TUTORIAL

A GUIDE ON **EXTRACTING AND TIDYING TWEETS** WITH R

Julia Bahia ADAMS

Instituto de Estudos da Linguagem – Universidade Estadual de Campinas (UNICAMP)

Carlos Augusto Jardim CHIARELLI

Faculdade de Engenharia Mecânica – Universidade Estadual de Campinas (UNICAMP)

OPEN ACCESS

EDITORS
– Miguel Oliveira, Jr. (UFAL)
– René Almeida (UFS)

REVIEWERS
– Bruno Rocha (UFMG)
– Heliana Mello (UFMG)

ABOUT THE AUTHORS
– Julia Bahia Adams
Conceptualização, Metodologia, Software, Escrita – Rascunho Original e Escrita – Análise e Edição.
– Carlos Augusto Jardim Chiarelli
Software.

DATES
–Received: 19/08/2021
–Accepted: 18/10/2021
–Published: 12/03/2021

HOW TO CITE
ADAMS, J.B.; CHIARELLI, C.A.J. (2021). A Guide on Extracting and Tidying Tweets with R. *Cadernos de Linguística*, v. 2, n. 4, e410.

ABSTRACT

Social media platforms represent a profuse resource for academic research and a wide range of untapped possibilities for linguists (D'ARCY; YOUNG, 2012). This rapidly developing field presents various ethical issues and unique challenges regarding methods to retrieve and analyze data. This tutorial provides a straightforward guide to harvesting and tidying Twitter data, focused mainly on the Tweets' text, by using the R programming language (R CORE TEAM, 2020) via Twitter APIs. The R code was developed in Adams (2020), based on the *rtweet* package (KEARNEY, 2018), and successfully resulted in a script for corpora compilation. In this tutorial, we discuss limitations, problems, and solutions in our framework for conducting ethical research on this social networking site. Our ethical concerns go beyond what we "agree to" in terms of use and privacy policies, that is, we argue that their content does not contemplate all the concerns researchers need to attend to. Additionally, our aim is to show that using Twitter as a data source does not require advanced computational skills.

RESUMO

As plataformas de redes sociais representam uma profunda fonte de dados para pesquisas acadêmicas e um amplo leque de possibilidades para linguistas (D'ARCY; YOUNG, 2012). Este campo em rápido desenvolvimento apresenta diversas questões éticas e desafios únicos no que concerne os métodos de coleta e análise de dados. Esse tutorial oferece um guia direto para extração e mineração de dados do Twitter, voltando-se principalmente para o texto dos Tweets, por meio da linguagem de programação R (R CORE TEAM, 2020) via os Twitter APIs. O código em R foi desenvolvido em Adams (2020), com base no pacote *rtweet* (KEARNEY, 2018), e resultou com sucesso em um *script* para compilação de *corpora*. Nesse guia, são discutidas limitações, problemas e soluções na nossa abordagem para a condução ética de pesquisa nessa rede social. Nossas preocupações éticas vão além daquilo com o que "concordamos" nos termos de uso e nas políticas de privacidade, isto é, argumentamos que seu conteúdo não abrange todas as questões a que pesquisadoras(es) devem responder. Ademais, nosso objetivo é demonstrar que utilizar o Twitter como uma fonte de dados não requer habilidades computacionais avançadas.

KEYWORDS

Data-Collection Methods; Social Media; Research Ethics.

PALAVRAS-CHAVE

Metodologia de Coleta de Dados; Rede Social; Ética em Pesquisa.

INTRODUCTION

The realm of social media presents a wide range of possibilities for linguistic research, which raises unique methodological challenges and ethical issues (D'ARCY; YOUNG, 2012). One of these challenges is the creation of guidelines, protocols, or standards for ethical conduction of social media research. This is due to various reasons, such as the distinctions between these digital platforms — for example, between Facebook, Instagram, and Twitter — and the specificities of their community standards, terms of use and privacy policies; the distinct use that virtual community members make of these platforms; and the variety of ethical questions that arise depending on the type of work being carried out and each context, as well as the norms that govern these virtual spaces.

