How does YouTube handle the COVID-19 conspiratorial content?

Master Thesis

presented by

ALBERT Léna

to obtain the degree of **Master 2 MIAGE IKSEM**

of the University Paris 1 Panthéon - Sorbonne

Academic Year: 2020 - 2021

Date of Master Thesis Defense: 2021-09-24

Master Thesis Director: Nicolas HERBAUT

Jury Member: Irina RYCHKOVA

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the pedagogical team of the UFR27 of the University Paris 1 - Panthéon Sorbonne, for their guidance and their teaching throughout my 5 years of MIASHS and MIAGE studies. I would like to especially thank Mr. Daniel Diaz who gave me, through his teaching, a great appetite for computer science since the L3.

In the context of this master thesis, I would like to thank in particular Mr. Nicolas Herbaut for his availability, his advice, and help given throughout the process of creation of this thesis. Moreover, his investment in the transverse project Streaming For Good allowed me to collaborate with other students.

I would also like to thank Mrs. Irina Rychkova and Mr. Camille Salinesi, who followed the tutoring sessions that accompanied this thesis, for their understanding and advice, but also and especially for the questions and concerns they raised, allowing me to step back and focus on the hard tasks early enough not to lose motivation.

I would also like to thank the L3 team composed of Sylvain Vocale, Pierre Rambert, and Isaïa Guille who realized the major part of the development allowing the collection and storage of the necessary data for this Proof of Concept. I would like to underline the investment and the help received from Sylvain and Pierre who continued to work with me once their "Common Project" was finished.

Finally, I would like to thank my apprenticeship supervisor and President of Sold-Out, Mr. Jean-Maxime Guedel, for allowing me to work as a project manager in his Start Up, for the time, benevolence, and advice regarding computer science and many other topics. I would also like to thank him for allowing me to take my first steps in the Cyber-Security domain and meet other more experimented developer who help me grow in this field of knowledge.

Table of Content

List of Tables

List of Figures

ABSTRACT

As social media platforms became one of the information‘s main sources in the past decade, the importance of content veracity grew. YouTube, as a platform of user-generated content has a great work to do regarding moderation. The platform is confronted to a double standard as, from an ethical viewpoint, recommending misinformation can be harmful in real life, and from a financial view, YouTube has an incentive in keeping the user hooked on their screen, which can easily be accomplished through promoting extreme and fringe content. The impact of such a content promotion can be especially harmful regarding health and was thereupon a great concern along the COVID-19 pandemic. With more than 70% of the watched videos being recommended via the recommendation algorithms, the impact of these machine learning algorithm is to be studied.

To fully understand the recommendation algorithm of YouTube and the personalization that derives from it, we automated a browser, and performed actions on the platform, just as regular humans would do. We created Google accounts and created initial watch histories to better see the personalization. We collected and stored all the data loaded on the pages while performing these runs. We parallelly collected and manually labeled about 6900 videos that were fed to a classifier in order to train and test it; the classifier yields an error rate of 0.151. We then labeled all the collected videos using our classifier.

We found the number of conspiratorial videos regarding flat earth bigger than the number debunking this theory, whereas the number of COVID-19 informative is larger than the number of conspiratorial content on this topic. Finally, we noticed that the proportion of conspiratorial content does not drastically change with the profile, but that the proportion of informative content is more impacted by the watch history (and therefore the profile) for COVID-19 topics. We can say that measures were taken regarding the recommendation of COVID-19 conspiratorial content, but that there is still a long road ahead.

I – INTRODUCTION

As social media platforms became one of the main sources of information in the past decade, the importance of the veracity of content grew. YouTube, with its more than 2 billion monthly active users, is the second largest search engine and the most popular video-sharing platform [3]. YouTube is “often” or “sometimes” used by 53% of the U.S adults as a source of news according to a Pew Research Center survey conducted in 2020 [6]. However, YouTube is a platform of user-generated content and therefore information available is not necessarily verified nor true. YouTube, as many other social media platforms, struggle to mitigate the inappropriate content, partly due to its scale.

More than the availability of misinformative, hateful, conspiratorial, or violent content, the way this kind of content can be recommended by YouTube has been a growing concern recently [28]. The recommendation algorithm used by YouTube is a black-box algorithm and is often called out to be promoting more and more extreme videos, creating radicalization pathways [7]. Due to its economic model, YouTube has a financial incentive in increasing the user watch-time. To do so, the whole point of the recommendation algorithm is to maximize user engagement. It tends to offer videos that endorse and reinforce the viewpoint of the user, creating addicting experiences closing other views and rewarding more extreme and controversial content [34]. This concern is even of bigger interest as more than 70% of YouTube watched content is recommended by this recommendation algorithm [8].

Yet, YouTube has contested these controversies and has announced to take action to reduce “harmful misinformation” and to “tackle hate” [9-13]. After trying to use only Machine Learning (ML) to flag and remove harmful video, due to too much censorship of the ML and skepticism from the Artificial Intelligence and moderation experts, YouTube reintroduced humans in the moderation process [14]. However, these efforts seem to have been concentrated toward only some topics (among which the COVID-19 pandemic), leaving many pseudo-science and conspiratorial content often recommended by YouTube [1,2].

Finally, the recommendation algorithm of YouTube is highly influenced by the viewer’s watch history. The possibility of creating a “filter bubble”, defined as a state of informational isolation where only a single point of view is available to a user, is still in debate in the academic community as some say YouTube does not create such an “echo chamber”. The possibility of creating a filter bubble of extreme content is however a huge concern as it can be harmful; however, the reality of this risk is debated [15].

This paper will focus on the following research question: **How does YouTube handle the COVID-19 conspiratorial content?** To answer this question, we formulated the following questions that we’ll try to answer through this paper.

RQ 1 - Can we effectively detect conspiratorial content on the YouTube platform?

RQ 2 - Does YouTube promote conspiracy theories over information?

RQ 3 - Are COVID-19 conspiracy theories handled differently than other conspiracy theories?

The remainder of this thesis is organized as follows. Section 2 defines and describes the concepts used. Section 3 presents, reviews, and compare related works to expose what we tried to add to the already existing research studying the same topic as this thesis. In the 4th Section we present the methodology followed. Section 5 presents the results, while Section 6 presents a discussion of the results. The 7th section presents the limitations of this master thesis along with the future works. Finally, Section 8 provides a conclusion to this Proof of Concept. In the 9th Section all References are listed while the last section is an Appendix.

II – BACKGROUND

In this Section we present the notions and concepts required for a full understanding of this thesis. We start by defining the terms of pseudoscience, conspiracy theories, and misinformation that are used interchangeably in this thesis. We then present the concept of recommendation algorithm, and then go on with its possible drifts toward filter bubble. Finally, we present the heart of this master thesis: the YouTube platform, and especially its Human-Computer Interaction (HCI)

1. Pseudoscience, Conspiracy Theories, Misinformation

1.1.1 Pseudoscience

Pseudoscience can be defined as a reasoning claiming to be scientific, based on facts and statements but that do not respect the scientific method [35]. Pseudoscience statements can be especially qualified by the use of vague, untestable, exaggerated claims; a closeness refutation and to evaluation by others leading to an over-reliance on confirmation; the accusation of critiquing groups as enemies, and the blaming of problems on little groups of persons or individuals [36]. Even though pseudoscientific beliefs do not necessarily aim to hurt, some pseudoscience content can be harmful, especially in topics regarding health.

This term is not to be bewildered with non-science which are not scientific beliefs, expressed as so [36]. Non-sciences are less hurtful as they are not presented as science, they have less impact on hard sciences (such as health).

1.1.2 Misinformation & Disinformation

Misinformation can be described as information mistakenly presented as facts, with no intention to deceive [37], whereas disinformation is often used in propaganda and presents the intention to deceive [38]. Even though there is no intention to mislead, misinformation is information that has already proven to be false. This type of false information can be referred to as “fake news” (but also verified facts can be referred to as fake news, especially in the political discourse) [37]. Common sense, education, media literacy, are a good ways to determine the factuality of an information. However, due to the huge spread of misinformation and disinformation online, the principle of fact checking can be irrelevant. Furthermore, once a misinformation is commonly accepted by a group of people, the diffusion of a corrective message can be ineffective, especially if the misinformation message was repeated before the correction [39]. Misinformation therefore appears as a vicious circle that spreads exponentially online, and can be harmful, once again in the health domain for instance. The best way to tackle misinformation seems to prevent, and the limit the diffusion of misinformative content.

1.1.3 Conspiracy theories

A conspiracy can be defined as a secret plan led by a group of individuals in order to proceed to unlawful activities, especially with political motivation. Conspiracy theories are the explanation of events with a conspiracy when other explanations are more plausible and that the conspiracy theory goes against the common consensus [40]. Conspiracy theories are hermetic to refutation and use circular reasoning (each argument used for contestation or absence of argument for refutation are reinterpreted and used as proof of the conspiracy) and sometimes pseudoscience; they rely on faith, and thus tend to persuade more than convince [41].

The paper written by E. Hussein et al [3] refers to misinformation while studying 9/11 conspiracy theories, chemtrails conspiracy theories, flat earth, moon landing conspiracy theories, vaccine controversies, M. Faddoul et al [2] refers to these same topics as conspiracy theories, and more precisely as “alternative science and history”. Finally, K. Papadamou [1] refers to flat earth, anti-vaccination, anti-mask as pseudoscientific content. We therefore can affirm that these terms are close enough in their definition and use in these different papers to be exchangeable in this master thesis.

1.1.4 Studied conspiracies

Covid-19: a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This disease led to a pandemic started in 2019 and currently still going on.

- Informative videos are the one related to this topic, presenting statistics, news on national TV channels, scientific findings, explanation of the disease and its symptoms.
- Conspiracy videos are the one claiming this disease does not exist, anti-mask movement, linking 5G and COVID-19, claiming the virus was man-made and/or used as a weapon.

Flat earth theory: a conception that the Earth’s shape is a plane or a disk. The explanation of the sunset, sunrises, moon rises, moon set, and many other astrological phenomena are explained with the presence of a “Dome” above the earth refracting light.

- Informative content regarding this topic is, in this master thesis, referring to videos debunking this theory.
- Conspiracy content are videos explaining the accuracy of the model, “proving” its truthfulness, explaining why “people” don’t want us to know that the earth is flat.

We also collected anti-vaxx content in our ground truth dataset, however we did not use them for reasons explained in Section 4.2.2. From these definitions we created the keywords later used in the search bar of YouTube to study the recommendation of conspiratorial on the search results (cf Section 5.4).

2. Recommendation Algorithm, Personalization & Filter Bubble

1.2.1 Recommendation Algorithm & Personalization

During any use of a search engine, the content is ordered and filtered in a certain way; this is called a recommendation system [42]. The content proposed is the one the most susceptible to interest the user. What is considered to be of interest to a user can be determined either by its profile (age, gender, location, …), by its history online, or by both [43]. The type of content-based filtering is called personalization.

Even without performing a search online, especially on social media platforms, content is filtered according to the user’s preferences and pasts interactions to increase user engagement and time spent on the platform. These recommendation systems are implemented thanks to recommendation algorithms. In the past few years, claims have been formulated, accusing social media platforms to spotlight sensational content to increase user engagement [44].

YouTube does not derogate from this accusation, as conspiratorial and harmful content can be easily found on the platform. YouTube’s recommendation algorithms are black boxes, and are therefore not easy to study from an external eye to audit the truthfulness of these claims. YouTube however did not contradict the fact that many conspiratorial and harmful content is available on its platform, and even announced to take measures to limit the presence and recommendation of harmful content on the platform [9-13].

