Analyzing Time-Dependent YouTube Spam

Brian Mahabir
Boston University
Boston, MA, USA
bmahabir@bu.edu

ABSTRACT

The advent of social media has created a space for abusers of the system to reap benefits with low effort. The social media platform YouTube has become a big host of rampant online spam from 2013-2023. In early April 2023, YouTube enacted a spam mitigation update that reduced the number of spam to less than 1 percent. This paper presents an empirical study of the differences in spam since the update and investigates the drivers of spam on YouTube since before the update. Overall, spam on YouTube has grown so large in part due to YouTube's own laziness and platform specific exploitation attackers use to gain capital easily. Since the update, spam has evolved dependent on the target audience under content creators. Spam is also utilizing AI-generated messages to sneak under current mitigation techniques. Spam has become increasingly more complex to classify based on language models alone thus account verification methods should be used at the forefront of mitigation.

KEYWORDS

Spam, Social Media, AI, YouTube

ACM Reference format:

Brian Mahabir. 2023. Analyzing Time-Dependent YouTube Spam. ACM Trans. Graph. 37, 4, Article 111 (May 2023), 6 pages. https://doi.org/10.1145/1234567890

1 Introduction

Social media has disseminated to almost every corner of the world. For many, social media acts as the only form of communication, the only method for getting public information, and the only source of entertainment. However, social media's weak infrastructure against malicious activity has resulted in a festering community of abusers preying on victims.

One social media platform in particular has garnered a nasty reputation as of late regarding the amount of spam abusers online. That platform is YouTube. Before explaining the problem of spam on YouTube, it is necessary to understand how YouTube works

YouTube is a platform that allows users to post videos for others to watch. YouTube also allows for live streaming, which is a form of live broadcasting over the internet. A content creator is a person who posts videos. Content creators can make a living on YouTube through company sponsorships once the creator gets enough views and subscribers. Subscribers are people who sign up to get notifications for new videos from a creator. Subscriber count has become a status symbol for how popular a certain content creator is.

Now with some of the terminologies explained let us move on to the problem of spam on YouTube. YouTube has a subsection underneath the video called the comment section where people can post messages. The comment section, by default, is sorted by most liked. YouTube spam consists of unwarranted bulk messages that clog the comment section. Spam on YouTube is especially troublesome as it breaks up community interactions and gives an easy place for phishing to happen. Spam can also be used as DDOS to freeze view counts for videos.

Since 2013, YouTube spam has steadily increased to an average of 30% per popular video before the update. It had gotten to the point where prominent content creators have shared their disdain for it; gaining massive attention. Regardless of the publicity, YouTube took an inexcusable amount of time before issuing a fix. In the meantime, creators had implemented their own measures against spam by utilizing bots with filtering models that systematically remove spam one by one. This did not work well at scale, however.

So a big question remains, why did YouTube take so long to fix it? If content creators can utilize bots to suppress spam with a good success rate, wouldn't YouTube have an easy time implementing what has already been tested as successful? The answer comes down to an economic analysis of YouTube's gross income. YouTube makes the majority of its money off ad revenue. A giant portion of YouTube's ad revenue comes from big content creators. YouTube relies on strict rules for advertiser-friendly content. Granted still, advertisers continued to retract sponsorships losing money for YouTube over time. Spam on YouTube has continued to play an indirect role in the demonetization of advertiser-friendly content. In spite of this, YouTube decided that spam contributed to a negligible amount of loss and would rather put resources into dividing income through different means. As a result, spam continued to flourish.

Author's address: Brian Mahabir, bmahabir@bu.edu, Boston University, P.O. Box 1212, Boston, Massachusetts, USA, 43017-6221

Author's address: Brian Mahabir, bmahabir@bu.edu, Boston University, P.O. Box 1212, Boston, Massachusetts, USA, 43017-6221.

