Dataset of coronavirus content from Instagram with an exploratory analysis

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, Gareth Tyson

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 6, 2021, accepted October 25, 2021, date of publication November 8, 2021, date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3126552

Dataset of Coronavirus Content From Instagram With an Exploratory Analysis

KOOSHA ZAREI[1], REZA FARAHBAKHSH[1], NOEL CRESPI[1], AND GARETH TYSON[2]
[1]Institut Polytechnique de Paris, Télécom SudParis, 91011 Evry, France
[2]Queen Mary University of London, London E1 4NS, U.K.

Corresponding author: Koosha Zarei (koosha.zarei@telecom-sudparis.eu)

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/S033564/1.

ABSTRACT The novel coronavirus (COVID-19) pandemic outbreak is drastically shaping and reshaping many aspects of our life, with a huge impact on our social life. In this era of lockdown policies in most of the major cities around the world, we see a huge increase in people and professionals' engagement in social media. Online Social Networks are playing an important role in news propagation as well as keeping people in contact. At the same time, social media is both a blessing and a curse as the coronavirus infodemic has become a major concern, and is already a topic that needs special attention and further research. In this study, we publish a multilingual coronavirus (COVID-19) Instagram dataset that we have continuously collected during the first wave of the pandemic from 5 January 2020 to 30 May 2020. The dataset contains 25.7K posts, 829K comments, and 3.2M likes in various subjects from different publishers such as 'public accounts', 'fake accounts (bots)', 'newsagencies', 'influencers', 'celebrities', 'business pages', etc. In addition to the dataset, this paper provides an analysis of the behaviour of the publishers. We study the behavioural aspects of the users in terms of their engagement, use of hashtags, activities, reactions as well as a full analysis of the published content related to the COVID-19. We believe this contribution helps the research community to better understand the dynamics behind this phenomenon in Instagram, as one of the major social media.

INDEX TERMS Coronavirus, COVID-19, Instagram, social network analysis, dataset, bot, fake content.

I. INTRODUCTION

A. THE APPEARANCE OF THE CORONAVIRUS

The novel coronavirus (COVID-19) was declared a pandemic by the World Health Organisation (WHO) on 11 March 2020.[1] Since then, the world has experienced almost 3 million cases. To mitigate its spread, many governments have therefore imposed unprecedented social distancing measures that have led to millions becoming housebound. This has resulted in a flurry of research activity surrounding both understanding and countering the outbreak [1]. As part of this, social media has become a vital tool in disseminating public health information and maintaining connectivity amongst people. Several recent studies have relied on Twitter data to better understand this [2]–[5]. These have primarily focused on health-related (mis)information, but there have also been studies into online hate [6]. Despite this, there has been only limited exploration of other social modalities, such as image content.

B. THE FIRST CORONAVIRUS WAVE

The World Health Organization published the first disease outbreak news on 5th January 2020 [7], containing risk assessment and advice [8]. On 9th March 2020, Italy declared a nationwide lockdown, which was the first in Europe. France reported over 10,000 coronavirus cases on March 19th, and then, on March 23th, Boris Johnson announced a UK-wide partial lockdown. On March 26th, the United States officially became the country hardest hit by the pandemic [9]. Table 1 summarizes the important events that happened during the first outbreak of the pandemic between January 2020 and March 2020. In this study, we mark and use these critical events in our analysis.

Our work is underpinned by a large-scale dataset from Instagram. We started collecting the COVID-19 content on Instagram on 5 January 2020, and we continued this task until 30 May 2020. During these months, we have seen a

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

[1]https://tinyurl.com/WHOPandemicAnnouncement

TABLE 1. Events of the 1st Coronavirus wave.

Date	Event	Ref.
Jan 5, 2020	WHO: first Disease Outbreak News	[7]
Jan 23, 2020	Wuhan City: the city is under quarantine	[10]
Jan 30, 2020	WHO: global public-health emergency	[10]
Feb 19, 2020	Iran: coronavirus outbreak begins	[10]
Mar 11, 2020	WHO declares the outbreak a pandemic	[10]
Mar 13, 2020	Trump: declares a National Emergency	[11]
Mar 14, 2020	France: The closure of public places	[12]
Mar 18, 2020	Italy: national lockdown	[10]
Mar16, 2020	Trump: uses term of Chinese Virus	[10]
Mar 23, 2020	UK: announces lockdown	[11]
Mar 26, 2020	USA: highest coronavirus cases	[12]
Apr 1, 2020	UN: worst crisis since World War II	[12]
Apr 14, 2020	Trump: halt in funding for the WHO	[10]

series of important events in the world. The virus emerged in Wuhan city in China, followed by high levels of hospitalisation. The virus then spread into middle-eastern countries such as Iran, and then the outbreak began in Europe. Later, the World Health Organization (WHO) announced a global pandemic, and many countries went into full or partial lockdown. Eventually, the situation eased and countries relaxed their lockdowns, albeit with some then experiencing a second wave. All these events spanned just five months.

C. ROADMAP & CONTRIBUTIONS

In this study, we present and explore an Instagram dataset, covering several major keywords and hashtags related to COVID-19 during the first wave of the pandemic (Table 2). We hope that this study can help support a number of use cases. Summary of the main contributions in this study are:

- We present a large dataset of COVID-19 related content published in Instagram including metrics such as posts content and reactions across various communities (Section III).
- We perform a preliminary analysis of posts and their publishers with the goal of distinguishing published content by humans and machines (Section IV). We observe that 6.9% of posts are distributed by bots.
- We present interesting observations across hashtags by publishers (Section V). We observe many geographical hashtags are associated with '*#quarantine*. We witness several off-topic and unrelated hashtags co-located with the main COVID-19 hashtags. We also observe 36 isolated islands in the graph of hashtags.

II. RELATED WORK

We break related work into two groups: (*i*) studies related to collecting COVID-19 datasets from social media, and (*ii*) studies related to analysing COVID-19 related contents in social media.

A. COVID-19 DATASETS

There have been a range of COVID-19 social media datasets released. To date, this predominantly covers textual data (*e.g.* Twitter). To assist in this, Kazemi *et al.* [13] provides a toolbox for processing textual data related to COVID-19. In terms of data, the first efforts in this direction were from authors in [2], which provides a large Twitter dataset related to COVID-19 (by crawling major hashtags and trusted accounts). Another similar study [3], provides an Arabic Twitter dataset with a similar data collection methodology. Lopez *et al.* [4] provide another Twitter dataset, including geolocated tweets. There are some further efforts on providing similar datasets from Twitter [14]–[16]. Melotte *et al.* [17] present a more controlled and compact dataset without requiring extensive preprocessing or tweet-hydration. The proposed dataset comprises tens of thousands of geotagged tweets originally collected over 255 days in 2020 across 10 metropolitan areas (in North America). Sharma *et al.* [5] also made a public dashboard[2] available, summarising data across more than 5 million real-time tweets. In another study [18], to help the evaluation of the determinants and impact of the COVID-19 at a large scale, the authors present a new dataset with socio-demographic, economic, public policy, health, pollution and environmental factors for the European Union. Akindtande *et al.* [19] present a dataset that investigates the magnitude of the misinformation content influencing scepticisms about the novel COVID-19 pandemic in Africa and the data is collected via an electronic questionnaire method from twenty-one Africa countries. In medicine, Sass et al introduce the "German Corona Consensus Dataset" (GECCO), a uniform dataset that uses international terminologies and health IT standards to improve interoperability of COVID-19 data [20].

