Digital Narratives of COVID-19: A Twitter Dataset for Text Analysis in Spanish.

Allés-Torrent, Susanna, del Rio Riande, Gimena, Bonnell, Jerry, Song, Dieyun y Hernández, Nidia.

Cita: Allés-Torrent, Susanna, del Rio Riande, Gimena, Bonnell, Jerry, Song, Dieyun y Hernández, Nidia (2021). Digital Narratives of COVID-19: A Twitter Dataset for Text Analysis in Spanish. , 7.

Dirección estable: https://www.aacademica.org/gimena.delrio.riande/184

ARK: https://n2t.net/ark:/13683/pdea/38D

Esta obra está bajo una licencia de Creative Commons. Para ver una copia de esta licencia, visite https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es.

Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de acceso abierto. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org.
**ABSTRACT**

*Digital Narratives of COVID-19* (DH Covid) offers a curated Twitter corpus of digital conversations about the Coronavirus pandemic. The dataset is collected through a script via Twitter’s Application Programming Interface (API) starting on April 24th, 2020, and stored on GitHub as an open access repository of tweet identifiers that can be consulted, downloaded, and reused by scholars interested in Natural Language Processing (NLP), topic modelling, and other quantitative methods. A stable version of the dataset has also been released through Zenodo. Twitter datasets are structured in three main collections: tweets in Spanish worldwide; geolocated tweets in six Spanish-speaking areas spanning North and Central America (Mexico, Columbia, Ecuador), South America (Argentina, Peru), and Europe (Spain); and geolocated tweets in English and Spanish from the greater Miami area in South Florida.
(1) OVERVIEW
REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.3824950

CONTEXT

Although Twitter could be understood as a massive corpus of texts in which many distant reading methodologies are deployed, interest in it as a resource for Digital Humanities (DH) projects has not been widespread. Its most popular use to date has been related to the movement “Twitter for scholarly networking”, in which digital humanists analysed how the DH community grew and established their particular networks (Grandjean, 2016; Williams, Terras, & Warwick, 2013; Quan-Haase, Martin, & McCoy-Peet, 2015). However, we can highlight relevant DH studies that relate Twitter data to Natural Language Processing (NLP) (Sinclair & Rockwell, 2016; Jockers & Underwood, 2016; Gelfgren, 2016) and the DHNow initiative that has been offering open tools and resources to work with Twitter. Moreover, we note other responses closely related to DH that performed qualitative and quantitative analysis on Twitter (Chew and Eysenbach, 2010; Fu, Liang, Saroha, Tse, & Fung, 2016).

Digital Narratives of COVID-19 (DHCovid) is a digital humanities project funded by the University of Miami (FL) and developed in collaboration with the National Scientific and Technical Research Council (CONICET, Argentina) that investigates the sociolinguistic and geographical trends and topics in Twitter conversations surrounding the COVID-19 pandemic. Since April 2020, this project has been collecting COVID-19 related tweets in Spanish as well as tweets in English and Spanish in specific geographic locations: Argentina, Colombia, Ecuador, Mexico, Perú, Spain, and South Florida.

(2) METHOD

STEPS

To assemble the Twitter corpus, a PHP programming language script mines the Twitter data streaming through Twitter’s Application Programming Interface (API) and recovers a series of specific tweet identifiers (IDs). Our data mining sampling strategy consists of four main variables: language, keywords, region, and date.

The corpus is available through three repositories:

(1) GitHub: Tweet IDs are stored in a MySQL relational database where they are “hydrated,” that is, all metadata associated with the tweets is recovered, including its body text. Then, an additional script organizes the tweet IDs in the database by day, language, and region, and creates a plaintext file for each combination with a list of corresponding tweet IDs. The script generates these files daily and organizes them into folders, where each directory represents one day. These are uploaded directly to our public GitHub repository (Table 1). The data collection began on April 24th, 2020, and new tweets are automatically uploaded daily, until May 2021.

(2) Project website endpoint. A free access endpoint for query and download of “hydrated” tweets can be accessed from DHCovid website. An additional script queries the database and recovers body text of tweets (see Quality Control section). The access to a tidied and structured Twitter corpus for on-demand querying is one of the most meaningful contributions of our project for data reuse and text mining activities.