1.2.2 Filter Bubble

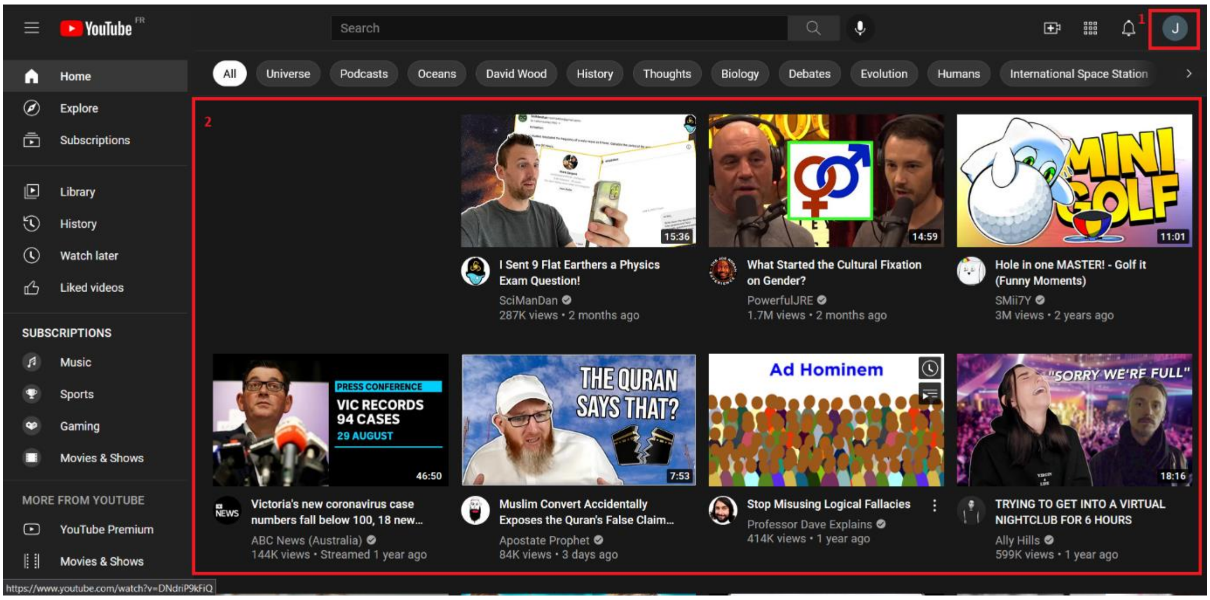
The filter bubble is a term proposed by the internet activist Eli Pariser in 2010 in his book *The Filter Bubble* to designate the state of informational isolation created by over-personalization of content proposed online by the recommendation algorithms [45]. The filter bubble can also be called an echo chamber, which was originally the term applied to news media. Both these terms describe a situation where only viewpoints that are the same as yours are presented to you, repeated, and amplified, leading to a reinforcement of one's beliefs [23]. This reinforcement of the opinion is made without factual support, as surrounded by people sharing the same beliefs. This partitioning of the internet can increase polarization and extremism. This can be a risk for democracy as people need to come across opinions that differ from their own opinions, to develop themselves fully. Otherwise, people might enter a spiral of attitudinal reinforcement and drift towards more extreme viewpoints [24]. Furthermore, it can close someone off to other viewpoints, ideas and interest, and also can create the impression that our interest and ideas are the only ones that exist.

However, this concept of filter bubble is widely disputed. First of all we don't know how much recommendation algorithms are responsible for the creation of such a filter bubble and how much is due to the confirmation bias, as people naturally tend to avoid information that challenges their viewpoints. In communication science, this behavior can be called selective exposure. Some studies therefore try to differentiate the self-selected personalization (the fact that people choose to go towards groups of link-minded people), called "explicit personalization" from pre-selected personalization called "implicit personalization" [24]. Second, debates are still going on whether beneficial or harmful this effect can be. Some studies acknowledge the fact that recommender systems present narrower content over time, but that the user experience is better as the user rates better the content recommended to them [26]. Lastly, in a more academic approach, the lack of clear and testable definition across disciplines often lead to several research addressing the same topic but in different ways, based on different definitions. The lack of empirical data across domains proving the existence of filter bubbles is also criticized [25].

3. The YouTube HCI

YouTube platform is composed of 4 pages: the homepage, the search result page, the page presented to the user while watching a video, and finally the channel page. The recommendation appears in all these pages except for the channel page.

During this master thesis, we will often refer to the homepage which present videos selected by YouTube especially for the user and following global tendencies on YouTube (see figure 1).



I Sent 9 Flat Earthers a Physics Exam Question!
SciManDan
287K views • 2 months ago

What Started the Cultural Fixation on Gender?
PowerfulJRE
1.7M views • 2 months ago

Hole in one MASTER! - Golf it (Funny Moments)
SMii7Y
3M views • 2 years ago

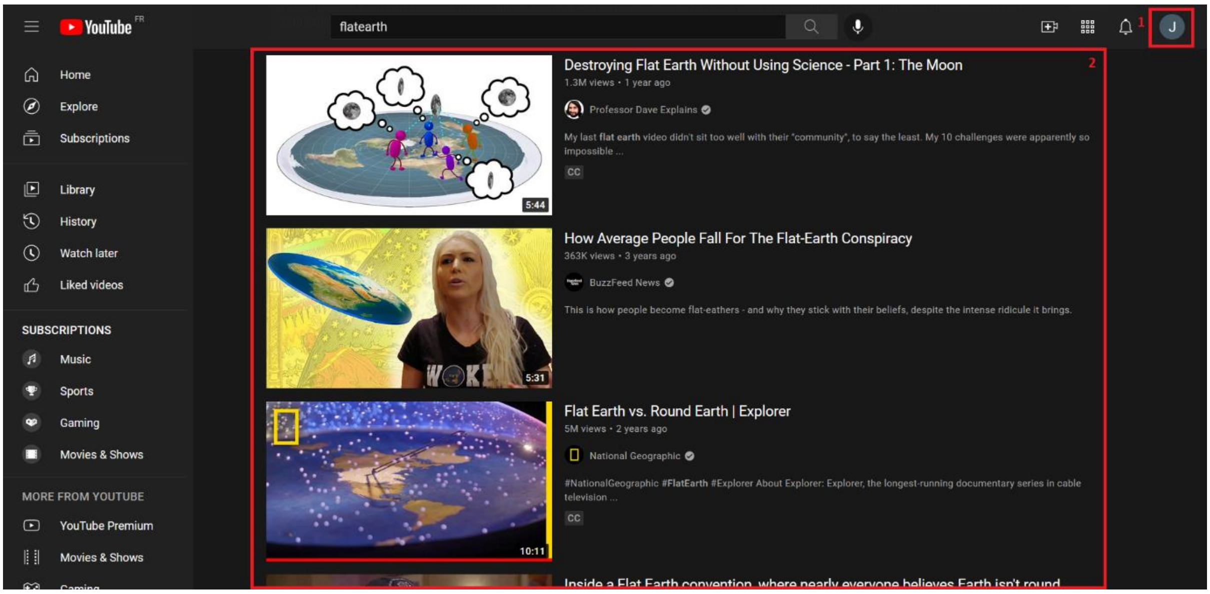
Victoria's new coronavirus case numbers fall below 100, 18 new...
ABC News (Australia)
144K views • Streamed 1 year ago

Muslim Convert Accidentally Exposes the Quran's False Claim...
Apostate Prophet
84K views • 3 days ago

Stop Misusing Logical Fallacies
Professor Dave Explains
414K views • 1 year ago

TRYING TO GET INTO A VIRTUAL NIGHTCLUB FOR 6 HOURS
Ally Hills
599K views • 1 year ago

Figure 1 - The Youtube HomePage. (1) A user being connected. (2) The videos recommended to the currently connected user.



Destroying Flat Earth Without Using Science - Part 1: The Moon
1.3M views • 1 year ago
Professor Dave Explains
My last flat earth video didn't sit too well with their "community", to say the least. My 10 challenges were apparently so impossible ...
CC

How Average People Fall For The Flat-Earth Conspiracy
363K views • 3 years ago
BuzzFeed News
This is how people become flat-eathers - and why they stick with their beliefs, despite the intense ridicule it brings.

Flat Earth vs. Round Earth | Explorer
5M views • 2 years ago
National Geographic
#NationalGeographic #FlatEarth #Explorer About Explorer: Explorer, the longest-running documentary series in cable television ...
CC

Inside a Flat Earth convention, where nearly everyone believes Earth isn't round

Figure 2 - The YouTube Search's Result Page. (1) A user being connected. (2) The videos recommended to the currently connected user according to the search performed

This master thesis also has interest in the recommended video while performing a search on YouTube. The results, and the order in which the videos are displayed are determined thanks to the recommendation algorithm (see figure 2).

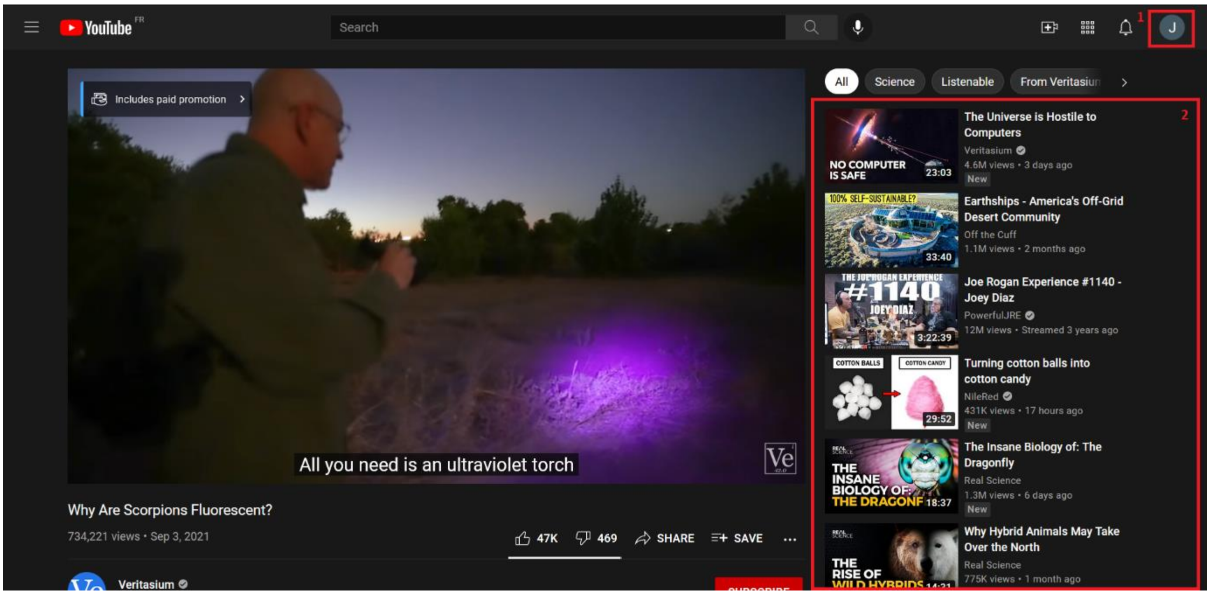

YouTube FR
Search
1
J
All
Science
Listenable
From Veritasium
2
Includes paid promotion
All you need is an ultraviolet torch
Why Are Scorpions Fluorescent?
734,221 views • Sep 3, 2021
47K
469
SHARE
SAVE
Veritasium
The Universe is Hostile to Computers
Veritasium
4.6M views • 3 days ago
New
NO COMPUTER IS SAFE
23:03
Earthships - America's Off-Grid Desert Community
Off the Cuff
1.1M views • 2 months ago
33:40
Joe Rogan Experience #1140 - Joey Diaz
PowerfulJRE
12M views • Streamed 3 years ago
#1140
JOEY DIAZ
3:22:39
Turning cotton balls into cotton candy
NileRed
431K views • 17 hours ago
New
COTTON BALLS
COTTON CANDY
29:52
The Insane Biology of: The Dragonfly
Real Science
1.3M views • 6 days ago
New
THE INSANE BIOLOGY OF: THE DRAGONFLY
18:37
Why Hybrid Animals May Take Over the North
Real Science
775K views • 1 month ago
THE RISE OF WILD HYBRIDS

Figure 3 - Video Playing Page. (1) A user being connected. (2) The recommended videos for the currently connected user according to the video already being watched

Finally, the last page on YouTube that we can see recommendation on, and that we also studied in this master thesis is the page when a video is currently played. On this page, on the right-hand side there is a list of recommended videos personalized according to the user previous interactions on the platform, and the video currently being watched. We will later refer to this part as “the sidebar recommendation”. The first video of the sidebar is considered as the Up-Next video, as this is the one that will automatically play if the user does not select a video himself (see figure 3). Other videos on the sidebar are the recommended videos to the user.

