Revisiting YouTube Algorithm's Allegation Of User Radicalization

Sangjin David Lee
Computer Science, NYUAD
sangjin.lee@nyu.edu

Advised by: Talal Rahwan, Yasir Zhaki, Marcin Waniek

ABSTRACT

Since its launch in 2005, YouTube has enjoyed ever-growing popularity as the largest video-sharing platform on the internet. Their marketability can mainly be attributed to the possibility of content monetization and advanced user experience. Alongside its expansion, however, there have been growing concerns on the effects of its veiled recommendation algorithm. Specifically, Hosseinmardi's 2021 paper *Examining the consumption of radical content on YouTube* contends that the growth of YouTube has prompted fears that viewers are being radicalized through a mix of subjective recommendations which have been alleged to divert focus to extreme political content. Due to YouTube's questionable recommendation amidst convenient accessibility to various demographics, a vast number of viewers are exposed to politically charged videos without much stoppage. In this paper, we explore ways to reverse engineer the network of YouTube to analyze recommendations customized to each user. We scrutinize the difference in algorithm's behavior by obtaining the viewing history for each user. Based on the set of channels and their political classification from Hosseinmardi paper, we create a set of videos belonging to six classes. Then we employ a heuristic that generates random users who fit into each class. By injecting varying numbers of different class videos into the viewing history of users from different classes, we determine the change in algorithm's behavior. Our research evaluates the reaction of YouTube's recommendation algorithm to disclose the degree to which the platform systematically feeds radical content to the viewing eyes.

This report is submitted to NYUAD's capstone repository in fulfillment of NYUAD's Computer Science major graduation requirements.

جامعة نيويورك أبوظبي
NYU ABU DHABI

Capstone Project2, Spring 2022, Abu Dhabi, UAE

KEYWORDS

Recommendation system, YouTube, NLP, Computational Social Science, Deep Learning, Political classification, Database systems

Reference Format:
Sangjin David Lee. 2022. Revisiting YouTube Algorithm's Allegation Of User Radicalization. In *NYUAD Capstone Project2 Reports, Spring 2022, Abu Dhabi, UAE.* 7 pages.

1 INTRODUCTION

The course of the capstone research had been changed at the end of Summer 2021. The author has decided to include the original research as part of this report as a record of the exploratory undertaking for Computer Science capstone. The following introductory paragraphs will outline both the initial and the modified approach to research.

1.1 Initial Approach

The initial research approach was to explore ways to reverse engineer the network of YouTube to determine where the algorithm takes children viewers and derive measures to escape it. We attempted to scrutinize the network by crawling through YouTube's recommended videos for a duration of different times in Chrome's incognito mode, as previous studies show that user history affects the algorithm. We intended to use the crawled results to delineate the network graph in which a node represents a video and edges represent the link between videos. Then employing a classifier to identify disturbing videos, we aimed at determining the number of hops and minutes elapsed between the initial children video and the disruptive video that includes but not limited to nudity, violence, profanity, and drugs. Following this finding, we planned to develop a method to inject noise in the user's trajectory so that children users could be safeguarded from being led towards inappropriate content by YouTube's algorithm.

Our intention was to disclose the close proximity between children videos and disturbing videos, indicating YouTube's

weak regulatory measures on content or even intentional lead towards provocative videos. We also wanted to raise awareness for parental guidance on the platform as more children gain access to YouTube and are often left unattended.

The recommendation algorithm of YouTube is deeply diffused into the vast number of contemporary lives worldwide. With the rampant rise in connectivity through smartphones, high-speed internet, as well as advanced video technology, YouTube has been able to acquire and retain the highest number of users among all existing video platforms. According to the statistics reported by YouTube, more than 2 billion logged-in users visit YouTube every month, which adds up to nearly one-third of the Internet. People also spend over billion hours watching videos and render billions of views on YouTube. As hinted before, over 70% of the platform's watch time is generated from mobile devices in 100 countries and more. Other noteworthy aspect is that YouTube's prime time market is 18 to 34 year olds, reaching greater number of individuals in the U.S. audience than any television network on mobile alone.

