Changes in Reading Behaviour for Dutch and Flemish Book Readers during Lockdown

Adapted from: Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.

0. Introduction

This research examined the effects of the pandamic on usage of the website Goodreads by members from the Dutch and Flemish book readers group. By scraping user data from this Goodreads user group, we aimed to find out if the pandameic and its lockdowns had an effect on the number of books that the group members read and added to their bookshelfs.

This documentation solely focuseess on answering the questions of the Datasheet for Datasets. In the 'anylsis' pdf file that is also presented together with this documentation (src\reporting\analysis.pdf), interested readers can find the actual analysis of the obtained dataset, in which we tried to find an answer to the main question of this research.

Due to the recent nature of the events, publicly available databases are still scarce. Therefore, we provide a source of valid data ready to be used in the academic field and create managerial insights relevant to the book industry's multiple intervenients. To the extent of our knowledge, there are no other datasets publicly available that examine the reading behaviour of Dutch and Flemish book readers during the pandemic.

The dataset created in our research could, for example, be interesting for (Dutch / Flemish) book publishers wanting to investigate what influence a lockdown could have on the behaviour of their customers, or, for example, for academic / governmental reserarchers that want to investigate the impact of Covid-19 lockdowns on the behaviour of the affected population.

1. Motivation

1.1 *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The dataset was developed to study the impact of Covid-19 lockdown on reading habits, focusing on Dutch and Flemish books readers. Due to social distance restrictions, people could not pursue their usual hobbies. Therefore, reading behavior could have changed during the lockdown period since people had more free time, which could be used to read. With the primary purpose to understand if there was an increase (or not) of books read during the outbreak, we offer a valid dataset for Dutch and Flemish books read at the daily level during the period of 2019, 2020, and 2021 on Goodreads.com. The high granularity of the data allows a further analysis at a weekly or even monthly level to understand if there were fluctuations in specific periods according to the researcher's needs.

1.2 *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

The current dataset was developed students of the Online and Data Collection Management course integrated into the Master's program of Marketing Analytics at Tilburg University.

1.3 *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

There is no funding behind the development of the current project.

1.4 *Motivation for choice of website*

The choice set was composed by wikibooks.org, Blinkist, bookdepository.com, amazon books, thestorygraph.com,

* https://arxiv.org/abs/1803.09010

wolnelektury.pl, and goodreads.com. All these possible resources provided individual book information (e.g., author name). Whereas most platforms (i.e. wikibooks.org, Blinkist, bookdepository.com and amazon books) show up as a marketplace, wolnelektury.pl offers an assortment of free books to download. The structure of these websites makes them less useful to investigate actual reading behaviour. Moreover, Blinkist is a mobile app that would not technically fit our objectives.Hence, our final choice set is between thestorygraph.com and goodreads.com since both added the social media component beyond the book information considered default criteria. Because Goodreads has a more extensive market coverage and access to more information either at user and book level, we choose to go for Goodreads. However, one of the major draw backs on goodreads.com in relation to thestorygraph.com is the fact that the book gender is defined by the users, consequently giving an absolute result.

2. Composition

2.1 *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

There are two main entities examined: all members in Goodreads Dutch & Flemish group and the respective books read. On 14-10-2021, there were 1.355 members, of which only 685 were active users between 2019 and 2021. Those 685 members read 55.844 books between 2019 and 2021. All members with a public profile can be accessed; therefore, their respective information can be retrieved.

For each member, the user id was retrieved, books read from 2019 until 2021, and the respective rating provided by the user. Secondly, for each book, the following information is presented: book title, author's name, book URL, when the user started to read and finished the book, and the average rating of the book.

2.2 *How many instances are there in total (of each type, if appropriate)?*

The data set contains 55.844 books read by 685 users.

2.3 *Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

Because the purpose of this study is to investigate the impact of the Covid-19 lockdown on reading habits amongst Dutch and Flemish books readers, we narrow our scope towards the research period from 2019 until 2021 (where 2019 serves as ‘control’ year). The first member of the "Netherlands & Flanders" group joined in July 2007. Therefore, many books were not scraped since there is no fit with our research purpose. Overall, the dataset contains all possible instances regarding the research period.

2.4 *What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.*

For each instance, we have access to a book read by a particular user. Simultaneously, each instance state when the user (user-id) started and finished the book (%Y-%mm-%dd format), the respective rating provided by the user, and the overall rating of the book on Goodreads.com. Finally, the book's title, URL, and the author's name are also given. Although we have 55.844 books but only 685 members, it shows that each user can be associated with one or more books.

2.5 *Is there a label or target associated with each instance? If so, please provide a description.*

Each user is labeled by a user-id and each book is labeled by its title.

