DORTMUND CENTER
FOR DATA-BASED
MEDIA ANALYSIS

DoCMA Working Paper #4

February 2021

# corona100d

German-language Twitter dataset of the first 100 days after Chancellor Merkel addressed the coronavirus outbreak on TV

Jonas Rieger and Gerret von Nordheim*

**Cite as:**
Rieger, J. & von Nordheim, G. (2021). "corona100d – German-language Twitter dataset of the first 100 days after Chancellor Merkel addressed the coronavirus outbreak on TV." *DoCMA Working Paper #4, Feb. 2021.* DOI: 10.17877/DE290R-21911.

Version 1.0, February 2021

*Jonas Rieger is a researcher at TU Dortmund University and DoCMA, Dr. Gerret von Nordheim is a postdoctoral researcher at the University Hamburg and DoCMA. The authors would like to thank Prof. Dr. Henrik Müller, Prof. Dr. Erich Schubert, and Prof. Dr. Carsten Jentsch for helpful comments.

# Abstract

In this paper, we present a German-language Twitter dataset related to the Covid-19 pandemic. We show how the R (R Core Team 2020) package rtweet (Kearney 2019) and a combination of keywords can be used to create the dataset and provide a way to rehydrate most of the tweets. The dataset consists of 3 699 623 tweets from 2020/03/19 to 2020/06/26 and was constructed from hourly API requests of 50 000 tweets. In a brief analysis, we give first insights into the dataset and provide approaches that can be refined in further research.

Key words: Covid-19, SARS-CoV-2, scraper, data, text, developer, dev, Twitter

# 1 Introduction

In search of information about the Covid-19 pandemic, many people turn to social media. Platforms such as Twitter are reporting record increases in users (Paul and Culliford 2020). However, platforms play an ambivalent role during the current crisis: they are channels of education and criticism used by traditional media, scientists, and public institutions, but they are also venues for the infodemic that spread alongside the pandemic, multipliers of conspiracy theories and disinformation.

Because of this diversity, the emerging data traces are an interesting starting point for scholars to examine different dimensions of the pandemic. It is therefore not surprising that many researchers have already collected and published data sets. For example, Ambalina (2019) presents 21 datasets, one of which contains over 150 million tweets about the Covid-19 pandemic, is multilingual and consists mostly of English, French, and Spanish tweets. Another multilingual collection of tweets about Covid-19 is provided by Dimitrov et al. (2020). Their dataset consists of over 8 million tweets from October 2019 to April 2020.

Most Twitter datasets are English-language or multilingual, with English-language tweets forming the bulk of the data. To date, however, there is no comprehensive collection of German-language tweets related to the Covid-19 pandemic as far as we know. Therefore, we present corona100d, a dataset consisting of 3.7 million German-language tweets from 2020/03/19 to 2020/06/26, the first 100 days after the then Chancellor Angela Merkel addressed citizens in a television speech on the outbreak of the Covid-19 pandemic.

In Section 2, we explain how we created the dataset and how it can be reproduced almost completely. In Section 3, we demonstrate with a small sample? how the data can be used to gain insights into the issues discussed at a particular time in the course of the Covid-19 pandemic. Our analysis is limited to basic descriptions of the dataset. All codes to create the dataset and a file of the status IDs to rehydrate the dataset are provided at the repository `https://github.com/JonasRieger/corona100d`.

# 2 Methodology

The introduced dataset was created using the statistical programming software R (R Core Team 2020) with the help of the packages rtweet (Kearney 2019) and lubridate (Grolemund and Wickham 2011). We started the API requests on 2020/03/18 at 8pm, which was when the then German Chancellor Angela Merkel spoke to citizens in a TV address on the outbreak of the Covid-19 pandemic. To ensure comparability between the days through full consideration, we used only tweets posted no earlier than 2020/03/19 for this corpus. We requested 50 000 tweets every hour using the keyword combination

```
coronavirusde  OR corona    OR coronavirus  OR covid19  OR covid
               OR pandemie  OR epidemie     OR virus    OR SARSCoV2.
```

