Analyzing YouTube Videos Shared on Whatsapp in the Early COVID-19 Crisis

Conference Paper · October 2020

DOI: 10.1145/3428658.3431090

6 authors, including:

Marisa Vasconcelos
IBM Research, Sao Paulo, Brazil
46 PUBLICATIONS **491** CITATIONS
SEE PROFILE

Erica Pereira
Federal University of Minas Gerais
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Manoel Ribeiro
École Polytechnique Fédérale de Lausanne
31 PUBLICATIONS **306** CITATIONS
SEE PROFILE

Philipe Melo
Federal University of Minas Gerais
22 PUBLICATIONS **154** CITATIONS
SEE PROFILE

Analyzing YouTube Videos Shared on Whatsapp in the Early COVID-19 Crisis

Marisa Vasconcelos
IBM Research, Brazil
marisaav@br.ibm.com

Erica Pereira
UFMG, Brazil
ericapereira@dcc.ufmg.br

Samuel Guimarães
UFMG, Brazil
samuelsg@dcc.ufmg.br

Manoel Horta Ribeiro
EPFL, Switzerland
manoel.hortaribeiro@epfl.ch

Philipe Melo
UFMG, Brazil
philipe@dcc.ufmg.br

Fabrício Benevenuto
UFMG, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Whatsapp is the most popular messaging app in the world. It is not only used as a one-to-one messaging app but also as a platform for group discussion. Recently, Whatsapp has gained the spotlight for its role in disseminating (often low-quality) information. Our study focuses on YouTube videos shared by political-oriented public groups on Whatsapp for a month during the COVID-19 pandemic. Through a careful analysis of the topical distribution and the lexicon present in the videos shared, we shed light on the COVID-19 debate happening in these groups. Moreover, we compare COVID-19 related videos with other political videos being shared in the groups. We observe that videos that discuss political themes have more emotional attributes as well as topics related to typical right-wing concerns such as family, work, and religion than videos discussing both pandemic and politics.

1 INTRODUCTION

Whatsapp is the main messaging app used in Brazil with almost 120 million users in the country, according to data released by the company in March 2020[1]. According to a government-related campaign, 79% of users use the platform as a news source to share and discuss information with friends and group members [12]. Focusing on that usage, we here investigate the use of YouTube links shared on Whatsapp.

Youtube is also broadly used for people as a news channel and understanding the links shared on Whatsapp groups give us an idea about what and how people are being informed. Our analysis extends existing research on Whatsapp during the 2018 Brazilian presidential election in which was observed the presence of misinformation in posts shared on the same groups [13, 14]. Here, we focused on the first month of the COVID-19 pandemics in Brazil examining the cross-platform aspect by investigating the links between the content shared on Whatsapp and YouTube.

In the last months, several studies focused on the dissemination of rumors and conspiracies in social networks during COVID-19 pandemics [3, 15]. A few of these studies [3, 6] focused on investigating the promotion of political content in contrast to health content during the pandemic period. Thus, current research analyzed the political presence under pandemics posts, we here go in the opposite direction, focusing on how political-oriented groups discuss pandemic topics.

[1]https://www.businessofapps.com/data/whatsapp-statistics/

Our main contribution is the investigation of content shared as YouTube videos on political-oriented Whatsapp groups during the first month of the pandemic. We characterized the channels and videos shared on Whatsapp groups, as well as the content of the videos. For that, we did a lexical and topical analysis in the video captions, identifying the frequent terms with the events that occurred in Brazil. We observe that videos that discuss political themes have more emotional attributes, as well as topics related to typical ring-wing concerns such as family, work, and religion than videos discussing both pandemic and politics.

2 RELATED WORK

Our study is aligned to studies that investigate publicly group chats, characterization of posts during the pandemic as well as YouTube video content analysis. Our main focus here is to investigate how the the pandemic is discussed in politically oriented chat groups.

Previous studies analyzed the messaging app to understand the patterns of the users when interacting in that enviroment. Garimella *et al.* [7] presented a methodology to collect such Whatsapp group and characterization of the collected groups. This methodology was followed by a several following studies [8, 13, 14]. These studies focused on analyzed misinformation messages shared on publicly accessible Whatsapp groups related to Brazilian politics.