2.6 *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

The only missing information occurs on the variable "date.read" which represents when the user finished the book. The information is either missing because the user did not provide the date when they finished the book (or it was never finished), or the information is incomplete (the user does not provide the exact day). Both date variables (date.read and date.added) are in a "%Y-%mm-%dd" format. Since we look at changes in reading habits at the weekly level, an absence of an exact day makes the instance obsolete. Consequently, for 19.647 instances we can not examine the correlation between reading period and the lockdown, but are limited to looking at the 'date added' variable (which represents the moment the reader added the book to his/hers book shelf).

2.7 *Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

The book URL on Goodreads.com with the respective overall book rating, author's name and book title. Regarding the users, there is a relationship between the individual Goodreads user id and books read (via books URL). For each member it is also possible to know when the user started and finished the book and the rating provided.

2.8 *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

There are no recommended data splits.

2.9 *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The data is self-contained.

2.10 *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

There is no data considered confidential. All the data retrieved belong to active users between 2019 and 2021 with a publicly available profile, that are only referred to by their user id (and not by name).

2.11 *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

The data set does not contain any offensive, insulting, threatening, or causer of anxiety.

2.12 *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

Yes, the dataset can relate with people via their user id, which leads to their Goodreads personal page where it is possible to view the information the individual decides to make public.

2.13 *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* The dataset does not identify any subpopulations.

2.14 *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

It is possible to identify individuals indirectly. The data set provides the user id, with which it is possible to connect with the personal user page on Goodreads, which is possible to infer in the majority of the cases the gender, country/state of origin, and in some particular cases, even the direct link to other social media platforms. However, only if the particular

user chose to share this information on its Goodreads user page.

2.15 *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

The dataset does not contain any data that might be considered sensitive.

3. Collection Process

3.1 *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-ofspeech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

All acquired data was directly observable on the Goodreads website.

3.2 *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

The data was collected by a webscraping program written in the probramming language Python. By comparing the collected data of a randomly selected number of users to the data that is visually present on their goodreads book shelf, it was validated if the scraper actually scraped all correct information.

3.3 *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

The data is not a sample of a larger set, it contains all available data from the Goodreads group "Dutch and Flemish book readers".

3.4 *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

Only the students, who are students from the MSc Marketing Analytyics, where involved in the data collection process. They were only compensated for their work by gaining in depth knowledge on web-scraping and did not receive any financial compensation.

3.5 *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

The final dataset was collected in about 8 hours. This was done from 00:00 to +/- 8:00 AM (Dutch time), in which, assumingly, the majority of the users of the Dutch and Flemish group were not activily changing their profiles. However, we do not certainly know if there did not change anything in the composition of the group's users or the users bookshelfs. Nonetheless, the scraper was run multiple different nights, to see if the datasets would significantly differ from day to day, and it was observed that the scraped entitities typically differed less than 1% between the different scraping moments.

3.6 *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

There were no ethical review processes conducted.

3.7 *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The dataset relates to people.

3.8 *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

The data of the individuals was collected indirectly via the website Goodreads (however, the data on this website was directly provided by the individual users themselves).

3.9 *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

The users were not notified that their data was being collected.

3.10 *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

The individuals were not asked for consent to scrape their data.

3.11 *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

There was no consent obtained from the individuals.

3.12 *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

No analysis on the potential impact of the dataset was conducted.

4. Preprocessing, cleaning, labeling

4.1 *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-ofspeech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

Yes, the following preporcessing / cleaning procedures were conducted:

1. The user's user id's were extracted from their bookshelve url's and these were used to identify them later on in the dataset.
2. Instances with private bookshelfs were removed from the list of 'users to scrape'.

3. Instances that did not have any books on their bookshelfs were removed from the list of 'users to scrape'.

4.2 *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

No, the "raw" data was not saved in addition to the preprocessed data.

4.3 *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

The code used to preporcess the data is incorporated in the 'collect.py' file that is provided together with this documentation (data package/src/collection/collect.py).

5. Uses

5.1 *Has the dataset been used for any tasks already? If so, please provide a description.*

The dataset has not been used for any tasks beyond the present statistical analysis conducted on R Studio.

5.2 *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

No, there is no repository that links to papers or systems that use the dataset.

5.3 *What (other) tasks could the dataset be used for?*

The dataset created in our research could, for example, be used by (Dutch / Flemish) book publishers wanting to investigate what influence a lockdown could have on the behaviour of their customers, or, for example, it could be used by academic / governmental reserarchers that want to investigate the impact of Covid-19 lockdowns on the behaviour of the affected population.

5.4 *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

There is nothing that might impact future uses.

5.5 *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset can be used for any researcher/private or public entity interested in analyzing (Dutch or Flemish) consumer behavior changes, with a focus on literature habits, during the Covid-19 lockdown.