Portuguese Twitter Dataset on COVID-19

Richard Adolph Aires Jonker
Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
Aveiro, Portugal
richard.jonker@ua.pt

Roshan Poudel
Research Centre in Digitalization and Intelligent Robotics
Polytechnic Institute of Bragança
Bragança, Portugal
roshan@ipb.pt

Olga Fajarda
Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
Aveiro, Portugal
olga.oliveira@ua.pt

Sérgio Matos
Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
Aveiro, Portugal
aleixomatos@ua.pt

José Luís Oliveira
Institute of Electronics and Informatics Engineering of Aveiro
University of Aveiro
Aveiro, Portugal
jlo@ua.pt

Rui Pedro Lopes
Research Centre in Digitalization and Intelligent Robotics
Polytechnic Institute of Bragança
Bragança, Portugal
rlopes@ipb.pt

***Abstract*—Over the last two years, the COVID-19 pandemic has affected hundreds of millions of people around the world. As in many crises, people turn to social media platforms, like Twitter, to communicate and share information. Twitter datasets have been used over the years in many research studies to extract valuable information. Therefore, several large COVID-19 Twitter datasets have been released over the last two years. However, none of these datasets contains only Portuguese Tweets, despite the Portuguese Language being reported as one of the top five languages used on Twitter. In this paper, we present the first large-scale Portuguese COVID-19 Twitter dataset. The dataset contains over 19 million Tweets spanning 2020 and 2021, allowing the entire pandemic to be analyzed. We also conducted a sentiment analysis on the dataset and correlated the various spikes in Tweet count and sentiment scores to various news articles and government announcements in Portugal and Brazil.**

The dataset is available at: https://github.com/bioinformatics-ua/Portuguese-Covid19-Dataset

***Index Terms*—COVID-19, Twitter, Dataset, Sentiment analysis**

TODO LIST

I. INTRODUCTION

COVID-19 has been so far the most influential crisis situation of the 21st century. According to the World Health Organization (WHO), there have been more than 530 million confirmed cases of COVID-19, including more than 6,3 million deaths [1], with all governments having enforced home-based quarantines and lockdowns to reduce the spread of the infection. These lockdowns have been shown to lead to various negative outcomes in youth, including social and psychological costs [2]. It has also been shown that a high level of stress, anxiety and depression amongst the general public occurred as a result of the pandemic [3]–[5].

During major events and crises, people tend to use more social media platforms to communicate, share information and express their state of mind. It has been shown that news often appears on social media before mainstream news reporters [6], with the addition of people reacting quickly to these events. Twitter[1] is a very popular social media platform worldwide and the primarily used platform to collect large amounts of social media data for research purposes, due to its accessible openly available Application Programming Interface (API). Twitter datasets have shown to be useful to detect influenza outbreaks [7], [8], help disaster management [9], for public health surveillance [10], and to identify the flow of misinformation [11]. Furthermore, sentiment analysis on Twitter datasets provides valuable insights into users' responses to specific events [12] which can be used to improve the public's perception of similar future events.

Over the last two years, several COVID-19 Twitter datasets have been released [13]–[21]. Most of these datasets are multilingual [13]–[15] or contain only English language Tweets [16]–[18]. Only very few language-specific datasets are non-English [19]–[21]. There are significant behavior differences across Twitter users of different languages [22] and therefore it is important to have non-English language-specific COVID-19 Twitter datasets for research purpose. Due to the Twitter API rate limitation, multilingual datasets will always have fewer Tweets from a specific language than language-dedicated collections.

Alshaabi et.al. [23] used 118 billion Tweets collected be-

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 European Union

[1]Twitter Social Media Platform: www.twitter.com

tween 2009 and 2020 to explore the daily use of languages on Twitter and reported the Portuguese language as one of the top five dominant languages on Twitter. Despite the prevalence of the Portuguese language on Twitter and therefore the importance to have Portuguese COVID-19 Twitter datasets to be used for research, only one Portuguese COVID-19 Twitter dataset [19] has been released so far. This dataset contains around 4 million Tweets collected over a 5-month period, in the early stage of the pandemic, from January until May 2020.

In this paper, we present a large-scale Portuguese COVID-19 Twitter dataset containing over 19 million tweets from the years 2020 and 2021. The dataset includes Tweet IDs and various summary statistics: daily aggregation of context annotations, hashtags and the geo-PlaceID aggregated by country. Additionally, we performed sentiment analysis on every Tweet and include the sentiment score in the dataset.