YouTube has been extensively studied and analyzed in diverse aspects. We based our methodology in two studies [1, 10] that also used video captions to analyze their content. Araujo *et al.* [1] used the video transcripts to analyze the content of the most recurrent videos for children. Political videos, more specifically, right-wing YouTube videos were studied by Ottoni *et al.* [10] in terms of hate, violence, and discriminatory bias presented both in users' comments and video contents.

In the last months, several studies analyzed the spread of misinformation during the pandemic. Cinelli *et al.* [5] performed a comparative analysis on misinformation dissemination in five different social media platforms during the COVID-19 period. They found that aspects of each platform and the interaction patterns of the users engaged with the COVID-19 topic drive the information dissemination in each system. Ferrara *et al.* [6] detected the presence of bots promoting political conspiracies in the US during the pandemic while Boberg *et al.* [3] analyzed the pandemic populism analyzing alternative news media posts on Facebook.

# Total Whatsapp msgs with YT links	59,397
# Total distinct video links	14,602
# Total distinct channel links	185
# Total Whatsapp groups	1,512
# Total distinct users	6,601
Collected period	March 1st to April 1st
# Total of unique links	14,787

Table 1: Whatsapp Dataset Description

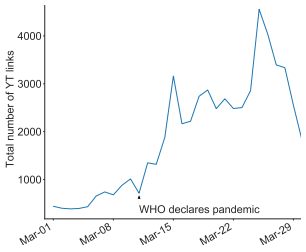


Figure 1: Daily Whatsapp messages with YT links

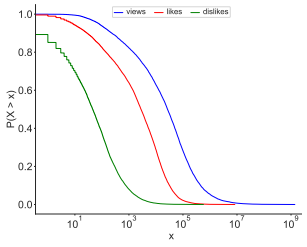


Figure 2: Likes, dislikes and, views distribution

3 METHODOLOGY

3.1 Data Collection

We now describe our methodology to collect YouTube (YT) links from Whatsapp groups. We use the same Whatsapp groups collected by Resende *et al.* [14]. Their approach focused on monitoring publicly accessible political groups since 2018. From those groups, we collected the YouTube links shared on these WhatsApp groups. Our goal is to use the same groups, to compare their activity during the first months of the COVID-19 pandemic. Table 1 shows a summary of this dataset.

From March 1^{st} to April 1^{st} 2020, we retrieve almost 60 thousand messages containing at least one Youtube link. Table 1 shows an overview of our Whatsapp dataset. Among the messages links, there were links not only to videos with a person speaking but to music videos and YouTube channels. Figure 1 shows the number of Whatsapp messages with YouTube links shared on a daily basis during the period of collection. Note two peaks in the number of links, one in March 15^{th} and the second one in March $25th$. The first peak was during the same day of the far-right demonstrations supporting the Brazilian president Bolsonaro[2]. The second peak happened one day after Bolsonaro's speech in which he dismisses isolation and blamed the press for hysteria. The videos on that day are related mostly to his speech from its total reproduction to other people praising his ideas. Using the links, we query the YouTube API[3] to download the metadata of each video, such as its number of likes, views, textual transcriptions, categories, title, description, and duration. Some videos could not have their metadata collected, we thus decided to filter them out leaving us with 12,631 videos.

[2]https://g1.globo.com/politica/noticia/2020/03/15/cidades-brasileiras-tem-atos-pro-governo.ghtml
[3]https://developers.google.com/youtube/v3/

Category	% videos	% views	% likes
News & Politics	37.16	6.18	17.14
People & Blogs	27.76	8.85	22.27
Entertainment	14.60	8.84	18.12
Education	6.23	1.88	4.93
Music	3.13	65.77	24.69
Comedy	2.01	1.25	3.99
Science & Technology	1.89	1.12	2.79
Nonprofits & Activism	1.88	1.29	1.47
Film & Animation	1.08	1.76	1.17
Other categories	1.72	1.72	1.94

Table 2: YouTube Categories Videos Description

3.2 YouTube Dataset Characteristics

The following discussion is based on the metadata of 12,631 distinct videos successfully collected using the YouTube API.

Video Popularity: As our first analysis, we focus on the popularity of the videos shared such as views, likes, and dislikes. Figure 2 shows the complementary cumulative distribution of the number of views, likes, and dislikes per YouTube video shared across the monitored Whatsapp groups. Note the log scale on the x-axis. We can see that 63% of the videos have at most 10,000 views. We observe that the top-10 most viewed videos were music video clips.

