Covid Data Analytics Repository: An interdisciplinary look into the COVID-19 pandemic in Brazil

Ramon A. S. Franco[1,2], Pedro Loures Alzamora[1], Janaína Guiginski[1], Evandro L. T. P. Cunha[1], Tereza Bernardes[1], Juan F. Galindo[3], Luana Passos[1,2], Raquel Schneider[1], Bruno Chagas[1], Kícila Ferreguetti[1], Luísa Cardoso[1], Pedro Moreira[1], Wallace Pereira[1], Ana Paula Couto da Silva[1], Wagner Meira Jr.[1]

[1] Universidade Federal de Minas Gerais, UFMG, Brazil
{evandrocunha,ana.coutosilva,meira@}dcc.ufmg.br;
{pedro.loures,janainaguiginski,tbernardesfaria,bruno.azevedo.chaga,
kicilaferreguetti,pedrovxm, luisacgs,walleceufmgbr}@gmail.com, raquelschneider03@hotmail.com
[2] Universidade Federal do Oeste da Bahia, UFOB, Brazil
{ramon.franco,lpsouz}@ufob.edu.br
[3] Universidade Estadual de Campinas, Unicamp, Brazil
jgalindoj@gmail.com

Abstract. This article describes the construction and deployment of the Covid Data Analytics Repository, a source for interdisciplinary studies about the impact of the COVID-19 pandemic in Brazil. We collected different types of data from official (IBGE, DATASUS) and non-official (Brasil.IO) sources, online social networks (Instagram, Twitter), and from a search engine analysis tool (Google Trends). We used these data to perform investigations aimed to understand the impacts of COVID-19 in the country, from economics to social behavior. At the moment of publication of this article, our repository contains 1,508 documents, classified into two main types: (i) databases and tables downloaded from the aforementioned sources; and (ii) papers, reports, maps and graphs resulting from the analyses that we performed. As a means to allow reproducibility and foster follow-up studies, we released our repository for public use.

Categories and Subject Descriptors: H.2.5 [**Database Management**]: Heterogeneous Databases; H.3.m [**Information Storage and Retrieval**]: Miscellaneous

Keywords: coronavirus, datasets, digital health, social networks

1. INTRODUCTION

Since the first months of 2020, the COVID-19 pandemic has been imposing several challenges worldwide. The lack of strong knowledge about the new virus, the challenges of a new life under (sometimes severe) restrictions and uncertainties, and the huge impact on the economy are some of the reasons that mobilize researchers across the world to work on the understanding of the impacts of this pandemic in society. Also, a consensus was created on the need for these studies to be carried out considering concepts and methods from different areas of knowledgeand extracting information from data coming from various sources.

Brazil was one of the most affected countries by the COVID-19 pandemic. In face of these challenges, several studies investigated the facets of the pandemic in this country through data. Some of them, for instance, characterized the evolution of the disease [Ranzani et al. 2021], dealing with the under-reporting of cases from official agencies [Veiga e Silva et al. 2020]. Other works, in turn,

proposed models to predict the evolution of COVID-19 by using data from the first months of the pandemic employing different methods [Bastos and Cajueiro 2020; Pereira et al. 2020], or even using geolocalization and population mobility data [Peixoto et al. 2020].

In this context, the Covid Data Analytics Project (CDA Project)[1] was designed and implemented by an interdisciplinary team based at Universidade Federal de Minas Gerais (UFMG), between June 2020 and March 2021. This team was composed by undergraduate and graduate students, as well as postdoctoral and senior researchers, from areas as diverse as Computer Science, Demography, Economics, History, Linguistics, Medicine, and Social Science. The main goal of the CDA Project was to better understand the impacts of the COVID-19 pandemic on public health, economics, and social behavior from the lens of structured and unstructured data gathered from official and non-official sources, as well as from the web. This article, which builds upon our prior work [Moreira et al. 2021], presents the methodology for constructing the Covid Data Analytics Repository (CDA Repository), created during the advancement of the CDA Project. The repository is publicly available at the Zenodo platform[2]. Moreover, we offer here a deeper analysis of this repository, since we also present the outlines of the following set of investigations on how COVID-19 affected the lives of Brazilian citizens: (i) a characterization of the population's perception on the impact of COVID-19 in the job market based on Instagram posts; (ii) a model to predict the variation of COVID-19 cases and deaths based on the search volume of specific query terms on Google Search; (iii) a spatiotemporal analysis of the evolution of COVID-19 in the country; and (iv) a lexical analysis of Twitter posts and the correlation of certain words with epidemiological indicators. All these studies were carried out using data contained in the CDA Repository.

The remainder of this article is organized as follows. Section 2 summarizes prior related work, while Section 3 describes our methodology to build the CDA Repository. An overview of the data included in the repository is presented in Section 4, and a set of analyses that used the data available in the repository is displayed in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

A large amount of research on COVID-19 was published in 2020 and 2021, and, consequently, several repositories containing pandemic-related data were released. The 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository [Dong et al. 2020], compiled by researchers from the Johns Hopkins University, might be considered one of the widest openly available data repositories[3] regarding the pandemic [Miller 2020]. This repository uses, as data sources, publications from institutions such as the World Health Organization (WHO) and the European Centre for Disease Prevention and Control (ECDC), among several others. It also includes Brazilian data, which come from the Ministry of Health and from Universidade Federal de Viçosa.