**Code 1:** R procedure to download tweets based on keywords

```
library(rtweet)
library(lubridate)
token = readRDS(".rtweet_token.rds")

repeat({
  nextscrape = Sys.time()
  hour(nextscrape) = hour(nextscrape) + 1
  minute(nextscrape) = 30
  second(nextscrape) = 0
  time = Sys.time()
  ##download tweets##
  rt = try(search_tweets(
    "coronavirusde OR corona OR coronavirus OR covid19 OR covid OR pandemie
    OR epidemie OR virus OR SARSCoV2", n = 50000, lang = "de",
    include_rts = TRUE, token = token, retryonratelimit = TRUE))
  ##repeat at a maximum of 3 times if download failed##
  trys = 1
  while(class(rt)[1] == "try-error" && trys < 3 && Sys.time() + 10*60 < nextscrape){
    time = Sys.time()
    rt = try(search_tweets(
      "coronavirusde OR corona OR coronavirus OR covid19 OR covid OR pandemie
      OR epidemie OR virus OR SARSCoV2", n = 50000, lang = "de",
      include_rts = TRUE, token = token, retryonratelimit = TRUE))
    trys = trys+1
  }
  ##save tweets##
  filename = paste0(gsub(" ", "", gsub("-", "", gsub(":", "", time))), ".csv")
  save_as_csv(rt, file_name = file.path("tweets", filename))
  ##wait for next scrape##
  wait = difftime(nextscrape, Sys.time(), units = "secs")
  if(wait > 0) Sys.sleep(wait)
})
```

The keywords are not case-sensitive. The query not only considers hashtags but also performs a pattern search, i.e., those tweets are considered that contain at least one of the words from the search query as a partial word. It is not possible to get all German-language tweets on the topic. Instead, the query provides a sample of tweets that satisfy the keyword search. This may result in duplicates being created during the procedure. In this case, our procedure selects the last status of the tweet so that the most recent status of, for example, the favorite and retweet numbers could be taken into account.

In this paper, we present a dataset consisting of data up to 2020/07/06. To create the corpus, we then considered all the tweets from this set of tweets that were posted by 2020/06/26, i.e., by the 100th day after Angela Merkel addressed the German public. This results in an average of about 37 000 tweets per day, although in Section 3 we will show that the numbers are not uniformly distributed over time.

Code 1 shows the R procedure for requesting the tweets. To use the code, it is necessary to create a token for the Twitter API (Twitter 2020), which in this example is saved under `.rtweet_token.rds`. The example script is suitable for getting recent tweets. Thus, the presented dataset cannot be reproduced in this way. However, based on the

**Code 2:** R procedure to rehydrate tweets based on status IDs

```
library(rtweet)
token = readRDS(".rtweet_token.rds")
status = gsub("x", "", readLines("status_id.txt"))
##lasts around 10 hours##
tweets = lookup_tweets(status[1:90000], token = token)
i = 90001
n = length(status)
while(i <= n){
  time = Sys.time()
  tweets = rbind(tweets, lookup_tweets(status[i:min((i+89999), n)], token = token))
  i = i+90000
  elapsed = difftime(Sys.time(), time, units = "mins")
  if(elapsed < 15 && i <= n) Sys.sleep(as.numeric(15 - elapsed)*60)
}
saveRDS(tweets, file = "rehydrated.rds")
```

status IDs we publish, most of the tweets can be retrieved. For this, it is necessary that a tweet is still online and has not been deleted by Twitter or the author. Using the function lookup_tweets from the rtweet package, 90 000 tweets can be downloaded every 15 minutes, so in about ten hours the entire dataset should be reproducible except for the deleted tweets. An example script for creating the dataset in this way is shown in Code 2. In an application of the procedure on 2021/01/22 we were able to rehydrate 86% of the tweets. We plan to investigate the characteristics of these deleted tweets in further research.

# 3 Insights

We performed the following brief analysis of the presented dataset using the R packages data.table (Dowle and Srinivasan 2020), urltools (Keyes et al. 2019) and tosca (Koppers et al. 2020). The word clouds are plotted using ggwordcloud (Le Pennec and Slowikowski 2019) in ggplot2 (Wickham 2016). For processing text data in R, we recommend the packages tm (Feinerer et al. 2008) and quanteda (Benoit et al. 2018). In addition, the packages topicmodels (Grün and Hornik 2011), stm (Roberts et al. 2019) and ldaPrototype (Rieger 2020) offer good possibilities for modeling text data using topic models.