The remainder of the paper is organized as follows: Section II discusses the methodology used to collect the dataset, including tools and software. Section III contains an in-depth breakdown of the dataset with various exploratory items investigated, including a breakdown of the peaks of the Tweet counts in the dataset, and various news articles that appear during the same time. Finally, Section IV concludes the paper.

II. Methodology

This section presents the approach and various tools used to collect the Tweets regarding the COVID-19 pandemic, as well as the techniques used to perform sentiment analysis on each Tweet.

A. Dataset collection

In order to collect the Tweets, the Python library Tweepy was used [24]. The library offers a flexible and easy-to-use interface in order to collect Tweets and dump them into a text file. The library works with Twitter's official API. A Twitter academic research account was used in order to access the Twitter Search API V2 [2]. This allows for a full archive search of Tweets since the first Tweet in 2006, retrieving all Tweets matching a query. Tweets were collected for the years 2020 and 2021, starting from January 1, 00:00 2020 and ending on December 31 23:59 2021. The keywords used to collect the Tweets can be seen in **Table I**. After collecting the Tweets for the year 2020, the queries were changed to retrieve more Tweets, however, due to Twitter API constraints we have not finished collecting the missing Tweets for 2020, so the expected Tweet count in the dataset for 2020 would be higher with these extra keywords.

The data collected excluded all retweets as they contain the same text and would add unnecessary data to the dataset. We added various search terms to the query after the first year in order to try and collect more data. All Tweets contain words relevant to COVID, the pandemic, confinements, and were tagged by Twitter as Portuguese. From the keywords used to collect the Tweets of this dataset, the keywords *#covid19pt*, *novocoronavirus*, *pandemia*, *confinamento*, and *desconfinamento* were not used by any of the multilingual datasets [13]–[15]. Therefore this dataset has different Portuguese Tweets than the Portuguese Tweets retrieved by these multilingual datasets.

[2]Twitter Search APIv2: https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

Year	Keywords
2020	covid OR covid19 OR #covid19pt
2021	covid OR covid19 OR #covid19pt OR #novocoronavirus OR coronavirus OR pandemic OR pandemia OR confinamento OR desconfinamento OR (corona virus)

TABLE I: Query terms used to collect the dataset.

Due to the large scale of the dataset ELK (Elasticsearch, Logstash and Kibana) stack was used to store, process and manage the data [25]. Using ELK allows the loading and processing of data efficiently as well as easily generating visualizations that summarize the data.

In order to perform sentiment analysis, and perform more complex data manipulations we used python and the elasticsearch library in order to connect to and stream the data from the ELK server.

B. Sentiment analysis

To perform sentiment analysis we used the model LeIA - Lexicon for Adapted Inference [26]. LeIA is a Portuguese lexicon fork for VADER (Valence Aware Dictionary for sEntiment Reasoning), which is used for sentiment analysis on social media [27]. VADER is a lexicon-based sentiment analysis tool, meaning that linguists have decided on a set of words with sentiment values which will act as a dictionary, from which the sentiments for documents are calculated. The main advantage of using a lexicon-based sentiment analyzer is that it is very fast and does not need to be trained. VADER is designed for use on social media, as it accounts for various internet slang-related acronyms, words and symbols. It uses various heuristics in order to further influence the sentiment. These heuristics include: punctuation - type and count, use of capitalization, degree modifiers, such as adverbs, the use of contrastive conjunctions, and the use of certain tri-grams. LeIA further builds upon this model by providing a lexicon that is based on Portuguese social media.

VADER outputs a dictionary of 4 values: positive, negative, neutral and compound. Positive, negative and neutral refer to the proportion of words in the text which are positive, negative and neutral, respectively. The compound score is the heuristic-adjusted cumulative score for the sentence. This score ranges from -1 to 1, with -1 being a negative sentiment and 1 being a positive sentiment. The dataset provides the positive, negative and compound scores for each Tweet. Neutral is not provided as it can be calculated with $1 - (pos + neg)$. In order to determine if a Tweet is positive, negative or neutral, the same

thresholds recommended by the original VADER authors are used:

- negative: $compound < -0.05$
- neutral: $-0.05 \leq compound \leq 0.05$
- postive: $compound > 0.05$

III. Dataset

The dataset is organized by day and in compliance with Twitter's Terms of Service, we only publish the Tweet IDs, sentiment scores and summary statistics. The dataset contains 19,306,166 tweets, coming from 2,988,098 different users. In the year 2020, we collected 9,324,186 and in 2021 we collected 9,981,980 Tweets. According to the entity tags assigned by Twitter (tags based on what content is written in the Tweet), around 1 million of the Tweets are tagged to be about Brazil and around 100,000 are tagged to be about Portugal, the rest of the Tweets do not have associated tags. Just under 700,000 Tweets are associated with a Place ID (a user-defined location assigned to their Tweet), most of which are associated with locations in Brazil. There are only 43,697 Tweets tagged with geo-location coordinates.

