Assessing Open Government Data Quality using Machine Learning: The Case of Satu Data Indonesia

Dwi Puspita Sari[1,*,†], Dimaz Cahya Ardhi[1,†], Mayra E. Santiago[1,†] and Bagus Jatmiko[2,†]

[1]College of Emergency Preparedness, Homeland Security and Cybersecurity, University at Albany, State University New York, 1400 Washington Ave, Albany, NY 12222

[2]Information Sciences Department, Naval Postgraduate School, 1 University Circle, Monterey, CA 93943

Abstract

Sharing government data with the public is a government initiative to increase accountability, transparency, and citizen participation. Through open government data (OGD) portal and for effective usage, the government needs to provide a good quality OGD, which is characterized by its accuracy, relevance, completeness, availability, and timeliness. This study aims to assess the quality of OGD published through the Indonesian national OGD portal named Satu Data Indonesia (SDI) Portal using a machine learning (ML) approach. The study begins by observing the OGD published on its portal and collecting the metadata that describes its published data. After collecting metadata, we test its quality using a ML approach. This study utilizes Orange, an open-source machine learning toolkit, to provide ML predictions with a scoring system.

Keywords

open government data, OGD, Satu Data Indonesia, machine learning, data quality

1. Introduction

The use of the OGD portal is one of the implementations of digital transformation and open government initiatives to provide data and information using technology. In 2019, the President of the Republic of Indonesia confirmed Regulation Number 39 on "Satu Data Indonesia", which acts as the legal umbrella for the Indonesian government's open data portal. SDI portal, as the official policy data management portal, aims to create good quality data that is easy to access and can be shared within government agencies, the public sector, non-government organizations, and the private sector. To benefit from the OGD, government agencies, as OGD providers, need to ensure that their data published to the public is of good quality. The good quality of the data needs to fulfill the criteria, including accuracy, relevance, completeness, availability, and timeliness. Additionally, government agencies must consider the data and open formats that

EGOV-CeDEM-ePart conference, September 1-5, 2024, Ghent University and KU Leuven, Ghent/Leuven, Belgium

*Corresponding author.

†These authors contributed equally.

dsari@albany.edu (D. P. Sari); dardhi@albany.edu (D. C. Ardhi); msantiago1@albany.edu (M. E. Santiago); bagus.jatmiko.id@nps.edu (B. Jatmiko)

0009-0002-7451-439X (D. P. Sari); 0009-0006-9306-0820 (D. C. Ardhi); 0009-0003-4713-4153 (M. E. Santiago); 0000-0002-1092-2602 (B. Jatmiko)

can limit public access [1]. The research question that guides our study is What is the quality of OGD published through the Satu Data Indonesia portal?

2. Methodology

The study focuses on a single case named SDI portal (https://data.go.id/), the official Indonesia OGD portal established in 2022. This case was selected to represent a developing country that started to develop a portal to disseminate government data, which includes different levels of agencies. SDI published 395,228 datasets by April 2024. First, we observed the SDI portal and collected the metadata of OGD published through SDI. The metadata for OGD describes dataset sources and details of dataset information [2] [3] [4]. Our preliminary study collected 1,000 dataset metadata from SDI through a random selection process. Second, we examined the dataset quality using an ML approach and supervised machine learning for binary classification. We used Orange software widget test, score, and predictions to evaluate our model.

3. Expected Findings and Future Work

In this study, we expect to build data training and test our model with different algorithms. We expect to compare those models, which can perform better in predicting OGD quality. The performance score should include the area value under the receiver operating characteristic curve, classification accuracy, F1, precision, recall, and correlation coefficient. The result of this study contributes insight into the quality of OGD published through the SDI portal for the Indonesian government to understand their data quality and keep improving their data quality to increase public accountability and participation. Furthermore, it will also raise awareness and engagement for the government and citizens on implementing and utilizing OGD to improve citizens' living standards by enhancing and implementing good governance through good quality OGD.

References

[1] M. Belhiah, B. Bounab, Pemerintah luncurkan portal satu data indonesia, in: MIT International Conference on Information Quality, UA Little Rock, 2017, pp. 1–13.

[2] A. Zuiderwijk, M. Janssen, I. Susha, Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators, Journal of Organizational Computing and Electronic Commerce 26 (2022). doi:https://doi.org/10.1080/10919392.2015.1125180.

[3] S. Saxena, Usage by stakeholders" as the objective of "transparency-by-design" in open government data, Information and Learning Science 118 (2017). doi:https://doi.org/10.1108/ILS-05-2017-0034.

[4] M. Kaasenbrood, A. Zuiderwijk, M. Janssen, M. de Jong, N. Bharosa, Exploring the factors influencing the adoption of open government data by private organisations, International Journal of Public Administration in the Digital Age 2 (2015). doi:https://10.4018/ijpada.2015040105.