This tutorial focuses on Twitter, a data-rich microblogging platform that was launched in 2006 and that reached 199 million daily active users in 2021. Our aim is to present a 'how-to' guide on harvesting Twitter data and compiling a corpus with R (R CORE TEAM, 2020). In doing so, we discuss limitations, problems, and solutions in our framework for conducting ethical research on this social networking site. Aside from corpora compilation, the R programming language is a free software environment that can be used for several computational tasks, such as statistical computing, graphics, among others (see BAAYEN, 2008; GRIES, 2009; LEVSHINA, 2015). Although this guide is not an introduction to R for linguists (see OUSHIRO, 2014) nor to data science (see WICKHAM; GROLEMUND, 2017) or to *tidyverse* (see WICKHAM *et al.*, 2019), it intends to show that collecting data via Twitter APIs is not as daunting and does not require advanced computational skills as it can initially seem.

Our *researchTwitter* script is a list of commands that provides several functions developed to extract and tidy Twitter data, all of which will be examined in detail in the following section. The *rtweet* package (KEARNEY, 2018), which allows a more approachable way to import data, was used conversely to the *twitteR* package (GENTRY, 2013), since the former package is up to date and actively maintained whereas the latter one is deprecated.

It is worth mentioning that this method design results from a variationist sociolinguistic project about stranded prepositions and syntactic variation in Brazilian Portuguese, that had to address the issue of data scarcity (ADAMS, 2020). This undergraduate research project focused a great deal on methods, which often goes unnoticed in some academic circles. Similarly to Schilling's considerations about sociolinguistic field methods (SCHILLING, 2013), we also evaluate that this other kind of methodology plays a crucial role in shaping our data, and, as a result, our findings and conclusions; hence the embrace of open access to our code (see EASTERBROOK, 2014; STODDEN, 2011), and the push towards open science — especially principles 'A' and 'R', which represent some degree of 'FAIRness' in our work (see WILKINSON *et al.*, 2016, p. 4-5).

A variety of approaches were taken to draw and reach this outline, so how we initially envisioned to obtain this text data thoroughly changed during the testing process. Moreover, this process allowed us to evaluate strengths and weaknesses of each approach (SCHILLING, 2013) and draw a more polished and precise methodological approach. The data-collection process in Adams (2020) resulted in a corpus of approximately ten million words, consisting of roughly 450,000 Tweets.

The last part of this tutorial offers a discussion on ethical issues that emerged when dealing with Twitter data and what strategies for data anonymization were developed to bypass these challenges (ADAMS, 2020). It approaches expectations of privacy (D'ARCY; YOUNG, 2012; ZIMMER, 2010) regarding modern technologies, the nature of informed consent, and the role of scholars when engaging in research within virtual social network platforms.

1. DATA HARVESTING

To compile the Twitter corpus through R (R CORE TEAM, 2020), a script was built based on the *rtweet* package (KEARNEY, 2018), that provides several functions designed to extract Twitter data. In outline, the script extracts Tweets through Twitter APIs[1], as demonstrated in Listing 1; then, as shown in Listing 2, it cleans the data to remove the variables that are lists, which are a type of object in R. If it is the first time running the script, it will also create a CSV text file where the new data from the following extractions will be attached to (Listings 3 and 4). The last part of this script contains a function that adds more data to the main file (Listing 4). All blocks of our R code have comments indicating what that specific command does as part of the script. Each comment line is identified by an initial "#" and ends in the same numbered line.

```
puxaTWEET <- function(){

  # Package library to extract the data
  biblio <- c("rtweet", "magrittr")

  # Installing and importing packages
  for(pacote in biblio){

    # Checking if packages are installed
    if( !(pacote %in% rownames(installed.packages())))
      install.packages(pacote)

    library(pacote, character.only = TRUE)
  }

```

[1] An application programming interface (API) is a set of routines or patterns to build software applications and integrate systems, as a bridge between applications to connect, communicate and share data.