The WHO manages a wide range of global data repositories related to healthcare and well being, according to the demands of its member states. Since the beginning of the pandemic, the WHO updates its Coronavirus (COVID-19) Dashboard[4] daily, including information concerning confirmed cases, deaths, and vaccines. This is an open-access dataset, which is intended to foster the Universal Health Coverage goal by helping to monitor the availability of healthcare resources. Also, a team of researchers affiliated to the Max Planck Institute for Demographic Research created the COVerAGE-DB, an open database containing data on cases and deaths from a large number of countries [Riffe et al. 2021]. The conforming the COVerAGE-DB database is standardized and harmonized, allowing the comparison of the effects of COVID-19 in different regions of the world, taking into account age

[1]CDA Project website: `https://covid.dcc.ufmg.br`
[2]URL of the CDA Repository: `https://zenodo.org/record/5176798`
[3]The 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository is available at: `https://github.com/CSSEGISandData/COVID-19`
[4]The WHO Coronavirus (COVID-19) Dashboard is available at: `https://covid19.who.int/`

structures and gender distributions. Besides, the authors provide a tutorial containing instructions regarding data extraction[5].

In Brazil, a research from the Universidade Federal de Viçosa published a dataset containing the reported number of COVID-19 cases and deaths at the municipal and state levels[6]. The author uses official (from the Ministry of Health) and non-official (from sources such as Brasil.IO) data [Cota 2020].

It is also possible to find datasets containing social networks data regarding the pandemic. Researchers from the Institut Polytechnique de Paris and from the Queen Mary University of London published (what they claim to be) the first Instagram dataset on COVID-19 [Zarei et al. 2020]. This dataset is multilingual and is publicly available[7]. It contains 18.5 thousand comments and 329 thousand likes from 5.3 thousand posts (posted by 2.5 thousand users) published between 05 January and 30 March 2020. Also, data from several online social networking systems (Twitter, Instagram, YouTube, Reddit, and Gab) on COVID-19 were analyzed to provide an assessment on the evolution of discourse on a global scale for each of these platforms [Cinelli et al. 2020]. The resulting dataset (that, according to the authors, is available upon request) is composed of 1.3 million posts and 7.5 million comments produced by 3.7 million users.

A public dataset[8] with data from Twitter about the coronavirus was collected since 28 January 2020, using the Twitter streaming API and Tweepy to follow specific keywords and accounts [Chen et al. 2020]: 72 million tweets were collected until 21 March 2020, constituting 600 GB of data. Finally, researchers from the Universidade Federal do Ceará and the Universidade Estadual do Ceará compiled a dataset of WhatsApp messages[9] related to the COVID-19 pandemic in Brazilian Portuguese and labeled messages promoting misinformation. Over 228 thousand messages were collected between April and June 2020, and 85% of them were considered fake [Martins et al. 2021].

In this section, we presented a small part of the data been made available, so far, related to the COVID-19 pandemic. Yet, despite the relevance of these works, the number of open datasets containing COVID-19 related data, gathered from Brazilian online social networking systems, and also from other sources, is still limited. To address the lack of large repositories with these characteristics, we decided to publish the dataset used during the development of the CDA Project.

3. DATA REPOSITORY CONSTRUCTION

The goal of the CDA Repository is to gather multi-source heterogeneous data in order to help researchers from different fields to comprehend how the COVID-19 swept through Brazil. Figure 1 depicts the repository construction methodology, which is composed of three main steps: (1) multi-source data identification; (2) data collection; and (3) data integration and release via a web portal.

3.1 Step 1: multi-source data identification

We first sought to identify the data sources to support the analyses that we were interested in. In order to guide such effort, we raised a set of initial interdisciplinary research questions, such as: *What are the spatiotemporal patterns of the spread of COVID-19 in Brazil? How does the debate about COVID-19 develop in online social media, for instance regarding the population's perception of the impact of the pandemic on the job market? How can telehealth technologies help facing the COVID-19 pandemic? What are the characteristics of the political and ideological use of medical-scientific*

[5]The tutorial for getting started with COVerAGE-DB datasets is available at: `https://timriffe.github.io/covid_age/GettingStarted.html`
[6]The website of the project is: `https://covid19br.wcota.me/en/`
[7]The dataset can be downloaded at: `https://github.com/kooshazarei/COVID-19-InstaPostIDs`
[8]Available at: `https://github.com/echen102/COVID-19-TweetIDs`
[9]Available at: `https://zenodo.org/record/5193932`

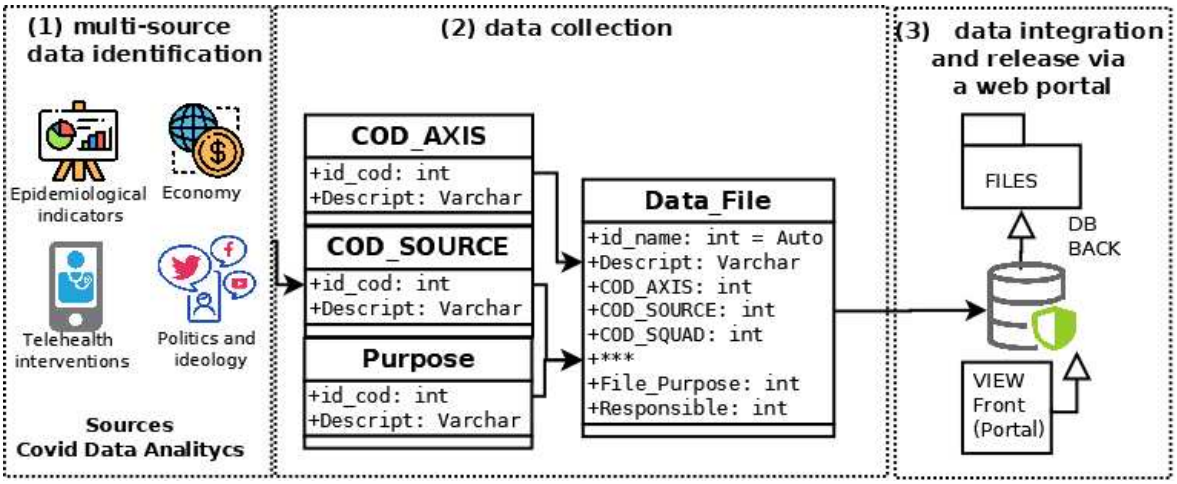

Fig. 1. Construction methodology of the Covid Data Analytics Repository.

information posted in social networks? Is it possible to predict patterns of variation in the numbers of cases and deaths based on data from the web? We then selected the following data sources in order to help us answer these and other questions: structured data from official sources (such as IBGE and DATASUS) and non-official sources (such as Brasil.IO), data from online social networking systems (Instagram, Twitter, YouTube), and data from a search engine analysis tool (Google Trends).

