Summer Ai

Runting Shao

Ningji Shen

Chang Xu

Joyce Zhang

DATA 590 AU-21

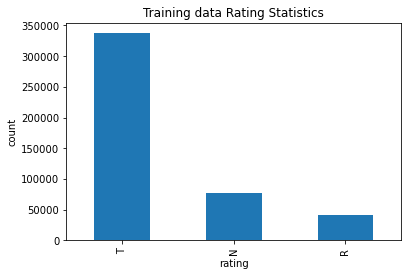
December 1, 2021

Data Exploration

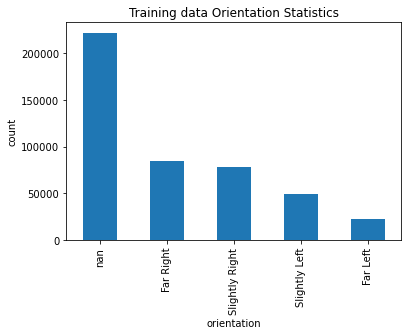
Our data consists of two parts: the original datasets and the datasets containing domains’ information extracted using our web crawler. The original datasets contain information about social media posts related to misinformation. There is a training dataset and a testing dataset, both containing 455,909 rows and the following attributes:

* id: An unique identifier for the instance
* date: The date and time when the social media post was posted
* headline: The headline of the news page to which the social media post links
* message: The message attached to the social media post, usually a comment of the new in the news link
* link: The hyperlink to the news page
* domain: The domain of the news page
* rating: Whether the domain is labeled as truthful(T), untruthful(N), or repeat defender(R)
* orientation: The political polarization of the post
* sourceEchochamber: Whether the news or the news site is conservative

All attributes are categorical. We found significant amounts of missing values only in message and orientation. The missing values in message indicate there is no message attached to the social media post. The missing values in orientation indicate a neutral orientation. Therefore we do not need additional cleaning steps for the missing values. The statistics for the rating, orientation and sourceEchochamber are the following:



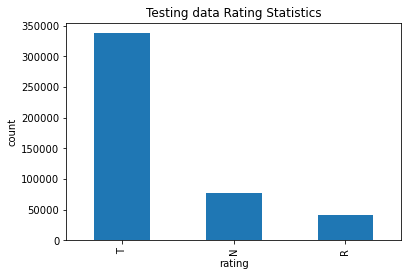
|  |  |
| --- | --- |
| **Training data - rating** | |
| Rating | Count |
| T | 337,414 |
| N | 77,562 |
| R | 40,933 |



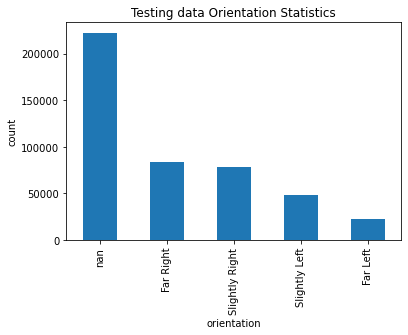
|  |  |
| --- | --- |
| **Training data - orientation** | |
| Orientation | Count |
| Far Left | 22,754 |
| Far Right | 84,195 |
| Slightly Left | 49,020 |
| Slightly Right | 77,878 |
| NaN | 22,2062 |



|  |  |
| --- | --- |
| **Training data - sourceEchochamber** | |
| sourceEchochamber | Count |
| Conservative | 207,380 |
| Liberal | 248,529 |



|  |  |
| --- | --- |
| **Testing data - rating** | |
| Rating | Count |
| T | 337,670 |
| N | 77,044 |
| R | 41,195 |



|  |  |
| --- | --- |
| **Testing data - orientation** | |
| Orientation | Count |
| Far Left | 22,705 |
| Far Right | 84,086 |
| Slightly Left | 48,506 |
| Slightly Right | 78,107 |
| NaN | 222,505 |



|  |  |
| --- | --- |
| **Testing data - sourceEchochamber** | |
| sourceEchochamber | Count |
| Conservative | 207,899 |
| Liberal | 248,010 |

The correlations of the rating of misinformation and other attributes are what we want to study in this project, and thus we are going to discuss this in the future. Meanwhile, we did not observe any outliers since our datasets contain mainly categorical variables. The datasets have potential bias. The rating, orientation, and sourceEchochamber are produced using existing models from The Carter Center Democracy Program, while the data used to train the models are classified by humans. Thus human bias may involve while determining the rating, orientation, and sourceEchochamber for the data used to train the models.