Correspondingly, my research provides not only a platform-wide measure of the amount of spam on YouTube but investigates the quirks of the platform that allows spam to fester so wildly. We see that the ability for spam to hide in the comment section, gather popularity from likes, manipulate view count, and tailor their messages to the type of content creator's audience, all promote spam's tenacity and ferocity.

Separately, in my research, I came across a dataset that categorized spam comments under the top 10 most viewed videos between 2013 and 2015. My plan was then to use common machine learning techniques to create a proof of concept for the implementation of spam mitigation at scale. From the given dataset, I used the methodology from the paper, "YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor" [1] to create my own model using Naïve Bayes algorithm. From here, I proposed a hypothesis, which stated that emojis and URLs both aid in the detection of spam and can help deal with overfitting. I then extracted the two features, preprocessed them, and ran a linear regression model to quantize the correlation. What I found is that URL detection had a greater correlation and aided in overfitting while emojis had no correlation.

During this research, YouTube updated the platform with spam mitigation methods that reduced the amount of spam to less than 1%. From here, I was in a unique opportunity to analyze the changes in spam before and after the update. Despite this opportunity, the lack of unbiased spam detection limited my data gathering to collecting data by hand. Over the course of a month, I gathered around 100 comments. From here, I categorized the spam by hand. Overall, what I have found is that

1. Spam no longer use emojis.
2. Spam phishing now incentivizes victims to click on their bio profile instead of embedding the link in the comment.
3. Spam livestream bots utilize inappropriate imagery in their bio profile pictures to incentives victims to click on the link.
4. Spam utilized AI generated messages to further tailor their attack to a specific group of audience.

At the very least, this paper provides a small windowed incite to a larger more in-depth analysis on the features of spam and its evolution.

2 Preliminary Study: Investigating the Problem

During my preliminary search, I came across a dataset that categorized spam from 2013-2015 as the top 10 most viewed videos. I also came across a research paper that used this dataset to create a prediction model. My initial goal was then to reproduce these results. Before creating the model, however, my first goal was to explore new features that could account for potential overfitting and expand the dataset.

The dataset together is composed of 1956 comments. Before I explored the messages themselves, I played around with the other given features to see if any correlated enough to add to the classification model.

First, I checked how much of the dataset spam vs ham was. Approximately, it was half-and-half perfect for training a model (as shown in Figure 1).

Next, I plotted the top five accounts with the most comments (shown in Figure 2). The account with the most comments only amounted to eight. This shows that the dataset is unbiased towards particular accounts. Next, I checked comment data segmentation. 2015 was the highest for ham followed by 2014 for spam.

Following this preliminary exercise, I then looked into the features of the messages. From observational analysis over the years, I have seen more spam with emojis than without so my next goal was to check the emoji segmentation for spam and ham.

This proved to be quite troublesome. Initially, I paid no attention to the dataset's format but quickly realized that the CSV format wrongfully encodes Unicode to Latin1 destroying any special characters like emojis. Instead of emojis, the characters convert into arbitrary strange characters followed by strange boxes. These boxes act as null characters and are destroyed during conversion. The process renders the arbitrary characters useless in trying to convert back to Unicode. This phenomenon is known as mojibaked. Luckily, I used a Python package called ftfy [9] to convert the text properly, which preserved the emojis. As stated before, the conversion from mojibaked back to proper Unicode is difficult because there is missing information. This library uses a heuristic approach to solve that problem.

From here, we see from Figure 4 that emojis don't play as big of a role in the detection of spam as I'd thought. We see that most comments in general don't have emojis. The majority of comments that do have emojis aren't classified as spam either.

Intrigued by the subversion of expectation, I then decided to segment the comments by detection of URLs as URLs in comments are also hypothesized to be linked with spam. I did this by text analysis checking if the comments had the common https:// string attached. Again, we see a similar trend that most comments don't contain the particular feature in mind. What is interesting here is that for the comments that do have URLs, the majority were labeled as spam (as shown in Figure 5).