3.2 Step 2: data collection

In the next step, we collected data from the previously selected sources. Due to the heterogeneous nature of the collected data, we adopted different approaches for collection and categorization. At the moment of the publication of this article, the CDA Repository contained:

—data obtained from IBGE[10] (Brazilian Institute of Geography and Statistics), including socioeconomic data from PNAD[11] (National Household Sample Survey);

—health-related data from DATASUS[12] (Department of Informatics of the Brazilian Unified Health System (SUS));

—weekly consolidated data containing epidemiological indicators, including the total number of COVID-19 cases and deaths in each Brazilian municipality, obtained from Brasil.IO[13], a project that aims to make Brazilian data of public interest more open and available;

—anonymized Instagram posts related to the job market during the pandemic;

—weekly total number of Twitter posts that include specific keywords related to COVID-19;

—information regarding the search volume of selected query terms related to the pandemic in Brazil, obtained from Google Trends[14], a search engine analysis tool that provides information on the popularity of search queries in Google Search across regions.

Additional details on the specific data available at the CDA Repository are presented in Section 4.1.

[10]Official IBGE website: `https://www.ibge.gov.br/`
[11]PNAD website: `shorturl.at/bdpBG`
[12]Official DATASUS website: `https://datasus.saude.gov.br/`
[13]Brasil.IO project website: `https://brasil.io/home/`
[14]Google Trends website: `https://trends.google.com/`

3.3 Step 3: data integration and release via web portal

Afterwards, in the last step, we cleaned, transformed and combined our data from different sources into a single, unified access platform, available through the CDA Project web portal[15]. This step included the following stages:

—*Data cleaning.* A cleaning process for all types of data collected from web sources was implemented for: missing data pre-filling, noise reduction, identification and removal of unneeded values, and resolution of inconsistencies. The pandas library for Python was used. The main challenge in data cleaning was the identification and filtering of incomplete or unwanted files. This situation arose due to the sending of information with incomplete data. Incomplete data was sent in most cases by researchers who were not familiar with this type of process. This was solved by constantly verifying the data and requesting its resubmission with the assistance of one of our specialized collaborators.

—*Data enrichment.* In this stage, the collected data files were enriched with the following metadata: name, creation date, description, last update, geolocation, status ("finished", "not finished", "under construction"), research topic[16], and data source. For structured data, we further added: technique of collection, date and time of collection, file purpose, and name of the person responsible for providing the file. The metadata description follows the Open Standards for Data released by the Open Data Institute[17].

—*Database implementation.* At first, the database was created on a local machine, by using MySQL Workbench and Python for file and metadata insertion. Then, the integration of this initial database with the CDA Project database was performed. To integrate our data collections, we used the free software phpMyAdmin. Data from different sources were integrated through their metadata.

—*Search interface implementation.* To provide a user-friendly interface for data access, we designed the search interface displayed in Figure 2.

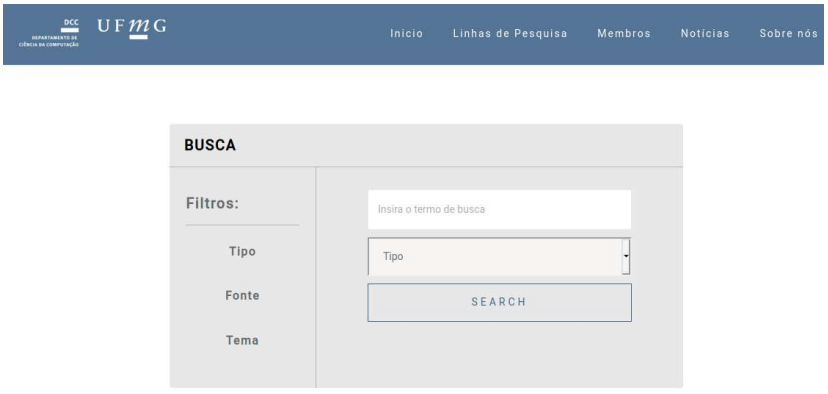

Fig. 2. Search interface at the CDA Project website. Translation of the filters (*filtros*): type (*tipo*), source (*fonte*), and topic (*tema*).

This interface retrieves files and metadata through search term oriented queries and/or filters (file type, source, and research topic). By using this search interface, different types of queries can be performed. For instance, one can type in the search field the term "number of cases" and select a

[15]The unified access platform is available at: `https://covid.dcc.ufmg.br/buscador.php`.
[16]The CDA Project was divided into four main research topics: (i) "Analysis of the behavior of the Brazilian economy during the pandemic"; (ii) "Telehealth intervention strategies in the COVID-19 pandemic"; (iii) "Epidemiological indicators and web behavior"; and (iv) "Politics, ideology and medical information in online social networks".
[17]More information on the Open Standards for Data is available at: `https://standards.theodi.org/`

filter, which identifies the target file type (map, image, text etc.). One can also enable searches using only the filter "type", resulting in a set of documents (data files and analyses) from a particular data source (Instagram, Google Trends, IBGE etc.). Figure 3 shows the result for the query "Covid-19", displaying a list of the most relevant files and their URLs for data downloading.