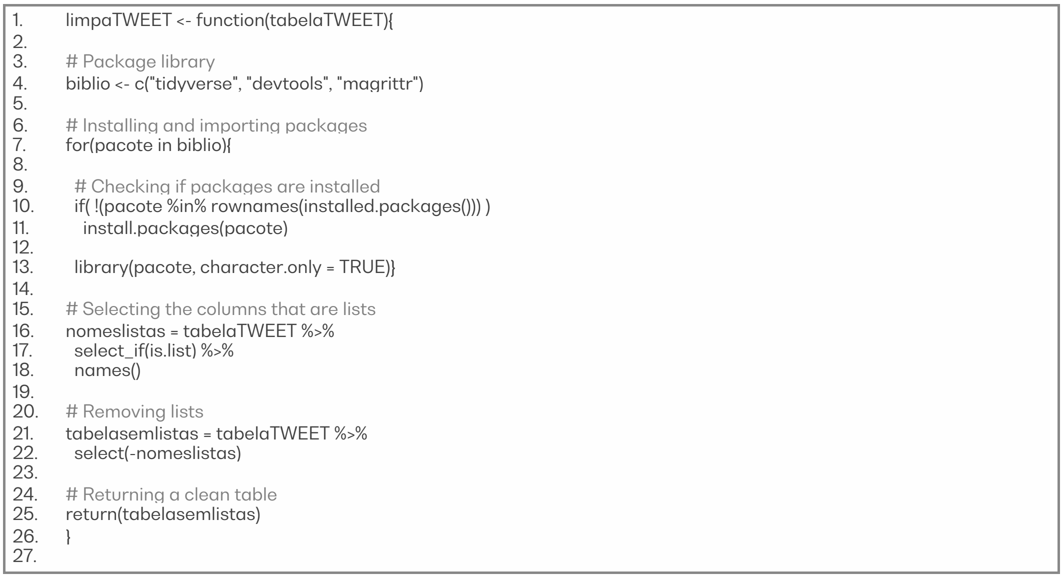
```
16.     # Access to the Twitter API
17.     create_token(
18.       app = "NameOfApplication",
19.       consumer_key <- "YourConsumerKey",
20.       consumer_secret <-"YourConsumerSecret",
21.       access_token <- "YourAccessToken",
22.       access_secret <- "YourAccessSecret"
23.      )
24.
25.     # Keys, tokens, and other information to complete the fields above can be found
26.     ## at Twitter's Developer Platform
27.
28.     # Looking up the coordinates to specify that only Tweets
29.     ## published in Brazil should be extracted
30.     br <- lookup_coords("brazil")
31.
32.     # Searching Tweets
33.     tweets <- search_tweets(
34.       q = "YourChoiceOfQuery",  # This is where the choice of search terms is indicated
35.       n = 1000,
36.       type = "recent",
37.       lang = "pt",
38.       include_rts = FALSE,
39.       geocode = br,
40.       max_id = NULL,
41.       parse = TRUE,
42.       token = NULL,
43.       retryonratelimit = TRUE,
44.       verbose = FALSE,
45.       show_col_types = FALSE
46.     )
47.
48.     # Returns the table with lists
49.     return(tweets)
50.   }
51.
```

Listing 1. R function that extracts Tweets

It is fundamental to highlight lines 16-23, regarding access to Twitter's API. The keys, tokens, and other credentials necessary in order to fill out those fields can be obtained through Twitter's Developer Platform, where applying for a developer account will be a requirement. There is an Academic Research application available, which gives access to higher levels of data than the standard application. As stated on the Developer Portal, the keys and tokens are unique identifiers that authenticate your request and a type of authorization to gain specific access to data, respectively.

Another essential aspect is line 34, since *q* establishes the query to be searched, which is used to filter and select Tweets to be returned. For further information on the arguments of the *search_tweets* function and on indicating multiple terms in the search query, see *rtweet*'s package description (KEARNEY, 2018).

The decision to remove variables that were lists, as shown in Listing 2, derived from this type of data structure being more versatile and not figuring out how to go around the complication of simplifying all these variables correctly — something similar to what the function *unlist(x)* does to produce vectors. However, most of these specific variables did not contain relevant metadata information for the purposes of the research in Adams (2020); additionally, we bypassed the few issues that arose.



```
limpaTWEET <- function(tabelaTWEET){

# Package library
biblio <- c("tidyverse", "devtools", "magrittr")

# Installing and importing packages
for(pacote in biblio){

  # Checking if packages are installed
  if( !(pacote %in% rownames(installed.packages())) )
    install.packages(pacote)

  library(pacote, character.only = TRUE)}

# Selecting the columns that are lists
nomeslistas = tabelaTWEET %>%
  select_if(is.list) %>%
  names()

# Removing lists
tabelasemlistas = tabelaTWEET %>%
  select(-nomeslistas)

# Returning a clean table
return(tabelasemlistas)
}

```

Listing 2. R function that removes lists from the object bound to tweets.