III – RELATED WORKS

In this chapter, we’re going to present the related works, what they studied and found, and how this master thesis places itself among those studies.

Social media platforms became a place of information and a way to access news. Extensive research has been led regarding the influence that a proposed content can have on one’s opinion, and how social media can polarize the user point of view. Traditional social networks such as Facebook have been investigated, especially after the Cambridge Analytica scandal [27]. Academic work studying the truthfulness of the online content on social media has then also been conducted, as the combination of misinformation and polarization can be harmful. The extremization of content and opinion on social media leading to violence has therefore been a great subject of study with twitter for instance. YouTube has been more recently called out for recommending extreme content, and misinformative content, especially since 2019 with several articles from The Times, The New Yorker [28]. Since then, several papers have been redacted regarding this topic.

In 2020, G Chaslot, a French former employee of YouTube denounced the YouTube recommendation algorithm declaring that it promoted extreme content, and therefore conspiratorial content. This led to the paper *A longitudinal analysis of YouTube's promotion of conspiracy videos* in collaboration with M. Faddoul published in 2020 [1]. This paper aims to study the proportion of recommended conspiratorial videos over time, more precisely from November 2018 to February 2020.

This study focuses on three main fields that are alternative science and history, prophecies and online cults, and political conspiracies and QAnon. A ground-truth dataset has been manually collected from books referencing conspiracy theories and websites such as 4chan and reddit (r/conspiracy, r/conspiracyHub …). The final dataset was composed of 1095 videos both conspiratorial and not conspiratorial. From these videos were retrieved the transcript of the video, the video snippet (concatenation of tags, title, and description of the video), the comment, and the perceived impact of the comment. Each one of this information is then labelled as conspiratorial or not by a FastText Classifier; the output labels for these four modules are then combined into a logistic regression layer to predict the global label for the video.

The study focuses on the Up-next recommendation of YouTube, without any watch history. The recommendation of conspiratorial videos consistently decreased from April 2019 to June 2019, by 50% and then 70%, which is consistent with the YouTube announcements.

However, the recommendation of popular conspiratorial videos increased back since its low point in June 2020. This study also finds that the conspiracy likelihood of the Up-Next video highly depends on the currently watched video. This finding highly suggests the importance of watch history, which was not studied in this paper. In this master thesis we worked especially on the effect of watch history on the recommendation of conspiratorial content. We also study the different pages we can see the recommendation algorithm instead of just focusing on the Up-Next recommendation.

On the other side of the spectrum, the work from E. Hussein et al called *Measuring misinformation in video search platforms: An audit study on YouTube* [3] emphasizes greatly on the personalization part of the recommendation system. Especially regarding misinformative content, the creation of filter bubbles has been subject of debate and should not even be possible with YouTube work to moderate “harmful” misinformative content. The topics studied in this paper are 9/11, moon landing, chemtrails conspiracy theories, vaccines controversy, flat earth theory. This study demonstrates that the information of the user’s account such as gender, age, and geolocation does not impact newly created accounts with no watch history. However, gender makes a difference with the existence of watch history with females receiving more conspiratorial content on the Up-Next recommendation whereas male receive more conspiratorial content on the top-5 recommended videos and in the search results.

This study concludes that watch history does impact the recommendation as a watch history composed of videos promoting conspiracy theories significantly increased the number of misinformative videos recommended. An interesting point is that watching videos of one of the five studied misinformative topics did not only increase the number of misinformative videos about this topic, but also the misinformative content about other subjects. The impact of the watch's history is especially notable in the search results, except for the vaccine controversies with the opposite happening: a conspiratorial watch history led to a higher number of videos debunking controversies than with other watch history. This finding might credibly be linked to the effort made by YouTube to reduce the recommendation of anti-vaccination content (as it can be considered harmful). In our study, we do not focus on the personalization made by YouTube regarding the profile, but we do take a further look on the personalization that derives from the watch history.

The third main notable related work is the one conducted by K. Papadamou et al [2] in 2020, entitled *"It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations*. This content especially focuses on the effect of watch history on the YouTube recommendation algorithm regarding several pseudoscience topics such as content link to COVID-19, Anti-vaccination, Anti-mask and flat earth theories

Taking the same approach as [1], they trained several FastText classifiers, able to label each video as pseudoscientific or not for the snippet (concatenation of title and description of the video), the tags, the transcript, and the top 200 comments. With an ablation study, they confirmed that the combination of all these inputs provided a better accuracy; the obtained labels are therefore merged with a dense neural network allowing an accuracy of 0.79. Using a base set of 6,6K videos reduced by about half once irrelevant videos removed, and sorted thanks to crowdsourcing, K. Papadamou obtained a low agreement score for the annotation task and a 0.74 F1-score, reflecting the subjectiveness of classifying a video as pseudo-scientific or scientific. However, it does provide a meaningful insight of the truthfulness of a video.

Using Selenium in headless mode and ChromeDriver, K. Papadamou developed some script to automate the watching of videos and therefore create 3 different profiles: Science Profile, Pseudoscience Profile and a Mixt Profile, each one with its own watching history of at least 22 videos (minimal number of videos estimated by them to notice personalized recommendation), watching 50% of each one (due to the fact that it is not clear how satisfaction score is calculated by YouTube). Once these profiles were made, experiments were run in parallel for the three of them, focusing on the recommendation on the homepage, in the search results, and the top-10 recommended videos.

Finally, this paper concludes that the user’s watch history highly impacts the recommendation of videos. Many pseudoscientific contents are still recommended by YouTube, especially in the search results (more than in the sidebar recommendation or in the Homepage). It also concludes that long standing pseudoscience content has a higher recommendation rate than recent pseudoscience content, especially for the COVID-19 where misinformation has been tackled by YouTube.

This paper was published in May 2021 and was only available as a Preprint before that. It's methodology greatly approaches the one we're conducting in this master thesis as we also developed automation tools with Selenium and ChromeDriver, in order to focus on the influence of the user's watching history on the recommendation algorithm. We also used FastText Classifier as the tool to label our videos. The main differences are that we're including a network dimension to our classifier by labelling the channels of the videos depending on their content and taking these labels into account when sorting the videos, allowing us to have a classifier with a better accuracy. Also, we did not only focus on the recommendation of conspiratorial content, but also on informative content, as comparing the proposition of both seems to be more relevant and has not yet been studied.

IV – RESEARCH METHODOLOGY

To conduct this experiment, several steps were needed to gather data to analyze. We developed a bot watching videos and scrapping information on YouTube. These videos are stored in a MySQL database. The link between the scrapper and the database is made thanks to an API developed in Symfony. A part of the data and metadata regarding the videos are obtained thanks to the YouTube API. This data is collected thanks to a RabbitMQ with Celery Worker, connected to the Symfony API to send back the data to the database. We ran the bot with several Google accounts with different watch histories to better approximate the personalization and recommendation algorithm of YouTube. Due to the huge amount of video to sort and analyze we developed a natural language classifier that can classify videos among several labels.

1. Constitution of a Ground Truth Dataset

4.1.1 Data collection

At the very beginning of the project, we manually collected some data in order to study the feasibility of the project and do some quick tests on the classifier in order to discover how it works, and how to maximize its performance. We watched and manually labelled all the videos to sort them with better accuracy. Due to how time consuming it was, this option was discarded.

The first set of seeds of videos was collected thanks to existing datasets of videos published along papers studying the same topic as this master thesis. We retrieve especially much data from [1] and its dataset partly available [29] and from [3] and its dataset available [30].

We then considered other sources such as team having studied the same topic, but that did not open source their data. The first team contacted was the one of the YouTube-Regrets projects led by the Mozilla Foundation [5]. The project consists of the creation of a plugin that enables the user to anonymously signal a video that they consider harmful. These videos are then manually labeled to verify the accuracy of the claim and can be later on used to conduct studies on the recommendation of harmful content on YouTube. Due to their strict policy, the collaboration did not happen, but they planned to do a follow up project with a more flexible policy and might therefore be able to share their data in a few months.

We then contacted the team of K. Papadamou [2]. They gave us access to their data in very short notice, and once the data was formatted to fit in our database, it was about 2500 videos of either informative or conspiratorial content that were added, and about 4000 irrelevant videos (still useful for the classifier).

4.1.2 Data Collection Challenges

During this step, one of the challenges was to find videos before YouTube limits their diffusion. Indeed, YouTube has tended to limit the recommendation of conspiratorial content since the announcement of new measures in January 2019. Some of the videos listed in the previous studies were not available at the time we collected them to train our classifier. Mainly COVID-19 related content has been limited according to [2]. At the end of this first step, the number of seed videos (video id) was 6900; while the number of videos still available online was 6557. Many these videos did not have a transcript available, or comments were deactivated either by the authors of the video, or by YouTube in its attempt to limit the functionalities of some videos considered harmful.

We manually reviewed all the videos constituting the ground truth dataset, as after a few tests we noticed that many of them were not well classified. K. Papadamou did notice that, as we found a field “author review” on the dataset we accessed, but not many videos were reviewed by the author. As the determination of conspiratorial content can be subjective, and sometimes the limit between pseudo-scientific and scientific content is blurry, as evocated in [2], we defined some inclusion criteria to objectivize as much as possible this step of classifying videos.

I1 - If the video expresses the belief that certain events or situations are secretly manipulated behind the scenes by powerful forces with negative intent.

I2 - The video contains ‘evidence’ that seems to support the conspiracy theory (often described as “simple” or “basic” proofs). These evidences are described as irrefutable.

I3 - The video exposes an alleged secret plot not acknowledged by the majority of people.

I4 - The video divides the world into good or bad.

I5 - The video scapegoats people and groups. (i.e. blames only one little group for everything bad that happened)

A video that meets at least two of these inclusion criteria is classified as conspiratorial. After that step we obtained the following dataset

Label	**Number of Videos**	**% of the total dataset**
Covid Conspiracy	869	12.56%
Covid Informative	710	10.27%
Flat Earth Conspiracy	332	4.80%
Flat Earth Informative	143	2.69%
Anti-Vaccine Conspiracy	339	4.91%

Anti-Vaccine Informative	268	3.88%
Irrelevant	4249	61.49%

Table 1 – Final Ground Truth Dataset

2. The Natural Language Classifier

4.2.1 The choice of NLP classifier

To classify all the YouTube videos, we trained a natural language processing classifier on different parts of the videos which are:

- the title of the video
- the top 200 comments (when available)
- the transcript of the video (when available)

To classify videos, we used an open-source library developed by Facebook AI Research for Natural Language Processing named FastText [46]. FastText is easy to use and very efficient for text classification: it can train a model with millions of examples in around 10 minutes [17]. Furthermore, we compared FastText with the Natural Language Classifier developed by IBM [47] and found out that the Facebook classifier enabled the processing for long text, whereas the IBM one did not.

	FastText	IBM Watson
Functionality	Word Embedding & NL Classifier	NL Classifier
Language	Shell / Python	Curl / Go / Java / Node / Python / Ruby
ML Type	Unsupervised & Supervised Learning	Supervised Learning

Author	Facebook AI research	IBM
Implementation	Bag of n-words, base of n-Grams and ML Algorithms	Automatic Learning Algorithms
Expected Inputs	Any Text	Short Text Entry (maximum of 1024 characters for training and 2048 for testing/classification)

Table 2 – Comparison between Watson Classifier and FastText Classifier

For the task of text classification, FastText represents the text as bags of words. As keeping the order is very computationally expensive but of great interest, FastText combines bags of words with bags of n-grams that allow to keep some information about the local word order. FastText creates low rank matrices to represent word-embedding, as the results of the factorization linear classification which are then fed to a linear classifier. [17]

4.2.2 The FastText Classifier

Before training our FastText classifier, we preprocessed the data as recommended in the documentation [48]. We replaced contractions with full words (for instance “can’t” is replaced by “can not”), we removed the html tags and URLs that might have been present, we removed the special characters (non-alphanumeric characters).

To adapt our classifier as close as possible to our needs, FastText allow to fine-tune several parameters among which:

- the learning rate, which corresponds to how much the model will change after processing each example.
- the epoch, which is how many times the model will see an example during training, this is useful when the dataset is relatively small
- the word n grams, which allows us to keep the order of the words, as explained before.

