Integrated Citizen Centered Digital Health and Social Care
A. Värri et al. (Eds.)

doi:10.3233/SHTI200705

Automatic Exploitation of YouTube Data: A Study of Videos Published by a French YouTuber During COVID-19 Quarantine in France

Gery LAURENT[a,b,1] Benjamin GUINHOUYA[a,b],
Marielle WHATELET[c] and Antoine LAMER[a,b,c]
[a] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, 59000, Lille, France*
[b] *Univ. Lille, Faculté Ingénierie et Management de la Santé, 59000, Lille, France*
[c] *CHU Lille, Public Health Department, F-59000 Lille, France*

Abstract. The objective of this study was to test the feasibility of automatically extracting and exploiting data from the YouTube platform, with a focus on the videos produced by the French YouTuber HugoDécrypte during COVID-19 quarantine in France. For this, we used the YouTube API, which allows the automatic collection of data and meta-data of videos. We have identified the main topics addressed in the comments of the videos and assessed their polarity. Our results provide insights on topics trends over the course of the quarantine and highlight users sentiment towards on-going events. The method can be expanded to large video sets to automatically analyse high amount of user-produced data.

Keywords. COVID-19, YouTube, citizen, Natural Language Processing

1. Introduction

YouTube is an online video-sharing platform, used by individuals, professionals or institutions. It provides a huge range of Health-Related Content and was used during COVID-19 outbreak as a source of information [1-3]. In most of the studies using YouTube, videos are watched and then analysed manually [4]. However, there exist automatic methods to extract useful information from online content, and in particular YouTube [5].

The objective of this study is to test the feasibility of automatically extracting and exploiting data from YouTube. For this, we will seek to identify the main topics addressed in the videos of a French well-known Youtuber, and the audience's reactions to these topics.

[1] Corresponding Author, Géry Laurent, ULR2694, Pôle Recherche 1, place de Verdun 59045 - Lille, France; E-mail: gery.laurent.etu@univ-lille.fr

2. Material and Methods

A YouTube video is characterized by a title, an author, the video itself, subtitles, a number of positive/negative rates (like/dislike), comments. In this work, we analysed the YouTube channel HugoDécrypte followed by more than 868 000 subscribers [6]. We collected and exploited a maximum of data from videos of the "daily news" playlist released between March 17th 2020 and May 04th 2020, as it provides a chronological overview of daily news regarding the COVID-19 outbreak in France.

YouTube Data API provides a way to collect metadata of a video: title, author, duration, publication date, number of like/dislike, number of views, number of comments, text description or its comments: author, publication date, text, number of like/dislike. The API can be implemented in Python. Some of the metrics retrieved by the API, namely number of views or like/dislike count, are representing a snapshot of current data at the time of the API call and thus cannot be used retroactively. Retrieval of number of views of a video overtime necessitates data collection stream. Thus, for the present study, data used were the video title, publication date, comments publication date and text.

In order to identify the main topics addressed in the videos, we have implemented the following steps. (i) We extracted comments published in the 24 hours following the release of the videos. (ii) For each video, we detected the 10 most frequent words from which topics were drawn. Data management followed Natural Language Processing (NLP) steps: accent and stopwords removal, tokenization, stemming. (iii) Topic cooccurrence was assessed by computing the frequency of comments sharing both topics. The two most co-occurring topics were linked in the network graph as well as the topics presenting at least half the co-occurring frequency. (iv) Data from all the videos were pooled to analyse the most discussed topics globally and compute the relative daily frequency of comments for each topic.

To evaluate the audience's reactions to these topics, we used the polarity score obtained using SAS Viya VisualTextAnalytics software [7]. This method is using a sentiment lexicon attributing polarity scores to individual words. The overall comment score is then computed based on each word polarity score as well as sentence structure, punctuation and emoticons, on a scale of -1 (extremely negative) to +1 (extremely positive). Average polarity, polarity distribution and number of positive, negative and neutral comments were calculated for each topic on a video per video basis.

The following python libraries were used for this study: pandas and numpy (data management), matplotlib.pyplot and seaborn (data visualization), nltk, sacremoses and sklearn (NLP).

3. Results

Between March 17th 2020 and May 04th 2020, 49 videos related to the COVID-19 outbreak were published on the playlist "daily news" of "HugoDécrypte" channel. These videos received 38 725 comments in the 24 hours following the release. For each video, the median [1st quartile; 3rd quartile] number of comments was 771 [682 ; 865]. After removal of stopwords, 135 unique topics were identified across the 49 videos. The top 10 discussed topics are presented in Table 1 and Figure 1.