I then decided to check if the two features I created mathematically correlated to the classification of spam. I ran both features through a linear regression model. I first sanitized and preprocessed the features to satisfy the parameters for the linear regression model using a log transform. Then, I ran the model and generated a summary for multiple tests for correlation. What I found is that URL links are a decent indicator of spam because we see a low P score followed by a high T and coeff score (as shown in Figure 6).

Finally, I vectorized the comments text using TF-IDF (Term Frequency - Inverse Document Frequency) algorithm, which produces word frequency. Following this, I ran the updated dataset through the Naïve Bays model, which is a classical model, tested to be the most accurate for the prediction of spam. Overall, the results are an accuracy of 88.78% with a false positive of 7.4% and false negative of 3.83%.

Figure 1: **Spam vs Ham**

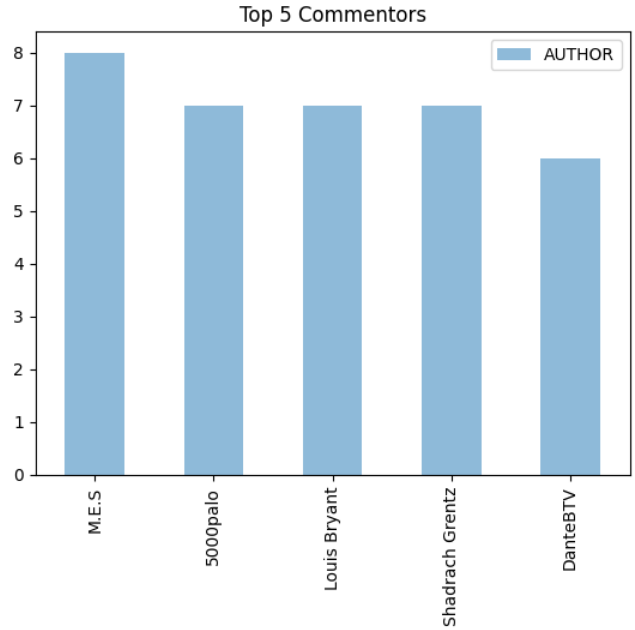

Figure 2: Top 5 Commenters

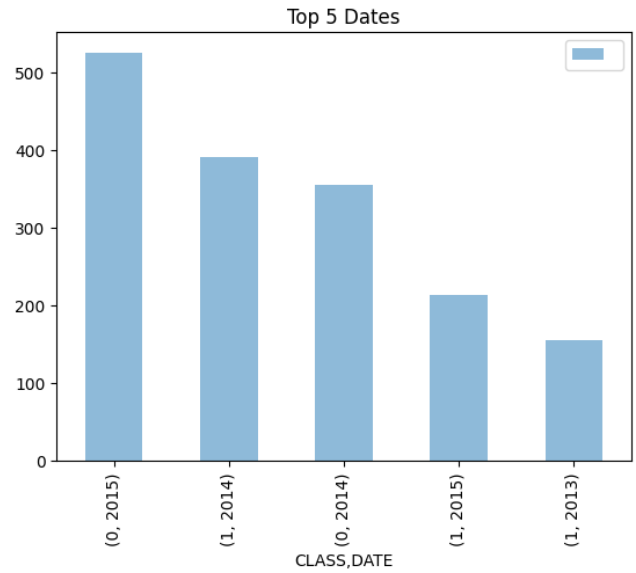

Figure 3: **Dates by Year**

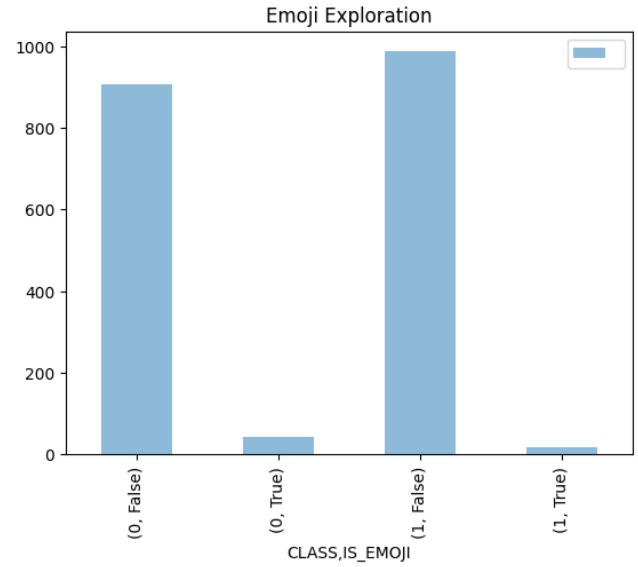

Figure 4: Emoji Segmentation

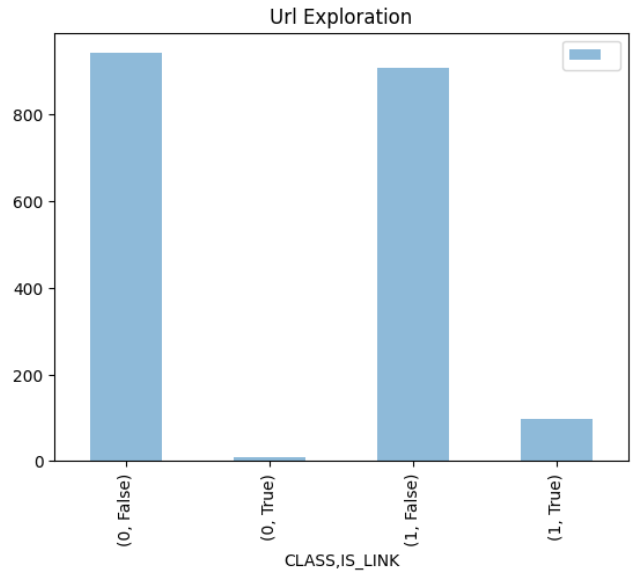