YouTube surely reaches the younger generation as well, in fact, children. According to a recent Pew Research Center survey of U.S. adults, YouTube plays a critical role in delivering entertainment for youth. 81% of all parents with children aged 11 and under claim they have allowed their child to watch YouTube videos at least once. And 34% of parents state their child regularly watches YouTube videos. It should be remembered that YouTube expressly notes that it is not meant for children below the age of 13.

Many users are turning to content on YouTube to help them understand the world and learn new things, yet large shares say they encounter negative experiences with content on the platform. Approximately two-thirds of users (64%) claim they at least occasionally come across videos that are blatantly fake or untrue when accessing YouTube, and 60% say they at least occasionally encounter videos that display individuals engaged in harmful or disturbing conduct. In addition, 61% of parents who allow their young child to watch content on YouTube claim they have come across content that they thought was inappropriate [3].

So is the YouTube algorithm naive after all? According to BBC Trending, thousands of videos on YouTube that seem to be adaptations of iconic cartoons, in fact, feature shocking and offensive material that is not safe for children. There are numerous YouTube videos containing cartoon characters from the Disney movie such as "Frozen, the Minions franchise, Doc McStuffins, and Thomas the Tank Engine" that can seemingly pass for the real cartoon. For instance, "Toys and Funny Kids Surprise Eggs," is one of the top 100 most viewed YouTube channel in the world, with over 5 billion views on its episodes. The channel's banner features a variety of child friendly cartoon characters. However, as mentioned in the writing, the titles of the videos on the site include "FROZEN ELSA HUGE SNOT," "NAKED HULK LOSES HIS PANTS," and "BLOODY ELSA: Frozen Elsa's Arm is Broken by Spiderman" and contain violence as well as explicit toilet humor [2].

1.2 Modified Approach

The need for the modified approach arose during the execution of the initial approach. We crawled video data based on YouTube's recommendation using a breadth-first search and looked into a total of 75 to 120 recommended videos – depending on the number of recommended videos rendered on the screen– per encountered video in the recommendation chain. After having stored all videos, we extracted the number of inappropriate content the crawler has encountered by looking at the 'forKids' tag as the first layer of inspection. Consequently, we found that there were 0 inappropriate video out of nearly 7000 videos. Therefore, even in this preliminary stage, we were able to find out that YouTube's algorithm was doing a fairly decent job at sieving out unwanted content for kids. At this point of the exploration we felt that it was unnecessary to proceed with the experiment as our hypothesis that YouTube's recommendation algorithm will inevitably provide content children must not watch is largely unsupported by the statistics.

The newly set orientation of the research now focuses on a different issue of YouTube's recommendation algorithm. The platform's popularity has exacerbated the latest worries about YouTube users becoming radicalized by a mix of biased recommended videos which have been said to steer attention to extreme political content. Therefore, we decided to study the effect of YouTube's recommendation on political orientation of users. Maintaining the unique aspect of our study, which is to set YouTube as the subject of the experiment, we aimed at analyzing the political orientation of recommended videos for user classes with varying viewing history.

2 RELATED WORK

Until today, there have been various efforts to analyze internet platforms such as audit studies on Amazon, detecting videos on YouTube, or even mapping out YouTube [13]. To outline some of the most relevant efforts to our research, one website called *Algotransparency* discloses what videos are promoted by Youtube's recommendation algorithm under certain search queries. It then lists the recommended videos in the order of how common they are with regards to the subjects [1]. Another research named *Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube* does an audit experiment to scrutinize whether personalization plays a role in intensifying misinformation. The paper concludes that a watch history developed for a user

determines the extent to which the person will be recommended misinformation and, in turn, one misinformative video further leads to more recommendations with misinformation, resulting in the creation of a filter bubble effect [5]. A more relevant research with respect to the network graph of YouTube could be the paper *Measurement and analysis of online social networks* [9]. The paper finds that there exists a trend in social networks where high-degree nodes tend to connect to other high-degree nodes, forming a core or a cluster of the network. It states that YoutTube's network is the exception to this trend due to the fact that YouTube has a more "celebrity-driven nature" in which numerous unpopular users connect to a few exceptionally popular users. In investigating a route that leads from one children video to eventually an inappropriate video, we will refer to the previous research which finds that nontransparent YouTube algorithm even leads users to gradually watch more radical and extreme videos [11]. The most important research is the *Examining the consumption of radical content on YouTube* [4] that claims they find no evidence that YouTube recommendations generate systematic interaction with far-right videos, nor do we discover distinct proof that anti-woke channels provide as a bridge to the far right. Instead, they say that consumption of political information on YouTube appears to reflect individual tastes that cover the entirety of the web.