In what concerns the users, the most active user in the dataset corresponded to a Portuguese news agency, which shared 30,944 Tweets, with an average of 43 Tweets per day. The cumulative percentage of the number of Tweets for the ten most active users is shown in **Figure 1**. In **Table II**, a summary of the Top 10 user types can be seen, most of which are news agencies. The accounts labeled "Unknown deleted account", are accounts which do not exist anymore so we are unable to properly classify them.

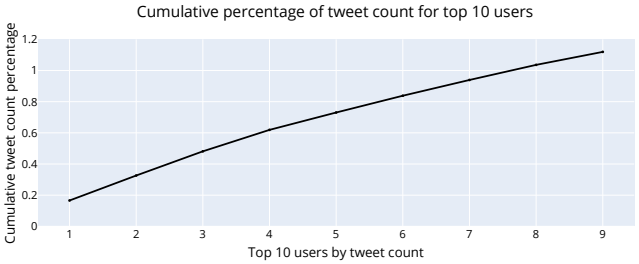


Fig. 1: Cumulative percentage of tweet count for top 10 users.

Around half the users, 1,402,092 out of 2,988,098 (47%), have only 1 Tweet, and 2,712,533 users (91%) have ≤ 10 Tweets which is shown in **Figure 2**. The users which have ≤ 10 Tweets make up for 90.78% (17,526,283) of the total number of Tweets in the dataset.

The top trending hashtags that are not COVID related can be seen in **Table III**. We can see that most of the top hashtags are related to Brazil, either being related to the Brazilian president Jair Bolsonaro (ForaBolsonaro, BolsonaroGenocida, ForaBolsonaroGenocida, BolsonaroTemRazao), news agencies (G1, NDmais) or events such as Big Brother Brazil (BBB21). The only hashtags not directly related to Brazil is "FiqueEmCasa" - stay at home, a campaign by the Portuguese government to incentive people to confine, and Saúde, the Portuguese word for health.

User type	Count of records
Portuguese News	30,944
Portuguese News	29,987
Portuguese News	26,395
Brazilian Bot	21,860
Portuguese News	21,510
Portuguese News	20,810
Unknown deleted account	19,524
Brazilian influencer	18,707
Brazilian Bot	16,102
Portuguese News	15,487
Portuguese News	15,393

TABLE II: Top 10 posting user types.

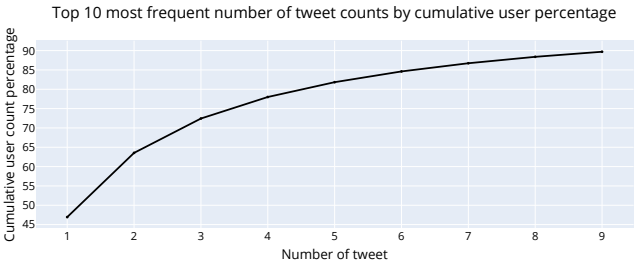


Fig. 2: Ten most frequent number of tweet counts by cumulative user percentage.

Top 10 hashtags not related to Covid	Count of records
ForaBolsonaro	41,060
G1	19,344
BolsonaroGenocida	18,210
BBB21	14,262
ForaBolsonaroGenocida	14,212
NDmais	13,044
FiqueEmCasa	11,973
BrasilContraAFome	10,792
Saúde	8,542
BolsonaroTemRazao	8,000

TABLE III: Top 10 hashtags not related to Covid.

The top ten context annotations for the dataset can be seen in **Table IV**. Context annotations are the people, places, products, and organizations mentioned in the Tweet, which are automatically tagged by Twitter. The most common context annotation is COVID-19 followed by Jair Bolsonaro and some general topics. We then see UOL being the 8th most common context annotation. UOL is a Brazilian News source. There is also a large following for South American football and finally João Dória, the mayor of São Paulo, Brazil.