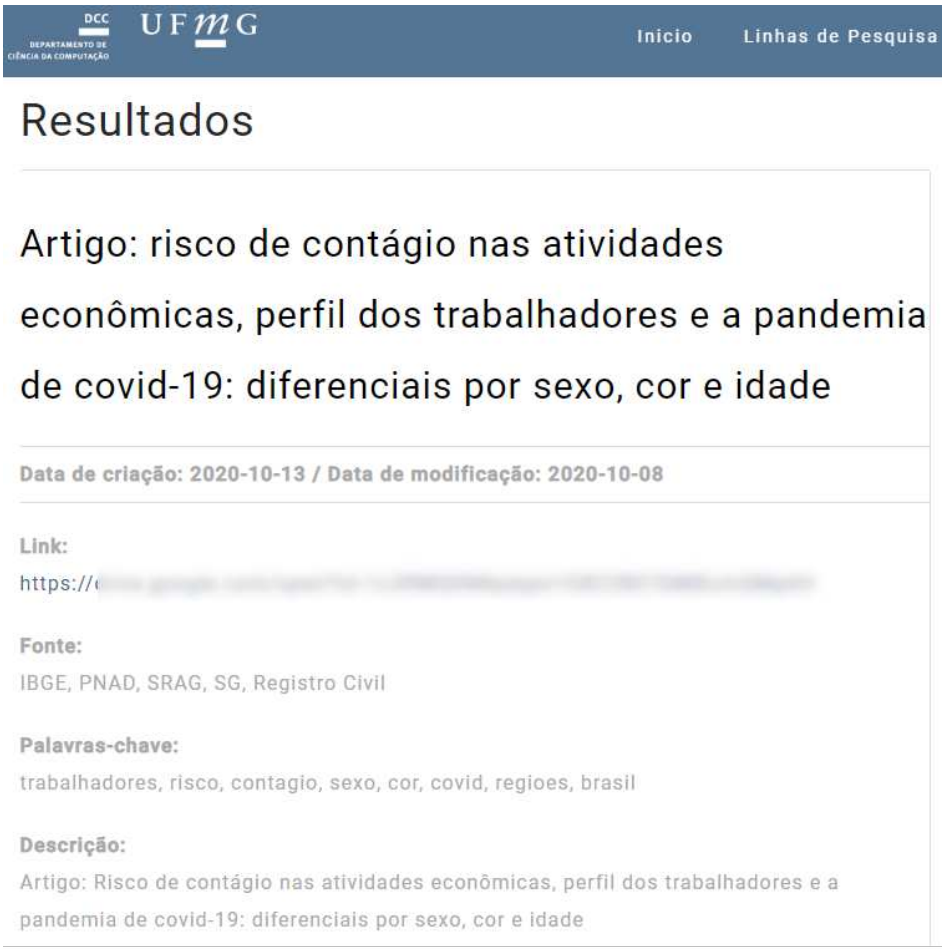

Fig. 3. Search results for the term "Covid-19".

4. COVID DATA ANALYTICS REPOSITORY OVERVIEW

In this section, we present an overview of the content at the publicly available CDA Repository at the moment of the publication of this article, which encompasses heterogeneous data generated from 01 January 2020 to 20 May 2021. In total, our collection includes 1,508 files, split into two main categories: (i) structured and unstructured data from the sources that we selected in the first step of our methodology (see Section 3.1); and (ii) papers, reports, maps, images and charts, created as outcomes of the analyses performed.

4.1 Structured and unstructured data

These files are categorized as follows:

—71 quarterly time series containing economic indicators in Brazilian states and in the country as a whole, in .csv format, with approximately 18,400 records;

—8 .csv files, one file for each hashtag we analysed, containing 91.680 Instagram posts, with anonymized user data. These posts included one or more hashtags from the following list: *#demissão* ("dismissal"), *#demitida* ("fired", fem.), *#demitido* ("fired", masc.), *#desempregada* ("unemployed", fem.), *#desempregado* ("unemployed", masc.), *#desemprego* ("unemployment"), *#falido* ("broke"), and *#reduçãodejornada* ("working hours reduction"). Posts published between 01 January 2020 and 13 September 2020 were collected. To scrape these data, we used a custom web crawler that relies on the Instaloader library for Python.[18]

—7 data processing scripts in Python programming language;

—4 .csv files containing the weekly total number of tweets and retweets that include specific keywords related to COVID-19 in the Brazilian scenario: *corona*, *covid*, *coronavirus*, *covid19*, *quarentena* ("quarantine"), *hidroxicloroquina* ("hydroxychloroquine"), *cloroquina* ("chloroquine"), *confinamento* ("lockdown"), *distanciamento social* ("social distancing"), *aglomeração* ("crowded area"), *aglomerações* ("crowded areas"), and *sars*. These files contain the daily volume of tweets and retweets, as well as the epidemiological indicators we calculated in our analyses (see Section 5.4);

—3 .csv files obtained from the Google Trends tool, containing information regarding the search volume of 124 selected query terms related to the pandemic in Brazil. We split the data collected into different files with the weekly total number of searchers performed in Brazil, in Minas Gerais State and the third one in which we categorize the selected query terms in major topics such as: vaccine, symptoms, prevention and treatments and medication.

4.2 Reports and results

The files containing analyses and reports represent 92% of all repository files. In addition to text documents, visual documentation, such as maps of the weekly lethality of the COVID-19 pandemic in Brazil and graphs of the indicators of pandemic progress in social networks under economic, health and political views, were also made available in different formats. The files are distributed as follows:

—23 comparative graphs of quarterly indicators with social and economic indicators, for example, indicators of the unemployment rate, the fiscal situation, layoffs, and percentages of hashtags related to the pandemic. Additionally, graphs of descriptive analysis of labor activities (essential and non-essential) divided by Federal, State, and regional levels have been added to the repository in (.svg format).

—409 charts of the total number of COVID-19 cases and deaths evolution in Brazil (.png format).

—365 charts of Instagram analytics, methodologies and indicators, in .png format;

—568 maps, line and bar charts with the accumulated COVID-19 cases and deaths in Brazil, between the 9^{th} and 32^{nd} epidemiological weeks of 2020 (.png format).

—15 .pdf files with provisional measures issued by the Brazilian Federal Government in 2020.