Video Category: Table 2 shows the distribution of YouTube videos per category in our dataset in terms of the number of videos, views, and likes. We can see that News & Politics and People & Blogs categories are most frequently shared on Whatsapp groups. This reflects the nature of the Whatsapp channels (e.g., Politics) where these links were collected. However, in terms of popularity (e.g., views and likes) the Music category is the most popular. Since the videos were uploaded to YouTube before the link sharing and there is no data about the timestamp of views and likes, we cannot imply that Whatsapp brought its users to YouTube.

Video Length and Upload Date: In our dataset, 27.4% of the videos' lengths have a short duration (e.g., up to 4 minutes) while 20.6% of the videos have a longer duration (e.g., up to 20 minutes). When we analyze when the videos were uploaded, we noticed that most videos were uploaded in 2020. The oldest uploaded video in our dataset was posted 14 years ago and it was a music video clip.

YouTube Channels: To analyze the political landscape of the YouTube videos we collected, first we needed to properly identify where in the political spectrum they fall. In order to classify the videos, we started by aggregating the videos by channel. Table 3 shows the top-10 channels with most of the videos. We observe that 6 out of 10 YouTube channels are right-wing channels. This is consistent with Bursztyn *et al.* [4] work that observed that right-wing groups are more abundant in Whatsapp.

4 VIDEO CONTENT ANALYSIS

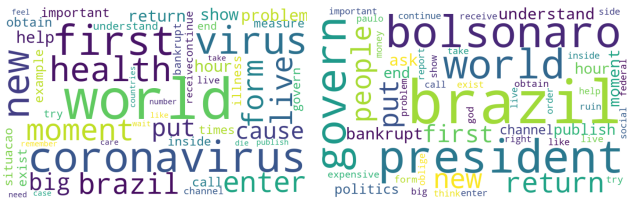
For the following analysis, we divide our dataset into two main subjects: pandemic and politics. For that, we use keywords related to each theme to filter videos using their tags. For the pandemics theme, the following words and variations were used: coronavirus,

Channel's name	# videos	Channel's name	# videos
Plantao 24 no ar	236	Band Jornalismo	118
Folha Politica	217	O Giro de Noticias	110
Os Pingos nos Is	167	Poder360	108
Noticias da Hora	123	BR Noticias	89
Gigante Patriota	122	Universo	81

Table 3: Top-10 Most shared YouTube Video Channels

COVID-19, pandemic, epidemic, contamination, health, hydroxychloroquine, China, and Wuhan. We also use the names of two popular Brazilian scientific disseminators, Drauzio Varella, Atila Iamarino. For the *Politics* dataset, we consider videos with politics names (e.g., Sergio Moro, Jair Bolsonaro) and political keywords (e.g., dictatorship, PT, STF). Thus, we produced three datasets using that strategy: one containing only videos about the *Pandemic* (902 videos), others containing only *Politics* (3,349), and a third one containing both politics and pandemic related videos (1,333). Note the lower percentage of the only pandemic related videos (16%) compared to only politics related videos (60%). That suggests that users in the Whatsapp groups were more interested in sharing political videos than videos related only to the pandemic crisis.

To analyze the content of the videos, we have collected their captions, which are the written version of what is said in the video. Using these captions, we analyze the lexicon and the topics content in the videos. Before performing the caption analysis, we lowercased and tokenized all words, removing punctuation and stopwords and we performed a lemmatization of all words in the captions. Figure 3 shows the word clouds of the top 50 most frequent words for the Pandemic and Politics datasets. We can observe the frequent presence of words related to the pandemic: illness, world, coronavirus for the Pandemics dataset while terms such as bankrupt, Brazil, world are frequently present in the Politics dataset.



(a) Pandemic (b) Politics

Figure 3: Word clouds (translated to English)

4.1 Topical analysis

We now perform a topics summarization approach in each dataset. For that, we employ the LDA algorithm [2] to infer the topics in a collection of text documents. We run the LDA implementation provided by the Gemsim package[4] varying the number of topics (k) from 5 to 50. To identify the best k, we use the coherence score metric [9], which was for $k = 30$. Table 4 shows the partial output of the LDA model for the top 2 topics for each dataset as well as the 15 words that best characterize that topic.