3 METHODOLOGY

3.1 Initial Approach

The uniqueness in our research lies in that YouTube is the subject of the experiment. In other words, our unique approach is that we provoke the algorithm itself to observe its effect. To form an environment to conduct our analysis, we need to construct a network of children videos. We initially do so by starting to crawl data from a single video of an input query. Data will be acquired from this parent video and its recommended videos that usually range from 100 to 120 in number. So, for instance, if crawling starts from one video, all its data will first be scrapped and then the video IDs of its 100 120 recommended videos will be put into a queue so that the crawler moves onto each of them in order. By repeating this process for every video, we could record the parent-child relationships between different videos. Every time a video undergoes a crawling process, its counter will be updated so that the video is not crawled no more than 10 times in repetition. The reason for this is that a parent video and a child video could keep recommending each other back and forth in a loop. Although this would indicate that these pair of videos are dominant in the network, we set a threshold to the loop to reduce the unnecessary computing power.

The data acquired for each video will be as the following: *video ID, title, description, tags, channel title, published time, duration, comment count, dislike count, like count, view count, for-Kids Boolean, recommended videos, parent video, counter.* Crawled data will be stored in a MySQL database where there will be two tables. The first table consists of all the features of a video as mentioned above. The second table consists of columns id, parent, and recommended video, in order to keep a separate record that would help delineate the networks graph of parent-child relationship between videos. Along with text data, a set number of frame images from each video will be stored so that we can employ image classification on these snapshots and determine whether or not the corresponding video is, in fact, a children video [6] [7].

Once the data is crawled, processed, and stored, we can identify the centroids because we need to have a set of dominant children videos. This is necessary to uncover what videos would be linked to all child friendly videos and so what videos are at the core of the clusters. We then have the user artificially get stuck in a cluster of this network. We render gravity to this cluster so that the user is not recommended disruptive videos outside the cluster of children videos he/she resides in.

So this begs the question: how would we then test and fine-tune the gravity we create? In order to do this, we need to take into account the agency problem, which is a conflict in interest between a service and a user.

So we will also design a heuristic that aims to escape the gravity. This is carried out by choosing a path that leads the user far away from the cluster and test whether or not the user is still recommended the same type of videos. Children don't just click randomly in real life so we will maximize the chance of escape by clicking videos that will most likely break out of the cluster.

As mentioned before, we discarded this initial approach and decided to change the course of our research upon the discovery that there was no extreme content on the periphery of children videos. This means that we were not able to identify a video within multiple hops from an arbitrary children video any significant content that could be deemed inappropriate. Acknowledging that there is currently little to no need to investigate further, we have decided to move on to our modified approach.

3.2 Modified Approach

3.2.1 Datasets. Hosseinmardi's research contains a set of channels, along with their classification into 6 categories: far left (fL), left (L), center (C), anti-woke (AW), right (R), and far right (fR). Given a channel with label (i.e., class) X, all videos in that channel have the same label, X. For any video that is no longer live on YouTube, the label was determined using a dedicated classifier. In order to revisit the allegations of YouTube's user radicalization, it was imperative that we

also build a dataset containing channels and videos and their respective political classifications. We learned that Hosseinmardi's dataset is, in fact, a product of merging two different datasets from Ribeiro [11] and Ledwich [8]. By merging the two datasets, we were able to obtain approximately 8330 channels and their political leanings.