We used a 60/40 split to both train and test our classifier. We trained our FastText classifier with 60% of the ground truth dataset, then tested and predicted labels for the remaining 40%. We use these predictions as the base dataset for the second part of the classifier responsible for aggregating the results.

4.2.3 Fine Tuning the NLC

We manually fine-tune a FastText classifier on the title feature top optimize its accuracy, using the following parameters the classifier yielded an accuracy of 0,780.

- a learning rate of 0.88
- an epoch of 25
- bigrams.

We then concatenated the top 200 comments of each video, which are the comments considered as “most relevant” by YouTube, and that appears automatically on top of the comments list. We obtain an accuracy of 0.734 with our classifier fine-tuned as follow:

- a learning rate of 1
- an epoch of 50
- bag of words of size 3

We proceeded in the same way for the transcripts of the video, and obtained an accuracy of 0.771 with:

- a learning rate of 1
- an epoch of 50
- and bags of 2 words

Each of these classifiers has a good accuracy (better than the one yielded in the classifiers of the ones in the Related Works). We expect to have better accuracy while combining all these inputs. We formulate the hypothesis that a channel publishing mainly conspiratorial videos, will keep on publishing conspiratorial videos, and same for informative videos. Based on this hypothesis and to improve the performance of the classifier, we computed the label of the channel as the most frequent label of the videos of the channel.

Once all the labels predicted for available features, we combined them to obtain the final classification of a video. To do so, we tried several well-known algorithms of machine learning and deep learning explained below with a split of 80/20 for the training and testing.

3. Combining the Inputs: The Final Classifier

Once the several Natural Language Classifier trained and the labels predicted for each feature, we need to compute the several inputs into one final output that will be the label of the video. To do so, we've selected several algorithms of Machine Learning and we performed a comparative analysis of those.

4.3.1 Distributed Random Forest (DRF)

Decision tree is a model composed of nodes (that can either be the root, an intermediate node, or a leaf) and branches. At each node, a feature is evaluated, and the dataset is split according to the value of the evaluated feature, creating a path to follow from the root to the leaf, where a label can be obtained. To build a good decision tree, different features are recursively evaluated, to determine which feature best splits the data at each node [49].

Random forest is a supervised algorithm of machine learning based on ensemble learning which is based on the idea of combining several algorithms (different algorithms or multiple times the same algorithm) to obtain better results. A random forest is the combination of several decision trees. [51]

We decided to use a decision tree for our classifier as this is one of the most naïve ML implementation for classification tasks. As the implementation of the of random forest in the scikit-learn library (later referred as sklearn) does not natively support string inputs as it considers the variables as continuous [50], we look for other solutions.

- Set a value to each label passed in input, but labels would be considered continuous, and therefore it would have introduced a sense of hierarchy in our labels that did not exist. This option was discarded.

- One workaround that we found was to one-hot-encode the data (or dummy the data). The idea is to represent each input as a vector composed of 0 and 1, indicating whether a categorical attribute is present or not.

Categorical value	One-Hot-Encoded Value
Label 1	[0,0,1]
Label 2	[0,1,0]
Label 3	[1,0,0]

Table 3 – “Translation” example of the categorical label to one-hot encoded label

However, with further research, we discover that one-hot-encoding is making the model worse by including sparsity and the gain of purity per split in the sub algorithm performed by any tree-based algorithm is very marginal [16]. According to this source, a tree with one-hot-encoding likely will look like the right image instead of the left one.

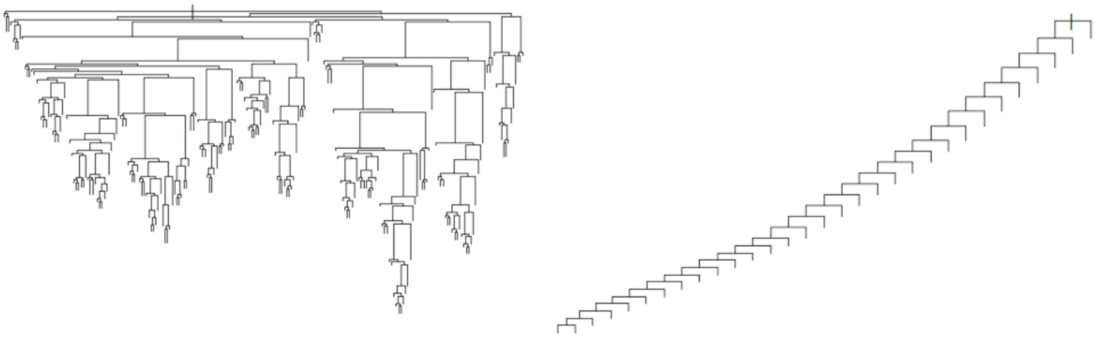


Figure 4 - Comparison between enum and one-hot encoding for trees (retrieve from [16])

We found a relevant Python module, called H2O, associated with the open-source Java-based software H2O developed by the H2O.ai company. This module grants access to the H2O JVM which provides a web server through a single active connection via REST calls. This dispenses a distributed, parallel, in memory process engine with already available learning algorithms which enable easy usage and a fast solution [22]. Furthermore, we found the algorithm already implemented has several options available such as the handling of unbalanced data as we’ve many more irrelevant videos than conspiratorial videos.

We decided to use the H2O module and its implementation of Distributed Random Forest (DRF) which natively supports categorical values [18], with several encoding available:

- One-hot-encode working as explained before
- Enum which maps input strings to integers and uses these integers to make splits. Each category is separate, and its number is irrelevant therefore keeping its categorical nature. For example, after the strings are mapped to integers for Enum, you can split {0, 1, 2, 3, 4, 5} as {0, 4, 5} and {1, 2, 3}.

The metrics shown in the Table 4 are the ones calculated by H2O itself. Even though the ML carry out a classification task, as the classification is multinomial instead of binomial, the precision and accuracy are not automatically calculated, and the metrics are the ones of regression tasks: Error Rate, Mean Square Error (MSE), Root Mean Square Errror (RMSE), Log Loss , Mean Errror Rate Per Class. As expected, the enum encoding yields better results as shown in the following table:

	DRF w/ enum encoding	DRF w/ one hot encoding
Error Rate	0.151	0.166
MSE	0.145	0.149
RMSE	0.381	0.387
Log Loss	0.548	0.583
Mean Per-Class Error	0.369	0.380

Table 4 – Performances of the DRF classifier with a comparison between enum encoding and one-hot encoding, trained and tested with our ground truth dataset

During the testing of both encoding, not only were the results different, but also the training time was about 1.2 to 1.5 times longer for the one-hot-encode.

4.3.2 Gradient Boosting Machine (GBM)

Gradient Boosting is a machine learning method based on ensemble learning, as in Random Forests. The weak learner used in GBM can be trees, once again as in the Random Forest. However, the way the trees are built is different, as in GBM, trees are built sequentially, with errors encountered during the training of the previous tree corrected in the new tree. We therefore tested to combine our inputs with GBM as it is a close method to DRF. Even though it can perform better than random forest with well-tuned parameters, the main drawbacks are that it might be longer to train, and that it is especially sensitive to overfitting (which can be especially bad for noisy data) [52].

Once again, we used the H2O module which had an implementation of GBM already available [20]. The classifier yields the following results:

	GBM w/ enum encoding	GBM w/ one-hot encoding
Error Rate	0.157	0.159
MSE	0.145	0.148
RMSE	0.383	0.384
Log Loss	0.590	0.663
Mean Per-Class Error	0.338	0.346

Table 5 - Performances of the GBM trained and tested with our ground truth dataset

4.3.3 Multinomial Logistic Regression (MLR)

As in the related work of M. Faddoul [1], the classifier combined the inputs with a Multinomial Regression Layer, we wanted to try this algorithm of ML to compare our results with theirs. Logistic Regression is a statistical model that determines the probability of an event with 2 possible outcomes (thus complementary outcomes) i.e. estimate a variable with two possible values. This model performs a binary classification based on independent variables either continuous or binary [54]. This classification is based on the logistic function which can be graphically represented by a S-shaped curve (called sigmoid curve).

The generalization of this model to multiclass problems is called Multinomial Logistic Regression. This model predicts the probability of the possible outcomes of categorically distributed dependent variables based on a set of independent variables of any kind (binary, categorical, continuous…) [53].

To implement and test this algorithm, we used Python and the H2O module which implements a Generalized Linear Model, performing both classification and regression [19] and obtained the following results:

	Multinomial Logistic Regression
Error Rate	0.185
MSE	0.173
RMSE	0.415
Log Loss	0.625

Table 6 – Performances of the Multinomial Regression Classifier trained and tested with our ground truth dataset

4.3.4 Deep Learning (DL): Neural Network

As in the related work from K. Papadamou, the classifier was based on a custom neural network [2], we tried to implement a Neural Network to compare our results to theirs. Deep Learning is based on artificial neural networks. There are several levels between the input and the output, with each layer extracting a higher level of information from its input i.e. each layer is transforming the input to a slightly more abstract representation of information [55].

We here again used the H2O implementation, which is a multilayer feedforward neural network with back-propagation and stochastic gradient descent [21]. A feedforward neural network is a network that does not form a cycle, meaning the layer can only be browsed in one direction. Gradient descent is an iterative optimization algorithm allowing to find a local minimum. The backpropagation allows to compute the gradient descent with respect to the weight one layer at a time. This classifier yields the following results:

	Deep Learning
Error Rate	0.173
MSE	0.146
RMSE	0.382
Log Loss	0.658
Mean Per-Class Error	0.378

Table 7 - Performances of the Deep Learning trained and tested with our ground truth dataset

4.3.5 Choosing the Machine Learning Algorithm

To determine which algorithm best fits our needs, we compared the several tested algorithms, and it appears that the Random Forest yields the best results. Maybe by better fine-tuning the GBM or having a more balance set of data, it would have performed better.

	Error Rate	MSE	RMSE	Log Loss	Mean Per-Class Error
DRF	0.151	0.145	0.381	0.548	0.369
GBM	0.157	0.145	0.383	0.590	0.338
MLR	0.185	0.173	0.415	0.625	N/A
DL	0.173	0.146	0.382	0.658	0.378

Table 8 - Comparison of the several ML algorithm for the classification

These results partly answer the RQ1, as it is clear that no matter the ML algorithm chosen, the classification has relatively a low error rate (about 85% of the videos classified will be attributed the correct label with the DRF model).

4.3.6 Handling of Missing Values

Handling missing values such as comments, captions, or labels of a channel was a concern. These values can be missing for several reason, among which:

- The author of the video deactivating comments
- YouTube deactivating comments in order to limit functionalities of some videos
- Captions that do not exist (auto-generate does not work, no human made subtitles, no voice in the video, …)

In the algorithms that we selected, the missing values in the training set are handled as if a missing value represents information (that is to say, they are absent for a reason). We then found several ways of handling the missing values:

- Do nothing and let the algorithm deal with the missing values
- Skip data with missing values, which will not use all the available information (which, in our case is very often, especially in the beginning where we did not had so much different channels labelled)
- Fill in missing values with the most frequent one in the other feature which might lead to only one input determining the whole label for a video.

	DRF (no treatment of missing values)	DRF (fill in missing values with the most frequent one)
Error Rate	0.151	0.159
MSE	0.145	0.150
RMSE	0.381	0.387
Log Loss	0.548	0.642
Mean Per-Class Error	0.369	0.346

Table 9 – Performances of DRF Classifier with and without pretreatment of missing values

Based on this result, we decided to let the random forest handle the missing values in its own way.

4.3.7 Ablation Study

Finally, we tested our classifier with all possible combinations of inputs to be sure that all features are useful and helpful to better the classifier. We found that, all available features were useful in the classifier even though it was not in the same proportion, as we discover that Title was about 40% of importance in the DRF model that we have trained and tested.