In Figure 1, we illustrate how the 3 699 623 tweets of the corpus are distributed over the 100 days of the observation period from 2020/03/19 to 2020/06/26. The curve of all the days of the week is shown in black, and the curves restricted to one day of the week are shown in the colors corresponding to the legend. Overall, a downward trend can clearly be seen. At the beginning of the observation period, almost 90 000 tweets are observed at the maximum on a Friday. By mid-April, this number drops to a level of about 45 000. It turns out that the days Monday to Thursday show a similar behavior over time, i.e., the number of tweets per day does not strongly depend on whether one considers a Monday or a Thursday. Conversely, the behavior of the blue curve, which belongs to Friday, varies quite strongly. There are weeks in which the curve is comparable to the curves of the weekdays from Monday to Thursday and other weeks in which Friday is more likely to

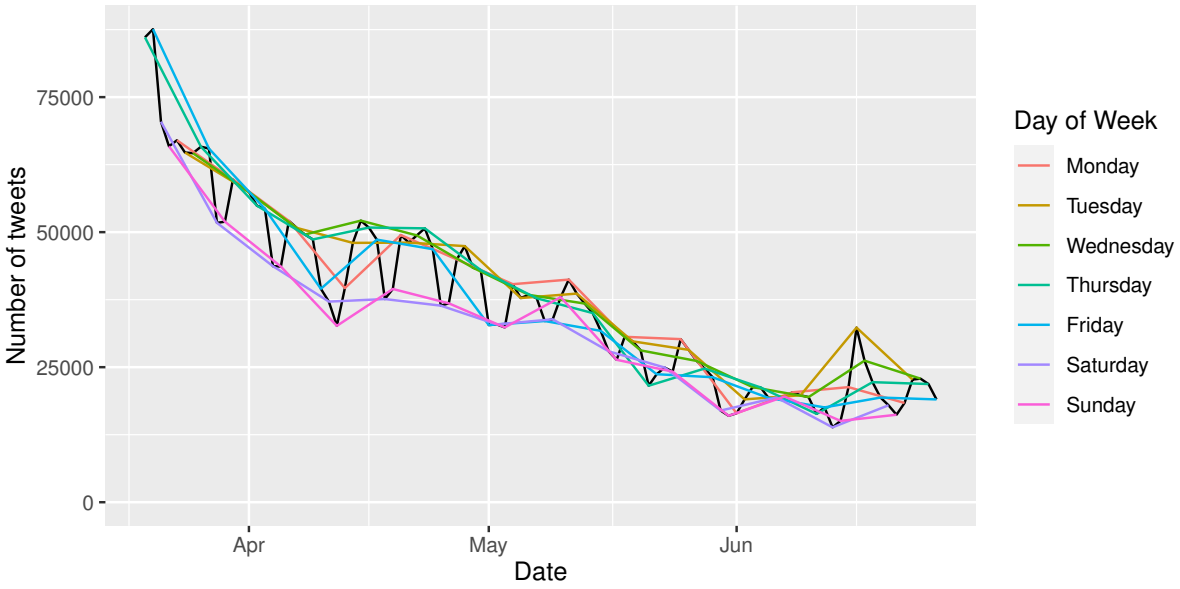


**Figure 1:** Number of tweets per day in the corona100d corpus

belong to the weekend group. Interestingly, the number of tweets per Saturday or Sunday tend to be underrepresented in comparison. The lowest number of tweets is observed with fewer than 14 000 tweets on Saturday, 2020/06/13. Shortly thereafter, on Tuesday, 2020/06/16, the greatest abnormality in the curves can be noticed. With more than 32 000 tweets in the dataset, this day clearly stands out compared to the previous day (21 000) and the following day (26 000).