(3) Zenodo. A first stable version of the dataset, published on May 13, 2020, was released through Zenodo as a compressed ZIP file containing folders of daily tweets made between April 24th, 2020 and May 12th, 2020. A second and final version will be uploaded by the end of the project in May 2021 with the complete collection of tweet IDs.

1 https://digitalhumanities.berkeley.edu/twitter-scholarly-networking.
2 Documenting the Now, https://www.docnow.io/.
3 https://covid.dh.miami.edu/.
4 Available at: https://github.com/dh-miami/narratives_covid19/tree/master/twitter-corpus.
5 Available at: https://covid.dh.miami.edu/get/.
SAMPLING STRATEGY

The recovery of tweets by language and keywords is straightforward: we only query tweets written in Spanish and English that contain one of the words from our user-defined keywords list. We delimited two lists of keywords in English and Spanish related to the COVID-19 pandemic. Consequently, only tweets with one of these words and/or hashtags are selected. The English keywords only apply to Miami area and include terms such as “covid,” “coronavirus,” “pandemic,” “quarantine,” “stayathome,” “outbreak,” “lockdown,” and “socialdistancing.” The Spanish keywords include “covid,” “coronavirus,” “pandemia,” “cuarentena,” “confinamiento,” “quedateencasa,” “desescalada,” and “distanciamientosocial (Allés-Torrent, 2020).

Our strategy is also shaped by Twitter API policies. First, in its free version, the API did not offer the possibility for querying tweets older than seven days. Second, Twitter allows users to publish georeferenced tweets (with exact location) but retrieving geotagged tweets is complicated due to the absence of a facility for querying by geographic region. A pragmatic approach led us to define “country” as a circle surrounding the area of interest, e.g., “Mexico” is defined as latitude 21.295658, longitude -100.291341, and a radius of 450 miles. Indeed, political and national borders will not always follow our selection criteria, so our area-specific corpus can sometimes contain tweets from a neighbouring country, e.g., a query for Argentina is conflated with parts of Uruguay, and Colombia with parts of Ecuador.

QUALITY CONTROL

The “hydration” of the collected tweet IDs undergoes an additional data tidying process before any body text data is returned to the user. We apply a set of rules to the tweet body text: enforce that all words are in lowercase, remove accents, punctuations, mention of users (@users) to protect privacy, and replace all links with a general “URL” term. While enforcing all text to be in UTF-8 encoding, a particular challenge unique to the Spanish corpus is accents and graphemes, such as the “ñ”, that can be difficult to process and preserve. Most of those cases were resolved through a script by detecting special entity codes and replacing them with the correct character (e.g. &ntilde as ñ). We have also transliterated emojis into its corresponding UTF-8 charset and eliminated them from our experiments as of now. This processing facilitates the application of NLP techniques.

(3) DATASET DESCRIPTION

OBJECT NAME

- http://doi.org/10.5281/zenodo.3824950
- https://github.com/dh-miami/narratives_covid19/tree/master/twitter-corpus
- https://covid.dh.miami.edu/get/

FORMAT NAMES AND VERSIONS

.txt

Table 1 Organization of plain text datasets in the GitHub repository.
CREATION DATES
2020-04-24 to 2021-05-31

DATASET CREATORS
Susanna Allés-Torrent: Conceptualization; Funding Acquisition; Project administration; Supervision; Writing
Gimena del Rio Riande: Conceptualization; Writing
Jerry Bonnell: Data curation; Software; Visualization
Dieyun Song: Data curation; Writing
Nidia Hernández: Data curation; Visualization; Writing

LANGUAGE
Spanish, English (See Table 2).

LICENSE
Creative Commons license Attribution 4.0 International (CC BY 4.0).

REPOSITORY NAME
GitHub for the continuously daily updated version. Zenodo for the stable DOI.

PUBLICATION DATE
First published to the GitHub repository on 2020-04-24. Afterwards it was released to Zenodo on 2020-05-13.

STATISTICS AND CONTENTS
In Table 2 and Figure 1 we offer the basic statistics of the dataset.