B. CONTENT ANALYSIS REVIEW

These Twitter datasets have been used for various lines of analysis. For example, Saire and Navarro [21] use the data to show the epidemiological impact of COVID-19 on press publications. Singh *et al.* [22] are also monitoring the flow of (mis)information flow across 2.7M tweets, and correlating it with infection rates. They find that misinformation and myths are discussed, but at a lower volume than other conversations. To the best of our knowledge, the only paper that has covered Instagram is by Cinelli *et al.* [23], who analyse Twitter, Instagram, YouTube, Reddit and Gab data about COVID-19. We complement this by making a public Instagram dataset available to the community. Kudchadkar *et al.* [24] observed trends in concurrent #PedsICU and #COVID19 usage which reflect evolving information, knowledge gained, and collaborations among the global pediatric critical care community in Twitter. In terms of sentiment analysis, Vijay *et al.* [25] analysed the tweets regarding COVID-19 from November 2019 to May 2020 in India and its effect. Most people started having Negative tweets but with increasing time shifted towards positive and neutral comments. We redirect readers to [1] for a comprehensive survey of ongoing data science research related to COVID-19. Another study showed that COVID-19 misinformation on Twitter was more

[2]https://usc-melady.github.io/COVID-19-Tweet-Analysis/

likely to come from unverified accounts, *i.e.,* accounts not confirmed to be human [26]. Pennycook *et al.* [27] showed why people believe and share misinformation related to COVID-19 and point to a suite of interventions based on accuracy nudges that social media platforms could directly implement. They believe such interventions are easily scalable and do not require platforms to make decision about what content to censor. The role of social media platforms in promoting COVID-19 conspiracy theories is also studied in [28]. Pang *et al.* [29] performed a comprehensive meta-analysis of publicly available global metabolomics datasets obtained from three countries (the United States, China and Brazil). They have implemented a computational pipeline to perform consistent raw spectra processing and conducted meta-analyses at pathway levels instead of individual feature levels. Mahmoudi et al [30] investigate the relationships between the counts of cases with COVID-19 and the deaths due to it in seven countries that are severely affected by this pandemic disease. In contrast to these prior works, we focus on image-based social media and explore the role that different account types play in disseminating COVID-19 related material.

III. DATA COLLECTION CAMPAIGN

A. CRAWLER ARCHITECTURE

In order to collect the Instagram public content, we develop a crawler that is able to handle various tasks simultaneously. This crawler connects to Instagram via multiple channels, downloads public data content concurrently, performs some NLP based pre-processing steps, and finally stores them in a NoSQL format database. In Instagram, a reaction to a post can be active (comment) or passive (like). The crawler relies on the official Instagram APIs described in [31]. To get the public content that is tagged with a specific hashtag or keyword, we use the Instagram Hashtag Engine which is available in [32]. This API returns public posts that have been tagged with particular hashtags. Our crawler runs on several virtual machines in parallel 24/7. Note, that we do not manually filter any posts and therefore we gather all posts containing the hashtags, regardless of the specific topics discussed within.

Figure 1 shows the complete architecture design of our crawler, which contains four different major parts to handle data crawling: (*i*) API Connection Layer, (*ii*) Proxy Layer, (*iii*) Main Body, and (*iv*) Database Layer. The process of receiving data is as follows:

1) The API Connection Layer (Block 1 in Figure 1) connects to the official Instagram platform [31], which is currently using the Graph API. The crawler is registered as an application to be able to perform user authentication [33]. Note, there are certain rate limitations for requesting information per hour [31].
2) Between the Connection Layer and the Main Module, the Proxy Layer (Block 2 in Figure 1) is responsible for handling multiple proxy IP addresses and creating multiple connection layers. This helps us to receive data at a faster rate from various IP addresses. Thus, these layers are working concurrently.
3) The main body of the crawler (Block 3 in Figure 1) contains several inner modules such as Post, Reaction, Profile, Social Connection, and Story or Live modules. These are responsible for getting parts that are associated with their names. For example, the Post module is programmed to get Instagram posts and metadata. These modules are directly connected to a scheduler that handles time management. For example, checking daily stories, updating reactions, looking for new posts, checking highlights, and revising new social connections. Last, the *Pre-Processing layer* is used to perform some basic pre-processing steps such as text cleaning, data management, language extraction, *etc.*
4) In the Database Layer (Block 4 in Figure 1), we store our data. We use MongoDB as the primary database and we keep each module in a separate corresponding collection. For example, post content is stored in the *post collection*.

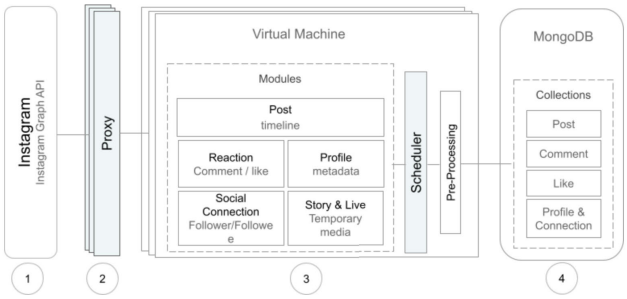


FIGURE 1. The architecture of our instagram crawler.

B. DATA COLLECTION AND PRE-PROCESSING

1) COLLECTION

On 5 January 2020, we prepared an initial list with '*#coronavirus*', '*#covid19*', and '*#covid_19*' keywords. We list the complete tracked keywords/hashtags in Table 2. Whenever a new keyword appears, we add it to our watch list. We continuously check new hashtags from [34] and [35] sources. For example, on 19 January 2020, we added '*#corona*', and '*#stayhome*'. By the end of January and beginning the lockdown in Europe, we also began to track '*#quarantine*', and '*#covid*' tags.

Using our above crawler, we continuously iterate over this list to collect associated posts. If any of the keywords exist in a post's *caption, hashtags, tagged users, location, or mentions*, we consider that post as COVID-19 related. In order to get post reactions, we revisit posts for two weeks after the initial posting to gather comments and likes.

2) GRAPHS

We later explore the relationships between hashtags. To achieve this, we induce a graph dataset whereby hashtags

TABLE 2. Tracking hashtags on instagram.

Hashtag	Post	Publisher	Reaction	Crawled Since
#coronavirus	12.7K	11K	7.3M	*January 5, 2020*
#covid19	8.0K	7K	6.5M	*January 5, 2020*
#covid_19	6.1K	5.9K	1.1M	*January 5, 2020*
#corona	2.9K	2.7K	1.9M	*January 19, 2020*
#stayhome	2.9K	2.7K	421K	*January 30, 2020*
#quarantine	2.3K	2.1K	322K	*January 30, 2020*
#covid	1.6K	1.4K	135K	*January 30, 2020*
#socialdistancing	0.7K	490	43.9K	*January 30, 2020*
#pandemic	0.7K	354	55K	*January 30, 2020*
#lockdown	1.3K	644	68K	*January 30, 2020*

that appear in posts are nodes, and edges indicate that two hashtags have appears in the same post (at least once). We only consider hashtags between 3 to 25 characters. We set the node weight as the frequency that a tag is used. We later plot graphs using [36].