	CHANNEL_TITLE	CHANNEL_ID	Ribeiro_Label	Ledwich_Label	Ribeiro_Label2
0	AltRight.com	UCSTy-H5iSiCcozas32sfJlQ	PartisanRight|WhiteIdentitarian	Alt-right	R
1	AmRen Podcasts	UCyZVnp-_owuoPlzNJNtaxZQ	PartisanRight|WhiteIdentitarian	Alt-right	R
2	AmRenVideos	NaN	NaN	Alt-right	NaN
3	Ayla Stewart Wife With A Purpose	NaN	NaN	Alt-right	NaN
4	Baked Alaska 2	NaN	NaN	Alt-right	NaN
...	...	...	...	...	...
8325	Joe Rogan University - Fan Channel	UCrm7RGOsstgF4qAG5mFGK-w	AntiSJW	NaN	L
8326	Political News Networks	UCmhT57vCaS4kPlalY_rrysg	NaN	NaN	L
8327	Philip DeFranco	UCIFSU9_bUb4Rc6OYfTt5SPw	NaN	NaN	C
8328	LastWeekTonight	UC3XTzVzaHQEd30rQbuvCtTQ	LateNightTalkShow|PartisanLeft	NaN	L
8329	NowThis World	UCgRvm1yLFoaQKhmaTqXk9SA	MissingLinkMedia|PartisanLeft|SocialJustice	NaN	L

8330 rows × 5 columns

Figure 1: Hosseinmardi dataset reproduced by merging two different datasets from Ribeiro and Ledwich.

Since Hosseinmardi did not have the information as to which videos inherit which political labels from either one of the sources, we had to decide which label to take from the two sources for every video. In case of conflicting labels, we followed the label from a more reliable source, Ribeiro. Other things to note for data processing were that some channel IDs were outdated or invalid and were simply no longer existing. We used YouTube API to obtain the appropriate channel IDs for a number of entries and filled up the gap. Moreover, in order to make the analysis more convenient, we translated each political classification into a integer representation (*[fL, L, C, AW, R, fR] = [1, 2, 3, 4, 5, 6]*). The preliminary prepossessing of data yielded a reduced number of rows at 6391 channels. The distribution of channels based on the labels were as the following figure.

LABEL	CHANNEL_TITLE
1	46
2	1028
3	527
4	628
5	1920
6	2241

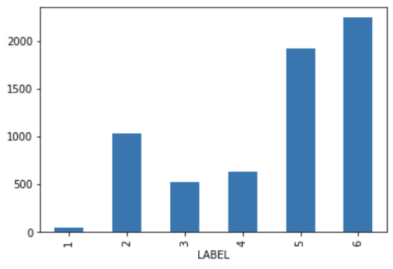

Figure 2: Distribution of channels based on their political labels

Once we had a list of channels and their political labels, we created a dataset of all videos that belong to each channel. For every channel we had, we used YouTube API to crawl and gather all the video IDs that it contained. In the end, we were able to crawl up to 407,847 video IDs. As a general rule of thumb outlined by Hosseinmardi, we made so that each video inherited its parent channel's label.

To gather as much relevant video features to feed the classification model later on, we crawled the following video metadata: *title, description, transcript, top 10 comments*. Other than video IDs, comments, and titles that were obtained using YouTube API, all descriptions were obtained using a different API called PyTube. Transcripts were obtained using YouTubeTranscript API and then uniformly truncated to 512 tokens using T5-Base Abstractive Summarization model. The reason why we tried as much as we could to use various APIs to acquire metadata was that YouTube API became the bottleneck of our experiment with its limited daily quota on requests. Also, we resorted to using an abstractive summarization model instead of an extractive one because the abstractive model generally has a better performance and reproduces essential parts of written text in a novel approach after interpretation using advanced natural language processing.

An important aspect of our research was that we had to replicate a real user experience on YouTube to render a clear evaluation on YouTube's behaviors. So in order to most effectively resemble the real-life experience on YouTube, we designed a heuristics that would randomly pick videos with a threshold on views. The threshold on views was necessary because it is unlikely for YouTube to recommend an arbitrary user a video with low views on a certain topic. We decided to make a pool of videos and select videos based on their weights, which in turn, are based on the number of views. It was necessary for us to inspect the distribution of views for each label to be confident that there is no video with an anomalous view. The distribution of views for each label is shown in Figure 3.