(4) REUSE POTENTIAL
The COVID-19 pandemic has motivated a plethora of ambitious digital research focusing on Twitter, including the role and importance of automated Twitter accounts, also known as bots (Ferrara, 2020), the increase of politically radical discourse in social media (Jiang, Chen, Yan, Lerman & Ferrara, 2020), and the general public perception (Abdo, Alghonaim & Essam, 2020).

| DATE    | FLSPA | FLENG | ECUADOR | PERU | COLOMBIA | SPAIN | ARGENTINA | MEXICO | SPANISH | TOTAL  |
|---------|-------|-------|---------|------|----------|-------|-----------|--------|---------|--------|
| 2020 Apr | 1.8k  | 5.7k  | 12.2k   | 12.8k| 39k      | 47.3k | 16k       | 93.7k  | 512.2k  | 740.7k |
| 2020 May | 6.1k  | 22k   | 48.6k   | 56.2k| 168.1k   | 182.7k| 68.5k     | 411.1k | 2.4M    | 3.4M   |
| 2020 Jun | 4.7k  | 18.9k | 34.7k   | 43.2k| 149.6k   | 124.7k| 76.6k     | 319.2k | 2.1M    | 2.9M   |
| 2020 Jul | 6.5k  | 28.6k | 34.7k   | 41.4k| 171.2k   | 127.3k| 79.8k     | 324.5k | 2.1M    | 2.9M   |
| 2020 Aug | 4.9k  | 16.4k | 22.5k   | 32k  | 114.7k   | 116.9k| 76.3k     | 225.5k | 1.5M    | 2.1M   |
| 2020 Sep | 4.3k  | 15.6k | 18.9k   | 25.5k| 86.8k    | 137.4k| 86.3k     | 184.5k | 1.5M    | 2.1M   |
| 2020 Oct | 5.5k  | 23.1k | 17.6k   | 21.8k| 85.2k    | 145.6k| 79.3k     | 205.3k | 1.5M    | 2.1M   |
| 2020 Nov | 4.6k  | 18.8k | 18.7k   | 18k  | 74.4k    | 134.8k| 66.9k     | 188.7k | 1.3M    | 1.7M   |
| 2020 Dec | 5.2k  | 21.1k | 20.2k   | 25.7k| 100.9k   | 116.9k| 72.6k     | 248.9k | 1.6M    | 2.2M   |
| 2021 Jan | 5.4k  | 16.4k | 27k     | 42k  | 125.9k   | 155.5k| 78k       | 304.4k | 1.9M    | 2.7M   |
| 2021 Feb | 3.9k  | 13k   | 20.7k   | 39.5k| 80k      | 112.3k| 50.3k     | 207.1k | 1.3M    | 1.9M   |
| 2021 Mar | 4k    | 14.6k | 21.6k   | 30.8k| 74k      | 90.3k | 67.2k     | 170.5k | 1.3M    | 1.8M   |
| Total   | 56.9k | 214.3k| 297.3k  | 388.9k| 1.3M     | 1.5M  | 817.7k    | 2.9M   | 19.2M   | 26.6M  |
Most of the efforts, however, mine and analyze tweets written in English (Banda et al., 2021; Kerchner & Wrubel, 2020; Lamsal, 2020). DHcovid bridges the gap by bringing Spanish Twitter narratives into the conversation.

We use a CC BY license to give full reuse of our dataset to the public as text or to apply combined technical and humanistic processing techniques, such as statistics, NLP, topic modelling, and sentiment analysis. Our datasets can be meaningful for teaching, research, and activism resources to be used across language, disciplinary, and professional boundaries. Scholars interested in the human experience of the pandemic, health professionals, policy makers, funding agencies, journalists, and active citizens are all welcome to engage with our project.

ACKNOWLEDGEMENTS
The authors would like to thank all collaborators of the project DHcovid: Mitsunori Ogihara for participating in all meetings and advising, Romina De León and Marisol Fila for contributing with data analysis and blog posts.

FUNDING INFORMATION
DHcovid was funded by the University of Miami under the program “Direct and indirect effects of COVID-19” (2020).

COMPETING INTERESTS
The authors have no competing interests to declare.