Figure 2 provides a clue to the cause of the significant outlier on 2020/06/16 in the number of tweets. The image shows a word cloud of the 35 most common secondary hashtags used in tweets that day. The size in which a single word is displayed depends on the square root of its frequency. Our understanding of secondary hashtags in this context includes all those that do not automatically result from the keyword as a parent topic. Specifically, we excluded the following hashtags:

| | | | | |
|---|---|---|---|---|
| `corona` | `covid19` | `sarscov2` | `coronavirus` | `coronavirusdeutschland` |
| `covid` | `covid_19` | `sars_cov_2` | `coronakrise` | `coronadeutschland` |
| `covid-19` | `covid_19de` | `covid2019de` | `viruscorona` | `coronavirusde` |
| `covid2019` | `covid__19` | `covid19de` | `deutschland` | `covid19deutschland` |

It is noticeable, on the day mentioned, the hashtag `coronawarnapp` clearly dominates. In fact, on that Tuesday, the Coronavirus warning app (German: Corona-Warn-App, The Federal Government 2020) was released. By tracking users in a privacy-compliant manner, the app is supposed to provide a reference point as to whether a person had been exposed to a high-risk encounter with a user later reported to be infected. Seemingly, this release was a topic that, on the one hand dominated the interest of the public on this day, and, on the other hand, it also seems to have generally led to a higher frequency and/or engagement of tweets on the topic of Corona, resulting in a significant increase in posted tweets for this day in our sample (cf. Figure 1).

Another topic on this day seems to deal mainly with the fate of self-employed people


**Figure 2:** Wordcloud of the top 35 secondary hashtags on Tuesday, 2020/06/16

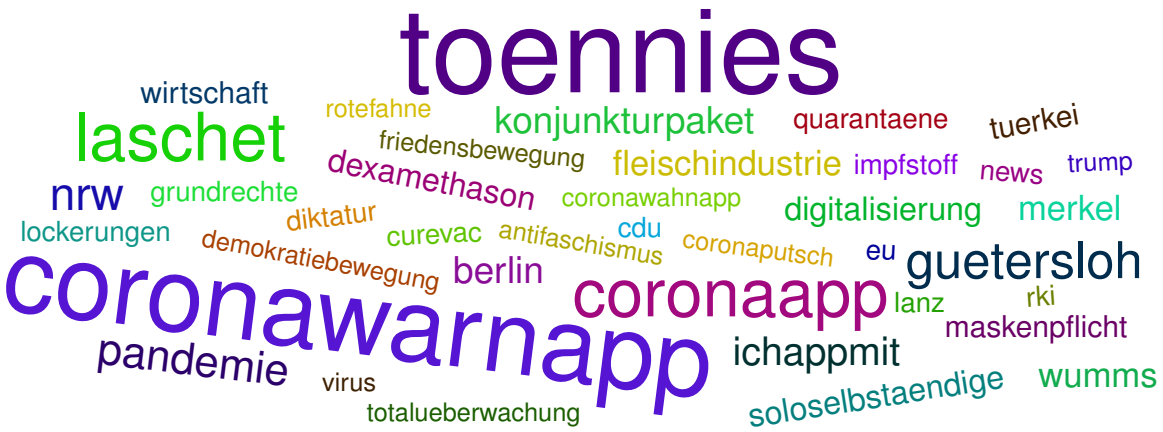


**Figure 3:** Wordcloud of the top 35 secondary hashtags on Wednesday, 2020/06/17

in the pandemic, indicated by the hashtags `konjunkturpaket` and `soloselbstaendige`. Examples of hashtags that are prominent over a longer period of time are `maskenpflicht`, `lockdown`, and `quarantaene`. In comparison, we illustrate the word cloud of the following day in Figure 3. On Wednesday, 2020/06/17, the warning app was still a prominent topic but was surpassed in popularity by the news that over 600 workers had become infected at a large German slaughterhouse owned by the meat processing company Tönnies Lebensmittel GmbH & Co. KG. This can be seen by the hashtags `toennies`, `fleischindustrie` and `guetersloh`. The comparison of the word clouds illustrates the possibility of using Twitter datasets to map this rapidly changing shift in the public's interests. In addition to hints about a vaccine (`impfstoff` and `curevac`), many hashtags suggesting criticism concerning government actions also show up on this day, such as `diktatur`, `totalueberwachung` and `grundrechte`.