A line chart summarizing the statistics of the Tweet count can be seen in **Figure 3**. The figure contains the total Tweet count aggregated per week, as well as a Moving average of the Total Tweets with a lag of 5. The 95% confidence interval

Top 10 context annotation entities	Count of records
COVID-19	18,537,431
Jair Bolsonaro	578,404
Services	240,967
Soccer	162,741
Entertainment	158,136
TV/Movies Related	138,233
South America - Soccer	124,727
UOL	119,726
João Dória	114,987
Drinks	95,941
CPG (Consumer Packaged Goods)	95,316
Other	109,558

TABLE IV: Top 10 Context annotation entities.

is generated from the moving average. Peaks are identified if the value is outside the previous 95% confidence interval. This approach is suggested by Brakel [28] for peak signal detection, with **lag** $= 5$, **threshold** $= 1.96$, and **influence** $= 1$. Using these parameters 12, separate peaks can be observed, which we will analyze, in conjunction with the number of positive, negative, and total tweets, as well as the number of COVID cases and deaths for both Portuguese and Brazil. The data regarding the number of cases and deaths were taken from the WHO [29]. The plots containing the deaths and cases use a percentage relative to the maximum value so that the graphs can be comparable. Due to the nature of the data, we investigated modeling the number of tweets with time series models, however, there did not appear to be any visible seasonal components and the models did not fit the data well.

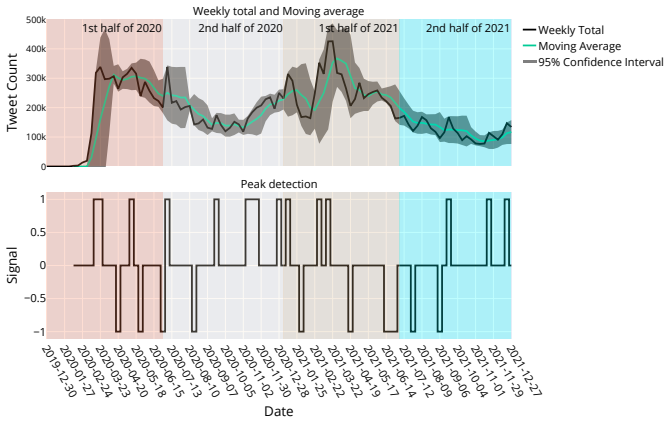


Fig. 3: Rolling weekly average of total tweets and peaks detected by peak detector.

Inspecting the first half of 2020 in **Figure 4**, the initial peak happened between 2020-03-16 and 2020-03-23. On 2020-03-16, the first death due to COVID-19 appeared in Portugal, and lockdown measures began taking place, namely restricting travel to Spain. On 2020-03-18, Portugal declared their first state of emergency. The next peak on 2020-05-11 could be attributed to the government announcement on 2020-05-07, where they announced that schools would resume presentially on 2020-05-18. In this announcement, they also mention that festivals and shows were banned until September 30. On 2020-05-15 the Portuguese government announced an extension to the state of calamity, with reduced restrictions in certain cases. The last peak for the first half of 2020 can be seen on 2020-06-07 the day after the Portuguese government defined a plan to recuperate from social and economic losses. The state of calamity in Portugal was further extended on 2020-06-12. Analyzing the sentiment values for this time period it can be seen that the number of positive tweets is relatively constant, with fluctuations in the number of negative Tweets matching those of the total Tweet counts. Generally, the number of negative Tweet counts were almost double that of the number of positive Tweet count. Looking into the statistics regarding the number of cases and deaths, we see a large peak of deaths around the first 2 weeks of April, and at the same time, we see Brazil's cases and deaths begin to increase, with a very large amount of relative deaths by the end of June.

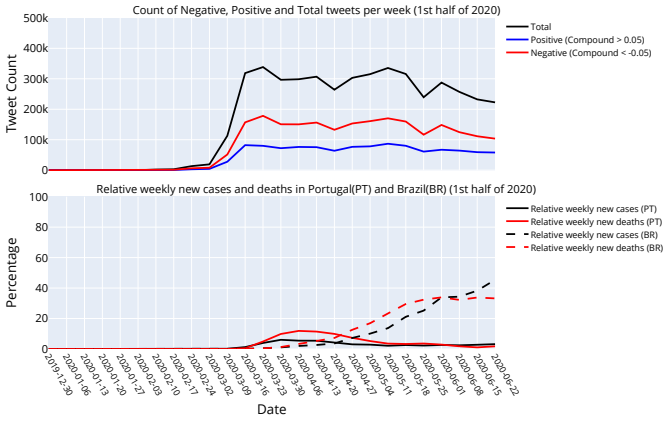


Fig. 4: Count of negative, positive and total tweets per week & Weekly new cases and deaths of PT and BR for the first half of 2020.