3) BOT DETECTION

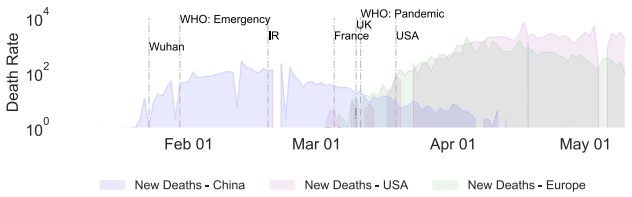
In order to identify bots, we extract and use features from [37]–[39] studies. Features are a combination of post and publisher metrics: *"profile image (image), biography text (text), account url (text), full name (text), number of followers (numeric), number of followee (numeric), account age (numeric), number of posts (numeric), avg. received like (numeric), avg. received comments (numeric), number of posts (numeric), number of issued like (numeric), number of issued comments (numeric), following/followee ratio (numeric), followers/post ratio (numeric), biography emoji count (numeric), biography hashtag count (numeric), biography length (numeric), verified (numeric), duplicated comments (numeric), number of followers that are bots (numeric), number of followee that are bots (numeric), post caption (text)"*. To build a training set, we randomly select 6K posts and manually label the profiles. Based on mentioned metrics, we examine each profile by hand and annotate it as "*bot*" or "*not bot*" identity. Metrics include profile-level features (*"full name, profile image, number of follower, verified, account age,* etc. *"*) and post-level features (*"received like, received comments, post caption,* etc. *"*). In the training set, each class has 2.1K validated samples. For all text-based features such as "biography", we remove all punctuation marks, stopwords and convert them to lowercase characters. Words are stemmed to reduce to their root forms. Numerical metrics are min-max normalised. Next, we train a Contextual LSTM Neural Network classifier with the same model architecture reported in [40]. In this model, both text and metadata metrics from posts and profiles are considered. First, we tokenize text metrics (*e.g.* biography) using Keras Tokenizer Class [41] and then the result is fed to the LSTM layer which outputs a 64-dimension vector. We attach numerical metadata to this vector and pass it through 2 ReLU activated layers of sizes 128 and 64. Finally, it connects to an output layer that predicts the label. We use a random split of 80% (training set) and 20% (test set), and to avoid over-fitting we use 10-fold Cross-validation. The Contextual model achieved a final accuracy of 88%, precision of 87%, recall of 87%, and F1 of 88%.

C. CHALLENGES AND LIMITATIONS

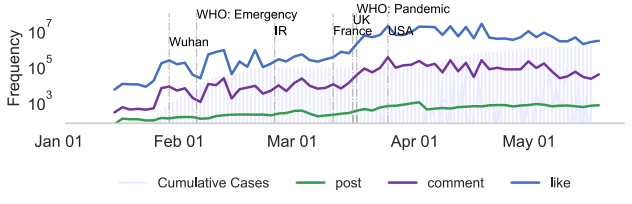
Note that as it is infeasible to collect all reactions. Hence, we define a limitation of 500 comments and 500 likes per post. We monitor reactions for up to two weeks to reach this limitation. In line with Instagram's Terms, Conditions, and Policies [42] as well as user privacy, we only gather publicly available data that is obtainable from Instagram. We also only rely on Instagram Posts and Reactions. We do not collect other data types such as Stories or Highlights.

D. DATA SUMMARY

Figure 2.a presents the death rate in different locations as reported by the World Health Organization (WHO). We compare that with the trends of total '*posts*', '*comments*', and '*likes*' in Figure 2.b. As reported by the WHO, the COVID-19 outbreak began several months before January 2020 in Wuhan (China), but at the time, it was not considered a global crisis yet. That is the reason why we see figures start with large numbers. Post content is published by various groups such as ordinary people, politicians, companies, news media, governments, fake identities, *etc.* The publishing rate increases continuously (post and reactions) during this period. However, there are some surges and fluctuations in numbers in several critical points: (*i*) The outbreak in the city of Wuhan and the first peak of death rate in China (Jan 2020); (*ii*) The announcement of a state of emergency by the WHO (Feb 2020); and (*iii*) The beginning of the first wave and the surging death rates in Europe and the USA (March 2020).



(a) World Health Organization Rep



(b) Content

FIGURE 2. (a) The death rate reported by the WHO; (b) The overall trends of the published content.

In total, we have collected 829K comments and 3.2M likes from 25.7K public posts. Posts are distributed by 13.3K publishers. Table 3 summarizes the general stats regarding the posts, profiles, and reactions. Each Instagram part may contain various data types. For example, in a post, there exist

'caption', 'location', 'date', 'hashtags', 'mentions', *etc.* In Table 4, we summarize and describe all data features. This covers four main data types: '*text*', '*numeric*', '*boolean*', '*date*', and '*binary*'. We store images in a binary format.

TABLE 3. General dataset stats.

Post count	25.7K	avg. follower per account (median)	2.3M (695)
Unique publishers	13.3K	avg. followee per account (median)	725 (276)
Total Comment	829K	avg. mediacount per account (median)	1.6K (127)
Total Like	3.2M	avg. received like per post (median)	10K (43)
Total reactions	4M	avg. received comment per post (median)	141 (2)

TABLE 4. List of data features.

Type	Attribute	Type	Description
Profile Metadata	follower count	*numeric*	*audience size*
	followee count	*numeric*	*friend size*
	media count	*numeric*	*published posts*
	is verified	*boolean*	*verified by instagram*
	is private	*boolean*	*public or private*
	full name	*text*	*full name text*
	biography	*text*	*biography text*
	username	*text*	*account username*
	id	*numeric*	*unique id*
	profile pic url	*text*	*picture url*
	external url	*text*	*if exists*
	is business account	*boolean*	*if exists*
Post Metadata	caption	*text*	*post caption*
	date	*date*	*publish date*
	like count	*numeric*	*number of likes*
	comment count	*numeric*	*number of comments*
	shortcode	*numeric*	*unique id*
	hashtags	*text*	*list of hashtags*
	mentions	*text*	*list of caption mentions*
	is video	*boolean*	*video or image*
	video url	*text*	*if video == true*
	location	*numeric*	*location tag*
	tagged users	*text*	*tagged users in photo*
	thumbnail	*binary*	*content thumbnail*
	id	*numeric*	*unique post id*
Like Reaction	username	*text*	*username*
	id	*numeric*	*unique user id*
Comment Reaction	username	*text*	*username*
	id	*numeric*	*unique user id*
	date	*date*	*publish date*
	text	*text*	*comment text*

E. ACCESS TO DATASET

This dataset is accessible through: https://github.com/kooshazarei/COVID-19-InstaPostIDs. We publish our dataset in agreement with Instagram's Terms & Conditions [42]. Thus, as it is not permissible to release the post content and reactions, we share the post IDs (known as *shortcodes*). Researchers can then use tools such as Instaloader [43] to dehydrate the dataset. For any further question, please contact Koosha Zarei (*koosha.zarei@telecom-sudparis.eu*).

IV. CHARACTERIZING PUBLISHERS

In this section, we categorize publishers, before inspecting how their publication rates differ.

A. PUBLISHER CATEGORIES

First, we strive to understand COVID-19 related publishers. We argue that this can offer insight into how this information is generated and distributed [44]. We particularly focus on understanding how much COVID-19 generated information can be considered reliable [45].

1) OVERVIEW OF PUBLISHER CATEGORIES

Overall, we identify approximately 13.3K unique publishers. We observe a range of account characteristics. For example, some accounts have a high number of followers, and some represent well-known figures such as celebrities or brands. We categorize publishers into the following groups, as summarized in Table 6:

a: NEWS AGENCIES

To identify News agencies, we make a list of English speaking agencies on Instagram using two sources [46], [47]. Then, we filter and verify more than twenty News media accounts in our dataset. While all these accounts are already verified and categorized as 'media/news companies' by Instagram, they usually have millions of followers. We list all existing News agencies in Table 5. We find that 12.2% of posts, 0.7% of unique publishers, and 26% of total reactions belong to News agencies.

TABLE 5. List of news agencies in our dataset.