Since removing particular variables that were lists meant we would lose information about the Tweets' language codes and location, such as coordinates and place, we included another *rtweet* function, that is, *lookup_coords( )*, which looks up latitude/longitude coordinate information for a specified location (KEARNEY, 2018, p. 36). This required a valid Google Maps API key, that can be obtained through the Google Cloud Platform Console (see KAHLE; WICKHAM, 2013).[2] Despite removing the language variable from the extracted data and in spite of not being able to fully guarantee all extracted Tweets are from native speakers of Brazilian Portuguese, it is important to clarify that by structuring the script this way, we still ensured it drew out Tweets written in Portuguese and published in Brazil (see Listing 1, lines 28-30 and line 39).

Further, Listing 3 elucidates the R function that creates the main file where the Tweets from all temporary tables with the extractions will be attached onto. One important part of this function is the selection of the first line of the data frame with *slice(1)*, where the variables' names are placed. This way, the columns remain named correctly according to those variables, such as USER_ID, CREATED_AT, and TEXT.

[2] If this type of information is not relevant for a certain research topic, lines 30 and 39 from Listing 1 can be commented out without any issue. This also indicates to the importance of reading packages' descriptions: modifications can be made in order to include or remove arguments and functions, which allows the assembly of something more suitable for a different research design.

```
This function should only be executed once, because it creates
 ## the file where all new Tweets will be attached onto later
 criaTABELA_principal <- function(){

 tabelaTWEET_dia <- puxaTWEET() # Extracts Tweets

 # Removes lists
 tabelaTWEET_dia <- tabelaTWEET_dia %>%
   limpaTWEET()

 # Selecting only one line in order to keep the columns
 ## as they are when the file is saved
 tabelaTWEET_dia <- tabelaTWEET_dia %>%
   slice(1)

 # Saving the file
 tabelaTWEET_dia %>%
   write_csv("tabelaTWEET_principal.csv")
 }

```

Listing 3. R function that creates the main file.

Moreover, Listing 4 shows how the main file is updated with the output of each search through temporary tables. The rows from the temporary tables are bound to the main file, then line 50 indicates to the removal of duplicate rows according to the TEXT variable, which is the Tweet's text, whilst preserving the remaining data. It is possible to specify that duplicate rows are removed according to a different variable.

```
1.    atualizaTWEETS_principal <- function(){
2.
3.    # Package library
4.    biblio <- c("tidyverse", "devtools", "magrittr")
5.
6.    # Installing and importing packages
7.    for(pacote in biblio){
8.
9.      # Checking if packages are installed
10.     if( !(pacote %in% rownames(installed.packages())) )
11.       install.packages(pacote)
12.
13.     library(pacote, character.only = TRUE)}
14.
15.   # First it lists all files in your directory
16.   ## If tabelaTWEET_principal exists, executes the next code;
17.   ## if not, it does not execute the code and moves to another function
18.
19.   arquivos_PASTA <- list.files()
20.
21.   existeTABELA <- arquivos_PASTA %>%
22.     str_detect("tabelaTWEET_principal.csv") %>%
23.     sum()
24.
25.   if(!existeTABELA) criaTABELA_principal()
26.
27.   tabelaTWEET_novo <- puxaTWEET() # Extracting new Tweets
28.
29.   # Removing lists from file with new Tweets
30.   tabelaTWEET_novo <- tabelaTWEET_novo %>%
31.     limpaTWEET()
32.
33.   # Saving the file with new Tweets and reading it again,
34.   ## in order to maintain compatibility between columns
35.   tabelaTWEET_novo %>%
36.     write_csv("tabelaTEMPORARIA.csv")
37.
38.   tabelaTWEET_novo <- read_csv("tabelaTEMPORARIA.csv")
39.
40.   # Deleting the temporary table
41.   file.remove("tabelaTEMPORARIA.csv")
42.
43.    # Reading the main Tweets file to then bind the Tweets from
44.    ## new extractions to it
45.    tabelaTWEET_principal <- read_csv("tabelaTWEET_principal.csv")
46.
47.    # Binding both data frames and removing duplicate rows
48.    tabelaTWEET_principal <- tabelaTWEET_principal %>%
49.      bind_rows(tabelaTWEET_novo) %>%
50.      distinct(text, .keep_all = TRUE)
51.
52.    # Saving the updated file
53.    tabelaTWEET_principal %>%
54.      write_csv("tabelaTWEET_principal.csv")
55.
56.    # To remove all objects present in the specified working environment, uncomment:
57.    ### rm(list = ls()) ###
58.
59.  }
60.
```

Listing 4. R function that attaches new Tweets to the main file.

