LIS 545 Term Paper

LIS 545 – Data Curation

Sam Solomon

Feb. 24, 2023

Sam Solomon
LIS 545
Data Curation I: Fundamentals
Feb. 24, 2023

Data and Metadata Profile

The data I examined are from "Yahoo! Answers Topic Classification" dataset, from the host site Zaggle, which uses the 10 largest main categories as a scope to gather data on questions and best answers posted on Yahoo! Answers. There are two data files, and both are in CSV format. Excel, or an equivalent program, is needed to open the dataset, but that appears to be the only software necessary to open and analyze the data. One file contains training data and another testing data.

The data comes from the Yahoo! Answers social network active from 2005-2021, in which users posted and replied to topical questions, and includes all questions and corresponding answers (totaling 4,483,032), and which represents the corpus up to October 25, 2007. Yahoo! Answers Research Alliance Webscope program allows researchers to obtain this dataset, *L6 - Yahoo! Answers Comprehensive Questions and Answers version 1.0 (multi part)*, and the group's website contains a brief description of the provenance. There does not appear to be a larger dataset with Yahoo! Answers' comprehensive questions and answers during the platform's entire history that is available publicly or for purchase. However, this data may exist for the organization's internal purposes and research.

According to the provenance metadata provided on Zaggle, the dataset is for approved non-commercial research purposes by recipients who have signed a Data Sharing Agreement with Yahoo!. However, the Zaggle webpage also includes metadata that identifies the licensing for "Yahoo! Answers Topic Classification" as public domain. It is unclear whether this small subset of data in Zaggle follows the guidelines described in the provenance note.

The key stakeholders of the data include the Yahoo! Company since the data obtained is from the company's platform. Users of Yahoo! Answers are stakeholders; although the data is anonymous, the questions and answers provided—particularly for classifications like "Health" and "Family & Relationships" may contain sensitive information or may simply be of a personal nature to the users. The researchers are stakeholders, as "researchers who practice open science benefit from increased citation rates, visibility, collaboration efficiency and ease of future work" (Gomes 2022). Sociology, information behavior, consumer behavior, AI and machine learning professionals are similarly all stakeholders, as they may find useful insights from the dataset. Studies that result from the dataset could hold significance to their professions.

The dataset comes with metadata on the webpage, including title, subtitle, summary/description of the data (available in an "about" section), and subject terms, which are linked. There is a link to download the dataset. Also listed on the webpage under the heading "Metadata" is the collaborator information, the "provenance" or a note on how the data was obtained and by who for what purpose. There are fields for authors, coverage, and DOI citation, but these are not filled out for this dataset. Metadata is provided on licensing, expected update frequency, and "usability" score. Code using the dataset is shared on the webpage, and this notebook has similar metadata displayed such as title and most recent update. In the CSV data

Sam Solomon
LIS 545
Data Curation I: Fundamentals
Feb. 24, 2023

file for the dataset, the question title, question content, and "best answer" content are further metadata elements. Class index is a metadata field in the data file as well, and the values (0-9?) for this element are based on the 10 topic classifications, listed in the about section:

- Society & Culture
- Science & Mathematics
- Health
- Education & Reference
- Computers & Internet
- Sports
- Business & Finance
- Entertainment & Music
- Family & Relationships
- Politics & Government

Despite the various metadata present, it is not very comprehensive. Some fields are available to input values, but those fields are left blank for this dataset. There is seemingly no metadata on the Yahoo! Answers users. Everything was anonymized through usernames on the Yahoo! Answers platform, but the dataset further does not have metadata on those users, their account life, number of previous posts/replies, etc. It may be helpful to have metadata on the other comments posted to the initial questions on Yahoo! Answers, rather than just the "best answer" chosen. The metadata also do not appear to be structured according to a particular metadata standard in the data file.

The current metadata could be enriched in multiple ways, for instance, to assist somebody unfamiliar with the data make use of the dataset for new purposes, or to improve users' ability to discover the dataset. Adding more linked data in general could serve both goals. For example, linked subjects in the metadata are lacking and do not do a full job of describing the data, but adding more tagged subjects could help users find the database through Kaggle and through search engines. "Online communities" is a subject tag available in Kaggle, which the dataset lacks. Additionally, tags can be created that are related to the topic classifications ('science and technology', etc.), which would encourage discoverability. Linked data could furthermore help someone unfamiliar with the dataset make use of the data in ways they may not have thought possible. Additional metadata should be created to show where the dataset was published, including the original paper for which the dataset was created, with links to those resources. These links would allow users to explore the possibilities for using the data.

The dataset was published as part of "Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28." This is stated in the about section clearly. There are other researchers who have utilized this dataset. The website PapersWithCode lists many related publications that use or refer to the dataset. I found this website by using Google and searching the dataset name, Yahoo! Answers Topic Classification. The website tracks papers that have used the dataset, and there is a total of 97 papers listed on the website that use the Yahoo! Answers dataset.

Sam Solomon
LIS 545
Data Curation I: Fundamentals
Feb. 24, 2023

Repository Profile

I chose the Social Computing Data Repository as a fitting repository to house the Yahoo! Answers Topic Classification dataset. This repository is a part of Arizona State University's Data Mining and Machine Learning Laboratory, and it hosts datasets from several different social media sites. The repository dictates a clear aim of contributing to research in machine learning, data mining, and social sciences. I chose this repository because Yahoo! Answers fits the profile of a social media network, and I had noted previously that many research projects that used the dataset were utilizing this resource for experimentation in AI and machine learning projects. Furthermore, most of the social media sites represented by datasets in the repository have some sort of blogging capacity. Yahoo! Answers can fit this description as a forum where exchanges occurred primarily through text. Similar social sites to Yahoo! Answers that are represented in the repository are Douban.com and Twitter.