@washingtonpost	@nbcnews	@abcnews	@dwnews
@skynews	@foxnews	@wsj	@france24_en
@bbcnews	@cnn	@time	@nytimes
@dailymail	@euronews.tv	@the.independent	@telegraph
@politico			

b: CELEBRITIES

We also witness the existence of posts from popular singers, actors, artists, sports players, and other figures. We compile a list of popular celebrities using [48], [49] and then search for them in our dataset. We find that these celebrity accounts tend to be verified public profiles, usually with millions of followers (avg. 80M) yet few followees (avg. 230). Some of the top figures that we see are *'@ladygaga', '@arianagrande', '@jlo', '@oprah', '@leonardodicaprio', '@christiano', '@leomessi', '@serenawilliams', '@davidbeckham', '@eltonjohn', '@jenniferaniston', '@theellenshow', '@kimkardashian', '@beyonce',* etc. The number of celebrity accounts is not as large as other groups, and they usually publish more Instagram stories or live broadcasts rather than posts. However, they obtain a large number of reactions, especially comments, which make them a valuable source (see Figure 5). This group holds 4.3% of all posts, 0.5% of unique publishers, and 45.2% of total reactions.

c: BUSINESS PAGES

These cover the official pages of companies on Instagram. To identify such accounts, we rely on [50], [51], and use the Instagram Category feature (as a company) [52]. Using these two resources, we extract all known business pages. We identify two types of business accounts: (*i*) profiles that

are already verified by Instagram as business profiles [53] such as '@Nike', '@google', '@chanelofficial', *etc.* with hundreds of followers. (*ii*) Profiles that represent small businesses that are not verified and have few followers. Business pages produce the longest caption length (average 628 characters) and tag the most people (1.5 on average). Business Pages hold 4.7% of the total posts, 26% of total reactions, and 2% of unique publishers.

d: INFLUENCERS

Some accounts are known as ''influencers''. These are refer to accounts that specifically attempt to influence public opinion, often in return for financial payments [40]. We filter and extract influencers based on feature set from [40]. Influencers utilized the highest number of hashtags within their posts (avg. 18). This group holds 4.8% of total posts, 1.3% of unique publishers, and 0.8% of total reactions.

e: BOTS

We further identify a set of bot accounts. These refer to accounts that are computationally operated [44]. We use a Contextual LSTM Neural Network classifier with 88% accuracy in order to train to identify bot accounts. The process of training the classifier, feature set, and results are explained in Section III-B in detail. Bot generate 6.9% of total posts, 2% of unique publishers, and 0.2% of total reactions.

f: PUBLIC ACCOUNTS

We refer to the rest of the publishers as ''Public Accounts''. In this category, profiles are non-verified public accounts that have a few to millions of followers. This group holds 67.1% of total posts (the most populated group), 93.5% of unique publishers, and 1.1% of total reactions.

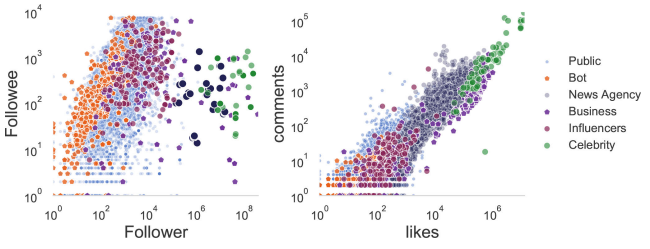


FIGURE 3. The Dot Plot of follower/followee count of categories (left), and received reactions across categories (right).

Validation: In order to validate categories, we manually check each one individually. For News agencies, Business pages, and Celebrities, we examine all samples and 100% of accounts are identified correctly. 86% of these accounts are already verified and approved by Instagram. For the influencer category, we randomly select 25% of samples and examine each by hand. 94.3% of influencers are identified correctly. To validate influencers we use the feature set from [40]. In the bot category, we randomly select 25% of samples and examine each manually. 94.3% of bots are identified correctly. To validate bots, we use [37]–[39] metrics. The process of bot detection is presented in Section III-B. Note, for the prior analysis, we remove incorrect samples from groups.

2) COMPARISON OF PUBLISHER CATEGORIES

We observe key differences among categories. Figure 3.a presents the Followers Friends Ratio (FFR) across account groups. This defines the social connectivity of an account [54]. Bot identities have ≤ 1 FFR, which means that they follow many other accounts, yet receive few followers in return. In contrast, News agency, Celebrity, and Business accounts tend to have ≥ 1 FFR. This ratio is considerably greater for News agency and Influencers as they have millions of followers.

We also inspect the attention generated by the posts of these accounts. To inspect this, Figure 3.b plots the number of comments *vs.* likes received by each group. Unsurprisingly, groups that have high FFR ratios receive considerably more reactions. The first notable category is Bots, which obtains considerably less attention than the other categories (avg. 26 likes and avg. 1.2 comments). In contrast, Celebrity (avg. 1.5M likes), Business pages (avg. 162K likes), and News agency (avg. 22.4K likes) get the most attention.

Arguably, the above results may be skewed as accounts with a large number of followers (as they are more likely to obtain reactions). To control for this, Table 6 presents the average engagement rate, as a percentage of the follower count. This actually shows that Bot accounts gain the largest engagement rate (20% compared to just 2% for business pages). This may, however, be a product of the type of accounts that follow bots. For instance, bots may follow each other and automatically like posts.

TABLE 6. General stats of publisher per categories.

Name	Post (%)	avg. like	avg. comment	avg. caption length	avg. #hashtag	avg. tagged user	avg. engagement rate (%)
Public Accounts	**67.1%**	110	7	375	16	0.5	17.0
News agencies	12.2%	22.4K	583	530	2.7	0.3	0.4
Bots	6.9%	26	1.2	344	17	0.5	20
Business Pages	4.7%	162K	731	**628**	7.5	**1.5**	4.0
Influencer	4.8%	425	52	411	**18**	1.1	9.0
Celebritis	4.3%	**1.5M**	**16.5K**	281	2.4	0.87	4.0

B. PUBLICATION RATE

Next, we explore the number of posts published by each category of account, presented in Figure 4 as a time series. Here, most of the posts are published by the 'Public' category (79% of posts) followed by 'News agencies' (12.2% of posts).

Overall, we see a growing number of weekly posts. Public publishers have the highest rate, thanks to the volume of accounts in this category (79%). Similarly, the Celebrity group publishes the fewest points (as they constitute just 0.3% of accounts). We find that these trends are also impacted by key events. The main surges occur, first, after Europe announces the pandemic on 14 March 2020 with 18.2%, followed by the USA on 26 March 2020 with 4.5%. News posts tend to be driven by dedicated coverage given to

COVID-19. For instance, 13 of the well-known agencies have launched a dedicated sections to cover the latest headlines. For example, BBC Coronavirus Stories [55], Euronews Special Coverage [56], Time COVID-19 Track [57], Skynews COVID-19 Section [58], and Foxnews Latest Coronavirus Headlines [59].

Perhaps most interesting is the Bot category. First, they publish a large volume of posts (4.9% of posts). This is the third most active category Second, they publish an almost fixed amount of posts during this pandemic, without the fluctuations seen in other groups. For example, compared to the News Agency group, we do not witness any noticeable peaks. This may be because of the computational manner in which such accounts generate content. The same trend is also reported in Twitter in [60] which shows an uptick in the frequency of bots' tweets referencing COVID-19 in the same period, and active bots sent 185K tweets and 1.4K retweets.