Figure 5: URL Segmentation

	coef	std err	t	P>|t|	[0.025	0.975]
IS_EMOJI	0.2612	0.087	3.009	0.003	0.091	0.432
IS_LINK	0.9017	0.066	13.642	0.000	0.772	1.031

Figure 6: Linear Regression Correlation Tests

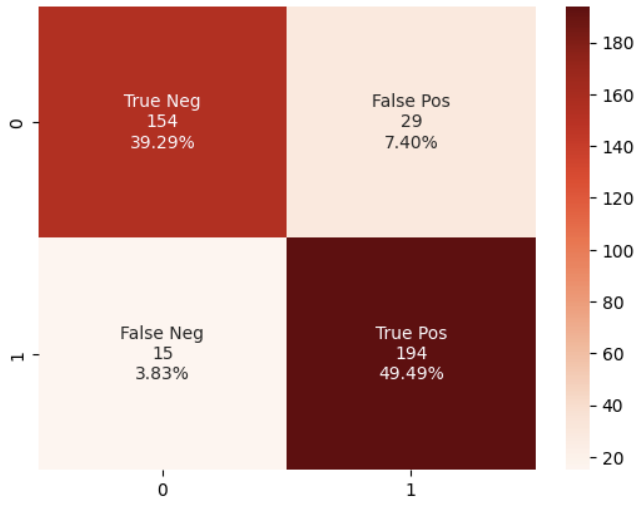

Figure 7: Naïve Bays Confusion Matrix

1.2 Platform Wide Comment Distribution

From the prior investigation, I wanted to see if my results could hold the same value by verifying the current distribution of comments on YouTube. From Statista, there are many statistics of YouTube comment distributions as close as the 4th quarter of 2022. We see that the top 10 most popular videos are still music videos [2]. Of the comments removed on YouTube, 91.1% were spam [3]. Of the number of comments removed from 2018 to 2022, we see two big spikes, one in q2 of 2020 and q4 of 2022 [4]. We also see a relatively steady increase (as shown in Figure 8). Assuming most of these removed comments are spam as suggested in the previous data metric, we can say that the distribution of spam has not changed significantly it has only increased. Generally, we can say that the small-scale analysis done prior can still be accepted for giving a small windowed view for exploring spam after 2015 at large.

