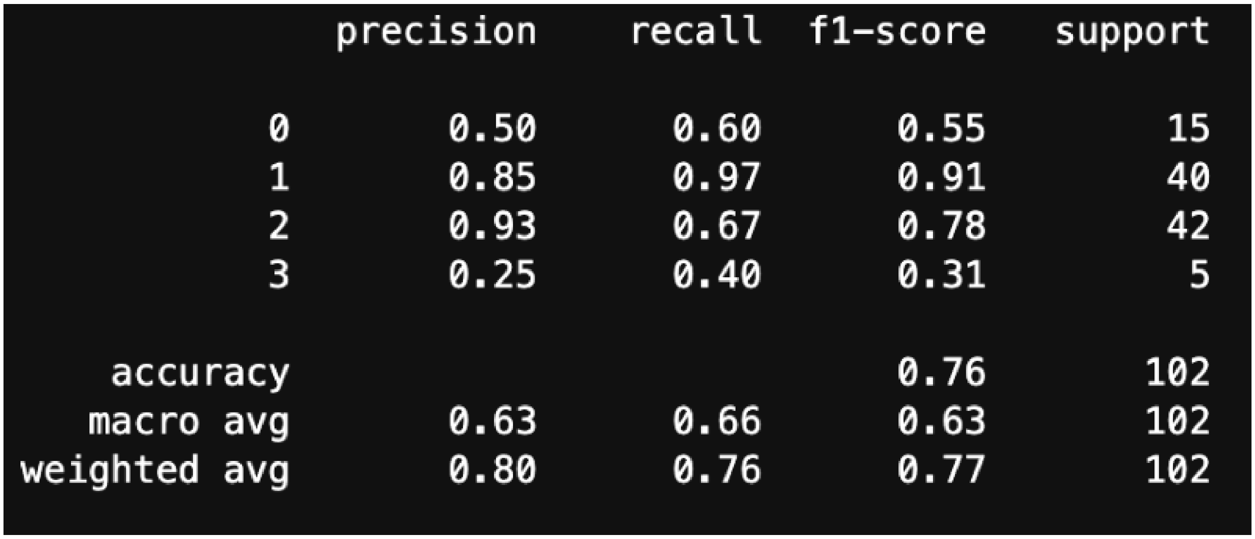
Figure 3: Performance of our baseline prompt

As you can see in Figure 3, our very first prompt did not encourage our model to consider context or related keywords. It also provided no instructions on return output or edge cases that may deviate from our intended task.

Binarzing our task results in a substantial increase in the performance of our model. Binarzing our baseline prompt resulted in an accuracy of 77%. Given more time, we would've likely binarized our final prompt as well. Binarzing allows the main problem to be broken down into subproblems. By asking it to predict if a video is or isn't a label, we can avoid missed predictions where traditional videos are classified as blockchain and vice-versa. A downside of binarizing is that token cost, the metric that Open AI uses to charge users of the GPT API, is doubled. Also, our classifier no longer considers both blockchain and traditional when predicting mixed, rather the predictions of the two prompts are summed together.

```
# Blockchain Binary Prompt
I am a YouTube video classifier. Provide me with a video snippet (title + summarized transcript + tags)
and I will analyze if the video recommends or teaches about blockchain investments(bitcoin, NFTs,
Ethereum, etc). I respond only with Yes and No.
```

```
Examples:
Snippet: Invest in Index Funds. You should invest in index funds. stocks investing
Answer: No

Snippet: Buy Crypto. You should invest in bitcoin. crypto invest
Answer: Yes

Here is the actual task:
Snippet:
Answer:

# Traditional Binary Prompt
I am a YouTube video classifier. Provide me with a video snippet (title + summarized transcript + tags)
and I will analyze if the video recommends or teaches about traditional investments (stocks, bonds,
commodities, real estate, etc). I respond only with Yes and No.

Examples:
Snippet: Invest in Index Funds. You should invest in index funds. stocks investing
Answer: Yes

Snippet: Buy Crypto. You should invest in bitcoin. crypto invest
Answer: No

Here is the actual task:
Snippet:
Answer:
```



```
              precision    recall  f1-score   support

           0       0.50      0.60      0.55        15
           1       0.85      0.97      0.91        40
           2       0.93      0.67      0.78        42
           3       0.25      0.40      0.31         5

    accuracy                           0.76       102
   macro avg       0.63      0.66      0.63       102
weighted avg       0.80      0.76      0.77       102
```

Figure 4: Performance of our binarized prompt model

As you can see in Figure 4, there is a substantial increase in the accuracy. Label 0 is unrelated, label 1 is blockchain, and label 2 is traditional. Label 3 is mixed, the sum of labels 1 and 2.