Moving onto the second half of 2020, seen in **Figure 5**, the first spike on 2020-07-06, is due to Jair Bolsonaro, the president of Brazil, testing positive for COVID-19 on 2020-07-07, verified by inspection of the data. Following this, the next peak is 2020-09-21. The Portuguese government extended the state of contingency on 2022-09-24, causing the number of cases and deaths still in Portugal to slowly decrease during October and November. There is then a large peak between 2020-11-09 and 2020-11-23. During this time, the Portuguese government issued a State of Emergency between 2020-11-09 and 2020-11-23. On 2020-11-08, the government added more restrictions, including travel bans, a curfew and the rapid testing of people in large spaces. The state of calamity was extended on 2020-11-12, with shops opening between 08:00 - 13:00 during weekends. On 2020-11-21, the government announced the mandatory use of masks in

workplaces and public roads, as well as further restrictions in restaurants. The final peak occurs on 2020-12-28, which is to be expected, as people are generally on holiday and will have more time to use social media. The Portuguese government also announced special restrictions for the holiday time on 2020-12-17, including the banning of movement on public roads and limiting the operational time of restaurants. All these restrictions did not manage to reduce the growing number of cases in Portugal. The negative sentiments also further increase during November. This could be due to either the rise in Portuguese cases or Brazilian cases during that time period.

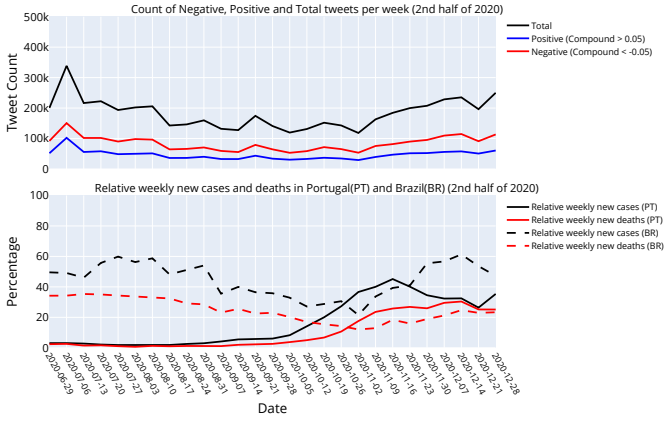


Fig. 5: Count of negative, positive and total tweets per week & Weekly new cases and deaths of PT and BR for the second half of 2020

Looking at the first half of 2021, **Figure 6**, the first spike of 2021 occurred during 2021-01-11. This can be attributed to the new daily maximum of cases and deaths caused by COVID and new cases, in Brazil during this time. The Portuguese government was heavily criticized, due to the alleviation of measurements during the Christmas period which could cause the high number of cases in Portugal. There was speculation that the government might impose a lockdown as a result of these high statistics. The peak on 2021-03-01 was difficult to attribute to any major event. There was a record for the number of deaths caused by the virus in Brazil, as well as very large moving averages for the number of deaths caused by COVID. Upon investigating the Twitter data during this time there are Tweets that mention the death of loved ones, and people grieving, as well as complaints about the Brazilian President. The high number of negative Tweets further supports this. This trend continues for the next two weeks where we see the next spike on 2021-03-15, which is also attributed to the rising number of deaths and cases in Brazil, with shortages of vaccines, and the eventual replacement of the Brazilian health minister on 2021-03-15. In Portugal, various alleviations to the state of emergency were announced on 2021-03-11 including the resuming of face-to-face schooling on 2021-03-15.

Moving onto the final part of our dataset, in the second half of 2021, **Figure 7**, the first spike occurred during 2021-09-

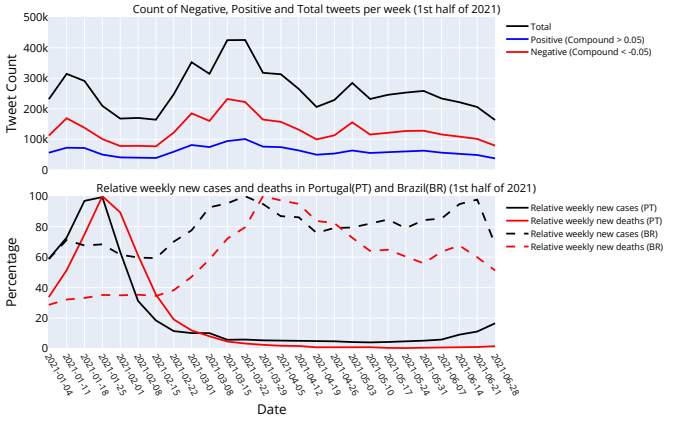


Fig. 6: Count of negative, positive and total tweets per week & Weekly new cases and deaths of PT and BR for the first half of 2021.