Table 1 is presenting the polarity distribution of the comments across the 10 most discussed topics based on SAS polarity score. The median [1st quartile; 3rd quartile] polarity of all the comments is 0.00 [-0.20 ; 0.00]. The two topics with the most percentage of positive comments were "thank you" and "Hugo" with respectively 30%

and 27%. The two topics with the most percentage of negative comments were “death” and “people” with respectively 87% and 62%. Overall, all the topics displayed a lower ratio of neutral comments and a higher ratio of negative comments compared to the global ratio for all the comments except for the two most represented topics “thank you” and “hugo”.

Figure 1 represents a heatmap of the relative occurrence for each topic, across the 49 videos. Amongst the topics we have: “thank you”, “Hugo”,“France”, “quarantine”, “masks”, “individuals”, “virus”, “deaths”, “people” and “country”. Main topics are evolving and fluctuating based on video content although most topics have a recurring pattern throughout the time of study.

Table 1. Number of comments and polarity distribution of the 10 most frequently mentioned topics. Polarity is expressed in median [1st quartile ; 3rd quartile].

Topic	Percentage of comments (Number)				Polarity
	Total	Positive	Negative	Neutral	
Total	38 725	15% (5 763)	33% (12 720)	52% (20 242)	0.00 [-0.20;0.00]
thank you	4 898	30% (1 473)	16% (800)	54% (2 625)	0.00 [0.00;0.00]
Hugo	3 837	27% (1 049)	19% (743)	53% (2 045)	0.00 [0.00;0.20]
France	2 652	13% (344)	53% (1 411)	34% (897)	-0.20 [-0.38;0.00]
quarantine	2 534	13% (323)	49% (1 233)	39% (978)	0.00 [-0.38;0.00]
masks	1 992	14% (273)	43% (866)	43% (853)	0.00 [-0.20;0.00]
individuals	1 993	11% (217)	60% (1 204)	29% (572)	-0.20 [-0.38;0.00]
virus	1 780	8% (147)	60% (1 061)	32% (572)	-0.20 [-0.38;0.00]
deaths	1 774	4% (63)	87% (1 540)	10% (171)	-0.38 [-0.54;-0.20]
people	1 850	10% (193)	62% (1 153)	27% (504)	-0.20 [-0.54;0.00]
country	1 432	11% (164)	59% (840)	30% (428)	-0.20 [-0.38;0.00]

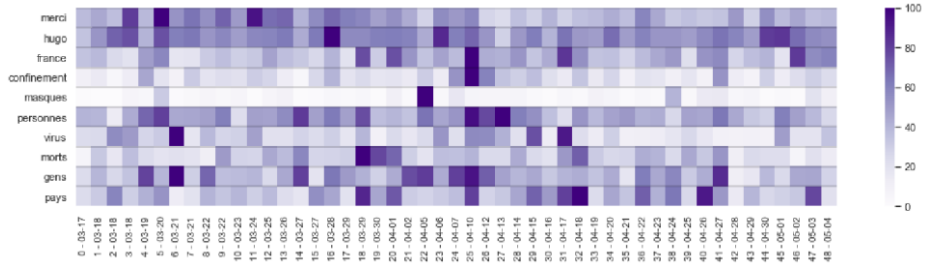

Figure 1. Frequency of the 10 most frequently mentioned topics in the comments across the 49 videos

Figure 2 represents the 10 topics discussed in the comments of the 9th video published on March 22th 2019, their average polarity and their co-occurrence. The three topics with a positive polarity are related to the “work” of “information” realized by the YouTuber Hugo, and the subscribers “thank” him for that. The other topics present a negative polarity, from -0.17 for “quarantine” to -0.33 for “deaths”, compared to the global video polarity of -0.07.

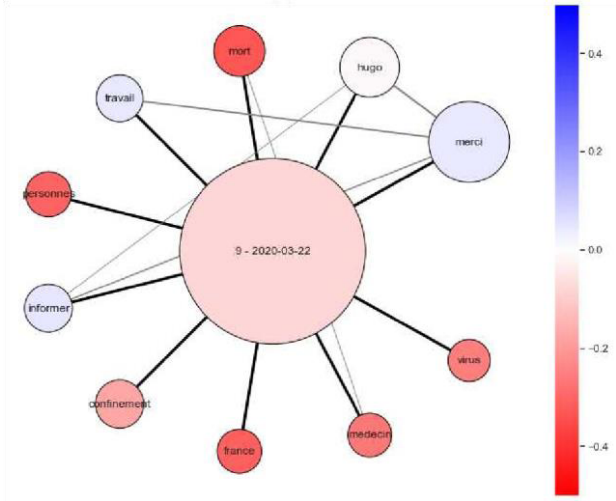

Figure 2. Topics discussed in the comments of the 9th video published on March 22th 2019, their polarity and their co-occurrence. Central circle represents the number of comments of the video. Outer circles represent the 10 most discussed topics in the videos, with a radius proportional to their frequency, and a color related to their average polarity. Edges are drawn between the most co-occurring topics.