Input features	Error Rate	MSE	RMSE	Log Loss
Title - Comments - Captions - Channel	0.151	0.145	0.381	0.548
Title - Comments - Captions	0.265	0.232	0.481	0.753
Title - Comments - Channel	0.181	0.191	0.437	0.641
Title - Captions - Channel	0.236	0.231	0.480	0.729
Comments - Captions - Channel	0.303	0.279	0.528	0.860
Title - Comments	0.194	0.182	0.427	0.705
Title - Captions	0.293	0.259	0.509	0.841
Title - Channel	0.183	0.168	0.410	0.593
Comments - Captions	0.328	0.313	0.559	0.991
Comments - Channel	0.263	0.246	0.496	0.802
Captions - Channels	0.363	0.335	0.579	1.015
Title	0.183	0.176	0.419	0.721
Comments	0.289	0.286	0.535	1.177
Captions	0.365	0.394	0.628	1.300
Channel	0.346	0.319	0.565	1.155

Table 10 - Ablation study for the chosen classifier (DRF without pretreatment of missing values, enum encoding)

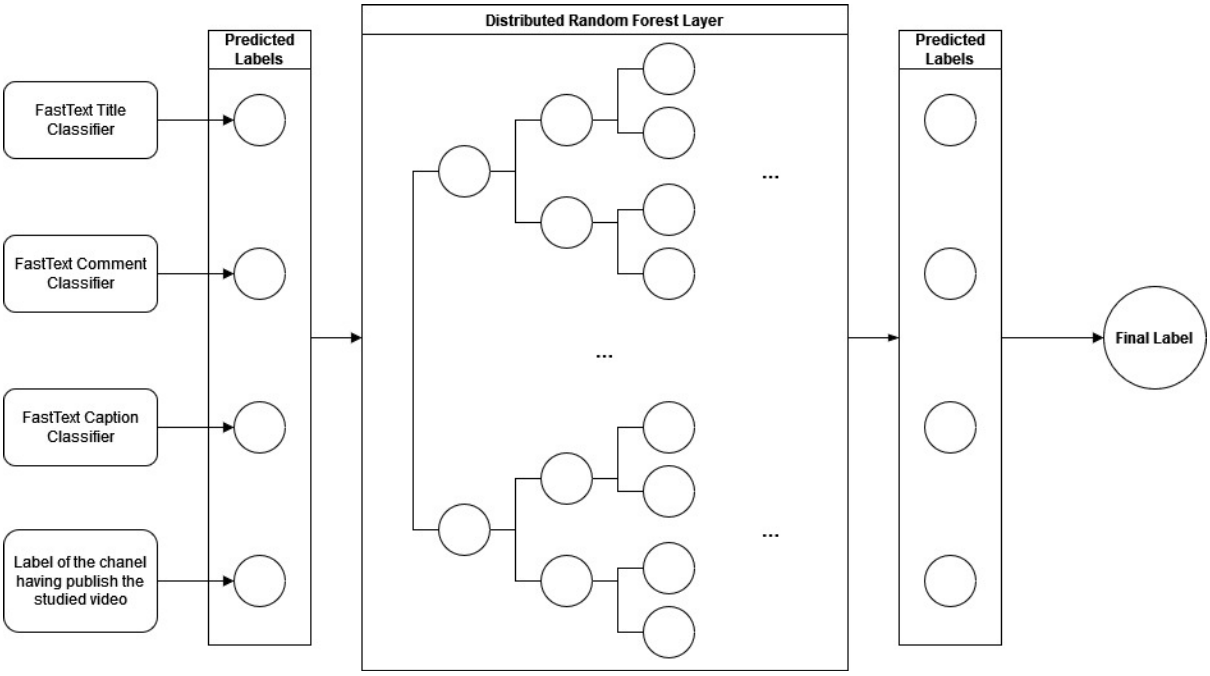

FastText Title Classifier
FastText Comment Classifier
FastText Caption Classifier
Label of the chanel having publish the studied video
Predicted Labels
Distributed Random Forest Layer
...
...
...
Predicted Labels
Final Label

Figure 5 - Schema of the final classifier: (1) Labels are predicted for each feature thanks to Fast Text Classifier. (2) They are gathered according to their video id. (3) Labels are fed as inputs to a Random Forest. (4) The DRF predicts several labels, each one with a probability. (5) The most probable label is considered as the final label for the video.

To fully answer the RQ1, with a H20 Distributed Random Forest, an enum encoding, the handling of unbalanced classes for the ground truth set, 200 trees of maximum depth of 50, our classifier yields an error rate of 0.151. For the proposed classifier the metrics are the following for the classification of conspiracy videos (by aggregation):

Precision: 0.649

Accuracy: 0.866

Recall: 0.713

F1-score: 0.679

During the testing, we saw a better metrics for the COVID topic, than for vaccine informative and flat earth informative content, due to the fact that there is not so many flat earth informative and vaccine informative videos in the ground truth dataset. The available option for unbalance dataset did not suffice to compensate such a discrepancy in our ground truth dataset, with some labels being less than 2% and other being almost 50% of the ground truth dataset.

To give a sense of evaluation of our classifier, we can compare it to the ones presented in the related works: we obtain a better accuracy (0.866 against 0.79), but lower precision, recall and F1-Score (respectively 0.649, 0.713, 0.679 for ours, against 0.77, 0.79, 0.64 for the one of K. Papadamou, and 0 78, 0.86, 0.82 for the one of M. Faddoul).

4. Conspiratorial Videos on YouTube Platform

4.4.1 Architecture

We used Python3 and Selenium browser automatization software, along with ChromeDriver to reproduce a human behavior on YouTube, and to scrap data, which we will refer as “the bot” in this master thesis. The figure represents the bot and the several services implicated in its functioning and in the retrieving of data. Following the figure 6, here are the steps performed:

- The YouTube scrapper is launch, with as argument the types of actions to perform (connecting with what kind of profile, how many videos to watch from an URL, how many videos to watch from the homepage, how many searches to perform and what type of search to perform…)
- (0-3): A call is made to a Golang API that convert the actions listed below to a more detailed json, thanks to data retrieve from the database (the mail and password to use, the URLs to watch, search to perform…).
- (4): The robot then performs the list of actions found in the json thanks to Selenium and ChromeDriver. While browsing on the YouTube platform, each action, id of video loaded on the page are sent to a Symfony API.
- (5.a-5.b): Part of the data such as action and video ids are directly stored in the database. The video ids are also sent to a RabbitMQ.
- (6-8): From the RabbitMQ, 3 messages are created, each one going to a celery worker, allowing to retrieve the title and metadata of the video, the top 200 comments, both with the YouTube API, and the third message launches an instance of the Robot responsible for collecting captions of the videos.

- (9): The responses to the different messages treated by the Celery Worker are then sent back to a second queue in the RabbitMQ
- (11-11): Every night a CRON task is launch and consumes the messages of the RabbitMQ in the response queue and data are sent back to the database

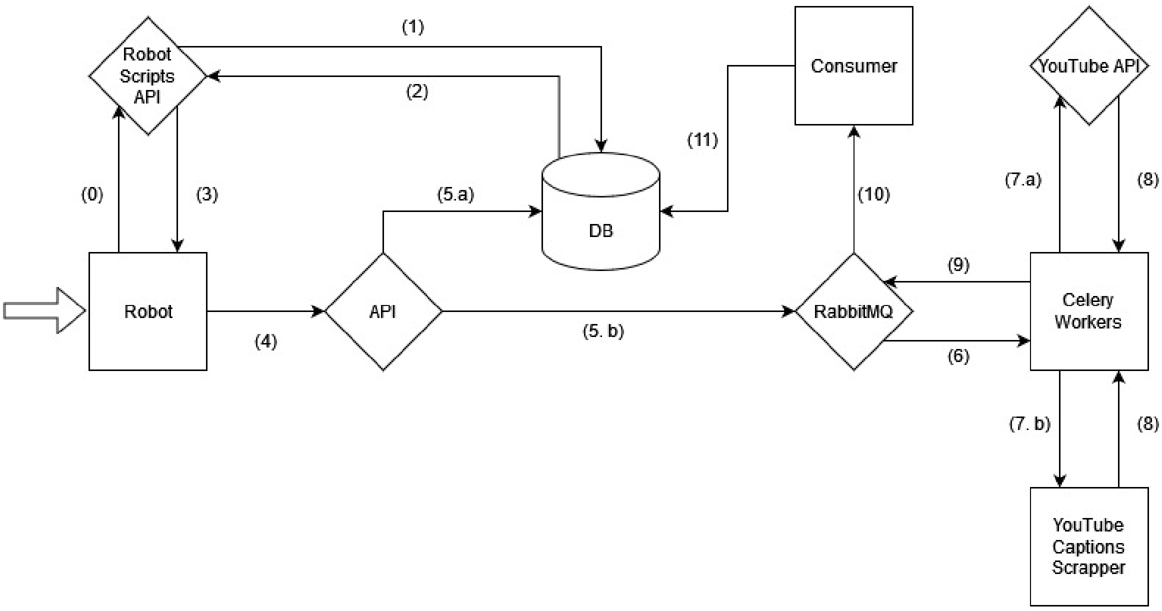


Figure 6 - High Level architecture of scrapper

This part was made in collaboration with three L3 MIAGE's students as part of their "Projet Commun"; more details in the Appendix Section

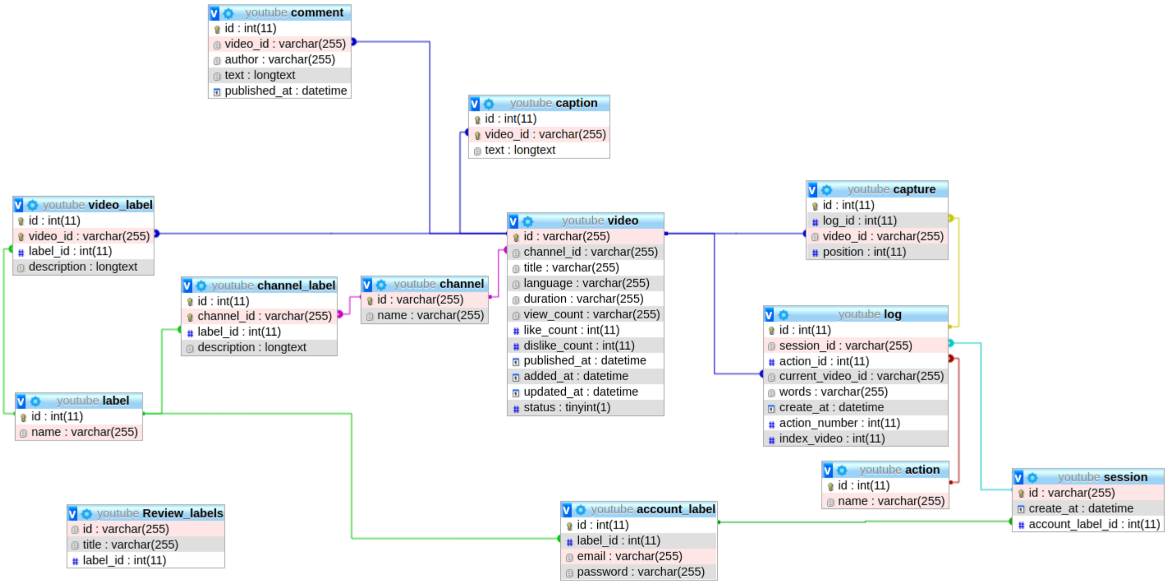


Figure 7 - Database UML

4.4.2 Usage

First, to establish several watch histories, we manually created several Google Accounts later referred as profiles, both male and female, with each one a unique name and surname. We decided not to watch vaccine related videos as this topic was too close to the development of the Coronavirus vaccine and that it would have been harder to determine if YouTube took action against vaccine controversies due to its relatedness to Covid-19 or because it can in itself be harmful. We then associated each account with a label from the ones we're studying (covid conspiracy, covid informative, flat earth conspiracy, flat earth informative, irrelevant). We then constitute watch histories. susceptible of creating a personalization for each one of these profile by watching 100 videos having the same label as the account in use. According to [2], only 22 videos are necessary to create personalization on YouTube. To analyze the pages where recommendation was met (cf figures 1 to 3 in the Section 2.3), we performed 1 to 4 runs for each account with one of the following lists of actions:

- Watch a video from the database labelled as the profile in use, then watch 25 Up Next.
- Search with keywords related to the profile and watch 25 videos
- Watch 25 videos from the Homepage

Each 2 to 4 runs of this type, we try to limit the drift of personalization by watching 50 videos of the label associated with the account.