These brief insights show the potential for more extensive analyses in each case. Table 1 displays statistics on the top 13 users' number of retweets and favorites. The user with the most tweets in the corona100d dataset is *FWWinterberg* (8151 tweets), with each tweet receiving an average of one favorite mark and half a retweet. The source *FWWinterberg* most often links to is *Der Spiegel*'s website. Almost a quarter of the links from this user direct to this source.

The entire dataset consists of 3 699 623 tweets that received a mean of 6 favorite tags and 1.37 retweets. In addition, 10% of the links in the tweets in turn lead to Twitter (cf. Table 2). Overall, just over 55% of the tweets in the corpus contain at least one link, with the presented 13 top users consistently linking in nearly all their tweets. Almost 6% of the tweets contain a quote, just under 40% of all tweets contain at least one hashtag,

**Table 1:** Mean number of favorite counts and retweets of tweets and favorite linked source in tweets from the 13 top users with the most tweets posted

| User | N | $\overline{\text{Favs}}$ | $\overline{\text{RTs}}$ | Favorite Source | Ratio |
|---|---|---|---|---|---|
| FWWinterberg | 8151 | 0.94 | 0.47 | spiegel | 0.23 |
| welt | 5266 | 15.62 | 6.77 | welt | 1.00 |
| FAZ_NET | 5194 | 1.59 | 0.79 | faz | 1.00 |
| Deutschland_BRD | 5077 | 0.07 | 0.03 | spiegel | 0.34 |
| AlfredHampp | 4490 | 0.25 | 0.74 | berliner-sonntagsblatt | 1.00 |
| FOCUS_TopNews | 4474 | 1.15 | 4.17 | focus | 1.00 |
| gnutiez | 4141 | 0.01 | 0.01 | gnutiez | 0.50 |
| Tagesspiegel | 3915 | 21.21 | 11.06 | tagesspiegel | 1.00 |
| Kleeblatt1977 | 3838 | 0.01 | 0.00 | nordbayern | 0.21 |
| focusgesundheit | 3635 | 0.18 | 0.12 | focus | 1.00 |
| BILD | 3630 | 10.42 | 4.27 | bild | 0.92 |
| focuspolitik | 3522 | 0.19 | 0.13 | focus | 1.00 |
| StN_News | 3480 | 0.44 | 0.14 | stuttgarter-nachrichten | 1.00 |
| $\Sigma$ | 3 699 623 | 6.00 | 1.37 | twitter | 0.10 |

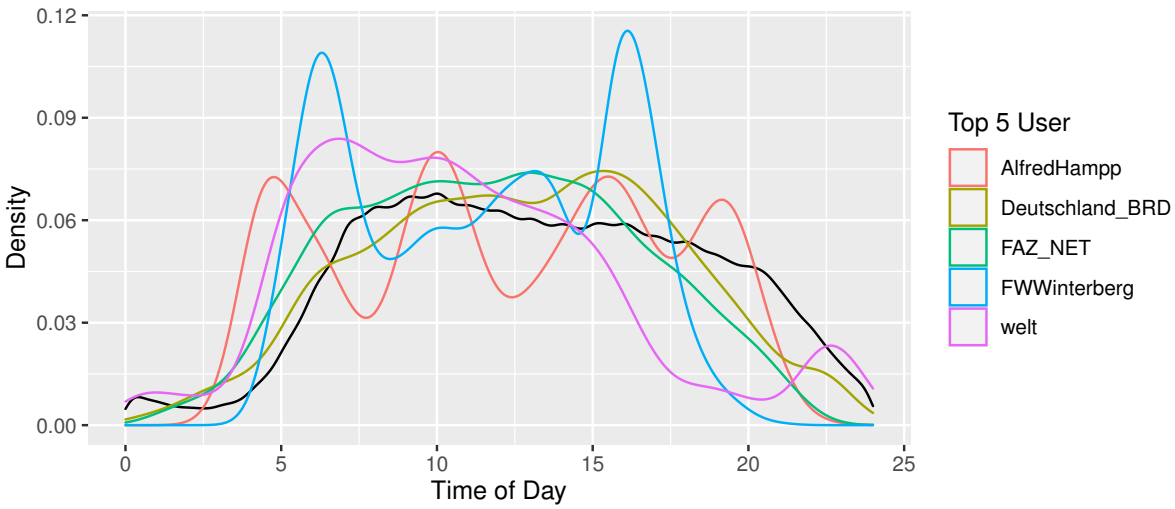


**Figure 4:** Density of the tweets' posting time for the top 5 users and in total (black)

and nearly 2 million tweets have not a single favorite tag. The most favorited tweet in the dataset is `x1261283547698036736` from *SabineLeidig* with 25 000 favorites and 7500 retweets. At the time of writing, this tweet has gathered over 35 000 favorites and more than 10 000 retweets.

