ODDPub Validation for SocArXiv and arXiv

SocArXiv

2019

Data

SocArXiv allows authors to input a link to their data source/repository upon submission of a pre-print (DOUBLE CHECK). This link can then be accessed via the API metadata. The presence of a data link was used as an indicator that a pre-print provides open data for the purposes of validating the ODDPub algorithm. When available, a data link is stored under the variable name "attributes.data_links". The data was manipulated using functions from the R package tidyverse [@tidyverse] to create a binary variable indicating data availability or lack thereof. We assume "attributes.data_links" to indicate the true availability of data for the purposes of validating the ODDPub algorithm. It is possible, however, that some authors failed to indicate their data availability in the proper field upon posting to SocArXiv, and thus some of the false positive may in fact be true positives. This was not verified directly in order to allow us to more easily consider the whole sample in our validation.

Against the data availability indicated by pre-print authors in our 2019 sample, the ODDPub algorithm performed with an accuracy of 93 percent, a sensitivity of 52 percent, and a specificity of 94 percent. In our 2020 and 2021 sample, the algorithm performed with an accuracy of 79 percent, a sensitivity of 29 percent, and a specificity of 92 percent. Specific predictions are broken down in Table @ref(tab:confusion-matrix-2019) and Table @ref(tab:confusion-matrix-2020).

It is unclear the precise inclusion criteria for data submitted to the data link field. It is possible that some of the links provided lead to data sets that are publicly available for reuse, which would not constitute "open data" by the ODDPub algorithm's definition, in which case the accuracy could potentially be higher in reality than 93 percent and 79 percent in the samples considered.

Table 1: ODDPub predictions for open data compared with data links provided by authors, 2019 sample

ODDPub algorithm	Data linked	No data linked
Open data detected	11	72
No open data detected	10	1107

Table 2: ODDPub prediction accuracy, 2019 sample

Metric	Value
Accuracy	0.93
Sensitivity	0.52
Specificity	0.94

Code

2020/2021

Code was not manually verified for COVID-19 related papers.

Table 3: ODDPub predictions for open data compared with data links provided by authors, COVID-19-related pre-prints sample

ODDPub algorithm	Data linked	No data linked
Open data detected	28	31
No open data detected	69	350

Table 4: ODDPub prediction accuracy, COVID-19-related pre-prints sample

Metric	Value
Accuracy	0.79
Sensitivity	0.29
Specificity	0.92

ArXiv

2019

We verified the accuracy of the ODDPub algorithm on a subset of our analyzed pre-prints from 2019 from arXiv. We took a simple random sample of 100 papers. In the original validation process, the annotators stratified by detection status prior to sampling to ensure relatively high representation of papers where open data/code was detected. Since the major concern for our manual verification is potential false negatives, this biased representation was unnecessary. Open data and code status were verified first via the "Code & Data" tab on each pre-print's page on the arXiv website, and then via checking for an explicit data availability section within the PDF and keyword search. Results were recorded manually in Excel (arxiv_2019_validation_sample.csv). This mimics the procedure outlined for the original validation of ODDPub.

Many of the pre-prints in arXiv did not use data or code, namely those from pure mathematics and physics. There were also several that reused other publicly or privately available data sets, and regardless of whether or not they were shared alongside the paper, these do not count as open data according to ODDPub standards.

```
## # A tibble: 2 x 3
##   Predicted             `Data available` `No data available`
##   <chr>                            <int>               <int>
## 1 Open data detected                   2                   2
## 2 No open data detected                1                  95

## # A tibble: 2 x 3
##   Predicted             `Code available` `No code available`
##   <chr>                            <int>               <int>
## 1 Open code detected                   3                   2
## 2 No open code detected                2                  93
```

```
## # A tibble: 3 x 2
##   Metric      Value
##   <chr>       <dbl>
## 1 Accuracy    0.97
## 2 Sensitivity 0.667
## 3 Specificity 0.979
```

```
## # A tibble: 3 x 2
##   Metric      Value
##   <chr>       <dbl>
## 1 Accuracy    0.96
## 2 Sensitivity 0.6
## 3 Specificity 0.979
```