The second part of our data contains information we extracted from the domains of the news pages in the original datasets using our web crawler. Our web crawler extracts two files from each domain in our original dataset: ads.txt and sellers.json, both introduced by IAB Technology Laboratory. Ads.txt provides a list of digital sellers who are authorized to sell a given set of impressions (publisher’s advertising inventory) via a bid request for each domain, and sellers.json stores the identity of the final seller of a bid request assuming that they are ads.txt authorized. We stored the record fo ads.txt for all the domains into adsDf.csv, which contains 967,406 rows and the following attributes:

* certificationId: The current certification authority used in the Trustworthy Accountability Group (TAG).
* ad\_domain: The advertising system, like an SSP or exchange, that the buyers connect to.
* Count of ad\_domains
* pubId: The identifier associated with the seller or reseller account within the advertising system listed in the domains
* domain: The platform which carries the advertisement. (change “source” back to “domain”)
* type: Whether the inventory is being sold directly by the owner or through a reseller.
  + "DIRECT" indicates that the publisher (content owner) directly controls the account indicated in pubId on the system in the domain.
  + "RESELLER" indicates that the publisher has authorized another entity to control the account indicated in pubId and resell their ad space via the system in domains.

We stored the record for sellers.json for all the domains into sellerDf.csv, which contains 4,431 rows and the following attributes:

* seller\_id: The identifier associated with the seller or reseller account within the advertising
* name: The name of the company (the legal entity) that is paid for inventory that is transacted under the given seller\_id.
* seller\_type: An enumeration of the type of account, either “PUBLISHER”, “INTERMEDIARY”, or “BOTH”.
  + A value of "PUBLISHER" indicates that the inventory sold through this account is on a site, app, or other medium owned by the named entity and the advertising system pays them directly.
  + A value of “INTERMEDIARY" indicates that the inventory sold through this account is not owned by the named entity or the advertising system does not pay them directly.
  + "BOTH" indicates that both types of inventory are transacted by this seller. Note that this field should be treated as case insensitive when interpreting the data.
* ad\_seller\_domain: The business domain name of the company (the legal entity) that is paid for inventory that is transacted under the given seller\_id.
* Count of ad\_seller\_domain
* is\_confidential: Indicates whether the identity of the seller is confidential.
* domain: The platform which carries the advertisement for this instance.
* comment: Any helpful description for this inventory.
* Count of all Ads on a domain
* Count of Seller id on a domain
* highest\_seller\_id - The seller id for the seller selling highest number of adds in a domain
* highest\_seller\_count – Count of ads from the highest seller id

Both datasets contain only categorical data. We observed missing values in many attributes. The detailed statistics about the missing values are the following:

|  |  |
| --- | --- |
| **adsDf.csv** | |
| Attribute | Missing counts |
| certificationId | 279,305 |
| domains | 997 |
| pubId | 1,035 |
| source | 0 |
| type | 1,057 |

|  |  |
| --- | --- |
| **sellerDf.csv** | |
| Attribute | Missing counts |
| seller\_id | 0 |
| name | 32 |
| seller\_type | 0 |
| domain | 210 |
| is\_confidential | 4,399 |
| source | 0 |
| comment | 4,416 |

TASK:

* Upload data:
  + Combined clean version of original data csv
  + Sellers csv
  + Ad csv
* Make changes in yellow suggested above.
* Drop red fields
* Join Ad and Seller to master (original cleaned csv)
* Domain financial related terms:
  + Donation/Donate - 1
  + Shop/Shopping - 2
  + Free/No Cost - 3
  + Pay/Money - 4
  + Subscribe/Subscription - 5
  + Newsletter - 6