Another analysis approach could be based on user activity at different times of the day. In Figure 4, we show as an example the distribution of all tweets over time of day from 12 am to 11:59 pm. In addition, the corresponding densities of the top 5 users are plotted. In particular, the individuals *FWWinterberg* and *AlfredHampp* show clear peaks at particular times. The density of *Die Welt*'s Twitter channel shows a clear peak at 10:30 pm with a curve that is generally similar to the overall curve. As expected, the main activity of users from the corona100d corpus is between 6 am and 9 pm. In this analysis, one could also

**Table 2:** Number of users and favorite user that linked the 13 most commonly linked sources in tweets

| Source | N | #Users | Favorite User | Ratio |
|---|---|---|---|---|
| twitter | 217 852 | 58 074 | luciamertins | 0.0056 |
| youtube | 101 978 | 32 614 | MichaelStawicki | 0.0085 |
| spiegel | 76 355 | 23 884 | SPIEGEL_alles | 0.0334 |
| focus | 63 032 | 7655 | FOCUS_TopNews | 0.0710 |
| faz | 54 072 | 12 547 | FAZ_NET | 0.0961 |
| sueddeutsche | 44 450 | 14 489 | SZ | 0.0591 |
| zeit | 43 550 | 15 485 | zeitonline | 0.0729 |
| welt | 43 199 | 10 331 | welt | 0.1216 |
| bild | 40 179 | 7483 | BILD | 0.0832 |
| tagesschau | 28 483 | 10 280 | tagesschau | 0.0796 |
| tagesspiegel | 26 769 | 9941 | Tagesspiegel | 0.1461 |
| n-tv | 24 382 | 6915 | ntvde | 0.0818 |
| nzz | 20 561 | 5451 | NZZ | 0.0905 |

imagine more complex groupings than those based on users. For example, one approach might be to investigate the topics discussed depending on the time of day, which may be an issue for future research.

Linked content in combination with engagement metrics can be used to gain a deeper understanding of the topics covered (cf. von Nordheim and Rieger 2020). Table 2 shows the 13 most frequently shared sources. The most common link is to Twitter itself: 217 852 tweets from 58 074 different users contain at least one such link. In addition to the two most shared pages of Twitter and YouTube, the other top 11 sources feature only online media sites. For each of these pages, most of the links to their websites are from their own Twitter channels. However, with almost 15% for *Tagesspiegel* and 12% for *Die Welt*, these two media stand out due to their high rate if self-reference. *Der Spiegel*'s channel apparently shares the fewest links to its own site compared to the rest of the users. This effect can have various causes. On the one hand, it may be due to the fact that *Der Spiegel* links are generally shared more frequently, and, on the other hand, it may be due to the fact that *Der Spiegel* shares a lower proportion of its own articles. Comparing the sources *Der Spiegel* and *Focus*, it is noticeable that URLs to *Der Spiegel*'s website are shared by a clearly broader group of people (23 884 users), while only 7655 different users share links to articles from *Focus*. This effect applies to a similar extent to *Bild*, the online presence of Germany's largest tabloid.

We were able to show in Section 2 how the corona100d dataset was created and how it can be made available to users. In this section, we then gave a brief insight into the dataset. We were able to observe a weekly seasonality in the number of tweets posted. We also looked at selected word clouds that indicated a rapidly changing topic agenda. In addition, we looked at typical statistics, such as favorite and retweet counts for a selection of users, the distribution of tweets across time of day, and the most frequently linked sources. All these brief analyses provide approaches and offer broad possibilities for further analysis. One example is the study of the sentiment of different topics over time.

