A first Instagram dataset on COVID-19

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, Gareth Tyson

A First Instagram Dataset on COVID-19

Koosha Zarei*, Reza Farahbakhsh*, Noël Crespi*, Gareth Tyson†
*CNRS Lab UMR5157, Télécom SudParis, Institut Polytechnique de Paris, Evry, France.
{koosha.zarei, reza.farahbakhsh, noel.crespi}@telecom-sudparis.eu
†Queen Mary University of London, United Kingdom. gareth.tyson@qmul.ac.uk

***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. Social media is playing an important role in news propagation as well as keeping people in contact. At the same time, this source 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 paper, we provide a multilingual coronavirus (COVID-19) Instagram dataset that we have been continuously collected since March 30, 2020. We are making our dataset available to the research community at https://github.com/kooshazarei/COVID-19-InstaPostIDs. We believe that this contribution will help the community to better understand the dynamics behind this phenomenon in Instagram, as one of the major social media. This dataset could also help study the propagation of misinformation related to this outbreak.**

***Index Terms*—Coronavirus; COVID-19; Instagram, Social Network Analysis; Dataset.**

I. Introduction

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 government have therefore imposed unprecedented social distancing measures that have led to millions become 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.

We argue this represents a limitation, particularly considering the importance of image-based content in the dissemination of news (and misinformation) [7], [8]. This paper introduces a COVID-19 Instagram dataset, which we make available for the research community. We have gathered data between January 5 and March 30 2020 (§III). The dataset covers **18.5K** comments and **329K** likes from **5.3K** posts. These posts have been distributed by **2.5K** publishers. The data predominantly covers English language posts, and we provide a number of important features covering both the content and the publisher (§IV). We hope that this dataset can help support a number of use cases. Hence, we conclude the paper by highlighting a number of potential uses related to COVID-19 social media analysis (§V). Details of how to access the data is presented in §VI.

II. Related Work

Most related to this work is the set of COVID-19 social media datasets recently released. To date, this predominantly covered textual data (*e.g.* Twitter). To assist in this, Kazemi et al. [9] provide a toolbox for processing textual data related to COVID-19. In terms of data, the first efforts in this direction was from authors in [2] which provide a large Twitter dataset related to Coronavirus (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 the geolocated tweets. There are some further efforts on providing similar datasets from twitter [10]–[12]. Sharma et al. [5] also made a public dashboard[2] available summarising data across more than 5 million real-time tweets.

These Twitter datasets are being used for various use cases. For example, Saire and Navarro [13] use the data to show the epidemiological impact of COVID-19 on press publications. Singh et al. [14] are also monitoring the flow of (mis)information flow across 2.7M tweets, and correlating it with infection rates to find that misinformation and myths are discussed, but at lower volume than other conversations. To the best of our knowledge, the only paper that has covered Instagram is by Cinelli et al. [15], 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. We redirect readers to [1] for a comprehensive survey of ongoing data science research related to COVID-19.

III. Data Collection

We have collected public posts from Instagram by crawling all posts associated with a set of COVID-19 hashtags presented in Table I.

Methodology. To be able to collect Instagram public content (in the shape of post), we use the official Instagram APIs

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

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

TABLE I: Tracked Hashtags on Instagram - Release v1.0

Hashtag	*Number of Posts*	*Crawled Since*
#coronavirus	4.4K	January 5, 2020
#covid19/covid_19	1.5K	January 15, 2020
#corona	1.0K	January 19, 2020
#stayhome	537	January 30, 2020

[16]. In particular, to get posts that are tagged with specific hashtags, the Instagram Hashtag Engine is used [17]. This API returns public photos and videos that have been tagged with particular hashtags. MongoDB is used as the core database and data is stored as JSON records. The crawler is responsible for gathering both posts and reactions. A reaction can be active (comment) or passive (like). As it is infeasible to collect all reactions, in this dataset, we define a limitat of 500 comments and 500 likes per post. Our crawler is running 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. The complete architecture of our crawler is described in this paper [18].

Release v1.0 (April 20, 2020). The first version of this data collection process started on January 5, 2020 and continued until March 30, 2020. The data gathering is still running as the lockdown has not been finished in many countries around the world (at the time of writing this paper). During this time **18.5K** comments and **329K** likes from **5.3K** public posts have been collected. These posts are distributed by **2.5K** publishers.

Ethics. In line with Instagram policies as well as user privacy, we only gather publicly available data that is obtainable from Instagram.