20, the same time we see a spike in the number of cases in Portugal. On 2021-09-23, the Portuguese government declared a situation of alert as of October 1, as well as the last stage of deconfinement, with the opening of bars. The next spike occurred on 2021-11-22. On 2021-11-25, the Portuguese government announced a state of calamity as of 2021-12-01, with teleworking between 2022-01-02 and 2022-01-09, as well as the mandatory use of masks in closed places. The final peak of the dataset occurred on 2021-12-20. As mentioned previously, it is expected to have an increase in the number of cases during the holiday times. Besides that, on 2021-12-22, the Portuguese government announced the closure of bars and nightclubs, as well as mandatory telework between 2021-12-25 and 2022-01-09. In Brazil, there was a cyberattack on the system whereby the states report their statistics for COVID. The attack was on 2021-12-10 and was still affecting certain states by 2021-12-20. This attack was preventing states from reporting their cases.

In **Figure 8**, we can see the average compound sentiment per week. Generally, we can see that the average sentiment is always negative ranging from -0.09 to around -0.21, reflecting the negative attitude to COVID-19 from the population as a whole. We do see some fluctuations in the sentiment which inversely correlate to that of the Tweet count. In general, when there is a peak in Tweet count, the average sentiment decreases. This is evident in the previous sentiment figures where generally the number of positive Tweets stays relatively constant with spikes in the number of negative Tweets. The first time this occurs is around the date 2020-03-23 which is during the time of the initial lockdown restrictions announced in Portugal. This tells us that initially, the Portuguese population had a negative attitude towards COVID-19. A relatively positive peak can be seen on 2020-07-06 the same time period when Jair Bolsonaro tested positive for COVID-19. We see a surprisingly negative dip around New Year's Eve 2021. This could be related to the Portuguese government restrictions

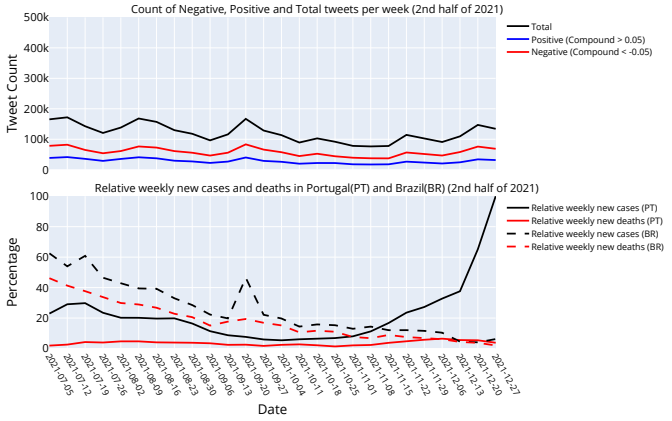


Fig. 7: Count of negative, positive and total tweets per week & Weekly new cases and deaths of PT and BR for the second half of 2021.

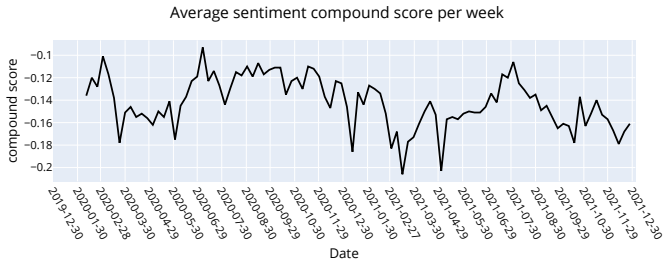


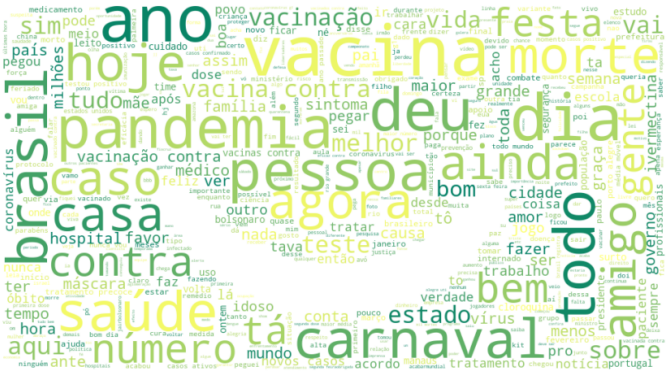
Fig. 8: Average sentiment compound score per week.

during this festive time. The next big dip can be seen on 2021-01-11 which corresponds roughly to the same time as in Portugal the new restrictions being announced at the beginning of January. This dip could also be attributed to people beginning to work after the New year's celebration. There is a large dip around 2021-03-15, which could be attributed to the suspension of the Astrazeneca vaccine. Finally, the last major dip can be seen on 2021-05-03. As mentioned earlier this could be related to the large number of deaths and cases in Brazil during this time period.