—2 interactive chart with the evolution of the COVID-19 mortality indicator in Brazil in 2020 (.html format);

—25 animated histogram graphs showing the evolution of lethality (accumulated deaths / accumulated cases) in percentage in all Brazilian Federate Units, each epidemiological week, from week 9 to week 31, in GIF format;

—8 reports analyzing the information available for collection in the Google Trends tool and of the information made available by the research topics, in .pdf format.

[18] Available at: `https://instaloader.github.io`

4.3 Data limitations

Our data has some limitations. First, the repository is a compilation of heterogeneous data used at different research topics of the Covid Data Analytics project. Our main contribution is to collect (using APIs or downloading data directly from government sites) and clean data from different sources (web and online social networks). Thus, this work shares a rich repository, with a set of heterogeneous data, which one with its own structure. Second, due to data privacy issues, data from online social networks were not fully made available in the repository. We only released summarized information, such as tweets and retweets with anonymized data and the set of analyses performed to answer the research questions mentioned in Section 3.

5. ANALYSES

To illustrate the potential of our data repository, we now turn our attention to the description of some analyses performed at the CDA Project. Section 5.1 delves deep into the Instagram data in order to unveil the perceptions about the COVID-19 negative impacts on the job market. Next, in Section 5.2, we evaluate possible correlations between queries on Google Search and Brazilian COVID-19 epidemiological indicators, through the use of Google Trends data. Section 5.3, in turn, focuses on a spatiotemporal analysis of the evolution of COVID-19 cases in the country, using data obtained from Brasil.IO. Finally, Section 5.4 presents a lexical analysis of the COVID-19 debate on Twitter and how the topics are correlated with epidemiological indicators in Brazil. All the data used in these studies is available at the CDA Repository. We refer the reader to the project's web portal to find more about further analyses.

5.1 Perceptions on the impact of COVID-19 on the job market using Instagram data

We characterized the Instagram posts related to COVID-19 and the job market in terms of the topics that they convey. We collected posts written in Portuguese and published between 01 January and 13 September 2020, covering the pre-pandemic period as well as its first months. Since Brazil ranks third in the total number of Instagram users (with 110 million users in July 2021)[19], we assume that the majority of posts in Portuguese are published by Brazilians, revealing what they are discussing. To ensure that the posts that we analysed were related to COVID-19, we applied a second filter (in addition to the one mentioned in Section 4.1, searching for the following strings: *coron*, *covid*, *quarent*, *home office*, *pand* and *vírus*. After this second filter, 20% of the collected posts were analysed.

To identify the most relevant topics discussed in this dataset, we used latent Dirichlet allocation (LDA) [Blei et al. 2003], a generative statistical model to automatically infer the topics in a collection of documents. We first applied LDA to all posts jointly, and then we compared the distributions of the identified topics in each group of posts, aiming at identifying differences between them. Before applying the LDA model, we cleaned the posts by removing characters and words with none or limited analytical value (stopwords), as well as hashtags. This cleaning allows for a better identification of what is being mentioned in the posts [20]. We ran the LDA algorithm using the Gensim Python library to perform topic analysis, and we established the number of topics $k = 3$. The resulting topics are presented in Table I, which shows the most representative words (according to the LDA output) for each topic.

Overall, the results suggest that Instagram users were affected both by the public health and economic crises. The most relevant words show people's concerns about unemployment and income loss. More specifically, topics 1 and 2 refer to the most important impacts of the pandemic on the job

[19] `https://www.statista.com/statistics/578364/countries-with-most-instagram-users/`

[20] Data collected from public databases was cleaned and enriched, following the procedure described at `http://covid.dcc.ufmg.br/linhas/economia/redes-sociais/`

Table I. Most representative words (translated into English) in the topics inferred by the LDA algorithm.

Topic	Most representative words (translated into English)
1	work, reduction, wage, pandemic, days, suspension, government, coronavirus, contract
2	pandemic, people, moment, crisis, companies, situation, all, health, many, job
3	you, home, make, money, people, work, day, life, earn, digital

market: the reduction of working hours, the economic crisis that affected workers and companies, as well as the government policies to fight this crisis. Topic 3 seems to be mainly characterized by words related to ways of maintaining or increasing people's income during lockdowns and social distancing.

This simple textual analysis discloses the worsening of working conditions due to the pandemic context. In the Brazilian case, these conditions reinforce the deleterious effects on workers' perspectives regarding maintenance of income and living conditions.

5.2 Correlations between web searches and COVID-19 epidemiological indicators

Previous work in literature relied on data from Google Trends as a proxy to understand the impacts of COVID-19 pandemic worldwide. For instance, authors in [Mavragani and Gkillas 2020] used Google Trends to explore the relationship between COVID-19 cases and deaths and online interest in the virus, focusing in the United States. Authors in [Brodeur et al. 2021], instead, used Google Trends data to test whether COVID-19 and the associated lockdowns implemented in Europe and America led to changes in well-being related topic search-terms. Here, similar to the work in [Mavragani and Gkillas 2020], we rely on Google Trends Data to understand the evolution of COVID-19 disease in Brazil.

A total of 124 web search terms were treated. Using the Spearman model we tried to find which keywords had the best correlations. For this work, we selected the terms that had the highest correlation with the number of COVID-19 cases and deaths. This was followed by calculating Pearson's correlation coefficient [Myers and Sirois 2004] between the frequency of the 124 terms and the values of the indicators. The calculations were performed between 0 and 5 weeks after the search for each term, the terms for which the correlation coefficient was greater than 0.7 were filtered out. After calculating the correlation, we were left with 23 terms (11 correlated with the number of cases and 12 with the number of deaths): With the 23 selected terms, we trained three different prediction models, which were: linear regression [Weisberg 2005], ARIMA [Box et al. 2015] and ARIMAX [Pankratz 2012]. Each selected model was used for both epidemiological indicators. For the linear regression model, the independent variables are the free frequency of the selected terms. The ARIMAX model considers both the time series and the frequency of the selected terms. [21]

Table II shows the models' accuracy using the RMSE (Root Mean Square Error) measure ARIMAX model diminishes the RSME by 13% and 8% w.r.t Linear Regression model for the cases and de prediction, respectively. Regarding ARIMA model, the results are even better: the RSME decreases by 22% and 40% for the cases and deaths prediction, respectively. These results show that using a model in which exogenous data is taken into account, i.e. the search frequency of the selected terms, increases the accuracy of the prediction task we are interested in.