With the appropriately labeled data acquired, processed, and stored, we could now move onto facilitating two things in a nutshell: constructing user profiles of different political leanings and classifying recommended videos as these users are exposed to videos of different levels of political spectrum.

At this stage, we have several channels from each class, and several randomly-generated users from each class. Now, we can do the following: For each video class, X in fL, L, C, AW, R, fR, we randomly select N videos from class X. Then, for each user class Y in fL, L, C, AW, R, fR, we pick M users from class Y to inject the N videos into each user's viewing history and quantify how the recommender reacts (e.g. by doing the following for a randomly-selected video V). Next, we record the generic recommendations displayed after viewing V, when there is no viewing history. We record the recommendations displayed for the M user of class Y. After that, record the recommendations displayed for the M

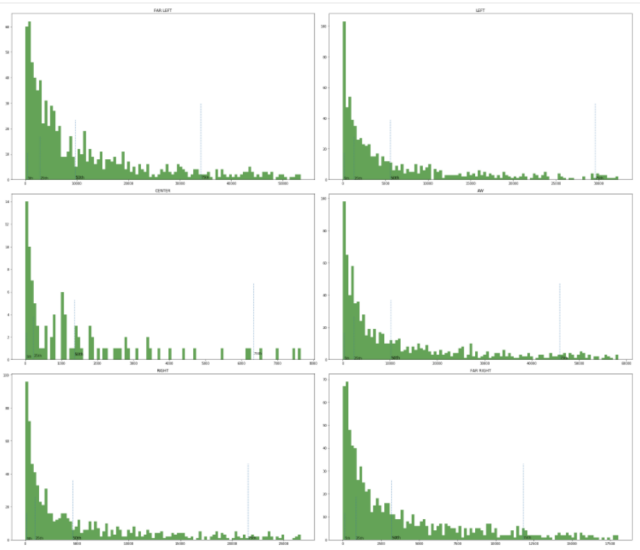

Figure 3: Distribution of views for each label

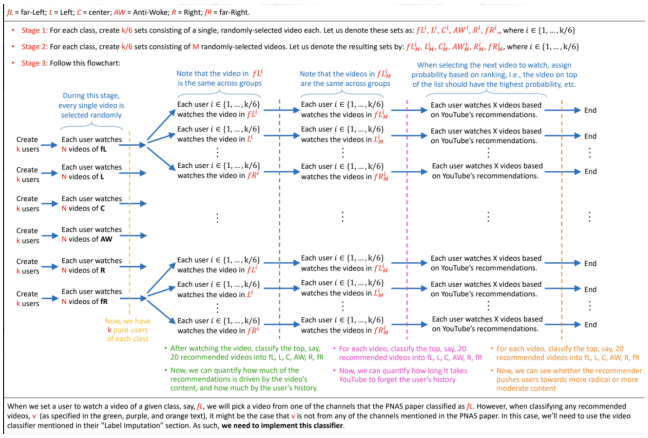

Figure 4: Pipeline for replicating building user profiles, user experience, and classifying recommendations

user of class Y after injecting N videos of class X. Compare the above three, to understand how the recommender reacts to users' viewing history. Notice how, in the above steps, we can have different combinations of classes. For example, we can inject fL videos into fL users, or into fR users, or AW users, etc. In the above step, we can set M to be, say, 100 users, and we can vary N, to see how the recommender reacts. Possible values of N could be, e.g., 1, 2, 3, 4... or if this is not enough see a visible change between the different values of N, then maybe 5, 10, 15, ... or maybe even 10, 20, 30, ... What matters is to see a difference in the recommender's behavior for different values of N [4].