[4]https://radimrehurek.com/gensim/models/ldamodel.html

Note that for the Pandemic dataset, we see words closely related to health such as *contaminate*, *mask*, *alcohol gel* as well as terms related negative sentiments such as *panic* and *fear*. Regarding the top-ranked topics for the Politics dataset, we can see terms like *democracy*, *Bolsonaro*, constitution, about the beginning of the demonstrations against the congress, and STF on March 15^{th}.

It is also possible to identify names of Brazilian right-wing politicians such as *Arthur Lira* congressman, Brazilian TV celebrities (e.g., Sikera Junior), and negative terms like *sue* and *attack*. The odd word, "superman", was found in the *politics* dataset. The term was associated with a military police officer and digital influencer Gabriel Monteiro, Bolsonaro's supporter, and right-wing militant. Lastly, terms such as "Dirceu" and "left", typically associated with left-wing were identified in the politics dataset. Checking the videos associated with these terms, we note that most of them were critics of left-wing groups.

Finally, it was observed in the Intersection dataset, as expected, many terms associated with politics and the pandemic at the same. Other terms related to the impact of the pandemic on the political/economic scenario, such as company, economy, employment, crisis, work, minister, death, health were connected to the minister of Health daily announcements about social distancing.

Dataset	Topic	Topic words
Pandemic	1	coronavirus, infect, pass, work, govern, protein, mask, fear, china, understand, research, alcohol gel, feed, panic, get the flu
Pandemic	2	water, avoid, medicate, health, care, people, live, supplement, immunity, vitamin, number, say, measure, city, reduce
Politics	1	noncompliance, paradigm, sikera junior, dirceu, truckers, president, bolsonaro, brazilian, constitution, senate, superman, left, economy, congressman, right
Politics	2	democracy, governor, journalist, justice, assistant, honest, father, minister, parlamentary, congress, politic, press, arthur lira, attack, sue
Intersection	1	company, money, economy, employment, crisis, buy, american countries, father, expensive, measure, president, person, test governor
Intersection	2	power, moment, minister, coronavirus, bolsonaro, Brazilian, people, health, quarantine, measure, now, speak, know, death, work

Table 4: Top 2 topics for each dataset (translated to English)

4.2 Lexical analysis

For the lexical analysis, we use the 2015 LIWC lexicon [11] that categorizes words into linguistic style, affective and cognitive attributes. To extract the lexical distribution of the terms used in the captions for each dataset, we used the Portuguese version of the lexicon. Thus, for each caption, we compute the value of each LIWC attribute, normalized by the number of words of each caption.

Given an input video caption, we compute the value of an LIWC attribute as the percentage of words in the caption that represents the given attribute. We thus normalize the attribute value by the size of each caption text. To compare the differences between the Pandemic and the Politics datasets, we calculate the relative differences between each correspondent LIWC attribute. Figure 4 shows that relative differences (up to 5%) only for attributes that were statistical significantly different ($p\text{-}value < 0.05$).

For instance, in Figure 4, a positive difference means videos in the Pandemic dataset had more of that attribute than the one in

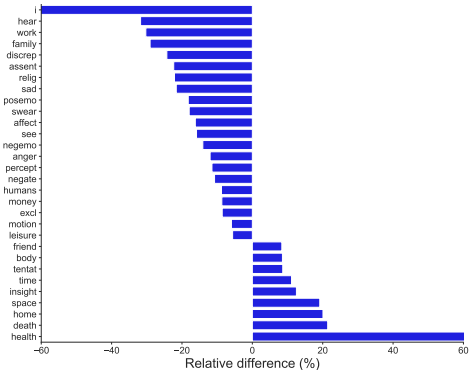


Figure 4: Relative difference between the Pandemic and the Politics videos

the Politics dataset and vice-versa. Regarding that comparison, we note that the attribute with the greatest presence in the pandemic videos is *health*, as shown in word cloud and LDA analysis in use of terms such as *infect*, *medicate*, *immunity*, and *care*. The second most frequent LIWC category was *death* which is also associated with the previous category and related to *illness* one of the most frequent terms in the pandemic word cloud. Moreover, we identify a significant presence of terms in the *insight* category, characterized by the terms related to analytical thinking. We manually examined some of the videos in that category and we observe that most of them provide information about the coronavirus. Note that for the same dataset, the "research" term was highlighted by the LDA model and in the word cloud some frequent terms associated with this category were "information", "think", and "fact", suggesting the informative nature of the videos.