The complete results of the research described in this section were already published [Locatelli et al. 2022]. Please refer to this article to see details and additional information regarding the methodology and findings of these analyses.

[21] Data collected from public databases was cleaned and enriched, following the procedure described at `http://covid.dcc.ufmg.br/linhas/politica-e-ideologia/`

Table II. Accuracy of the prediction models. Best accuracy values are highlighted.

Model	RMSE (Cases)	RMSE (Deaths)
Linear Regression	29.31	29.89
ARIMA	32.72	46.30
ARIMAX	**25.31**	**27.48**

5.3 Spatiotemporal analysis of COVID-19 cases in Brazil

Geoprocessing and spatial analysis are powerful tools to better understand complex, heterogeneous and self-correlated phenomena, such as the spread of viruses and diseases across time and space [Szwarcwald et al. 2000; Guimarães et al. 2020]. In this section, we introduce a work in which we performed an Exploratory Spatial Data Analysis (ESDA) by investigating the spatial patterns of COVID-19 cases across small areas composed by clusters of municipalities of Brazil. For this, we used the public available data gathered from the Brasil.IO project, which includes the daily number of COVID-19 cases and deaths in all Brazilian municipalities, from April 2020 to March 2021.[22]

To that end, we calculated, for each week, the Local Indicators of Spatial Association (LISA), which correlates a specific variable (in our analysis, the total number of COVID-19 cases/deaths) to a location, taking into account the values of the same variable in the neighboring locations [Rey S. J. 2020]. Figure 4 shows two maps of Brazil with the main clusters extracted from LISA data regarding the new cases of each immediate region through the weeks collected. The map on the left depicts the results from April to August 2020, and the map on the right shows the results from November 2020 to March 2021. We note a change on the geographical distribution of the intensity of new cases: in the early stages of the COVID-19 pandemic, new cases were strongly concentrated in the North and Northeast Regions (high-high cluster types). In the second period, we note that high-high cluster types spread through Brazil, reaching the Central-West, Southeast and South Regions. The Northeast Region and parts of the North Region (especially the state of Pará), in turn, presented only low-low cluster types.

We also investigated the virus diffusion patterns shown in the maps. We found that the regions with the highest incidence of COVID-19 cases overlap with the exportation route of soybeans and corn produced in Brazil. This finding corroborates previous results [Silva et al. 2020], which highlight the role of the transport system as a COVID-19 transmission vector and its effects on the "ruralization" of the disease.

5.4 Lexical analysis of Twitter posts and their correlation with epidemiological indicators

In recent years, online social networking systems have become a forum used by the population to debate a wide range of topics. Twitter is one of the most popular social media applications used to this end [Cunha et al. 2014; Du et al. 2017; Sultana et al. 2021; Kang et al. 2017; Marques-Toledo et al. 2017; Aiello et al. 2020; Li et al. 2020]. It is not surprising that the platform is an effervescent channel where people talks about the COVID-19 pandemic. In this last analysis presented here, we aim to investigate whether the topics discussed in the tweets follow the evolution of the Brazilian epidemiological indicators, i.e the number of COVID-19 cases and deaths.

We gathered a corpus of Portuguese-language tweets that would be informative of the online debate on COVID-19 pandemic. To that end, we used the Twitter API Search to collect tweets based on specific keywords related to COVID-19. We followed the list of keywords proposed in [Brum et al. 2020]. In total, we gathered over 97 million tweets, from March 2020 to January 2021.

[22]Data collected from public databases was cleaned and enriched, following the procedure described at `htpps://github.com/CDA-EPCWeb/Indicadores-Epidemiologicos`.