As a result, our CSV file contained one header row and the following variables as columns: TEXT, SCREEN_NAME, CREATED_AT, SOURCE, DISPLAY_TEXT_WIDTH, REPLY_TO_SCREEN_NAME, IS_QUOTE, PLACE_NAME, PLACE_TYPE, LOCATION, FOLLOWERS_COUNT, FRIENDS_COUNT, and FAVOURITES_COUNT. The remaining rows are all of extracted Tweets and their information.

Following the overview of these four functions that form our *researchTwitter* script, we highlight precisely what a researcher would have to do to reach their own CSV main file. That is, firstly one must open the script — preferably in a software application like RStudio —, insert their information in *create_token( )* (Listing 1, lines 16-23), and run each function separately, so that these objects are saved and made available in the environment. Whenever the query argument (Listing 1, line 34) or any other part of the functions are changed, it is necessary to save the script file and run this command line (Listing 5):

```
source("researchTwitter.R")
```

Listing 5. R command line to be run after modifying the *researchTwitter* script.

Finally, after loading all four functions, *atualizaTWEETS_principal( )* can be run several times — every single time it will call *puxaTWEET* and *limpaTWEET* to, respectively, extract more Tweets and remove lists. In other words, since all three other functions are embedded in *atualizaTWEETS_principal( )* (Listing 4, lines 15-25, 27, and 29-31), there is no need to run them individually as command lines after they have already been saved in the environment. If the working directory is the same and *criaTABELA_principal( )* is used as a command line after the start of a corpus compilation process, a user's main file will be overwritten. It is important to stress that there is a rate limit for requests under a specific time interval and the command line in Listing 5 has to be run in case of, for example, modifications in the search query.

2. ETHICS OF SOCIAL MEDIA RESEARCH

Initially, we also intended to build a corpus of Facebook data (ADAMS, 2020), but recent changes in Facebook Platform Terms and Developer Policies restricted what could be done with the *Rfacebook* package (BARBERA *et al.*, 2017) — something similar to the *rtweet* package (KEARNEY, 2018) —, thus we used Twitter as our primary source of data. Recent data scandals, such as the breach involving Facebook and Cambridge Analytica, have led social media platforms to review and limit what type of data can be extracted through their APIs. Along with advances in facial recognition, fingerprint sensors, tracking and other emerging technologies, it is even more crucial and imperative to openly (re)discuss our research conducts and practices on virtual platforms.

According to Sobo and De Munck, "[r]esearches are also expected to ensure that participants' rights and interests are always protected" (SOBO; DE MUNCK, 1998, p. 23) and, especially concerning linguists, "[r]esearch on language always involve human agents" (ECKERT, 2014, p. 13). Here we argue that it is not enough to guide ourselves by what is stated on terms of use and privacy policies, which we "consent to" before beginning to use any application, technological device, or social media. Assuming that those terms focus greatly on legal issues and that ethics are not necessarily taken into account in the foundation of companies' policies, we advocate that researchers need to carefully and critically consider the overlapping ethical and methodological issues in social media research. It is arguable how much academic research can rely on what terms and policies cover considering "a survey once found as few as 18 per cent of users may actually read terms conditions agreements" (ZIMMER; PROFERES, 2014 *apud* AHMED *et al.*, 2017, p. 18), which brings into question if agreeing to those conditions actually constitutes informed consent.

Beyond institutional and Research Ethics Committees requirements, "[i]n all aspects of research, transparency is critical" (D'ARCY; YOUNG, 2012, p. 538), which all researchers should support, encourage and respect in regards to doing science. Directly related to the notion of transparency, Eckert (2014) offers a definition of what embodies free and informed consent, one of the principles of ethical research: "consent should not be a matter of getting a signature on paper, but the establishment of an informed working relationship" (ECKERT, 2014, p. 14). Adding that "[i]nformed consent assumes the ability to grasp the implications of participation in the research and to make decisions for oneself" (ECKERT, 2014, p. 16-17).