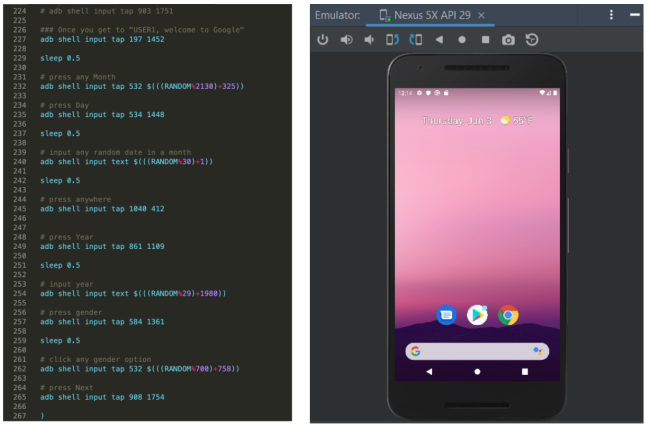

Figure 5: An example of an ADB script and how it can be run on a mobile phone to automate the user experience of watching videos on YouTube

3.2.2 Emulating a real-life user. A very crucial starting ground of our experiment was that the user experience should be replicated on a blank slate. What this means is that we want to create an environment where our virtual user's profile does not get spoiled and influenced by the cache or existing device information. To eradicate this problem, we decided to run our experiment on Android phones (the number of mobile phone would vary based on the number of k mentioned in the pipeline) that have gone through factory reset with an Android Debug Bridge (ADB). ADB allows for an automation of screen activities including screen touching, typing, swiping, and practically anything a human would physically do on a mobile phone. We made a new YouTube account for every user involved in the experiment. To ensure the blank slate starting point, we purchased the corresponding number of virtual sim cards and signed up using ADB without any real human intervention. After this point, a user would click on video recommendations or open video links so that their user profile could be built up from scratch.

When we set a user to watch a video of a given class, say, fL, we will pick a video from one of the channels that the Hosseinmardi classified as fL. However, when classifying any recommended videos, v (as specified in the green, purple, and orange text in Figure 4), it might be the case that v is not from any of the channels mentioned in Hosseinmardi's paper. In this case, we'll need to use the video classifier. Therefore we built a Single Variable Multi-Class Text Classification model to classify recommended videos at each stage.

3.2.3 Video Political Classifier. In designing a video political classifier, we tried to look for a multivariate multi-class classification model that could take in various text inputs and output a prediction from multiple target classes. However, the reality was that there is no multivariate classification

model developed thus far, hence, we decided to settle with a single-input model. We achieved this by concatenating string values from multiple features and making it into a single feature.

Due to the fact that we have multiple features for every video, we had to make a decision as to which combination of features would render the best accuracy of the model. Concatenating all features was not the ideal option because it simply made the argument verbose and confuse the model upon training. After running 15 different measures of model fits to find the best accuracy, we found that concatenating title, description, and transcript into one input gave the best accuracy.

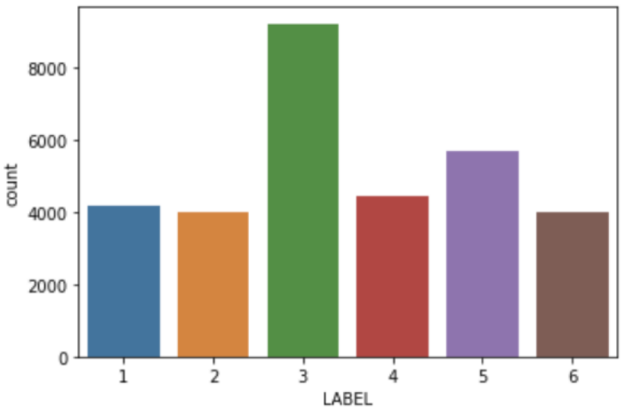

Figure 6: Number of total videos per label

When setting the number of videos to train for our classifier, we limited the number of videos per class to 4000 so that we avoid putting unnecessary weight on a certain class. For example, it just happened so that the video crawler gathered too many *center* videos presumably since this category contains news videos and the new channels on YouTube had the most number of content.

Starting from a 1D CNN layer, leaky ReLu, 1D MaxPool, and finally to a BSTM layer, our text classification model was designed so that it accommodates as much text as possible, especially when different text chunks are glued together. Eventually, the model showed the accuracy of 84%. The research is yet to use the developed model in carrying out the user simulation and classifying YouTube's recommendations. Our next goal is to raise the accuracy to over 95% by manipulating a variety of parameters and even trying out different activation functions within the model. With the more accurate model, we will be able to run the automated pipeline of user experience and acquire statistics on YouTube's behavior based on individual watch history.