Due to time constraint, we did not test the impact of liking or disliking a video, nor did we test watching a video only partially. We therefore don't know how YouTube determines the satisfaction score of a user watching a video. Therefore, all videos were fully watched, and no social interaction was made. Watching videos fully did make us lose some time as we encountered videos with up to 5 hours of content, and the watching of several videos was therefore long.

V – SOLUTION & RESULTS

1. Final dataset

To answer both RQ2 and RQ3, we ran the bot several times. It watched about 1400 videos, and loaded 77271 videos, with 12503 distinct videos during the several runs of the bot. We therefore obtained a final set of 18666 distinct videos. We classified all these videos thanks to our classifier presented in Section 4.2. and 4.3. Due to technical issues, we were not able to collect captions and therefore used the classifier without this feature. The final set of videos was the following one:

Label	Number of Videos	% of the total dataset
Covid Conspiracy	2008	10.77%
Covid Informative	2025	10.88%
Flat Earth Conspiracy	420	2.25%
Flat Earth Informative	163	0.87%
Anti-Vaccine Conspiracy	382	2.05%
Anti-Vaccine Informative	306	1.64%
Irrelevant	13345	71.56%

Table 11 - Final set of videos

As we stored the action performed, the order in which they were performed, the profile used, we were able to analyze the data and have the following results.

2. Up-Next Recommendations

The results of the Up Next are inspired by your watch history, and most importantly the video currently playing. For the flat earth topic, the informative profile maximizes the recommendation of overall flat earth content. However, for the different profiles, the proportion of informative video compared to conspiratorial videos is 3 to 5 times more conspiratorial videos.

For the COVID-19 topics, videos are globally way more recommended than flat earth videos, with at least 27% of the Up Next video being related to COVID-19 with no profile. When not logged in, the Up Next videos are more often informative than conspiratorial, whereas the contrary for logged in users. The factor of recommendation of conspiratorial videos compared to informative videos is however way smaller than the one of flat earth, as it is only around 1.4 more conspiratorial than informative. (cf Table 12 and Figure 8).

	Covid-19		Flat Earth	
	Conspiratorial	Informative	Conspiratorial	Informative
Conspiratorial Profile	19.61%	13.40%	11.11%	2.61%
Informative Profile	18.80%	14.17%	18.60%	5.81%
No Profile	11.59%	15.64%	6.94%	1.39%

Table 12 - Probability type of the Up-Next Video with different profiles

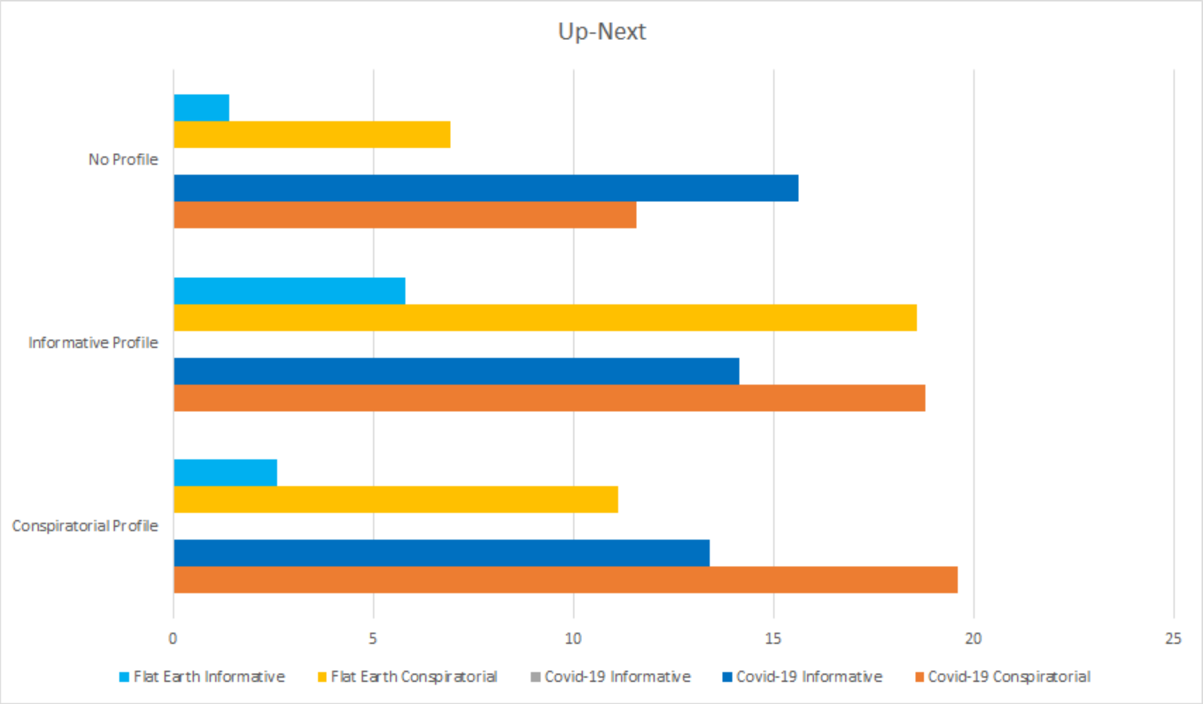


Figure 8 - Probability type of the Up-Next Video with different profiles

3. Sidebar Recommendations

The recommendation of videos while playing a video is based on the currently played videos, and the global watch history of the user. (cf Figure 3 in the YouTube HCI subsection). From the results presented in Table 13 and Figure 9, we can see that COVID-19 content is more available in the recommended videos while watching a video on the same topic (with 45% to almost 62% offered being COVID-19 related), against only 25% to 35% for flat earth related content.

For the flat earth topic, there is about two to three times more conspiratorial content recommended than informative content, no matter the profile and the currently played video. Even though the number of videos conspiratorial is bigger with a certain profile or current video type, the proportion of informative/conspiratorial videos stays approximately the same.

For the COVID-19 topic, there is often more informative videos, except when the current video is conspiratorial, where with an informative profile, the user is exposed to the same proportion of conspiratorial and informative content, while with no account or with a conspiratorial profile, the number of conspiratorial recommended videos is higher than the number of informative videos.

	Current video	Covid-19		Flat Earth	
		Conspi.	Info.	Conspi.	Info.
Conspi. Profile	Conspi.	26.95%	22.30%	21.92%	7.69%
	Info.	23.35%	33.53%	22.63%	13.16%
Info. Profile	Conspi.	28.44%	28.44%	18.84%	10.25%
	Info.	23.30%	38.69%	19.72%	10.73%
No Profile	Conspi.	27.59%	17.05%	17.38%	8.55%
	Info.	18.86%	35.40%	15.73%	9.09%

Table 13 – Types percentages of recommended videos while watching different types of videos

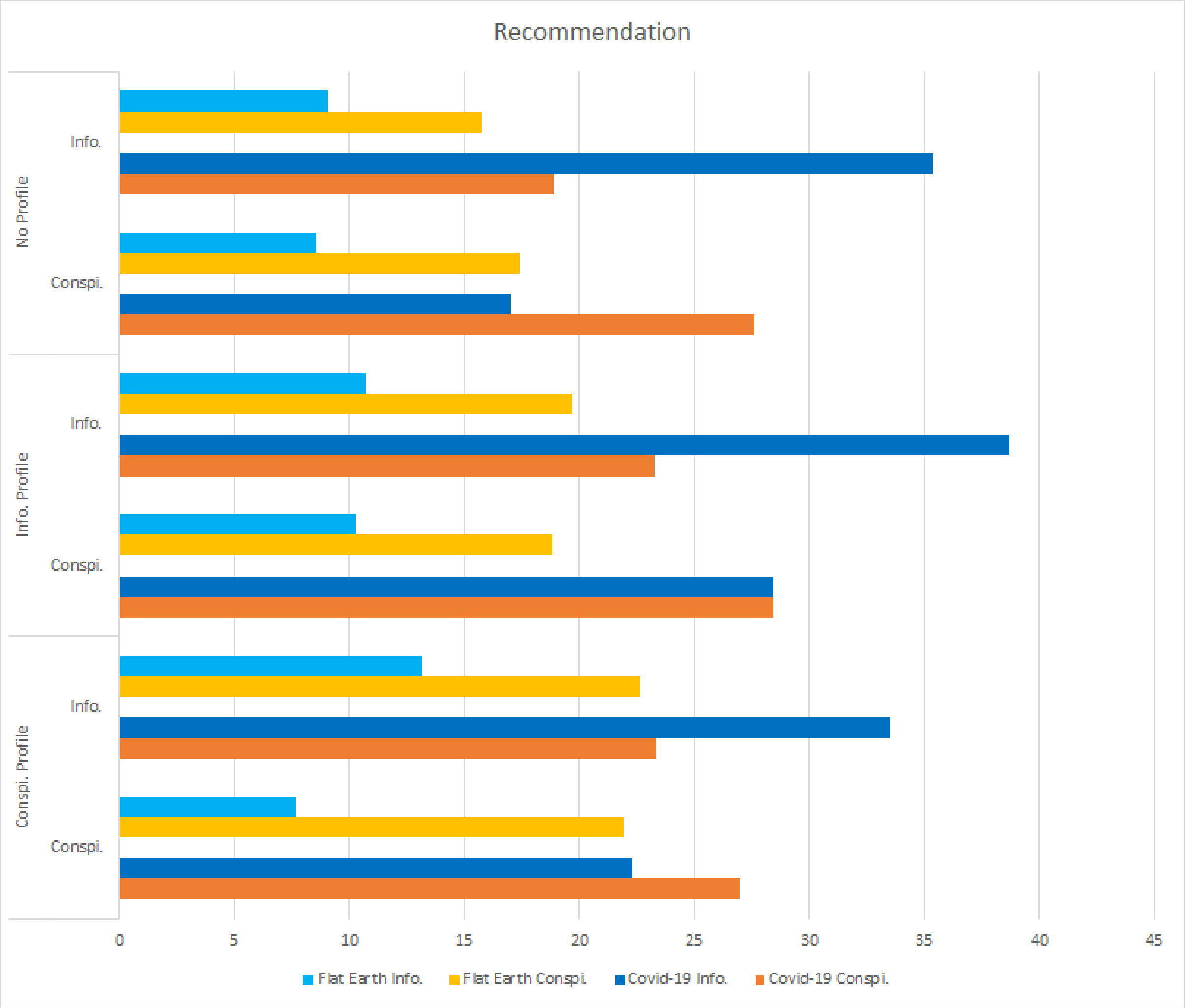


Figure 9 - Types percentages of recommended videos while watching different types of videos

4. Recommendation on Search Results

	Search type	Covid-19		Flat Earth	
		Conspi.	Info.	Conspi.	Info.
Conspi. Profile	Conspi	24.73%	17.08%	26.89%	10.42%
	Info.	19.58%	37.17%	27.39%	16.80%
Info. Profile	Conspi.	23.79%	28.86%	27.79%	11.90%
	Info.	17.31%	33.20%	28.73%	13.80%
No Profile	Conspi.	17.27%	22.58%	21.31%	10.50%