The data repository appears to be open for data submissions from anyone as long as the data meet the repository's collection scope; data should be derived from social media sites. The repository also has a particular emphasis on blogging sites. There is a separate page on the repository website where people can "donate" a dataset, and this page includes options to input relevant metadata publishable on the site. Submitters can also directly upload their data files on this page, provided the necessary metadata is included. There do not appear to be strict restrictions on the data files themselves. There is a 200MB limit on data files, and the repository requests that multiple files be included as a zipped file. Datasets in this repository include primarily observational data. There are no set domains for data, but the datasets currently compiled share social sciences, technology, and entertainment domains.

The repository also includes information for contacting the people who work on the repository with questions or comments, enabling users to get feedback and insight from an actual human who has had hands-on experience with the repository. In the case of dataset submitters with files over 200MB, the repository requests that users contact its office, with several methods of communication provided on the repository website. These resources are also available to any submitter or user with questions and comments, and a "contact us" tab is clearly marked on the main website menu. A mailing address, physical address, telephone/fax number, and email are all provided on the page, as well as a direct email form that will allow the user to send a message to the people behind the repository from the website. The email form is finally followed by the email of an actual person at the repository, who likely monitors the emails that come in from the form.

The fields where users must input metadata related to the dataset represent guidance that the repository provides as part of a Submission Information Package. There are required fields that must be filled to donate a dataset, which are listed in this order: the name of the data set, abstract, source of the data set, information on missing values, and the total number of instances/nodes and number of attributes/edges in the dataset. There are also sections to

include relevant information, attribute information, relevant papers, and citation requests/acknowledgments in that order. The relevant papers section is not filled out for every dataset, however. There is also optional metadata to add a "graphics file" that is representative of the dataset. This schema does not appear to be modeled after a specific metadata standard, but the metadata appears in much the same order on the webpage for each dataset as it appears when the submitter inputs it, which illustrates that a specific structure is utilized.

Additionally, there is no log-in information needed to download datasets, allowing users to openly access and use datasets in the repository. There appears to be only one way to access the datasets, which is through a direct file download. The repository's Dissemination Information Package, though not explicitly stated, is an on-demand download of the dataset. Many of the dataset downloads contain a zipped file folder with two or more CSV files of the data and a README document to describe the datasets, attributes, etc. A citation policy is available for users who want to publish material based on the datasets and/or software obtained from the repository, in which the repository creators suggest a certain citation format. This statement is available as a separate webpage on the repository website and describes the creators' preference for what they call a "pseudo-APA reference format" of citation. The format lists first authors/contributors, then date, dataset title, link, and finally, publishing institution information. The link used in the citation format is broken, however. There is also a version of the citation in BiBTeX format. Datasets can also set additional citation requests, which a submitter of a donated dataset can enter details about in the metadata field, "Citation Requests / Acknowledgements."

Recommended data citation:

Considerations for long-term preservation: The dataset creators may have the option to download a set of the data with the files already split due to the size of the data. Excel does not allow all rows of the file to be available at once, hindering the ability of some users to access to full data. Furthermore, many repositories (including the Social Computing Data Repository) and other sites like GitHub have file limits, so having this option accessible to users could be beneficial.

The data file could also be available in more formats outside of .CSV to ensure it can be downloaded and easily access suing different software/applications. To ensure long-term preservation, the repository must make sure the file format is updated should technology standards change.

Yahoo! Answers Research Alliance may consider releasing more of the raw data obtained from Yahoo! Answers following October 25, 2007.

Sam Solomon
LIS 545
Data Curation I: Fundamentals
Feb. 24, 2023

Statement about human subject considerations:

While the data may have some personally identifiable information, steps were taken to anonymize the subjects used in the data. The researchers do not include any identifying information, like an avatar or a username.

References

- Ardeshna, Bhavik. (2022). *Yahoo! Answers Topic Classification*. Web. https://www.kaggle.com/datasets/bhavikardeshna/yahoo-email-classification

- Borgman, C.L. (2015). *What are data?* [Chapter 2 from *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.]

- Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System (OAIS)*. https://public.ccsds.org/Pubs/650x0m2.pdf

- Gomes, D. G. E., et al. (2022). *Why don't we share data and cod*e? Perceived barriers and benefits to public archiving practices. *Proceedings of the Royal Society B: Biological Sciences*, 289 (1987)

- R. Zafarani and H. Liu. (2009). Social Computing Data Repository at ASU [http://datasets.syr.edu/pages/home.html]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering.

- Split CSV. (n.d.) https://www.splitcsv.com/index.html

- *Yahoo! Answers*. (2022). PapersWithCode. Web. https://paperswithcode.com/dataset/yahoo-answers#:~:text=for%20Text%20Classification-,The%20Yahoo!%20Answers%20topic%20classification%20dataset%20is%20constructed%20using%2010,samples%2060%2C000%20in%20this%20dataset.

- Yahoo! Answers Research Alliance Webscope. (n.d.). https://webscope.sandbox.yahoo.com/catalog.php?datatype=l