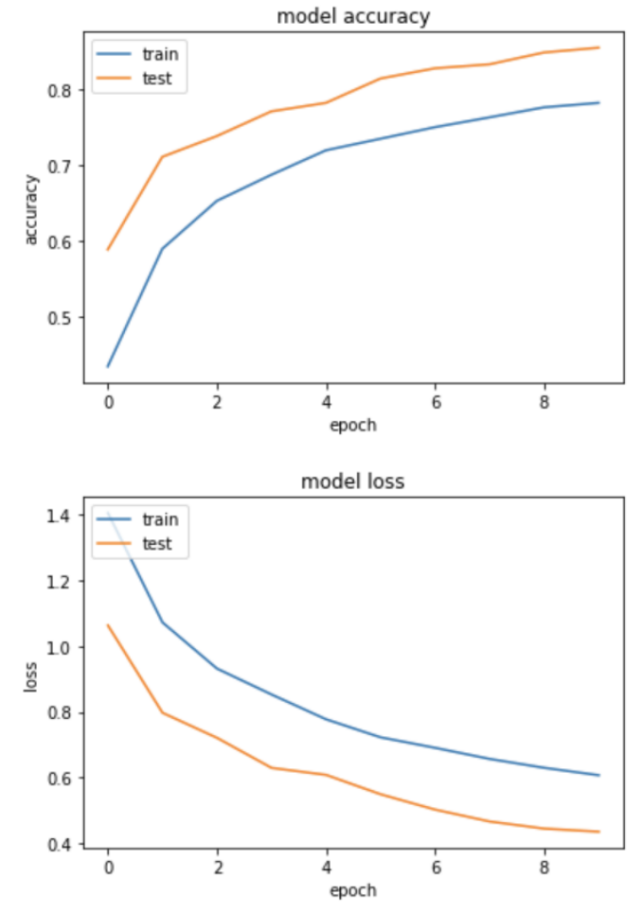

Figure 7: Model accuracy and loss of the text classification model

4 DISCUSSION

The research aims to disclose the effect of YouTube recommendation and how the platform could potentially and systematically influence the political orientation of the viewership. Today, it has become a norm for the majority of internet users to consume video content on YouTube in their daily lives. The contemporary digital audience is exposed to and adopt content far quicker than earlier generations. A sequence of videos could have a more serious impact on one's view of the world than we normally assume. While we are grateful for the wide distribution of modern information technology, it raises the question of whether YouTube recommendations are beneficial to the development of our political perspective that both willingly and unwillingly consume various digital content. Especially, if we are unknowingly driven by the algorithm to only consume certain content to serve some entity's needs, the finding would have a huge potential

to open up a whole new discourse on the infringement of rights, privacy, and also free will in our digital world. The research will be able to render a more clear delineation on the extent to which our political attention is systematically steered by YouTube's recommendation algorithm.

REFERENCES

[1] 2021. Algotransparency. http://www.algotransparency.org/

[2] 2021. *The Disturbing YouTube Videos That Are Tricking Children.* https://www.bbc.com/news/blogs-trending-39381889

[3] 2021. *YouTube Press.* https://www.youtube.com/intl/en-GB/about/press/

[4] Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences* 118, 32 (2021).

[5] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27. https://doi.org/10.1145/3392854

[6] Gurjyot Singh Kalra, Ramandeep Singh Kathuria, and Amit Kumar. 2019. YouTube Video Classification based on Title and Description Text. *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (2019). https://doi.org/10.1109/icccis48478.2019.8974514

[7] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014). https://doi.org/10.1109/cvpr.2014.223

[8] Mark Ledwich and Anna Zaitsev. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211* (2019).

[9] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.* 29–42.

[10] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 522–533. https://ojs.aaai.org/index.php/ICWSM/article/view/7320

[11] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020). https://doi.org/10.1145/3351095.3372879

[12] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 131–141.

[13] Bernhard Rieder, Òscar Coromina, and Ariadna Matamoros-Fernández. 2020. Mapping YouTube. *First Monday* (2020). https://doi.org/10.5210/fm.v25i8.10667