On the other hand, we observe a higher frequency of the attribute *I* representing personal pronouns in the first-person singular. This suggests the use in phrases that represents the personal opinion of their interlocutors. Finally, we note that affective attributes, such as positive (posemo), negative (negemo), anger, and sad emotions as well as negations (negate) were more frequent in the Politics dataset. Moreover, work, family, and religion (relig) attributes have a greater presence in the Politics dataset. Recall that the LDA model also identifies terms such as "company", "money", "employee", "crisis" and "bankrupt" word in the word cloud for the same dataset. This can be explained by the large presence of right-wing groups in Whatsapp. These groups had shown a major concern with negative economic consequences of the pandemic.

5 CONCLUSIONS

In this paper, we present an analysis of YouTube videos shared in Whatsapp groups during the first month of the COVID-19 pandemics in Brazil. Our investigation was centered on understanding what type of engagement Brazilian political groups were engaged during the pandemic crisis. Our results show that the volume of YouTube videos is correlated with external events such as demonstrations and presidential announcements. Examining the captions transcribing the videos, we observed that affective attributes, as well as attributes related to family, work, and money, were more common in political videos. On the other hand, videos that only discuss themes related to the pandemic (no political) were more neutral compared to the political ones as well as more focused on health and death themes. This analysis is the first step to understand to which extend these shared videos may impact people during highly discussed periods. Finally, further research will target more specific behaviors, such as misinformation dissemination, analysis of toxicity as well as comparison with other social networks.

ACKNOWLEDGMENTS

This work was partially supported by the project CNPq, CAPES, and FAPEMIG.

REFERENCES

[1] C. Souza Araújo, G. Magno, W. Meira, V. Almeida, P. Hartung, and D. Doneda. 2017. Characterizing Videos, Audience and Advertising in Youtube Channels for Kids. In *Social Informatics*.
[2] D. Blei, Andrew Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[3] S. Boberg, T. Quandt, T. Schatto-Eckrodt, and L. Frischlich. 2020. Pandemic Populism: Facebook Pages of Alternative News Media and the Corona Crisis – A Computational Content Analysis. arXiv:cs.SI/2004.02566
[4] V. S. Bursztyn and L. Birnbaum. 2019. Thousands of Small, Constant Rallies: A Large-Scale Analysis of Partisan WhatsApp Groups. In *ASONAM*.
[5] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. Valensise, E. Brugnoli, A. Schmidt, P. Zola, F. Zollo, and A. Scala. 2020. The COVID-19 Social Media Infodemic. arXiv:cs.SI/2003.05004
[6] E. Ferrara. 2020. What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday* (May 2020).
[7] K. Garimella and G. Tyson. 2018. WhatsApp Doc?: A First Look at WhatsApp Public Group Data. In *ICWSM*.
[8] A. Maros, J. Almeida, F. Benevenuto, and M. Vasconcelos. 2020. Analyzing the Use of Audio Messages in WhatsApp Groups. In *TheWebConf*.
[9] D. Newman, J. Han, K. Grieser, and T. Baldwin. 2010. Automatic evaluation of topic coherence. In *ACL*.
[10] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr., and V. Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *WebSci*.
[11] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
[12] G. Preta. 2019. WhatsApp é principal fonte de informação do brasileiro. https://olhardigital.com.br/noticia/whatsapp-e-principal-fonte-de-informacao-do-brasileiro/94143. Accessed in 2020-07-01.
[13] G. Resende, P. Melo, J. Reis, M. Vasconcelos, J. Almeida, and F. Benevenuto. 2019. Analyzing Textual (Mis)Information Shared in WhatsApp Groups. In *WebSci '19*.
[14] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *WWW'19*.
[15] S. Shahsavari, P. Holur, T. Tangherlini, and V. Roychowdhury. 2020. Conspiracy in the Time of Corona: Automatic detection of Covid-19 Conspiracy Theories in Social Media and the News. arXiv:cs.CL/2004.13783