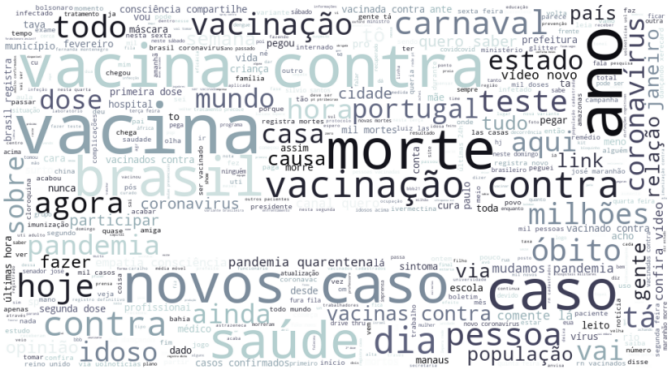
Wordclouds were generated from the dataset for each sentiment category, as seen in **Figure 9**. These exclude covid, covid19 and various Portuguese abbreviations. Generally, all the wordclouds show the same most frequent words. This is to be expected as the Tweets should speak about the same topics, namely, vaccination, friend, people, and death. Words associated with the positive wordcloud include God, hospital, world and father. Words associated with the negative wordcloud include governor, Bolsonaro, and virus. This tells us that people had negative feelings toward the virus as well as governors and the Brazilian president Jair Bolsonaro. In the neutral wordcloud no word of interest appears.

IV. Conclusions

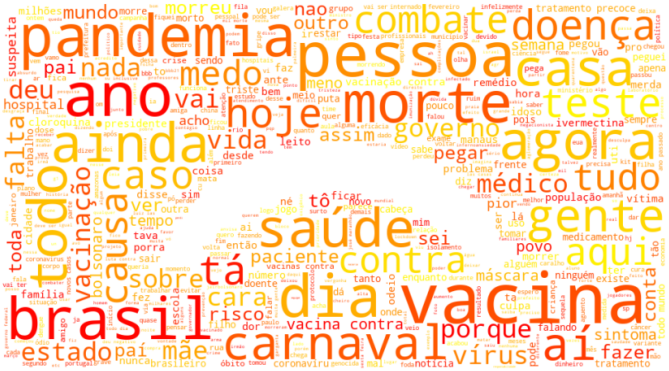
This paper describes a large-scale COVID-19 dataset, containing 19 million Portuguese Tweets from the years 2020 and



(a) Positive



(b) Neutral



(c) Negative

Fig. 9: Wordclouds for different sentiment categories (covid and covid19 are excluded).

2021. Sentiment analysis has been performed on every Tweet and the scores were included in the dataset. Additionally, an in-depth analysis of the dataset was performed, where peaks of Tweet counts and sentiment scores could be correlated with government announcements and relevant events related to Covid-19 happening in Portugal and Brazil. This study shows how language-specific Twitter datasets can be useful for social analysis. This dataset can be used for further knowledge extraction. It can be used, for example, to identify COVID-19 symptoms or to gather insights on how restrictions affect populations which can help governments to better manage

crises in the future.

ACKNOWLEDGMENT

This work was supported by FCT – Fundação para a Ciência e Tecnologia within project DSAIPA/AI/0088/2020.

REFERENCES

[1] "WHO Coronavirus (COVID-19) Dashboard." [Online]. Available: https://covid19.who.int (Accessed 2022-05-31).

[2] M. Owens, E. Townsend, E. Hall, T. Bhatia, R. Fitzgibbon, and F. Miller-Lakin, "Mental Health and Wellbeing in Young People in the UK during Lockdown (COVID-19)," *International Journal of Environmental Research and Public Health*, vol. 19, no. 3, p. 1132, Jan. 2022.

[3] A. P. Association, Mar. 2020. [Online]. Available: https://www.psychiatry.org/newsroom/news-releases/new-poll-covid-19-impacting-mental-well-being-americans-feeling-anxious-especially-for-loved-ones-older-adults-are-less-anxious (Accessed 2022-05-30).