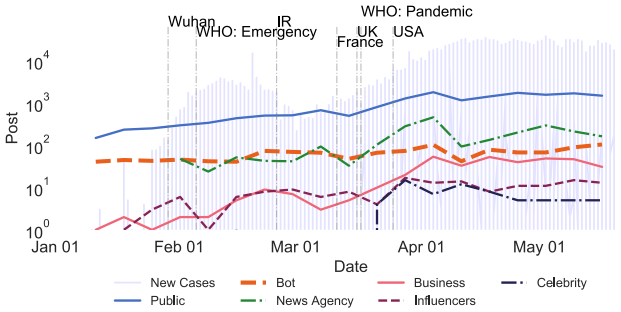


FIGURE 4. The trends of published posts by categories.

The Business category also exhibits unique trends. Due to the national lockdowns (largely introduced in April 2020), many businesses released information via Instagram and shifted activities online. Thus, we see noticeable growth in posts during this period. We also find that Influencers, ranging from 'nano' to 'mega' [40], increase their number of posts during this period. Similar trends can be seen amongst Celebrity accounts. Note that most in our dataset are American English-speaking figures. Therefore, there is almost no content until the start of the pandemic in Europe (March 2020). The authors of this study [61] also reported the same trends in Twitter and Instagram posts associated with COVID-19 content.

C. REACTIONS

We next look at the number of obtained reactions, as measured by comments and likes. Considering both, we witness some unique points: (*i*) In the Public category, we see a high number of comments (147K) and likes (2.2M), consistently across the whole time period. (*ii*) The Bot category receives the lowest number of reactions than other categories in both metrics (33K likes, 1.6K comments). Notably, both trends are constant during the time, regardless of events. (*iii*) The News agency category receives the largest number of comments (1.8M) and likes (71M) among all. Both figures nearly follow the same fluctuations. These figures peak in March 2020 where the virus reaches Europe, UK, and then the USA. (*iv*) In the Business group, we see user reactions peak in two spots: 1) the first outbreak in Iran in February 2020, and then 2) after declaring the lockdown in Europe and the USA in March 2020. Afterwards, it continues steadily (72M likes vs 32K comments). (*v*) In the influencer category, we see fluctuations. However, the overall trends remain steady (73K like and 8K comments). (*vi*) In celebrities, we see a huge surge in March 2020, where the outbreak starts in Europe and UK (122M likes vs 131K comments). Note that the reaction rate is zero before March 2020 as there is no COVID-19 related published post by celebrities. Most of the celebrities are from English-speaking countries.

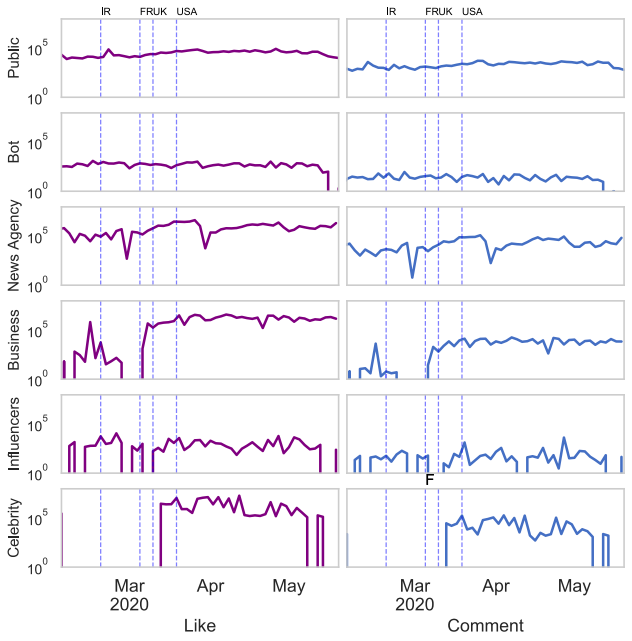


FIGURE 5. Reaction per category in a weekly scale.

In our study, another notable point is that during the first outbreak, three categories of News agency with 50%, Celebrity with 36%, and Business pages with 1.5% of total comments, obtain the highest written comments (total of 3.5M comments) through categories. Respectively, they also receive 26%, 45%, and 27% of total likes. In other words, these trusted groups potentially attract more people and target more audiences in Online Social Networks to distribute information, especially in critical moments such as the COVID-19 health crisis. More investigation through the comment text can lead us to understand the behaviour of people during the global crisis.

V. CHARACTERIZING THE USE OF HASHTAGS

As a proxy for post content, we next explore the hashtags employed by publishers.

A. DISTRIBUTION

A considerable part of the content (62% of posts) is tagged with '*#coronavirus*'. So, we consider this hashtag as the main hashtag. The '#coronavirus' tag is used by nearly 11K unique accounts and receives more than 7M reactions (comments and likes). We plot the hashtag graph of the most used keywords in this dataset in Figure 6. Note, the process of generating graphs is presented in Section III-B.

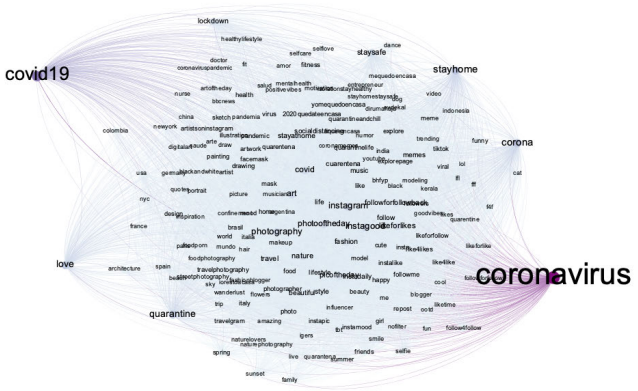


FIGURE 6. The graph of the top most used hashtags.

In order to observe the usage behaviour of hashtags in posts, we randomly select 5K posts (20% of data) that contain all categories. We manually check captions, images, hashtags, and publisher profiles. Among main hashtags (Table 2), we witness some important hashtags that are connected in numerous posts:

(1) We observe many geographical hashtags associated with '*#quarantine*, such as '*#spain*', '*#italy*', '*#usa*', '*#china*', '*#colombia*', '*#nyc*', '*#trip*', and '*#travel*' in 25.2% of posts. These tend to talk about lockdown-related situations, the spreading of the virus throughout the world, health crises in big cities, and the contamination rate in various locations. Similarly, in 31% of posts, with the '*#staysafe*' hashtag, we see '*#selflove*', '*#selfcare*', '*#fitness*', '*#love*', '*#mentalhealth*', '*#psitive vibes*', '*#motivation*', and '*#healthylifestyle*' hashtags. These tend to be tagged with general content during the first lockdown.

(2) On the other hand, we also observe a number of off-topic and unrelated hashtags (co-located with the main hashtags presented in Table 2). For example, with '*#coronavirus*', hashtags such as '*#likeforlike*', '*#f4f*', '*#lol*', '*#liikeforliike*', '*#followforfollow*', '*#followeme*', '*#tiktok*', and '*#instalike*' are tagged within 27.2% of posts. These posts are largely distributed by Public (73% of posts) and Bot categories (14% of posts).

(3) Similarly, in 11.7% of posts, we find the '*#covid19*' hashtag is associated with other non-related ones such as '*#portrait*', '*#picture*', '*#drawing*', '*#design*', '*#photooftheday*', '*#fashion*', and '*#hair*' that are used with. These tags are distributed by Public (91%) and Bot (9.2%) categories, and tend to promote fashion goods and photography materials.

We next look into their temporal trends during the first wave, presented as a time series in Figure 7. This figure shows what percentage of posts are tagged with the key hashtags (Table 2), as shown in the legend. We also plot the number of deaths and new cases (on a weekly scale) as reported by the WHO. Due to space constraints, we limit ourselves to the top five COVID-19 related hashtags.