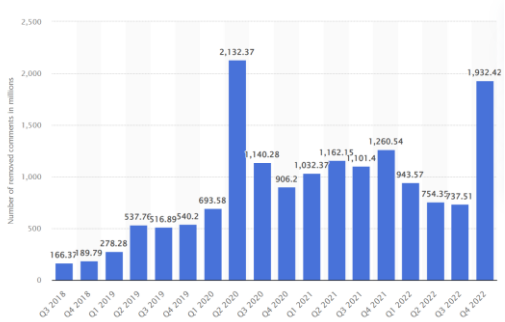

Figure 8: Number of YouTube video comments removed worldwide from 3rd quarter 2018 to 4th quarter 2022

1.3 Spam Categorization: Why Spam is so large on YouTube

At the same time, I also researched the categorization of YouTube spam and the platform-specific qualities that allow spam to grow so quickly. First exploring the UI design of the comment section, the UI promotes the user to keep scrolling to view more comments similar in style to the ranking system of search results in Google. By default, the comments are sorted by the most liked. This given quality allows spam to be hidden under normal circumstances where the spam isn't liked. So most users don't actually see the spam unless there is an unimaginable amount. This quality can be exploited for spam that freezes views.

In recent memory, a K-pop group by the name of Blackpink released a single for their comeback. Fueled by hate and jealousy from other fandoms, attackers utilized spam bots to ddos the video. The bots were able to like each other's comments as well, which made the comment section unreadable. The messages were created specifically to attack BlackPink so it was easy for the bots to see which comments were spam to like. YouTube froze the views on the video for verification which hindered the total amount of views the band could get because most views of a music video happens within the first 24 hours . Along with this, subjectively, seeing a view count for a certain video that is lower than expected given the same parameters could make a user deterred from clicking on the video.

We also see phishing spam utilizing the platform's customization features to impersonate the user name, verification check, and profile pic of a content creator in order to trick victims into thinking the account is real. The attacker spams comments with links to a WhatsApp chat. Once a

victim enters the chat, the attacker will try to persuade the victim into giving sensitive information through various rhetorical techniques.

Given the relative ease at which these attacks happened, it's easy to say YouTube has ironically given a safe space for spam to prosper.

1.4 Economic Analysis: How Spam effects YouTube's Revenue indirectly

The resulting spam on YouTube is largely due to YouTube's unwillingness and or laziness. The results here give an economic background to explain YouTube's decision against taking action for spam. According to Alphabet, the parent company of YouTube, the company saw ad revenue fall 2.6% on average from 2022 to March 2023 [7]. YouTube reports that advertisers have pulled back from the platform due to economic uncertainty [5].

The story goes a bit deeper, however. The first ad revenue crash in late 2015 took place because advertisers deemed the current content on YouTube not family-friendly despite YouTube having a separate app for family-friendly content. The idea was that advertisers could market themselves more if the content had the widest possible audience. From here, YouTube enacted a series of strict rules content creators should follow if they wanted to be sponsored. Despite the changes, YouTube has continued to lose advertiser trust from not being able to control the variety of content on the platform. The underlying consensus, therefore, is that spam has contributed indirectly to the unpredictable nature of the content. A big portion of spam consists of speech that is not family-friendly. Phishing can also contribute to the lack of trust in the platform. Overall, lately, it seems that YouTube has acknowledged these ideas and has pushed an update to reduce spam.

In addition, there is a lot of money to be made on YouTube through phishing. From the top 10 most views on YouTube, we can see just how astronomical those view counts are. Given comments are sorted by most liked, spam can exploit this system to increase the attention that they wouldn't otherwise get on other social media platforms. Given this system exploit there's no wonder so much of spam on YouTube consists of phishing.