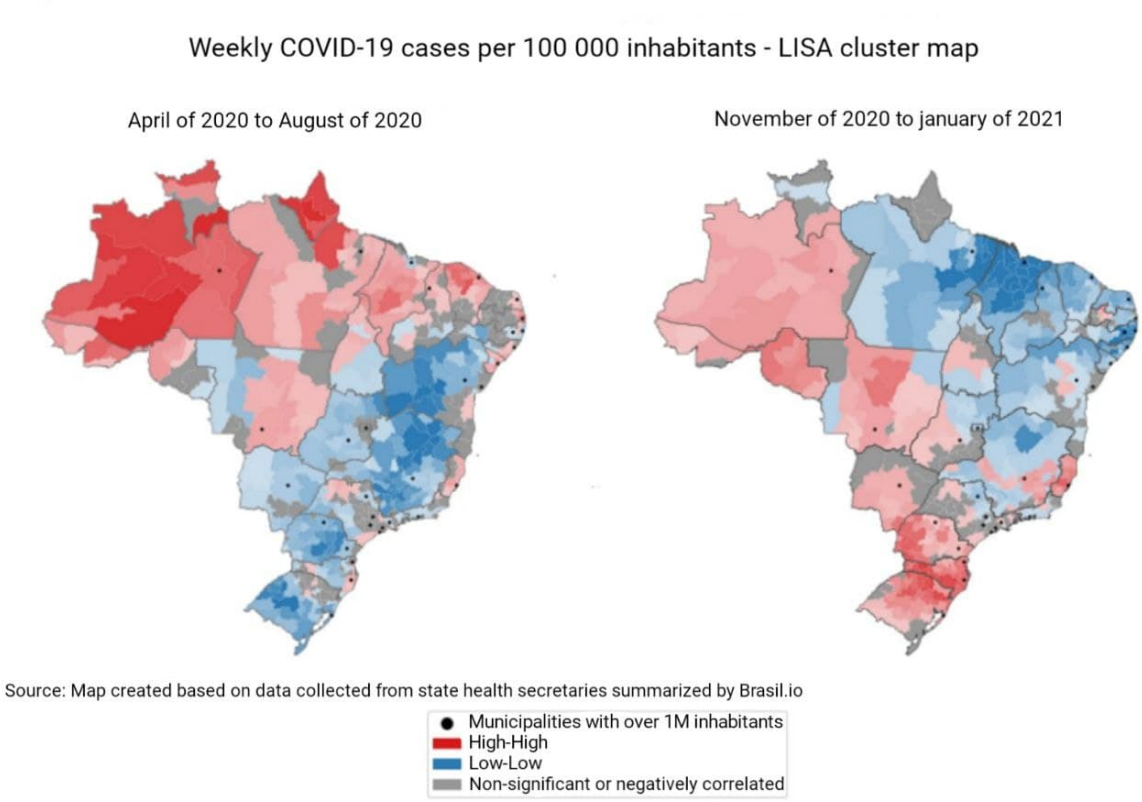

Fig. 4. Cluster maps of COVID-19 cases per 100,000 inhabitants.

In this analysis, Twitter data and two epidemiological indicators (growth factors of cases and deaths in Brazil) were used. The criteria for the conformation of Twitter keywords were defined from the work of [Brum et al. 2020]. Tables with these indicators, among others, are available in the data repository. There are also tables with counts of words present in Twitter posts related to the COVID-19 pandemic from March 2020 to January 2021 in the data repository.

From these datasets (indicators and tweets), it was possible to identify the discussion topics in Twitter more frequently related to the increase or reduction of cases and deaths. As an example, when considering discussions made two weeks before the increase or decrease on the number of deaths, the words most frequently found reveal mixed feelings: *laugh, joke, happy, tested, or peak*.

6. CONCLUDING REMARKS

In this work, we present the Covid Data Analytics Repository, a source of data for interdisciplinary studies about the impact of the COVID-19 pandemic in Brazil. At the moment of publication of this article, our data repository contains around 1,500 files, summarizing data from official (IBGE, PNADs, DATASUS) and non-official (Brasil.IO) sources, online social networks (Instagram, Twitter, YouTube), and a search engine analysis tool (Google Trends). This repository also contains a series of reports, documents and studies made publicly available by teams composed of students and researchers in fields as diverse as Computer Science, Demography, Economics, History, Linguistics, Medicine, and Social Science. As a result, we provide a large variety of interdisciplinary data and analyses concerning the COVID-19 pandemic in Brazil, from its beginning in 2020 until mid-2021.

We envisage some potential applications of our repository. Firstly, we believe that it is a valuable

source for new scientific studies. The implemented user-friendly web interface allows users to search among a set of charts and reports. Also, we facilitate the reuse of enriched data for interdisciplinary researches about the pandemic in Brazil, since the long-term availability of our datasets provides reusability for several purposes, including in visualization, query and analysis tools.

Secondly, the integrated data can be further used to support future research work that mainly approaches the overlapping of data from different sources. We believe that the data coming from the Covid Data Analytics Repository might play an important role, becoming more relevant if used by interdisciplinary collaborative groups, allowing to create a baseline for future research. Several research questions may arise from it, such as: Do the ideas shared by online social media users reflect the growth of deaths and cases? Does the number of unvaccinated people reflects the strength of anti-vaccine movements in online social networks? Does the profile of active users in the web correlate with the actual age structure? How is the correlation of web searches and COVID-19 epidemiological indicators in the country? Is it possible to estimate and quantitatively recognize the damage caused by misinformation in social media during the Covid-19 pandemic? How can the spatiotemporal analysis of COVID-19 cases in Brazil be replicated in other countries, especially in the Global South? What are the incidence rates of Covid-19 in Brazil and its relationship with the manifestations of opinion in online social networks?

Due to the complexity and gravity of the consequences of the COVID-19 pandemic in the entire country, we believe that, by storing and making all the information available in a unified data repository, we improve the understanding of the current situation and the planning of future actions. In that way, we expect that the implementation of this user-friendly search engine may contribute to increase the usage of data, artifacts, and analyses, benefiting new interdisciplinary research promoted by our team and by other researchers.

Finally, we hope that these datasets and analyses will assist in the tasks of investigating, researching and tracking the evolution of COVID-19 in Brazil and in the region, thus helping to understand this fast moving and particularly challenging and relevant scenario.

Acknowledgements: The research leading to these outcomes has been supported by CNPq, FAPEMIG, CAPES, the Covid Data Analytics Project (PRPq/UFMG; Sesu/MEC), MASWEB, INCT CYBER and ATMOSPHERE. We would like to thank and acknowledge the work of all undergraduate students involved in the Covid Data Analytics Project.

REFERENCES

Aiello, A. E., Renson, A., and Zivich, P. N. Social media– and internet-based disease surveillance for public health. *Annual Review of Public Health* 41 (1): 101–118, 2020.

Bastos, S. B. and Cajueiro, D. O. Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil. *Scientific Reports* 10 (1): 19457, 2020.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *the Journal of machine Learning research* vol. 3, pp. 993–1022, 2003.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: forecasting and control*. John Wiley & Sons, Washington, USA, 2015.

Brodeur, A., Clark, A. E., Fleche, S., and Powdthavee, N. Covid-19, lockdowns and well-being: Evidence from google trends. *Journal of public economics* vol. 193, pp. 104346, 2021.

Brum, P. V., Teixeira, M. C., Miranda, R., Vimieiro, R., Meira Jr, W., and Pappa, G. L. A characterization of portuguese tweets regarding the covid-19 pandemic. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, SBC, Online, October 4-8, 20210, pp. 177–184, 2020.

Chen, E., Lerman, K., and Ferrara, E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance* 6 (2): e19273, 2020.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. The COVID-19 social media infodemic. *Scientific Reports* 10 (1): 16598, 2020.