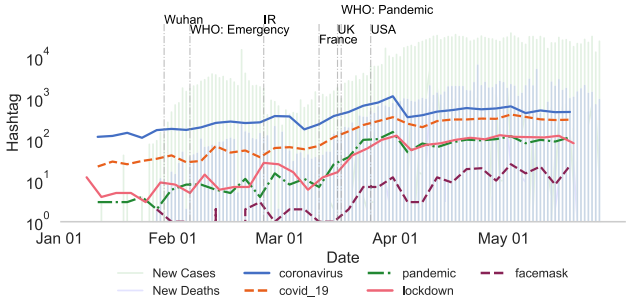


FIGURE 7. Trends of top most used keywords. Trends are compared with daily new positive cases and daily new deaths, reported by the WHO. (bar charts show new Cases|Deaths).

The '*#Coronavirus*' hashtag is the most used, as we crawl our dataset based on this keyword (14K). The second most used one is the '*#covid_19*' (11.2K) which is a more academic naming version of this virus. '*#lockdown*' (1.4K) and '*#pandemic*' (1K) hashtags follow nearly the same pattern from the very beginning days. Both tend to be tagged with posts that are talking about situations around lockdowns, the economical consequences, inviting people to work and study from home, and how to slow down the contamination rate. '*#facemask*' is another important hashtag that is used to encourage people to wear face masks in 9% of posts. This trend begins to grow from April 2020 when the first outbreak started in Europe and then in the USA. We can contrast these hashtags with the death and case rate. We note that these peak on two dates, in the middle of February 2020 (China, Iran, and Italy) and in April 2020 (Europe, USA, UK, Middle east). We witness all hashtags surge with these fluctuations especially in middle March (by 12%) where the virus reached Europe.

B. HASHTAG ISLANDS

Figure 8 presents the graph of all hashtags. We colour code the hashtags based on the category of accounts (each colour represents one category). If a hashtag belongs to two groups, that node gets the colour of the category that is greater in size. In this network, main hashtags such as '*#Coronavirus*' and '*#Covid19*' (Table 2) are located at the centre, surrounded by more connected nodes.

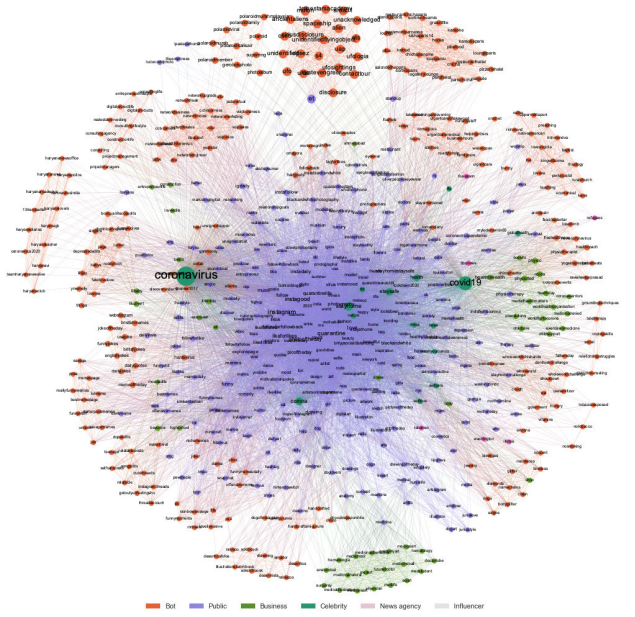
Nodes with larger sizes (frequency) are located closer to each other. We see this characteristic in hashtags used by 'Public', 'Celebrity', 'News agency', and 'Influencer' categories. That is the reason why they are located closer to the main nodes. Furthermore, they are more topic related,

and hold higher connection rates with others. In contrast, interestingly, hashtags primarily used by bots (red nodes) are located far from the centre with smaller sizes and fewer node connections. We also see the same behaviour in 'Business' nodes (green nodes).

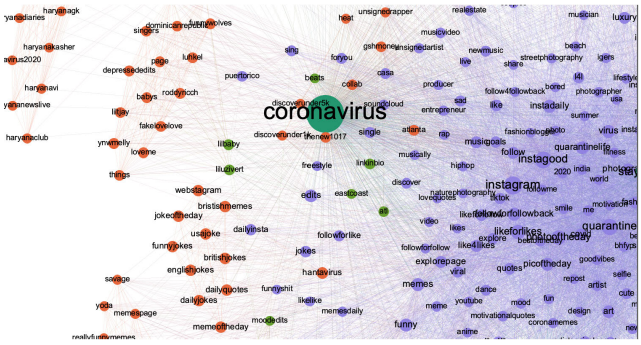
We witness 36 isolated islands. Islands are where there are a set of inter-connected hashtags that are disconnected from the main network of hashtags. Each island contains between 5 and 39 nodes. These islands have several characteristics: (*i*) In each island, all nodes are well-connected to each other (avg. 21 internal connections), but there is a weak connection to external nodes (avg. 6). (*ii*) Island nodes are used together in the same posts. So, node sizes are equal. (*iii*) Islands connect directly to the main network node or through a few nodes (max 4 connections). In this case, some islands are connected directly to the '*#Coronavirus*' node, but others with some extra nodes. (*iv*) Individual islands are disconnected from each other. In 21 islands (out of 36), we see no direct connection between them. This behaviour can be seen in the bot category (with 36 islands), which is presented in red nodes, and two green islands from business accounts. The same mechanism is also reported from authors of this study [44] which investigate the anatomy of online misinformation networks.

In Social Media, one strategy to gain more visibility is to tag posts with trending topics. This has been widely reported to be used by fake identities, spam, for-sale accounts, markets, some influencers, impersonators [62]. We witness suspicious behaviours among categories. For example, bots and some small business pages are using the Coronavirus topic to be in the Instagram Explorer and get more attention. As result, they appear like isolated islands with no connections to other nodes (or topics).

We look into two islands which are shown in Figure 8.b. The first island, from bot category (in red), is using 10 hashtags of '*#britishmemes*', '*#britishjokes*', '*#dailyjokes*', '*#webstagram*', '*#jokeoftheday*', '*#dailyquotes*', '*#usajoke*', '*#funnyjokes*', '*#memoftheday*', and '*#englishjokes*' to talk about jokes. These hashtags are used in 432 posts together in the same individual posts, and as results, the node sizes are the same. Each node is connected to all other island's internal nodes (9 nodes). This island connects to the main network node directly (through the '*#coronavirus*' hashtag), and indirectly (*e.g.* through '*#dailyinsta*', '*#jokes*', and '*#funnyshit*'). In the other island from bots, 14 hashtags of '*#depressededits*', '*#babys*', '*#fakelovelove*', '*#things*', '*#singers*', '*#loveme*', and others are used together. We also see that all internal nodes are connected to each other (each to 13 nodes). The node sizes are the same as they appear together in posts, and the island is connected to the network graph directly to the main node ('*#coronavirus*') and indirectly through some other nodes ('*#puertorico*', '*#lilbaby*', '*#sing*', '*#freestyle*', and others). There is no connection between this island and other identified islands. These hashtags are used in 561 posts together.



(a) Hashtag Graph



(b) Zoom preview

FIGURE 8. The use of hashtags across categories.