It is necessary to state that no information that was not already public is collected through the script made with the *rtweet* package functions (KEARNEY, 2018). In other words, we do not have access to private information from the users whose Tweets are part of our Twitter corpus. No participants were directly approached by the researchers. Furthermore, we took other measures to ensure the participants' privacy, which meant removing identifiable information from the corpus data set, as explained in detail subsequently (ADAMS, 2020).

As disclosed in Twitter's Privacy Policy and Help Center, Tweets are searchable by anyone around the globe through search engines and other third parties, which can retain copies of public information, even if that is deleted from Twitter services or if an account is deactivated. Users are given tools and settings to object, restrict or withdraw consent where applicable for the use of data provided to Twitter; they can also choose to share additional information, like e-mail addresses, phone numbers, address book contacts and public profiles.

When debating the possibility of acquiring written and informed consent from each one of the subjects that produced the vast amount of Tweets that are part of our sample, it was clear that an undergraduate research project (ADAMS, 2020) would not be able to reach out to all those users with consent forms (see AHMED *et al.*, 2017) for several different reasons. In consideration of the aspects mentioned above and specially to address the issue of written and informed consent being impracticable, it was decided that certain measures would be taken for data anonymization regarding the dissemination of research findings, for example in giving oral presentations or publishing in journals. When linguistic data from Tweets part of our corpus is cited as examples of the structures under analysis, any information that could lead to pinpointing a user has been erased, like profile pictures, usernames, and display names. No screenshots of any posts, comments or Tweets are used as well.

Moreover, users could be identified by the search of precise strings from Tweets — for example, by enclosing an entire phrase in quotation marks on Google Search. To avoid this from occurring, no exact-quoted content is used in any material, by cutting excerpts or by the substitution of secondary lexical items for synonyms; this way, we aim to avoid reverse searches and to maintain users' identities anonymous. This strategy results in not even the only people with access to the corpus being able to trace the original Tweet afterwards.

Also, there is no use of data with controversial content, like any form of intolerance, discrimination or prejudice regarding gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, ethnicity, or religion. This type of content was not removed from the corpus; if a part of text was coded in the process of analyzing variants of our variable of interest (ADAMS, 2020), those Tweets were not considered in any way as potential examples in material with description of our research findings. This decision was made considering that sensitive content, such as political discourses, could lead to someone attempting to find the author of a certain Tweet, which would compromise their privacy and break their anonymity.

Although an ideal solution would have been to reach out to those thousands of Twitter users individually with consent forms, the research project (ADAMS, 2020) took the ethical standpoint of not disclosing our Twitter corpus, among other measures previously described, to tackle the ethical and methodological challenges of doing research in a social networking site. In agreement with D'Arcy and Young (2012), in spite of the fact that users are tweeting in a public space, the "content is networked between actors with different privacy expectations" (D'ARCY; YOUNG, 2012, p. 542). As scholars, this expands our concerns over consent and privacy.

At last, this project aimed to contribute to a critical open dialogue between researchers regarding the emerging and unique challenges of engaging in research within rapidly evolving online social network platforms: "[t]hese include challenges to the traditional nature of consent, properly identifying and respecting expectations of privacy on social network sites, developing sufficient strategies for data anonymization [...]" (ZIMMER, 2010, p. 323). This shift from physical to virtual spaces also requires that institutional review boards have a better understanding of these other spheres (see D'ARCY; YOUNG, 2012; ZIMMER, 2010), to make headway and avoid potential shortcomings when overseeing research projects that retrieve data from social media.

3. ACKNOWLEDGMENTS

The development of the data-collection method detailed here was possible due to São Paulo Research Foundation's grant (Process number 2018/24511-8) and Livia Oushiro's academic advising. Thanks to Gabriel Catani for his (online) feedback on this tutorial and to Natasha Mourão for reviewing the manuscript.

REFERENCES

ADAMS, Julia Bahia. *Um estudo sobre preposition stranding e orphaning em falantes de português brasileiro*. Relatório final do processo no. 18/24511-8, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), 2020.