For that purpose, topic models can be helpful. It is likely that phases of the pandemic can be derived from this. In addition, it could be a promising approach to split the dataset into tweets from private individuals and others and then compare their topical patterns with the coverage in print and online media. These analyses can be framed under the heading “Mapping the German Corona debate on Twitter”.

## Acknowledgments

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at TU Dortmund University (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

# References


Ambalina, Limarc (2019). *Top Twitter Datasets for Natural Language Processing and Machine Learning*. Last accessed 2021/02/08. URL: https://lionbridge.ai/datasets/top-20-twitter-datasets-for-natural-language-processing-and-machine-learning/.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). "quanteda: An R package for the quantitative analysis of textual data". In: *Journal of Open Source Software* 3.30, p. 774. DOI: 10.21105/joss.00774.

Dimitrov, Dimitar, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze (2020). "TweetsCOV19 - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic". In: *Proceedings of the 29th CIKM conference*. ACM, pp. 2991–2998. DOI: 10.1145/3340531.3412765.

Dowle, Matt and Arun Srinivasan (2020). *data.table: Extension of 'data.frame'*. R package version 1.13.6. URL: https://CRAN.R-project.org/package=data.table.

Feinerer, Ingo, Kurt Hornik, and David Meyer (2008). "Text Mining Infrastructure in R". In: *Journal of Statistical Software* 25.5, pp. 1–54. DOI: 10.18637/jss.v025.i05.

Grolemund, Garrett and Hadley Wickham (2011). "Dates and Times Made Easy with lubridate". In: *Journal of Statistical Software* 40.3, pp. 1–25. URL: https://www.jstatsoft.org/v40/i03/.

Grün, Bettina and Kurt Hornik (2011). "topicmodels: An R Package for Fitting Topic Models". In: *Journal of Statistical Software* 40.13, pp. 1–30. DOI: 10.18637/jss.v040.i13.

Kearney, Michael W. (2019). "rtweet: Collecting and analyzing Twitter data". In: *Journal of Open Source Software* 4.42, p. 1829. DOI: 10.21105/joss.01829.

Keyes, Os, Jay Jacobs, Drew Schmidt, Mark Greenaway, Bob Rudis, Alex Pinto, Maryam Khezrzadeh, Peter Meilstrup, Adam M. Costello, Jeff Bezanson, Peter Meilstrup, and Xueyuan Jiang (2019). *urltools: Vectorised Tools for URL Handling and Parsing*. R package version 1.7.3. URL: https://CRAN.R-project.org/package=urltools.

Koppers, Lars, Jonas Rieger, Karin Boczek, and Gerret von Nordheim (2020). *tosca: Tools for Statistical Content Analysis*. R package version 0.2-0. DOI: 10.5281/zenodo.3591068.

Le Pennec, Erwan and Kamil Slowikowski (2019). *ggwordcloud: A Word Cloud Geom for 'ggplot2'*. R package version 0.5.0. URL: https://CRAN.R-project.org/package=ggwordcloud.

Paul, Katie and Elizabeth Culliford (2020). *Twitter shares rise on record yearly growth in daily users*. Last accessed 2021/02/08. URL: https://www.reuters.com/article/us-twitter-results-idUSKCN24O1EB.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rieger, Jonas (2020). "ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations". In: *Journal of Open Source Software* 5.51, p. 2181. DOI: 10.21105/joss.02181.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley (2019). "stm: An R Package for Structural Topic Models". In: *Journal of Statistical Software* 91.2, pp. 1–40. DOI: 10.18637/jss.v091.i02.

The Federal Government (2020). *The Corona-Warn-App: Helps us fight the Coronavirus.* Last accessed 2021/02/08. URL: https://www.bundesregierung.de/breg-de/themen/corona-warn-app/corona-warn-app-englisch.

Twitter (2020). *Twitter Developer API.* Last accessed 2021/02/08. URL: https://developer.twitter.com/en.

von Nordheim, Gerret and Jonas Rieger (2020). "Distorted by Populism – A computational analysis of German parliamentarians' linking practices on Twitter [Im Zerrspiegel des Populismus – Eine computergestützte Analyse der Verlinkungspraxis von Bundestagsabgeordneten auf Twitter]". In: *Publizistik* 65, pp. 403–424. DOI: 10.1007/s11616-020-00591-7.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.