Cota, W. Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level. *SciELO Preprints* 20 (x): 1–13, 2020.

Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C. N., and Almeida, V. He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLOS ONE* 9 (1): e87041, Jan., 2014.

Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20 (5): 533–534, 2020.

Du, J., Xu, J., Song, H., Liu, X., and Tao, C. Optimization on machine learning based approaches for sentiment analysis on hpv vaccines related tweets. *Journal of biomedical semantics* 8 (1): 1–7, 2017.

Guimarães, R. B., CatÃO, R. D. C., MARTINUCI, O. D. S., Pugliesi, E. A., and Matsumoto, P. S. S. O raciocínio geográfico e as chaves de leitura da covid-19 no território brasileiro. *Estudos avançados* vol. 34, pp. 119–140, 2020.

Kang, G. J., Ewing-Nelson, S. R., Mackey, L., Schlitt, J. T., Marathe, A., Abbas, K. M., and Swarup, S. Semantic network analysis of vaccine sentiment in online social media. *Vaccine* 35 (29): 3621–3638, 2017.

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., and Chen, H. Retrospective analysis of the possibility of predicting the covid-19 outbreak from internet searches and social media data, china, 2020. *Eurosurveillance* 25 (10): 10, 2020.

Locatelli, M. S. et al. Correlations between web searches and COVID-19 epidemiological indicators in Brazil. *Brazilian Archives of Biology and Technology* 65 (x): 00–7, 2022.

Marques-Toledo, C. d. A., Degener, C. M., Vinhal, L., Coelho, G., Meira, W., Codeço, C. T., and Teixeira, M. M. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting dengue at country and city level. *PLoS neglected tropical diseases* 11 (7): e0005729, 2017.

Martins, A. D. F., Cabral, L., Mourão, P. J. C., de Sá, I. C., Monteiro, J. M., and Machado, J. COVID19.BR: a dataset of misinformation about COVID-19 in Brazilian Portuguese WhatsApp messages. In *III Dataset Showcase Workshop (DSW)*. SBC, Online, October 4-8, 2021, pp. 138–147, 2021.

Mavragani, A. and Gkillas, K. Covid-19 predictability in the united states using google trends time series. *Scientific reports* 10 (1): 1–12, 2020.

Miller, M. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository: Johns Hopkins University Center for Systems Science and Engineering. *Bulletin - Association of Canadian Map Libraries and Archives (ACMLA)* 164 (2020): 47–51, 2020.

Moreira, P., Fonseca, R., Alzamora, P. L., Franco, R. A. S., Guiginski, J., Cunha, E. L. T. P., Bernardes, T., Chagas, B., Ferreguetti, K., Passos, L., Cardoso, L., Schneider, R., Pereira, W., da Silva, A. P. C., and Meira Jr., W. Covid Data Analytics: repositório de dados provenientes de múltiplas fontes sobre a pandemia de COVID-19 no Brasil. In *III Dataset Showcase Workshop (DSW)*. Vol. 03. SBC, Online, October 4-8, 2021, pp. 107–116, 2021.

Myers, L. and Sirois, M. J. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* vol. 12, pp. 138–147, 2004.

Pankratz, A. *Forecasting with dynamic regression models*. Vol. 935. John Wiley & Sons, Washington, USA, 2012.

Peixoto, P. S., Marcondes, D., Peixoto, C., and Oliva, S. M. Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. *PLOS ONE* 15 (7): e0235732, 2020.

Pereira, I. G., Guerin, J. M., Silva Júnior, A. G., Garcia, G. S., Piscitelli, P., Miani, A., Distante, C., and Gonçalves, L. M. G. Forecasting Covid-19 dynamics in Brazil: a data driven approach. *International Journal of Environmental Research and Public Health* 17 (14): 5115, 2020.

Ranzani, O. T., Bastos, L. S., Gelli, J. G. M., Marchesi, J. F., Baião, F., Hamacher, S., and Bozza, F. A. Characterisation of the first 250 000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. *The Lancet Respiratory Medicine* 9 (4): 407–418, 2021.

Rey S. J., Arribas-Bel D., W. L. J. Geographic data science with pysal and the pydata stack, 2020.

Riffe, T. et al. Data resource profile: COVerAGE-DB: a global demographic database of COVID-19 cases and deaths. *International Journal of Epidemiology* 50 (2): 390–390f, 2021.

Silva, R. J., Silva, K., Mattos, J., et al. Análise espacial sobre a dispersão da covid-19 no estado da bahia. *SciELO Preprints* vol. 15, pp. 1–10, 2020.

Sultana, A., Tasnim, S., Hossain, M. M., Bhattacharya, S., and Purohit, N. Digital screen time during the covid-19 pandemic: a public health concern. *F1000Research* 10 (81): 81, 2021.

Szwarcwald, C. L., Bastos, F. I., Esteves, M. A. P., and Andrade, C. L. A disseminação da epidemia da aids no brasil, no período de 1987-1996: uma análise espacial. *Cadernos de Saúde Pública* vol. 16, pp. S07–S19, 2000.

Veiga e Silva, L., de Andrade Abi Harb, M. d. P., dos Santos, A. M. T. B., de Mattos Teixeira, C. A., Gomes, V. H. M., Cardoso, E. H. S., da Silva, M. S., Vijaykumar, N. L., Carvalho, S. V., Ponce de Leon Ferreira de Carvalho, A., and Frances, C. R. L. COVID-19 mortality underreporting in Brazil: analysis of data from government internet portals. *Journal of Medical Internet Research* 22 (8): e21413, 2020.

Weisberg, S. *Applied linear regression*. Vol. 528. John Wiley & Sons, Washington, USA, 2005.

Zarei, K., Farahbakhsh, R., Crespi, N., and Tyson, G. A first Instagram dataset on COVID-19. *arXiv preprint: 2004.12226* 10 (x): 0–13, 2020.