AHMED, Wasim; BATH, Peter; DEMARTINI, Gianluca. Chapter 4 Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In: WOODFIELD, Kandy, (ed.) *The Ethics of Online Research*. *Advances in Research Ethics and Integrity* (2). Emerald, pp. 79-107, 2017. Disponível em: http://eprints.whiterose.ac.uk/126729/. Acesso em: 30 jun. 2021.

BAAYEN, R. Harald. *Analyzing Linguistic Data*. New York: Cambridge University Press, 2008.

BARBERA, Pablo; PICCIRILLI, Michael; GEISLER, Andrew; VAN ATTEVELD, Wouter. *Rfacebook: Access to Facebook API via R*. Comprehensive R Archive Network, 2017. Disponível em: https://CRAN.R-project.org/package=Rfacebook. Acesso em: 11 mar. 2019.

D'ARCY, Alexandra; YOUNG, Taylor Marie. Ethics and social media: Implications for sociolinguistics in the networked public. *Journal of Sociolinguistics*, v. 16, n. 4, p. 532-546, set. 2012. DOI 10.1111/j.1467-9841.2012.00543.x.

EASTERBROOK, Steve. Open code for open science?. *Nature Geoscience*, v. 7, n. 11, p. 779-781, nov. 2014. DOI 10.1038/ngeo2283.

ECKERT, Penelope. "Ethics in linguistic research". *In:* PODESVA, Robert J.; SHARMA, Deyvani. (Eds.). *Research Methods in Linguistics*. New York: Cambridge University Press, 2014. p. 11-26.

GENTRY, Jeff. *twitteR: R Based Twitter Client*. Comprehensive R Archive Network, 2013. Disponível em: https://CRAN.R-project.org/package=twitteR. Acesso em: 4 jan. 2019.

GRIES, Stefan Thomas. *Quantitative corpus linguistics with R: a practical introduction*. 1. ed. New York: Routledge, 2009.

KAHLE, David; WICKHAM, Hadley. ggmap: Spatial Visualization with ggplot2. *The R Journal*, v. 5, n. 1, p. 144-161, 2013. Disponível em: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf. Acesso em: 30 mar. 2018.

KEARNEY, Michael W. *rtweet: Collecting Twitter data*. Comprehensive R Archive Network, 2018. DOI 10.5281/zenodo.2528481.

LEVSHINA, Natalia. *How to do Linguistics with R: Data exploration and statistical analysis*. 1. ed. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2015.

OUSHIRO, Livia. "TRATAMENTO DE DADOS COM O R PARA ANÁLISES SOCIOLINGUÍSTICAS". *In:* FREITAG, Raquel Meister Ko. (Orgs.). *Metodologia de Coleta e Manipulação de Dados em Sociolinguística*. São Paulo: Editora Edgard Blücher, 2014. p. 134-177. DOI 10.5151/BlucherOA-MCMDS-10cap.

OUSHIRO, Livia. *Identidade na pluralidade: avaliação, produção e percepção linguística na cidade de São Paulo*. 2015. Dissertação (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015. DOI 10.11606/T.8.2015.tde-15062015-104952.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. Disponível em: https://www.R-project.org.

SCHILLING, Natalie. *Sociolinguistic Fieldwork*. 1. ed. New York: Cambridge University Press, 2013.

SOBO, Elisa J.; DE MUNCK, Victor C. "The Forest of Methods". *In:* DE MUNCK, Victor C.; SOBO, Elisa J. (Eds.) *Using Methods in the Field**: a practical introduction and casebook***. Walnut Creek: Altamira Press, 1998. p. 13-37.

STODDEN, Victoria. Trust Your Science? Open Your Data and Code. *Amstat News*, Alexandria, 1 jul. 2011. Disponível em: https://magazine.amstat.org/blog/2011/07/01/trust-your-science/. Acesso em: 22 set. 2021.

WICKHAM, Hadley; GROLEMUND, Garrett. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1. ed. Sebastopol: O'Reilly Media, Inc., 2017.

WICKHAM, Hadley *et al*. Welcome to the tidyverse. *Journal of Open Source Software*, v. 4, n. 43, p. 1686, 2019. DOI 10.21105/joss.01686.

WILKINSON, Mark D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, v. 3, n. 1, 2016. DOI 10.1038/sdata.2016.18.

ZIMMER, Michael. "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology*, v. 12, n. 4, pp. 313-325, 2010. DOI 10.1007/s10676-010-9227-5.