	Info	16.47%	28.24%	33.27%	22.66%

Table 14 – Types percentages after performing a conspiracy search

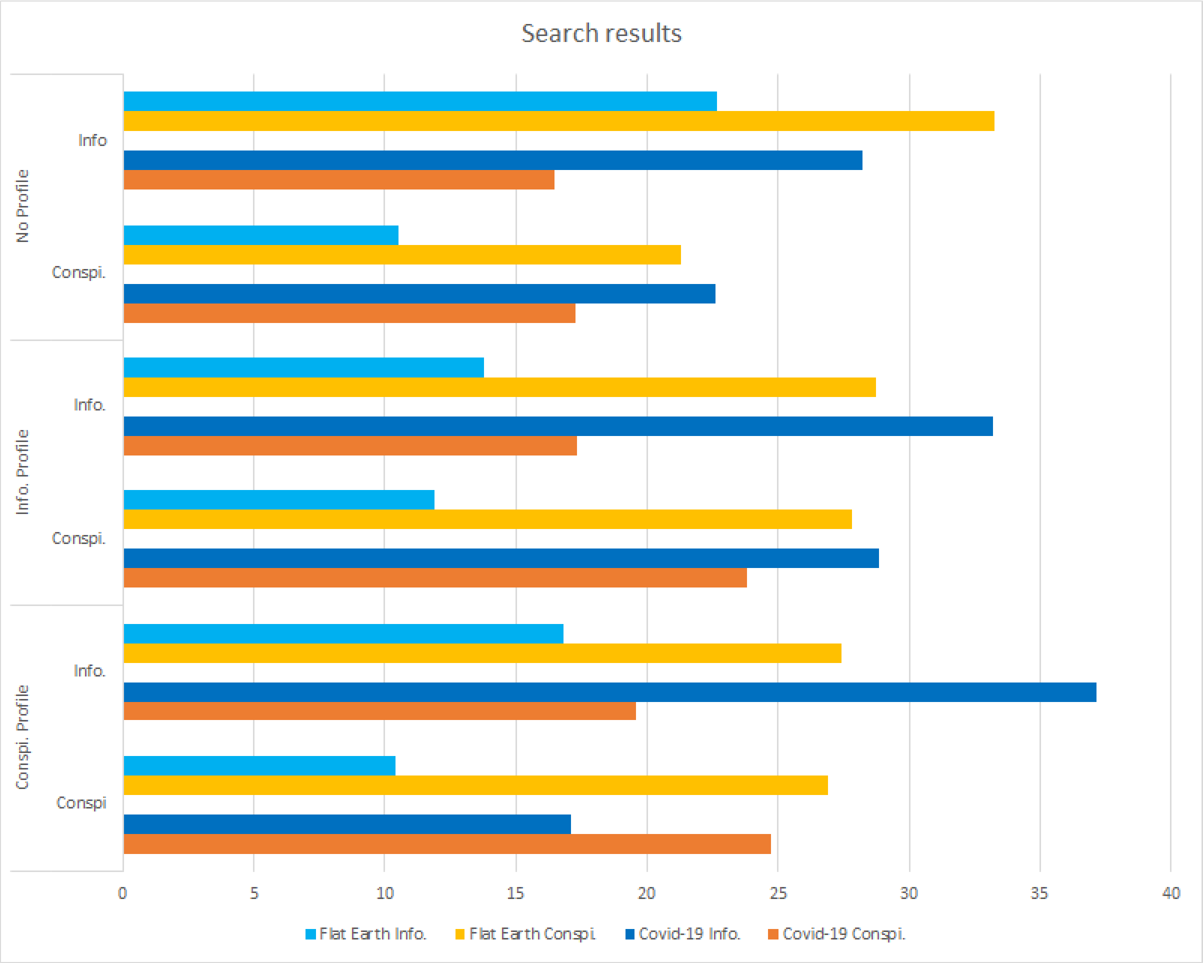


Figure 10 - Types percentages after performing a conspiracy search

From the results presented above in Table 14 and Figure 10, we can see that COVID-19 content is more available in the search results (with between 40% and almost 57% of the video resulting from a search being COVID-19 related), against only 31% to 56% for flat earth related content.

Strangely, for the flat earth conspiracy, when not using profile, the search results are the most relevant and the least relevant. Even though there is more conspiratorial content recommended, when looking for flat earth conspiratorial videos, not many flat earth videos are proposed, while when looking for informative videos regarding flat earth (i.e. debunking the theory), many videos are available to the user. For the flat earth topic, there is about two times more conspiratorial content recommended than informative content, no matter the profile and the search performed. Even though the number of videos conspiratorial is bigger with a certain profile or search type, the proportion of informative/conspiratorial videos stays approximately the same.

For the COVID-19 topic, while performing an informative search, we observe the opposite results, with twice as many informative videos than conspiratorial ones. While performing a search using conspiratorial keywords about COVID-19, there still are globally more informative videos presented to the user, than conspiratorial one, except for the conspiratorial profile. With this difference of treatment, we can see the filter bubble effect: a user with a conspiratorial watch history will probably look up information regarding what they already have been presented, and therefore will be offered more conspiratorial content. The type of profile does, once again, change the number of videos regarding covid, but the proportion of informative/conspiratorial is more often impacted by the search keywords used.

5. Recommendation on the Homepage

For the recommendation on the home page, we clearly see the impact of the personalization of the profile (cf Table 15 and Figure 11) as:

- without any profile, the proportion of informative and conspiratorial Covid related video is approximately the same.
- with a Covid Conspiratorial Profile the user is offered 1.5 times more conspiratorial video than informative video.
- with a Covid Informative Profile the user is presented about 2 times more informational videos than conspiratorial ones.

For the flat earth topic, the number of conspiratorial videos is in any case above the number of informational videos, however, with an informative profile, the number of conspiratorial videos decreases, while the number of informative videos increases.

	Covid-19		Flat Earth	
	Conspiratorial	Informative	Conspiratorial	Informative
Conspiratorial Profile	15.28%	9.76%	6.27%	3.10%

Informative Profile	5.27%	10.55%	5.02%	3.71%
No Profile	9.76%	9.61%	0%	0%

Table 15 - Types percentages of recommended videos on the homepage

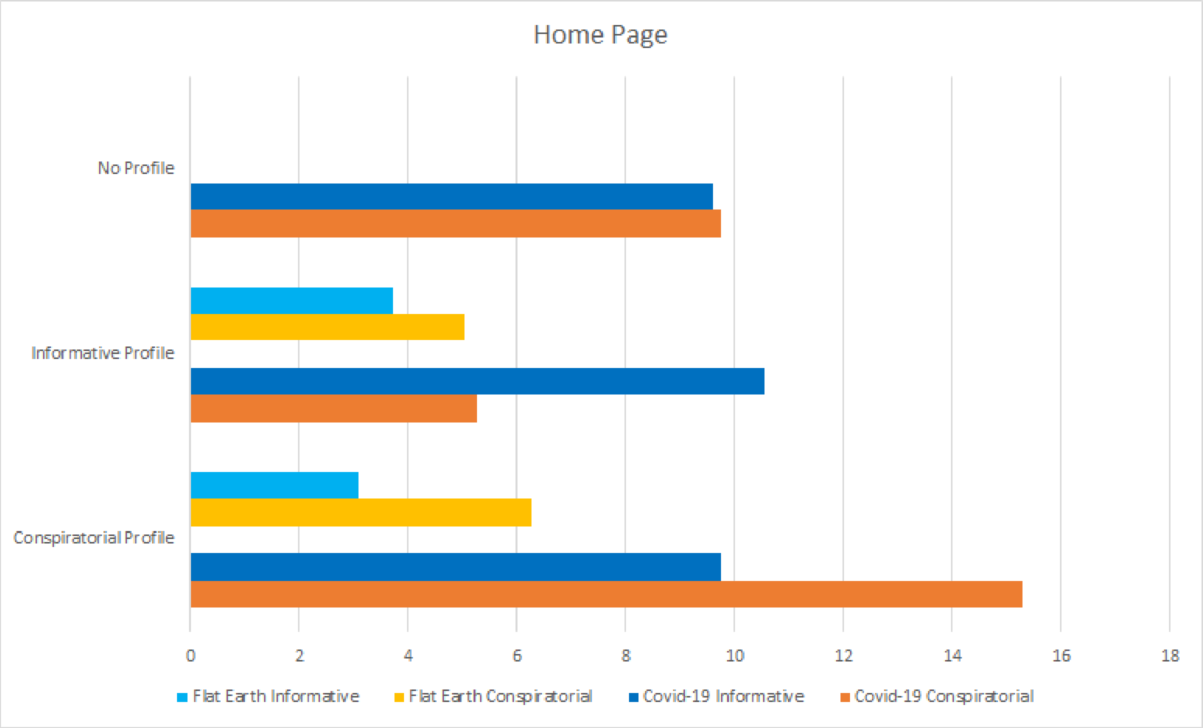


Figure 11 - Types percentages of recommended videos on the homepage

VI – DISCUSSION

The results presented in the last section show that the COVID-19 related videos are dealt with differently than the flat earth related videos. First of all, videos related to the Coronavirus Pandemic are more recommended to the user as they represent between 10 to 40% of the video presented to the user, even on the homepage without being logged in. This is easily understandable as the pandemic is a hot topic that drastically changed our everyday life. Furthermore, it is highly likely that there are less flat earth videos than COVID-19 videos available on YouTube, which also explain the difference of suggestion from YouTube for these two subjects.

Secondly, for every page of YouTube where recommendations are made and that we studied (homepage, page while watching a video, search results), there is as much or more informative content offered for COVID-19 related content, except when performing a conspiratorial search with a conspiratorial profile, or in the recommendation while playing a conspiratorial video with a conspiratorial content. On the contrary, for the flat earth theory, no matter the profile, and the place on YouTube, there is always more conspiratorial content than informative content (i.e. is content debunking of the flat earth theory). The impact of the profile is not clear regarding the recommendation of COVID-19 conspiratorial content (except on the home page), as the proportion stays approximatively the same between different profiles. This proportions do drastically change with the search type or video currently playing type. On the other hand, the proportion of COVID-19 informative content does vary with the profile type on all pages.

The third main finding is that the personalization does happen and is particularly visible on the homepage, especially regarding the COVID-19 pandemic. On the other pages of YouTube, having a conspiratorial watch history mainly impacts when we are already watching a conspiratorial video or performing a conspiratorial search. This is not particularly a good thing as people looking for or watching conspiratorial content are highly susceptible to already have a conspiratorial watch history. We therefore can understand how having watched several conspiratorial videos will lead to being offered more conspiratorial videos, and therefore will likely lead to watching more conspiratorial videos.

A filter bubble might be created, even for conspiratorial and harmful content such as health related topics on YouTube. However, the informational isolation is not complete as there are still informative videos available, but in a lower proportion than other videos.

Finally, we’ve noticed that having an informational profile does not diminish the number of conspiratorial videos presented to the user, and can even increase this number, yet the number of informational videos increases more, creating a bigger proportion of informational content compared to conspiratorial content.

To answer RQ2 and RQ3, from all these findings we can assume that YouTube did take measures regarding the recommendation of conspiratorial content for the Covid-19 pandemic, as there is globally more informative content than conspiratorial content, whereas it is the contrary on flat earth topics. The number of conspiratorial COVID-19 related videos is still high, and even if many more informational videos are available, there is still some work to do to limit the number of conspiratorial contents recommended to the user in the first place. Moreover, the possibility of creating a filter bubble of conspiratorial content is a concern as it can be harmful.

VII – FUTURE WORK / LIMITATIONS

Some limitations appeared regarding this master thesis. First of all, regarding the classifier, the fact that the ground truth dataset is unbalance, even with the option to deal with that proposed by the H2O framework, does show some limited performances for some type of video (such as vaccine informative and flat earth informative). Furthermore, even if we found better performances with the captions, due to YouTube changes in the handling of cookies, the captions were not collected and therefore not used for the classifier. In some future work, it might be interesting to include them to maximize the classifier accuracy. Finally, regarding the classifier, even though we did train and test the classifier on several train/test splits we did not save the results, and therefore did not perform a cross-fold validation in a scientific way.

For the runs of the bot, it might have been interesting to use a profile with irrelevant content, to see if different results are proposed when logged in than with no account. Another great test would have been, as found by as E. Hussein, to test whether watching conspiracy on a specific topic increases the proportion of conspiratorial content regarding other topics proposed to the user, and more particularly if that finding applies to COVID-19 related videos. Moreover, we did not a recreate conspiratorial or informative watch history between each, run, and as watching 22 videos might shift the watch history, some results might have been slightly different than if we recreated it between each run. We did not explore the way YouTube scores the satisfaction of the user and did not explore the impact of social interactions (like, disklike, comment). The fact that we always watch full videos was really time consuming and one of the main concerns, as having only a such short amount of time once the bot was developed.

Regarding the results, we only studied flat earth conspiracy theories as the reference for the way YouTube handles “long-standing” conspiracy theories. It is important to compare with other long-standing conspiracy theories that are not related to health and are therefore less incline to be considered as “harmful”. We did see more content related to COVID-19 than flat earth (informative or not) and see the impact of watch history, however we don’t know how the recommendation works for other non-conspiratorial topics.