On a separate note, I also researched how much phishing attackers make on YouTube. A content creator by the name of JerryRigsEverything did an empirical investigation on a phisher he found in his comment section. He paid the attacker money to get more information out of him. The attacker states that he makes $1000 on average every day [8]. In addition, the attacker states that he was a 21-year-old located in the US impersonating Jerry. The attacker makes his money by tricking the victim into paying shipping costs for a free giveaway. This study also puts into perspective the damage phishing can do to a company's reputation in the eyes of advertisers and the damaging of a user's experience.

2 Data Gathering and Methodology

As mentioned previously, YouTube released a spam mitigation update that reduced the amount of spam drastically. The results in this section quantify the reduction and explore methods for comparing spam before and after the update.

In my research, I discovered software written by the GitHub user ThioJoe that removed spam off YouTube through dictionary lookups of common features and more importantly account verification [6]. I decided to use this software instead of the machine learning models because of the account verification feature, which catches more spam. Through this software, I was able to get a larger view of how much of the comments section was spam before the update. I tested the software under a few music videos and popular content creators, which gave me on average around 30% of spam. I redid this test and got less than 1% after the update. What was interesting was that the software missed some spam that was under new comments on videos released after the update. I also saw that the spam I observed on videos before was removed after a few days. It seems that YouTube's new mitigation techniques are resource intensive given the lack of immediate comment termination. Another interesting fact is most of the spam the software classified had to do with account verification.

After getting a feel for the new landscape, I scoured through popular YouTube videos, live streams, and my personal subscriber list to gather spam comments by hand for one month. I didn't want to rely on any classification models or the software to reduce bias. I did not want to miss any spam that the models couldn't catch through text analysis. To not waste time for the analysis, I also categorized the spam as I was gathering them. From this, I gathered around 100 spam comments.

3 Results

I ended up categorizing spam based on specific features or lack thereof.

1. Spam no longer have emojis
2. Phishing accounts now have links in their bio instead of in the comment
3. Spam regularly tries to impersonate the verification mark
4. Spam bots have inappropriate profile pictures (despite no being allowed on YouTube) to garner more likes and attention (some are censored some aren't)
5. Spam bots use AI generated messages to hide from current mitigation techniques and now further fine-tuning comments for the audiences of specific content creator

Out of this categorization, the most interesting point is spam now using AI-generated messages. With the emergence of ChatGPT, generating human-like messages has become easier than ever. Attackers can abuse this software to take popular comments on a selected video, slightly change the contents, and post the comment to gain more likes. What I found interesting is that there is a big majority of AI-generated spam as the most liked comments under a video of content creators who post regularly. The message itself isn't harmful nor adds any adware, so it seems the only motivation for these types of spam is testing the waters for human activity.

Currently, Snapchat deployed a feature for users to interact with an AI bot that slowly learns about the user. Initial experience shows an incredible depth to the AI chatbot's knowledge only after a few conversations. YouTube attackers can utilize a similar technique to train AI spam bots on content creators' most liked comments to improve the human-like qualities of the message.

Another interesting point from the categorization is the spam bots inappropriate profile pics. The majority of this type of spam was spotted during live streams. This type of spam relies on incentivizing quick responses from the victim due to comment impermanence caused by the nature of live streams. Once the profile pic is clicked, the victim then is automatically taken to a company advertisement. It's theorized that nefarious companies wanting to maximize ad view activity use these types of bots. The profile pics themselves are either censored or not. The non-censored profile pics are shown in a way where certain parts of the body are removed likely confusing the digital image-processing algorithm.

Overall, we are getting to a point where it's no longer feasible to classify spam solely based on the message. That is another reason why I ended up taking the harder, less academic, approach of gathering spam by hand. Under these circumstances, it feels as if hope is lost in the fight against spam. Coincidentally, since this type of spam is inside of a social media platform one can utilize fingerprinting techniques to verify accounts. We see this type of technique in its infancy through ThioJoe's software, which verifies accounts by checking various activity status features.