VI. LIMITATIONS

There are some limitations in this study that we would like to highlight. (*i*) First, despite the large amount of information on Instagram regarding COVID-19, we were only able to collect a small fraction in the first wave. Similarly, it was not possible to collect all reactions, particularly for popular accounts that receive thousands of comments (*e.g.* BBC). (*ii*) Furthermore, many accounts (*e.g.* News agencies) use Instagram Stories, which we did not collect. This means that we can only offer a lower-bound on activity. (*iii*) We note that the 'Public Accounts' category contains the largest number of posts and accounts. This group contains all accounts that did not fall into one of the other categories. Therefore, it is likely that this group can be further subdivided and may contain a diversity of users that we do can capture. We believe further exploration can be carried out such as identifying account types, clustering publishers, tracking misinformation, *etc.* to understand the behaviour of various entities. (*iv*) We also believe there are some profiles in the public category that can

be considered as fake identities (*e.g.* bots, spammers, fake protective equipment stores for COVID-19, fake health news distributor) which may require more investigation. This may naturally impact our insights.

VII. CONCLUSION & FUTURE WORK

In this study, we have targeted the first wave of the COVID-19 pandemic that occurred between 5 January 2020 and 30 May 2020. We present a multilingual Instagram dataset containing 25.7K posts, 829K comments, and 3.2M likes. We summarize our key findings as follows:

- The majority of bots in this study publish off-topic content (with regards to COVID-19). They exploit the COVID-19 hashtags to spread their content (4.9% of posts). That is why we see many isolated islands in the graph of hashtags (Section V-B). This behaviour is also reported in [26], [39], [63]
- The number of reactions to trusted publishers (Celebrities, News Agencies, and Business Pages) is 110x times higher than unreliable publishers. This highlights the importance of trustworthy accounts in critical moments (Section IV).
- Celebrities received the greatest attention from people. Noticeably, they published the smallest number of posts (0.3%) but received the most likes (avg. 1.5M per post) and comments (avg. 16.5K per post).
- In contrast to what we expected, Influencers received very little engagement through their posts. 27.3% of influencers stick to the Coronavirus trend to gain more attention.
- In this study, 17 News companies covered the latest news of this health crisis. Despite having a larger number of followers (avg. 4.7M), we observe limited engagement (avg. 22K likes and avg. 582 comments per post).

We further believe there are a number of potential lines of future work. First, COVID-19 has triggered a number of safety measures have been such as social distancing and working from home. We believe exploring people's behaviors through the lens of social media may shed light on how people have reacted to these measures. Second, we believe it important to study the dissemination of misinformation (which our dataset could unerpin). Developing techniques to overcome this problem is another potential direction. Finally, we have observed a number of bot accounts discussing and engaging on COVID-19. The identification of invalid profiles is another effective strategy in preventing untrustworthy content.

REFERENCES

[1] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, and J. Crowcroft, "Leveraging data science to combat COVID-19: A comprehensive review," *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 85–103, Aug. 2020.

[2] E. Chen, K. Lerman, and E. Ferrara, "COVID-19: The first public coronavirus Twitter dataset," 2020, *arXiv:2003.07372*.

[3] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large Arabic Twitter dataset on COVID-19," 2020, *arXiv:2004.04315*.

[4] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset," 2020, *arXiv:2003.1035*.

[5] K. Sharma, S. Seo, C. Meng, S. Rambhatla, and Y. Liu, "COVID-19 on social media: Analyzing misinformation in Twitter conversations," 2020, *arXiv:2003.12309*.

[6] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, "'Go eat a bat, chang!': An early look on the emergence of sinophobic behavior on web communities in the face of COVID-19," Tech. Rep., 2020.

[7] WHO. (Nov. 2020). *Pneumonia of Unknown Cause—China*. [Online]. Available: https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/

[8] WHO. (Jun. 2020). *Who Timeline—COVID-19*. [Online]. Available: https://www.who.int/news-room/detail/27-04-2020-who-timeline—COVID-19

[9] *Nytimes*. (Jun. 2020). *The U.S. Now Leads the World in Confirmed Coronavirus Cases*. [Online]. Available: https://www.nytimes.com/2020/03/26/health/usa-coronavirus-cases.html

[10] Business Insider. (Nov. 2020). *A Complete Timeline of the Coronavirus Pandemic*. [Online]. Available: https://www.businessinsider.com/coronavirus-pandemic-timeline-history-major-events-2020-3?IR=T

[11] AJMC. (Nov. 2020). *A Timeline of COVID-19 Developments in 2020*. [Online]. Available: https://www.ajmc.com/view/a-timeline-of-COVID19-developments-in-2020

[12] COVID Reference. (Nov. 2020). *The COVID Textbook*. [Online]. Available: https://COVIDreference.com/timeline

[13] S. K. Rashed, J. Frid, and S. Aits, "English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19," 2020, *arXiv:2003.09865*.

[14] C. Jacobs. (2020). *Coronada: Tweets About COVID-19*. [Online]. Available: https://github.com/BayesForDays/coronada

[15] Smith. (Mar. 2020). *Coronavirus (COVID19) Tweets*. [Online]. Available: https://www.kaggle.com/smid80/coronavirus-COVID19-tweets

[16] C. Jacobs, "Coronada," GitHub Repository, 2020. [Online]. Available: https://github.com/BayesForDays/coronada

[17] S. Melotte and M. Kejriwal, "A geo-tagged COVID-19 Twitter dataset for 10 North American metropolitan areas over a 255-day period," *Data*, vol. 6, no. 6, p. 64, Jun. 2021.

[18] H. Omrani, M. Modroiu, J. Lenzi, B. Omrani, Z. Said, M. Suhrcke, A. Tchicaya, N. Nguyen, and B. Parmentier, "COVID-19 in europe: Dataset at a sub-national level," *Data Brief*, vol. 35, Apr. 2021, Art. no. 106939.

[19] O. Akintande and O. Olubusoye, "Datasets on how misinformation promotes immune perception of COVID-19 pandemic in Africa," *Data Brief*, vol. 31, Aug. 2020, Art. no. 106031.

[20] J. Sass, A. Bartschke, M. Lehne, A. Essenwanger, E. Rinaldi, S. Rudolph, K. U. Heitmann, J. J. Vehreschild, C. von Kalle, and S. Thun, "The German corona consensus dataset (GECCO): A standardized dataset for COVID-19 research in university medicine and beyond," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, Dec. 2020.

[21] J. E. C. Saire and R. C. Navarro, "What is the people posting about symptoms related to coronavirus in bogota, colombia?" 2020, *arXiv:2003.11159*.

[22] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at COVID-19 information and misinformation sharing on Twitter," 2020, *arXiv:2003.13907*.

[23] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The COVID-19 social media infodemic," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, Oct. 2020.

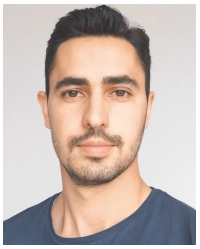
[24] S. R. Kudchadkar and C. L. Carroll, "Using social media for rapid information dissemination in a pandemic: #PedsICU and coronavirus disease 2019," *Pediatric Crit. Care Med.*, vol. 21, no. 8, pp. e538–e546, Aug. 2020.

[25] T. Vijay, A. Chawla, B. Dhanka, and P. Karmakar, "Sentiment analysis on COVID-19 Twitter data," in *Proc. 5th IEEE Int. Conf. Recent Adv. Innov. Eng. (ICRAIE)*, 2020, pp. 1–7, doi: 10.1109/ICRAIE51050.2020.9358301.

[26] R. Kouzy, J. A. Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. Akl, and K. Baddour, "Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter," *Cureus*, vol. 12, no. 3, pp. 1–9, Mar. 2020.

[27] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychol. Sci.*, vol. 31, no. 7, pp. 770–780, Jul. 2020.