The prompt below is our first prompt using GPT-3.5. It's very similar to our baseline prompt. It also performed quite similarly, with a 1-2% minor improvement in accuracy.

```
Labels:

- Blockchain: Cryptocurrency, NFTs, or other Web3 investments
- Traditional: Stocks, Bonds, Real Estate, Commodities, or other traditional investments
- Mixed: Videos that recommend or educate about both blockchain and traditional investments
- None: Videos that do not recommend or educate about either blockchain or traditional investments

Instructions:

- If a video snippet mentions keywords related to blockchain or traditional investments but does not
recommend or educate about the label, it will be classified as None.
- Please return predictions in the format" {Label} : {50 word or shorter rationale}
```

The biggest struggle our classifier had was dealing with edge cases. Our prompts often would hone in on keywords rather than the context of our video snippets. There were many prompts that led up to the final prompt, however rather than provide each prompt, the snippet below shows how asking GPT for rationale provided valuable insight on increasing prompt accuracy.

Prior to adding the following line to our instructions:

```
A video that is only downplaying an investment or discussing it negatively should be
classified as Unrelated.
```

Our model classified the following video (only the title is given):

Charlie Munger: Bitcoin is Worthless Rat Poison

With the following label and rationale:

```
"Blockchain: The snippet mentions Bitcoin, blockchain, and cryptocurrency, which are
all related to blockchain investments. Charlie Munger's negative comments about
Bitcoin also suggest that the snippet is related to blockchain investments."
```

Although the prompt emphasizes that the video should recommend blockchain to receive a blockchain label, the classifier struggles to do so, looking only for keywords. When adding the new instructions, the following output is received instead:

```
“Traditional: Charlie Munger downplays Bitcoin's value as an investment, stating that
it has no intrinsic value and is not needed as a payment system. He compares it to
"rat poison" and believes that it encourages speculation."
```

By inspecting our misclassified seed videos, we can use GPT’s rationale to modify our prompt for higher accuracy. There are still faults with this method. Sometimes GPT’s rationale does not match its prediction at all, i.e it predicts blockchain with a rationale discussing why it predicted traditional. Other times GPT’s rationale may result in a prompt that overfits the seed videos. This can be avoided by validating the end results of the classifier on the audit data through manual validation. In the future, given more time, we hope to be able to further tune the prompt and collect enough seed videos to both ‘train’ our prompt using the rationale strategy and ‘test’ our prompt using the remaining seed videos.

Results

RQ1 Can we effectively identify if a video is discussing or about crypto or traditional investments?

Using GPT-3.5, we were able to achieve up to 91% accuracy in classifying finance videos.

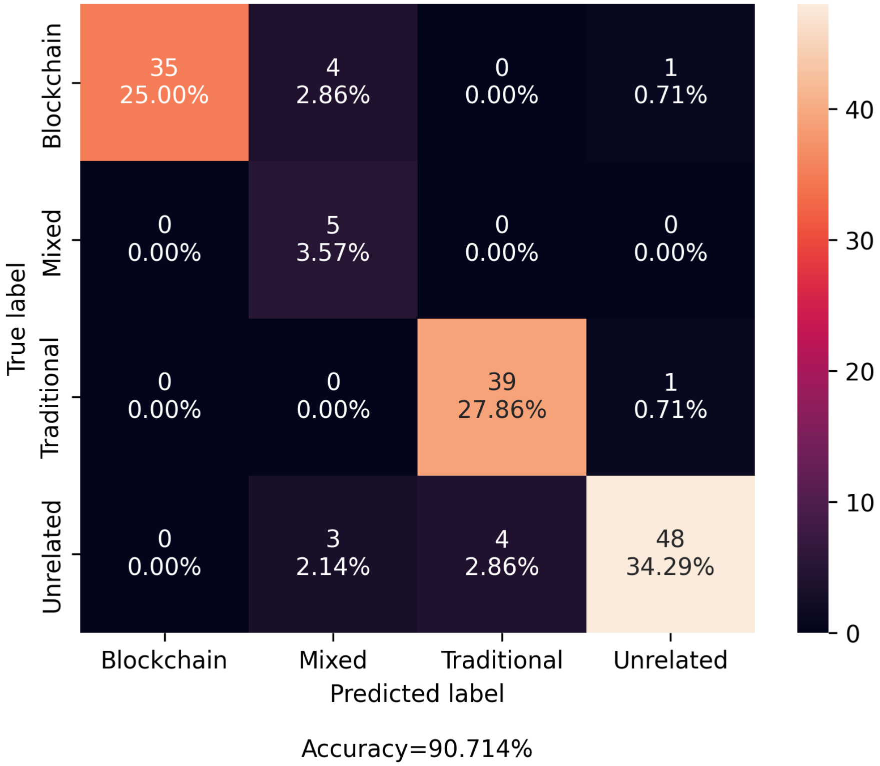
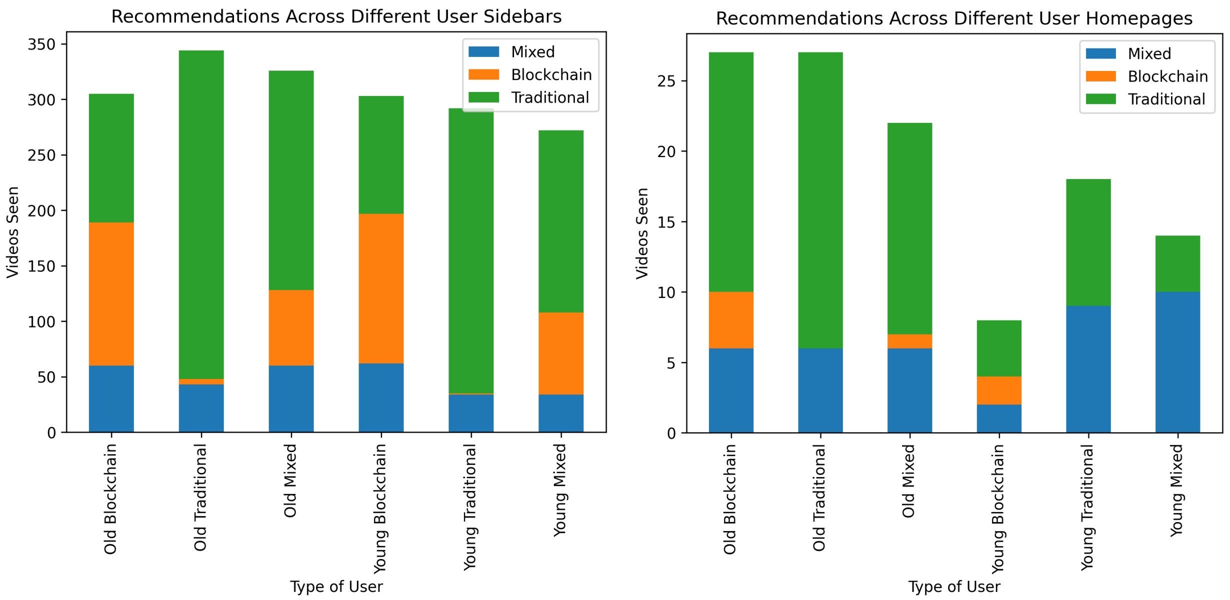