4. Discussion

In this study, we have automatically extracted and exploited the comments of 49 videos of a French youtuber, HugoDécrypte, who produced videos on current events during the French lockdown of COVID-19 outbreak. From this automatic analysis came out the main topics discussed in relation to the videos, and their polarity. The work of the YouTuber was received positively by subscribers, while the topics discussed have a negative polarity.

From a methodological point of view, the main difference with previous works is that we were able to perform an automatic exploitation of data from the YouTube platform [1]. It presents some advantages compared to the manual method: (i) the study can cover a larger number of videos, (ii) it can be replicated several times over time (iii) the methodology can be applied to any video to retrieve main topics and polarity.

Even if the YouTube API provides an easy way to automatically extract data from the YouTube platform, it also presents some limitations. First, all data are not available: subtitles can only be extracted by the owner of the channel. Secondly, when submitting retrospective queries, the API returns a limited amount of content. In order to have completeness, the query may be submitted in real time in streaming. Furthermore, the YouTube API is returning a non-exhaustive sample of the videos and comments that may vary from a query to another and based on the time between query and publication date. Last, the NLP methods for the treatment of comments used in this study delivered decent results with the selected videos, but it depends very much on the community, the vocabulary and language used as well as the topics discussed. This has yet to be tested in other contexts. Some parts of the extraction and cleaning process of the video content may depend of the context and need to be updated for each study.

Authors have to be cautious when interpreting results from polarity score. Indeed, we experimented with another French lexicon besides SAS sentiment analysis, from the library TextBlob_fr [8], which returned different raw polarity score. The first method is

biased towards negative score while the second method is biased towards positive score. Sentiment analysis studies on French content is lacking compared to English corpus. The development of a more complete and up to date lexicon for French content, especially focused on social network corpus, is required to improve the results and reliability. To go further, emotion analysis (happy, sad, angry, fearful, excited, bored) can also enhance results by providing a more precise picture of the community feeling towards the different topics [9].

While the study is aimed at studying the community interaction with the main discussed topics, there is no current way to retrieve topics discussed in the video without watching it and manually analysing audio and video content. Subtitles retrieval by other means than using the YouTube API could be considered. Besides, YouTube recently released a new feature allowing content creators to timestamp their video and split it in several chapters based on the topic discussed at that point. This could provide an easy way to automatically extract the different topics mentioned in the video [10].

5. Conclusions

Social Media and YouTube represent a novel and fast-growing way to share information and discuss about trending topics worldwide. With the explosion of video content, we proposed an automatic method to collect and exploit citizens produced data by highlighting main discussed topics in the comment section of a video and user sentiment towards it.

References

[1] Khatri P, Singh SR, Belani NK, Yeong YL, Lohan R, Lim YW, et al. YouTube as source of information on 2019 novel coronavirus outbreak: a cross sectional study of English and Mandarin content. Travel Med Infect Dis. March 20th 2020;101636.

[2] Kocyigit BF, Akaltun MS, Sahin AR. YouTube as a source of information on COVID-19 and rheumatic disease link [published online ahead of print, May 23th 2020]. Clin Rheumatol.

[3] Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, et al. The COVID19 Social Media Infodemic. ArXiv200305004 Nlin Physicsphysics [Internet]. March 10th 2020

[4] MacLeod MG, Hoppe DJ, Simunovic N, Bhandari M, Philippon MJ, Ayeni OR. YouTube as an information source for femoroacetabular impingement: a systematic review of video content. *Arthroscopy*. 2015;31(1):136-142.

[5] Liu J, Siegel L, Gibson LA, et al. Toward an Aggregate, Implicit, and Dynamic Model of Norm Formation: Capturing Large-Scale Media Representations of Dynamic Descriptive Norms Through Automated and Crowdsourced Content Analysis. *J Commun*. 2019;69(6):563-588.

[6] HugoDécrypte. Actus du jour [Internet]. [cited July, 10th 2020]. Available on : https://www.youtube.com/playlist?list=PLKDC6DUkHXj28ptnJ44sveHIKFgJ50U_v

[7] SAS Visual Text Analytics [Internet]. [Cited on July, 15th 2020]. Available on: https://www.sas.com/en_us/software/visual-text-analytics.html

[8] Loria S. textblob-fr: French language support for TextBlob. [Internet]. [Cited on July, 14th 2020]. Available on: https://github.com/sloria/textblob-fr

[9] Abdaoui A, Azé J, Bringay S, Poncelet P. FEEL: a French Expanded Emotion Lexicon. Lang Resour Eval. Sept 2017;51(3):833-55.

[10] Add chapters to a progress bar - YouTube Help [Internet]. [cited on July, 14th 2020]. Available on: https://support.google.com/youtube/answer/9884579?hl=en