IV. Dataset Description

To provide context for potential users of our dataset, we next brifely summarise the dataset and describe the characteristics of the content.

Hashtags. Recall that we gather the data by querying certain hashtags. Figure 2 presents the top hashtags tagged within the posts. Figure 1 also presents a wordcloud of the hashtags in our dataset. This naturally includes hashtags outside of our seed set used for crawling. There are intuitive examples, such as *corona*, *covid19*, *covid_19*, *stayathome*, *quarantine*, *love*, *covid*, *virus*, and *instagram*. The are therefore the most repeated hashtags that appear with #coronavirus. Note that this means will might miss posts that mention these concepts in other languages.

Fig. 1: Wordcloud of the most related hashtags in our dataset.

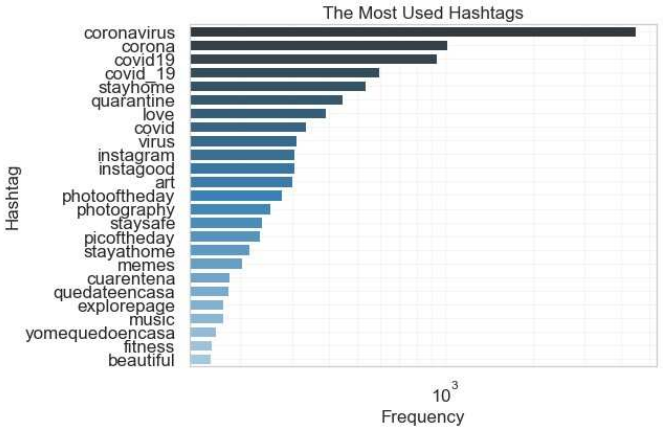

Fig. 2: Bar plot of the 25 most used hashtags.

Post Language. In order to identify the language of the post, we use spaCy library [19] and we apply it on the text of the caption. The language distribution is displayed in Table II. The dataset is dominated by English language content, making up almost 60% of posts. This is driven by our choice of an English-language hashtag seed set used for data collection. That said, we have broad coverage of other widely spoken languages too, *e.g.* Spanish (9.9%). Notice that there is no official metric to determine the post language. Therefore, we highlight that this analysis could mis-classify certain posts, such as those solely hashtags or emojis.

TABLE II: The Most Popular Languages

language	*code*	*of. #post*	*total %*
English	en	3.1K	58.3%
Spanish	es	530	9.9%
Portuguese	pt	378	7.1%
Italian	it	199	3.7%
French	fr	120	2.2%
Russian	ru	98	1.8%
Farsi	fa	96	1.8%
Arabic	ar	79	1.4%
Turkish	tr	68	1.2%
Other & non-detected	-	643	12.1%

Features. To keep data organized, the dataset is divided into four parts: (*i*) post content, (*ii*) publisher information, (*iii*) comment metrics, and (*iv*) like features (Table III). Posts contain key attributes such as a caption, list of hashtags, image/video, number of likes, number of comments, location, date, tagged list, *etc.* A post is published by a public account (or public Instagram page) and in our dataset, it can be *individual*, *fan page*, *news agency*, *influencer*, *blogger*, *etc.* Each post receives reactions in the form of comment and like that are issued by the audience/followers. The full feature list is presented in Table III. Furthermore, Table IV presents Post and Profile characteristics in detail.

V. Potential Research Topics

We hope that the dataset can support diverse research activities. Below we list a subset of potential topics, we believe the dataset could support:

TABLE III: Summary of the Feature Set - Release v1.0

Category	*Features*
Post	*caption (text), caption_language (text), shortcode (number), thumbnail (binary image), is_video (bool), video_url (text), viwer_has_liked (bool), location (tuple), hashtag (list), tagged (list), mentioned (list), date*
Publisher Profile	*username (text), id (number), follower (number), followee (number), media_count (number), biography (text), full name (text), is_verified (bool), is_private (bool), profile_picture_url (text)*
Like	*user_id (number), username (text)*
Comment	*text, date, username (text), user_id (number)*

TABLE IV: Post & Profile Characteristics

Post		***Profile***	
name	***value***	***name***	***value***
avg. caption len	388	*avg. follower*	2.6K
avg. received like	106	*avg. followee*	925
avg. received comment	7	*avg. mediacount*	385
is video (%)	0.2%	*avg. biography len (char)*	94
avg. hashtag	16	*unverified (%)*	99%
avg. mention account	0.6	*unique profiles*	2.5K
avg. tagged account	1		
has location (%)	1%		