4 Challenges and Limitations

My study faced two key limitations. First, the data collection process was extremely tedious and somewhat inconclusive. Part of the reason data collection was so tough was that I started collecting data as soon as the update occurred which wouldn't have given me enough time to see an evolution for large amounts of spam. Also, after the update, there was no way to go back and get older spam on YouTube to do a more conclusive comparative analysis. I was lucky enough to have the dataset that I did.

Secondly, time restrictions. With only less than 3 months to do the entire project, managing what aspect I should focus on was quite difficult. The research of spam is such a wide topic that includes not only theoretical analysis of spam itself but also includes grander subjects like the economic pipeline of spam.

5 Conclusion

I have presented an empirical study on the drivers of spam on YouTube and investigated spam evolution against increasingly rigid updates. First, we see that platform-specific qualities allow spam to prosper based on the type of service social media platforms are designed for. In the case of YouTube, which is a for-profit company based on ad revenue, spam prospers from the continued community activity. Separately, we also see how the UI and platform customization features plays a part in spam obfuscation. We dived into the economics of YouTube in an effort to explain why YouTube was unwilling to fix spam despite the overwhelming community grievances.

Apart from this, we also see the categorization of spam before and after YouTube's mitigation update. Initially, spam was mostly segmented from original comments and could be dealt with using a prediction model. After the update, we get a glimpse of new spam obfuscation that would render classical models obsolete. At the same time, we also see how account verification plays an important step in mitigating spam which would be otherwise undetectable by text analysis alone.

Finally, we get an overall glimpse of what the future holds for spam mitigation from the emergence of AI. Overall, despite the lack of time needed to fully develop the analysis it is interesting how features from the categorization of spam under the platform can provide a baseline for spam evolution.

ACKNOWLEDGMENTS

A big thank you to Professor Stringhini of Boston University for allowing the opportunity to design and research this project. Another special thanks to UCI's machine learning repository for the YouTube spam dataset.

REFERENCES

[1] Aziz, A. & Mohd Foozy, Cik Feresa & Shamala, Palaniappan & Suradi, Zurinah. (2018). Youtube spam comment detection using support vector machine and K–nearest neighbor. Indonesian Journal of Electrical Engineering and Computer Science. 12. 607-611. 10.11591/ijeecs.v12.i2.pp607-611.

[2] Ceci L. (2023, Feb 7). Most popular YouTube videos based on total global views as of February 2023. Statista. Retrieved May 5, 2023, from https://www.statista.com/statistics/249396/top-youtube-videos-views/

[3] Ceci L. (2023, April 14). Distribution of removed YouTube comments worldwide as of 4th quarter 2022, by removal reason. Statista. Retrieved May 5, 2023, from https://www.statista.com/statistics/1133165/share-removed-youtube-video-comments-worldwide-by-reason/

[4] Ceci L. (2023, April 14). Number of YouTube video comments removed worldwide from 3rd quarter 2018 to 4th quarter 2022. Statista. Retrieved May 5, 2023, from https://www.statista.com/statistics/1132989/number-removed-youtube-video-comments-worldwide/

[5] Forristal, L. (2023, April 25). YouTube continues to see ad revenue decline, 2.6% drop yoy. TechCrunch. Retrieved May 5, 2023, from https://techcrunch.com/2023/04/25/youtube-q1-2023/

[6] Jospeh.J. YT-Spammer-Purge, (2021), GitHub repository, https://github.com/ThioJoe/YT-Spammer-Purge

[7] MOUNTAIN VIEW, Calif. – April 25, 2023 – Alphabet Inc. (NASDAQ: GOOG, GOOGL) today announced financial results for the quarter ended March 31, 2023.

[8] Nelson, Z. (2023, March 9). I caught the YouTube scammer - $1000 dollars every day?! YouTube. Retrieved May 5, 2023, from https://www.youtube.com/watch?v=iROF9Dd7FXA

[9] Robyn Speer. (2019). Ftfy (Version 6.0). Zendo. https://doi.org/10.5281/zendo.2591652