[4] A. S. Sameer, M. A. Khan, S. Nissar, and M. Z. Banday, "Assessment of Mental Health and Various Coping Strategies among general population living Under Imposed COVID-Lockdown Across world: A Cross-Sectional Study," *Ethics, Medicine and Public Health*, vol. 15, p. 100571, Oct. 2020.

[5] K. M. Sønderskov, P. T. Dinesen, Z. I. Santini, and S. D. Østergaard, "The depressive state of Denmark during the COVID-19 pandemic," *Acta Neuropsychiatrica*, vol. 32, no. 4, pp. 226–228, Aug. 2020.

[6] M. Imran, F. Ofli, D. Caragea, and A. Torralba, "Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions," *Information Processing & Management*, vol. 57, no. 5, p. 102261, Sep. 2020.

[7] B. Alkouz, Z. Al Aghbari, M. A. Al-Garadi, and A. Sarker, "Deepluenza: Deep learning for influenza detection from twitter," *Expert Systems with Applications*, vol. 198, p. 116845, 2022.

[8] S. Molaei, M. Khansari, H. Veisi, and M. Salehi, "Predicting the spread of influenza epidemics by analyzing twitter messages," *Health and Technology*, vol. 9, no. 4, pp. 517–532, 2019.

[9] M. Karimiziarani, K. Jafarzadegan, P. Abbaszadeh, W. Shao, and H. Moradkhani, "Hazard risk awareness and disaster management: Extracting the information content of twitter data," *Sustainable Cities and Society*, vol. 77, p. 103577, 2022.

[10] S. E. Jordan, S. E. Hovet, I. C.-H. Fung, H. Liang, K.-W. Fu, and Z. T. H. Tse, "Using twitter for public health surveillance from monitoring and prediction to public response," *Data*, vol. 4, no. 1, p. 6, 2018.

[11] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[12] A. Saroj and S. Pal, "Use of social media in crisis management: A survey," *International Journal of Disaster Risk Reduction*, vol. 48, p. 101584, Sep. 2020.

[13] E. Chen, K. Lerman, and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set," *JMIR Public Health and Surveillance*, vol. 6, no. 2, p. e19273, May 2020.

[14] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, and G. Chowell, "A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration," *Epidemiologia*, vol. 2, no. 3, pp. 315–324, Sep. 2021.

[15] C. E. Lopez and C. Gallemore, "An augmented multilingual Twitter dataset for studying the COVID-19 infodemic," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 102, 2021.

[16] S. A. Memon and K. M. Carley, "Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset," *arXiv:2008.00791 [cs]*, Sep. 2020, arXiv: 2008.00791.

[17] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021.

[18] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, May 2021.

[19] T. de Melo and C. M. S. Figueiredo, "A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese," *Data in Brief*, vol. 32, p. 106179, Oct. 2020.

[20] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, "ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks," *arXiv:2004.05861 [cs]*, Mar. 2021, arXiv: 2004.05861.

[21] S. Allés-Torrent, G. del Rio Riande, J. Bonnell, D. Song, and N. Hernández, "Digital narratives of covid-19: A twitter dataset for text analysis in spanish," *Journal of Open Humanities Data*, vol. 7, 2021.

[22] L. Hong, G. Convertino, and E. Chi, "Language matters in twitter: A large scale study," in *Proceedings of the international AAAI conference on web and social media*, vol. 5, no. 1, 2011, pp. 518–521.

[23] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds, "The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020," *EPJ data science*, vol. 10, no. 1, p. 15, 2021.

[24] "Tweepy Documentation — tweepy 4.10.0 documentation." [Online]. Available: https://docs.tweepy.org/en/stable/ (Accessed 2022-05-30).

[25] "The ELK Stack: From the Creators of Elasticsearch." [Online]. Available: https://www.elastic.co/what-is/elk-stack (Accessed 2022-05-30).

[26] R. J. Almeida, "Leia - lexicon for adapted inference," https://github.com/rafjaa/LeIA, 2018, (Accessed 2022-05-30).

[27] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014.

[28] J. v. Brakel, "Robust peak detection algorithm using z-scores," https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/2264036222640362, 2014. [Online]. Available: https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data/22640362#22640362 (Accessed 2022-04-12).

[29] "WHO COVID-19 Dashboard," publisher: Geneva: World Health Organization, 2020. [Online]. Available: https://covid19.who.int/data/ (Accessed 2022-09-24).