AUTHOR CONTRIBUTION
Allés-Torrent: Conceptualization, Methodology; Writing – original draft; Writing – review & editing
Allés-Torrent et al.
Journal of Open Humanities Data
DOI: 10.5334/johd.28

AUTHOR AFFILIATIONS

Susanna Allés-Torrent
Department of Modern Languages and Literatures, University of Miami, Miami, USA

Gimena del Rio Riande
Instituto de Investigaciones Bibliográficas y Crítica Textual, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, AR

Jerry Bonnell
Department of Computer Science, University of Miami, Miami, USA; Department of History, University of Miami, USA

Dieyun Song
Department of Computer Science, University of Miami, Miami, USA

Nidia Hernández
Centro Argentino de Información Científica y Tecnológica, Consejo Nacional de Investigaciones Científicas y Técnicas, Buenos Aires, AR

PUBLISHER’S NOTE

This paper underwent peer review using the Cross-Publisher COVID-19 Rapid Review Initiative.

REFERENCES

Abdo, M. S., Alghonaim, A. S., & Essam, B. A. (2020). Public perception of COVID-19’s global health crisis on Twitter until 14 weeks after the outbreak. Digital Scholarship in the Humanities, fqaa037. DOI: https://doi.org/10.1093/llc/fqaa037

Allés-Torrent, S. (2020). A Twitter Dataset for Digital Narratives. Digital Narratives of COVID-19. Published May 23, 2020. Retrieved April 05, 2021 from https://covid.dh.miami.edu/2020/05/23/twitter-dataset-for-digital-narratives/

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. (2021). A large-scale COVID-19 Twitter chatter dataset for open scientific research—An international collaboration [Data set]. Zenodo. DOI: https://doi.org/10.5281/zenodo.4460047

Chew, C., & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS One, 5(11). DOI: https://doi.org/10.1371/journal.pone.0014118

Documenting the now. (n.d.). Retrieved April 05, 2021, from https://www.docnow.io/

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots? First Monday, 25(6). DOI: https://doi.org/10.5210/fm.v25i6.10833

Fu, K. W., Liang, H., Saroha, N., Tse, Z. T. H., Ip, P., & Fung, I. C. H. (2016). How people react to Zika virus outbreaks on Twitter? A computational content analysis. American Journal of Infection Control, 44(12), 1700–1702. DOI: https://doi.org/10.1016/j.ajic.2016.04.253

Gelfgren, S. (2016). Reading Twitter: Combining Qualitative and Quantitative Methods in the Interpretation of Twitter Material. In G. Griffin (Ed.), Research Methods for Reading Digital Data in the Digital Humanities (pp. 93–100). Edinburgh: Edinburgh University Press.

Grandjean, M. (2016). A Social Network Analysis of Twitter: Mapping the Digital Humanities Community. Cogent Arts & Humanities, 3(1). DOI: https://doi.org/10.1080/23311983.2016.1171458

Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political Polarization Drives Online Conversations about COVID-19 in the United States. Human Behavior and Emerging Technologies, 2(3), 200–211. DOI: https://doi.org/10.1002/hbe2.202

Jockers, M., & Underwood, T. (2016). Text-Mining the Humanities. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), A New Companion to Digital Humanities (pp. 291–306). New York: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118680605.ch20

Kerchner, D., & Wrubel, L. (2020). Coronavirus Tweet Ids [Data set]. Harvard Dataverse. DOI: https://doi.org/10.7910/DVN/LW0B7B
Lamsal, R. (2020). Coronavirus (COVID-19) Tweets Dataset [Data set]. IEEE. DOI: https://doi.org/10.21227/sj81-wf42

Quan-Haase, A., Martin, K., & McCoy-pee, L. (2015). Networks of Digital Humanities Scholars: The Informational and Social Uses and Gratifications of Twitter. Big Data & Society, 2(1). DOI: https://doi.org/10.1177/2053951715589417

Sinclair, S., & Rockwell, G. (2016). Text Analysis and Visualization: Making Meaning Count. In S. Schreiberman, R. Siemens & J. Unsworth (Eds.), A New Companion to Digital Humanities (pp. 274–290). New York: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118680605.ch19

Williams, S. A., Terras, M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. Journal of Documentation, 69(3). DOI: https://doi.org/10.1108/JD-03-2012-0027

TO CITE THIS ARTICLE:
Allès-Torrent, S., del Riade, G., Bonnell, J., Song, D., & Hernández, N. (2021). Digital Narratives of COVID-19: A Twitter Dataset for Text Analysis in Spanish. Journal of Open Humanities Data, 7: X, pp. 1–7. DOI: https://doi.org/10.5334/johd.28

Published: XX Month 202X

COPYRIGHT: © 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.
1. If possible, could you please provide Orcid ID for the authors?