[28] A. Bruns, S. Harrington, and E. Hurcombe, "'Corona? 5G? or both?': The dynamics of COVID-19/5G conspiracy theories on Facebook," *Media Int. Aust.*, vol. 177, no. 1, pp. 12–29, Nov. 2020.
[29] Z. Pang, G. Zhou, J. Chong, and J. Xia, "Comprehensive meta-analysis of COVID-19 global metabolomics datasets," *Metabolites*, vol. 11, no. 1, p. 44, Jan. 2021.
[30] M. R. Mahmoudi, D. Baleanu, S. S. Band, and A. Mosavi, "Factor analysis approach to classify COVID-19 datasets in several regions," *Results Phys.*, vol. 25, Jun. 2021, Art. no. 104071.
[31] Instagram. (Sep. 2021). *Official API Graph Instagram*. [Online]. Available: https://developers.facebook.com/docs/graph-api/
[32] Instagram. (Feb. 2020). *Instagram Hashtag Search*. [Online]. Available: https://developers.facebook.com/docs/instagram-api/guides/hashtag-search
[33] Instagram. (Sep. 2021). *API Access Tokens*. [Online]. Available: https://developers.facebook.com/docs/facebook-login/access-tokens/#usertokens
[34] (Jan. 2020). *Hashtagify*. [Online]. Available: https://hashtagify.me/hashtag/coronavirus
[35] (Jan. 2020). *Best Hashtags*. [Online]. Available: https://best-hashtags.com/hashtag/coronavirus/
[36] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," *ICWSM*, vol. 3, no. 1, pp. 361–362, Mar. 2009.
[37] Z. Gilani, R. Farahbakhsh, G. Tyson, and J. Crowcroft, "A large-scale behavioural analysis of bots and humans on Twitter," *ACM Trans. Web*, vol. 13, no. 1, pp. 1–23, Feb. 2019.
[38] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018.
[39] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, "Detection of bots in social media: A systematic review," *Inf. Process. Manage.*, vol. 57, no. 4, Jul. 2020, Art. no. 102250.
[40] K. Zarei, D. Ibosiola, R. Farahbakhsh, Z. Gilani, K. Garimella, N. Crespi, and G. Tyson, "Characterising and detecting sponsored influencer posts on Instagram," in *Proc. ACM/IEEE ASONAM*, 2020.
[41] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io
[42] Data Policy. (Mar. 2020). *Instagram Data Policy*. [Online]. Available: https://help.instagram.com/519522125107875
[43] Instaloader. (Jan. 2020). *Instaloader*. [Online]. Available: https://github.com/instaloader/instaloader
[44] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Anatomy of an online misinformation network," *PLoS ONE*, vol. 13, no. 4, pp. 1–23, 2018.
[45] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of low-credibility information on Twitter during the COVID-19 outbreak," 2020, *arXiv:2004.14484*.
[46] (Jan. 2020). *The Top Ten Most-Followed News Accounts on Twitter*. [Online]. Available: https://www.pressgazette.co.U.K./the-top-ten-most-followed-news-accounts-on-twitter/
[47] (Jan. 2020). *Similarweb, Top Websites Ranking for News and Media in the World*. [Online]. Available: https://www.similarweb.com/top-websites/category/news-and-media
[48] Statista. (Nov. 2020). *Instagram accounts With the Most Followers Worldwide 2020*. [Online]. Available: https://www.statista.com/statistics/421169/most-followers-instagram/
[49] Social Book. (Nov. 2020). *Top 100 Most-Followed Instagram Accounts*. [Online]. Available: https://socialbook.io/instagram-channel-rank/top-100-instagrammers
[50] Trackalytics. (Jan. 2020). *The Most Followed Instagram Profiles*. [Online]. Available: https://www.trackalytics.com/the-most-followed-instagram-profiles/page/1/
[51] (Jan. 2020). *20 Most Followed Brands on Instagram in 2019*. [Online]. Available: https://blog.unmetric.com/most-followed-brands-instagram
[52] Instagram. (Jan. 2020). *Stand our With Instagram*. [Online]. Available: https://business.instagram.com/getting-started
[53] Instagram. (Sep. 2020). *Instagram Business Account*. [Online]. Available: https://www.facebook.com/business/profiles
[54] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, Apr. 2020, pp. 1096–1103.
[55] BBC News. (Feb. 2020). *Coronavirus Pandemic*. [Online]. Available: https://www.bbc.com/news/coronavirus
[56] Euronews. (Feb. 2020). *Special Coronavirus*. [Online]. Available: https://www.euronews.com/special/coronavirus
[57] Time. (Feb. 2020). *Time COVID-19 Track*. [Online]. Available: https://time.com/tag/COVID-19/
[58] Skynews. (Feb. 2020). *COVID-19*. [Online]. Available: https://news.sky.com/topic/COVID-19-8518
[59] FoxNews. (Feb. 2020). *Latest Coronavirus Headlines*. [Online]. Available: https://www.foxnews.com/category/health/infectious-disease/coronavirus
[60] A. Al-Rawi and V. Shukla, "Bots as active news promoters: A digital analysis of COVID-19 tweets," *Information*, vol. 11, no. 10, p. 461, Sep. 2020.
[61] T. K. Mackey, J. Li, V. Purushothaman, M. Nali, N. Shah, C. Bardier, M. Cai, and B. Liang, "Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram," *JMIR Public Health Surveill.*, vol. 6, no. 3, Aug. 2020, Art. no. e20794.
[62] Y. Roth and N. Pickles. (May 2020). *Bot or Not? the Facts About Platform Manipulation on Twitter*. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2020/botornot.html
[63] M. Himelein-Wachowiak, S. Giorgi, A. Devoto, M. Rahman, L. Ungar, H. A. Schwartz, D. H. Epstein, L. Leggio, and B. Curtis, "Bots and misinformation spread on social media: Implications for COVID-19," *J. Med. Internet Res.*, vol. 23, no. 5, p. e26933, May 2021.



KOOSHA ZAREI is currently pursuing the Ph.D. degree in computer science with the Institut Polytechnique de Paris, France. He is an Active Member of the "Data Intelligence and Communication Engineering" Laboratory working on fake content and fake activities on online social networks. His research interests include online social networks, deep learning, NLP, fake news, and big data.

REZA FARAHBAKHSH received the Ph.D. degree from Paris VI (UPMC) jointly with the Institut-Mines Telecom, Telecom SudParis (CNRS Lab UMR5157), in 2015. He is currently an Adjunct Associate Professor with the Institut-Mines Telecom, Telecom SudParis, and a Data Scientist at TOTAL SA. He is actively involved in international collaborative projects. His research interests include NLP, language modeling, online social networks, and the IoT.

NOEL CRESPI received the master's degrees from the Universities of Orsay and Canterbury, the "Diplome d'ingénieur" from Telecom ParisTech, and the Ph.D. and Habilitation degrees from Paris VI University. He joined the Institute Mines-Telecom, in 2002, where he is currently a Professor and the M.Sc. Program Director, leading the Service Architecture Laboratory. He coordinates the standardization activities with Institute Telecom at ETSI, 3GPP, and ITU-T. He is also an Adjunct Professor at KAIST, South Korea, an Affiliate Professor at Concordia University, Canada, and a Guest Researcher at the University of Goettingen, Germany. He is the Scientific Director the French-Korean Laboratory ILLUMINE. He is the author/coauthor of 400 articles and contributions in standardization. His current research interests include data analytics, the Internet of Things, and softwarization.

GARETH TYSON received the Ph.D. degree from Lancaster University, in 2010. He is currently a Senior Lecturer with the Queen Mary University of London and a fellow with the Alan Turing Institute. His research interests include internet measurements, web, and social computing. He has won almost £2 million in external grants and received numerous awards, including the Best Student Paper Award at WWW 2020 and Outstanding Reviewer Awards at ICWSM.