VIII – CONCLUSION

In this master thesis, we manually collected and retrieve from similar research papers the IDs of about 6900 YouTube videos already labelled among several conspiratorial topic as either informative or conspiratorial. From these videos ID's we collected the metadata's, the comments, the captions, and the channel's ID of each video. We attributed a label to the channel as the most frequent label of the videos published by the channel. We trained some Fast Text natural language processing for the first 3 input, and then fed all the predicted labels to a distributed Random Forest responsible of determining the final label of each video.

Thanks to this methodology and the available ground truth, we obtained a classifier with an error rate of only 0.151. We then developed a bot with Selenium browser automatization tool and created several watch histories with Google Account profiles. From there, we performed several actions, and collected all the videos proposed to the user on the YouTube platform. We used the classifier to label encountered videos and study the proportion of conspiracy and informative videos recommended while considering the watch history.

After analyzing the results, we conclude how easy it is to see personalization on YouTube and how the recommendation of the same kind of content is bigger with a conspiratorial profile. We also found that no matter what profile used, there is more conspiratorial videos than debunking videos regarding the flat earth theory. However, it was also clearly seeable that YouTube did take measures regarding the COVID-19 pandemic, as even if there were more COVID-19 related videos, the proportion of informative videos was way bigger than for the flat earth topic. Finally, it is also important to note that, for the COVID-19 related content and flat earth conspiratorial content, a paragraph of information linked to Wikipedia is displayed on the top of the result of a search, or directly under a video while watching one. More than its recommendation algorithm, YouTube also tried other ways of tackling misinformative and conspiratorial content.

IX – REFERENCES

Scientific Papers

[1] - Faddoul, M., Chaslot, G., & Farid, H. (2020). A longitudinal analysis of YouTube's promotion of conspiracy videos. *arXiv preprint arXiv:2003.03318*.

[2] - Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2020). "It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. *arXiv preprint arXiv:2010.11638*.

[3] - Hussein, E., Juneja, P., & Mitra, T. (2020). Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1-27.

[4] - Li, H. O. Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: a pandemic of misinformation?. *BMJ global health*, *5*(5), e002604.

[7] - Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira Jr, W. (2020, January). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 131-141).

[15] - Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.

[17] - Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[24] - Zuiderveen Borgesius, F., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles?. *Internet Policy Review. Journal on Internet Regulation*, *5*(1).

[25] - Bruns, A. (2019). Filter bubble. Internet Policy Review, 8(4).

[26] - Nguyen, T. T., Hui, P. M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014, April). Exploring the filter bubble: the effect of using recommender systems on content diversity. In Proceedings of the 23rd international conference on World wide web (pp. 677-686).

[27] - Wilson, R. (2019, July). Cambridge analytica, Facebook, and Influence Operations: A case study and anticipatory ethical analysis. In European conference on cyber warfare and security (pp. 587-XX). Academic Conferences International Limited.

[35] - Curd, M., & Cover, J. A. (1998). Philosophy of science: The central issues. (pp 1-82)

[36] - Hansson, S. O. (2008). Science and pseudo-science.

[37] - Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096.

[38] - Woolley, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. *International Journal of Communicat ion*, *10*.

[39] - Ecker, U. K., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 1-25.

[40] - Barkun, M. (2015). Conspiracy theories as stigmatized knowledge. *Diogenes*, *62*(3-4), 114-120.

[41] - Keeley, B. L. (1999). Of conspiracy theories. *The journal of Philosophy*, *96*(3), 109-126.

[42] - Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Springer, Boston, MA.

[43] - Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. *Knowledge-Based Systems*, *157*, 1-9.

[44] - Ghaisani, A. P., Handayani, P. W., & Munajat, Q. (2017). Users' motivation in sharing information on social media. *Procedia Computer Science*, *124*, 530-535.

[47] - Packowski, S., & Lakhana, A. (2017, November). Using IBM watson cloud services to build natural language processing solutions to leverage chat tools. In *Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering* (pp. 211-218).

[49] - Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European journal of operations research*, *26*(1), 135-159.

[51] - Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

[52] - Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In *The elements of statistical learning* (pp. 337-387). Springer, New York, NY.

[53] - Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression.

[54] - Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167-179.

[55] - Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Others

[5] - https://foundation.mozilla.org/fr/campaigns/youtube-regrets/

[6] - Shearer, E., & Mitchell, A. (2021). News use across social media platforms in 2020.

[8] - Solsman, J. E. (2018). YouTube's AI is the puppet master over most of what you watch. *URL: https://www. cnet. com/news/youtube-ces-2018-neal-mohan*

[9] – The YouTube Team (2019). Continuing our Work to Improve Recommendations on YouTube. *URL: https://blog.youtube/news-and-events/continuing-our-work-to-improve*

[10] – The YouTube Team (2019). Our Ongoing Work to Tackle Hate. *URL: https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate*

[11] – The YouTube Team (2019). The Four Rs of Responsibility, Part 1: Removing harmful content. URL: https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove

[12] – The YouTube Team (2019). The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. *URL: https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce*

[13] – The YouTube Team (2020). Managing harmful conspiracy theories on YouTube. *URL: https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube*

[14] - J. Vincent. (2020). YouTube brings back more human moderators after AI systems over-censor. *URL: http://bit.ly/youtube-moderators*

[16] - https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769

[18] - https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html

[19] - https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html

[20] - https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html

[21] - https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html

[22] - https://docs.h2o.ai/h2o/latest-stable/h2o-py/docs/intro.html

[23] - Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[28] - Tufekci, Z. (2018). YouTube, the great radicalizer. The New York Times, 10, 2018.

[29] - https://github.com/youtube-dataset/conspiracy

[30] - https://github.com/social-comp/YouTubeAudit-data

[31] - https://github.com/kostantinos-papadamou/pseudoscience-paper

[32] - https://github.com/lenalbert/scriptGenYoutube

[33] - https://github.com/ultrafunkamsterdam/undetected-chromedriver

[34] - Hao, K. (2019). YouTube is experimenting with ways to make its algorithm even more addictive. *MIT Technology Review, September, 27*, 2019.

[45] - Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK

[46] - Bhattacharjee, J. (2018). *fastText Quick Start Guide: Get started with Facebook's library for text representation and classification*. Packt Publishing Ltd.

[48] - https://fasttext.cc/docs/en/support.html

[50] - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

X – APPENDIX

The bot – YouTube Scrapper

To create the scrapper, responsible of watching videos and collecting data, we collaborate with a team of three L3 students. With one to two meetings each week, after the first task of formalizing the requirements for the project, my role was the one of project manager with 4 main tasks:

- Defining SMART objectives to reach for the next meeting
- Helping (if needed) with the development
- Making team members communicate with each other
- Correcting the bugs that we encountered and that the L3 MIAGE students failed in resolving.

In order to detect the number of conspiratorial contents proposed to a user during a regular watching session on YouTube, and to observe the impact of the watch history on the recommendation algorithm, the main idea was to automate a browser. To reproduce as closely as possible a human behavior and, to add actions that might influence the recommendation, the requirements were the following:

	Requirement	*Completed*
Json of Actions	*R1. The bot shall send a request to an API with a payload of expected actions*	*Yes*
	R2. The bot shall receive a json file containing actions to execute.	*Yes*
Accounts	*R3. The bot shall be able to connect to Google Accounts.*	*Yes*
	R4. The credentials for the Google Account used by the bot shall be safely stored.	*No*
Performing	*R5. The bot shall be able to watch a video from an URL.*	*Yes*

Actions	*R6. The bot shall be able to watch a video by clicking on a thumbnail.*	*Yes*
	R7. The bot shall be able to like a video currently playing.	*Yes*
	R8. The bot shall be able to dislike a video currently playing.	*Yes*
	R9. The bot shall be able to select a video by its index (the nth video on the page)	*Yes*
	R10. The bot shall be able to go to the channel of a video currently playing	*Yes*
	R11. The bot shall be able to activate autoplay	*Yes*
	R12. The bot shall be able to deactivate autoplay	*Yes*
	R13. The bot shall be able to watch a video for a specified duration.	*Yes*
Storing of Data	*R14. The bot shall create sessions.*	*Yes*
	R15. The bot shall send the actions made and the order in which they were made (along with the sessionID) to an API.	*Yes*
	R16. The bot shall send the id of the watched videos (along with the sessionID) to an API.	*Yes*
	R17. The bot shall send the id of all the videos loaded on each page (along with the sessionID) to an API.	*Yes*
Deployment	*R18. The bot shall be deployed on a server.*	*No*

Table 16 - Requirements for the collection of Data

Due to time constraints, two of the requirements were not met during this project (R4 and R18). We however found solutions that are respectively:

- storing the credentials along with the email in the bot DB
- running the bot locally

The other main task I was in charge of was developing an API that will create JSON files with the list of actions to perform by the bot (cf R1 and R2) (called “bot scripts” in this master thesis). The API was developed in GoLang using Go Modules to ease deployment. This API can receive a POST request with a payload specifying the number of each action to perform, the profile type that will be used, the percentage of social interactions, the order in which to perform the actions. (cf the annex for an example and documentation of the API; cf [32] to access the full code)

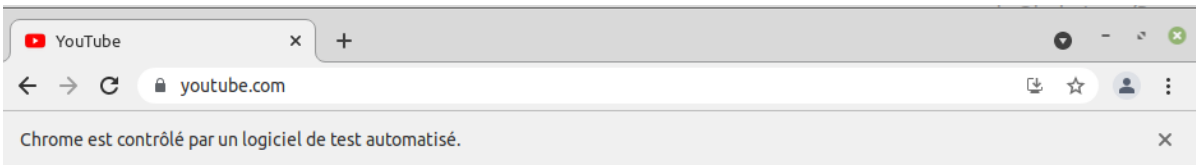
The bot in itself was created as a finite state machine using Python3 and Selenium which is originally a software to automate functional testing on websites. Before choosing Selenium, we studied the different browser automation software:

	Selenium	*Puppeteer*	*PlayWright*
	Open-Source	*OPen-Source; Developed by Google*	*Open-Source; developped by Microsoft*
Supported Language	*Java, Python, C#, Ruby, Perl, PHP, and JavaScript*	*Node.JS (an unofficial portation in Python called Pyppeteer exists)*	*JavaScript, Java, Python, and .NET C#*
Supported Browser	*Chrome, Firefox, IE, Edge, Opera, Safari, and more*	*Chrome & Chromium*	*Chromium, Firefox, and WebKit*
Community	*Commercial support for its users via its sponsors in Selenium Ecosystem along with self-support*	*Smaller community than Selenium*	*As fairly new, the support from the community is limited (compared to Selenium)*

	documents. Strong community support from professionals across the world		
Device Support	*Supports real device clouds and remote servers*	*N/A*	*Does not support real devices but supports emulators*

Table 17 - Comparision between Selenium, Puppeteer and PlayWright

Selenium is however often diverted to automate and scrap information which led to custom Selenium drivers such as undetected-chromedriver, (available [33]) allowing selenium to pass some mitigation bot systems. We used this custom selenium driver to be considered as a regular user by YouTube, making this pop up disappear:



Chrome est contrôlé par un logiciel de test automatisé.

Figure 12 - Selenium being detected by Chrome

While watching specified videos in order to create a watch history, the bot sends the ids of all the videos loaded on the page to an API and stored in the database. This API is made with the API Platform library of Symfony with a HATEOAS (Hypermedia As The Engine Of Application State) architecture, allowing good flexibility, but keeping the classic REST protocol.

This API was coupled with a RabbitMQ to prevent any loss of data in case an end of the API was unavailable and several Celery Workers to collect more precise data. The RabbitMQ can be described as an independent server with a queue, in charge of storing data as long as they have not been consumed by the Celery Worker. There are 3 Celery worker that are launch for each message received in the RabbitMQ:

- one using the YouTube API to collect the video metadata
- one using the YouTube API to collect the top 200 comments
- one using the bot to scrap caption

The collected data are then stored in a second queue of the RabbitMQ and are then consumed and sent to the database. The database stores all the information about a video that is useful to us (metadata, comments, caption, channel), along with the data relative to their watching by the bot, and with the data created from the classifier. The database has the following scheme