1) *Fake news, misinformation and rumors spreading:*. Several researcher have started to inspect COVID-19 misinformation. As an example, an infodemic observatory have analyzed more than 100M public messages to understand the digital response in online social media to COVID-19 outbreak.https://covid19obs.fbk.eu In another study, Sharma et al. [5] made a public dashboard available summarising data from real-time tweets in in https://usc-melady.github.io/COVID-19-Tweet-Analysis/ with a focus to misinformation spread analysis. We believe that our Instagram data could be used to evaluate the flow of misinformation (*e.g.* memes) on Instagram.
2) *Bot Population and bot generated content:* It is well known that bot content plays a prominent role in social media data [20]. These have the capacity of amplify misinformation or even act against public health policies (*e.g.* encourgaging a breakdown in social distancing). We posit that the data could be used to explore the role of bots in this dissemination.
3) *Behavioral change analysis during the pandemic:* The social distancing measures are created an unprecedented change to millions of people's lives. Understanding the behavioral consequences of this is vital for understanding things like adherence to social distancing policies and mental health consequences.
4) *Information sharing related Covid-19:* Information flow is vital during periods of emergency. We posit that the dataset can be used to understand the flow of information, as well as people's reactions to such information.

VI. Dataset Access

The presented dataset is accessible in this address on Github platform: https://github.com/kooshazarei/COVID-19-InstaPostIDs.
This is the first version of the dataset and we are still collecting data. Hence, we hope to make further versions available in the coming weeks and months. We publish this dataset in agreement with Instagram's Terms & Conditions [21], and as it is not possible to release the post content and reactions, we just distribute the post ID's. These are known as the *shortcodes*. Researchers can then simply retrieve post content through IDs by the help of some open-source projects such as Instaloader [22] that have been developed for such purposes. For any further question, please contact Koosha Zarei at *koosha.zarei@telecom-sudparis.eu*.

References

[1] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, and Jon Crowcrroft. Leveraging data science to combat covid-19: A comprehensive review. http://www.eecs.qmul.ac.uk/~tysong/files/COVID_19_Review_v1.pdf, 2020.

[2] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset, 2020.

[3] Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. Large arabic twitter dataset on covid-19, 2020.

[4] Christian E. Lopez, Malolan Vasu, and Caleb Gallemore. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset, 2020.

[5] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, Aastha Dua, and Yan Liu. Coronavirus on social media: Analyzing misinformation in Twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020.

[6] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. " go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19. *arXiv preprint arXiv:2004.04046*, 2020.

[7] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, 2018.

[8] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of IMC*, 2017.

[9] Salma Kazemi Rashed, Johan Frid, and Sonja Aits. English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19. *arXiv*, pages arXiv–2003, 2020.

[10] C. Jacobs. Coronada: Tweets about covid-19. https://github.com/BayesForDays/coronada, 2020.

[11] Smith, "coronavirus (covid19) tweets", mar 2020. [online]. available: www.kaggle.com/smid80/coronavirus-covid19-tweets.

[12] Cassandra Jacobs. Coronada, 2020.

[13] Josimar E Chire Saire and Roberto C Navarro. What is the people posting about symptoms related to Coronavirus in Bogota, Colombia? *arXiv preprint arXiv:2003.11159*, 2020.

[14] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv preprint arXiv:2003.13907*, 2020.

[15] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *arXiv preprint arXiv:2003.05004*, 2020.

[16] Instagram. Official api graph instagram. https://developers.facebook.com/docs/instagram-api, January 2020.
[17] Instagram. Instagram hashtag search. https://developers.facebook.com/docs/instagram-api/guides/hashtag-search, February 2020.
[18] Koosha Zarei, Reza Farahbakhsh, and Noel Crespi. Deep dive on politician impersonating accounts in social media. In *2019 IEEE Symposium on Computers and Communications (ISCC) (IEEE ISCC 2019)*, Barcelona, Spain, June 2019.
[19] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
[20] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, and Jon Crowcroft. A large-scale behavioural analysis of bots and humans on twitter. *ACM Transactions on the Web (TWEB)*, 13(1):1–23, 2019.
[21] Data Policy. Instagram data policy. https://help.instagram.com/519522125107875, March 2020.
[22] Instaloader. Instaloader. https://github.com/instaloader/instaloader, January 2020.