Figure 5: A confusion matrix showing the performance of our classifier on seed videos.

In Figure 5, the performance of our classifier on seed videos is shown. Our classifier had extremely high accuracy, struggling the most with mixed videos. Precision and recall on both blockchain and traditional videos were extremely high, with blockchain predictions being identified for every video.

RQ2 What is the proportion of videos that are being recommended to different age buckets for each of the labels we identified above on the homepage and video recommendations section? How does the age bucket our user falls into affect this and does it differ from what we expect given a user's predetermined watch history?



Figures 6 & 7: Classified video recommendations for different populations of audit users

In Figures 6 & 7 we plot the results of video classification on recommendations seen by our different populations of audit users. Note how blockchain videos are recommended when blockchain/mixed seed videos are used. This serves to affirm the efficacy of our audit and classification pipelines. In general, we find that blockchain-related videos are not frequently recommended unless prompted with blockchain seed videos. However, when we get blockchain video recommendations, they occur at similar rates across age groups, falling in line with what we hypothesized. Interestingly enough, older individuals actually saw more blockchain videos on their homepage than younger individuals, but fewer blockchain videos on their sidebars.

We performed Chi-Squared tests to statistically examine the differences between video recommendations across audit users. For our Chi-Squared tests, we focused on the mixed users, as they represent the average user. The intent of the blockchain and traditional users was to observe if an individual who was actively seeking one type of video (say blockchain) would receive more of the other label due to their age. For example, if we had rejected our null hypothesis, we would've expected younger people to get more blockchain videos. Under that assumption, it would be odd if our young users who watched a lot of traditional videos got significantly higher blockchain recommendations while watching these traditional videos compared to their older counterparts. This was not true, however.

We also discard mixed videos and unrelated videos, only comparing the frequency of traditional and blockchain videos to reduce the degrees of freedom of our test. Separate tests were performed for recommendations found on the YouTube homepage (p=0.339), and recommendations found on the sidebar of the video (p=0.201), neither being statistically significant. Given more time in the future, we would like to run more statistical tests with higher degrees of freedom.

Conclusion

We were able to successfully implement a browser automation and video classification pipeline to perform our audit on the opaque YouTube recommendation system. We did not find statistically significant results that would suggest that the YouTube recommendation system recommends one type of investment more to one age group as opposed to another.

Given our limited resources and timeframe, our audit was limited in scope. Expanding the scope of our audit would be a great next step, across dimensions such as more runs of the audit, audits with greater “depth” using more seed videos, expansion to other platforms, etc.

We encourage the reproduction of our audit, not only to verify our results but also because it is reasonable to expect that YouTube recommendations are a constantly changing system. The code for our auditing pipeline is public and was written with the goal of reproducibility and reproduction in mind. With the possibility of explicit changes to the system by Google, as well as the nature of a live recommendation system that continues to collect interaction data from users, it is important to regularly audit these systems.