Machine Learning Driven Aid Classification for Sustainable Development

Junho Lee[1,2*] , **Hyeonho Song**[1,3*] , **Dongjoon Lee**[1] , **Sundong Kim**[4] , **Jisoo Sim**[1] , **Meeyoung Cha**[3,1] and **Kyung-Ryul Park**[1]
[1]Korea Advanced Institute of Science and Technology (KAIST), South Korea
[2]Inter-American Development Bank (IDB), United States
[3]Institute for Basic Science (IBS), South Korea
[4]Gwangju Institute of Science and Technology (GIST), South Korea
{lelias, hyun78, djlee1, index, meeyoungcha, park.kr}@kaist.ac.kr sundong@gist.ac.kr

Abstract

This paper explores how machine learning can help classify aid activities by sector using the OECD Creditor Reporting System (CRS). The CRS is a key source of data for monitoring and evaluating aid flows in line with the United Nations Sustainable Development Goals (SDGs), especially SDG17 which calls for global partnership and data sharing. To address the challenges of current labor-intensive practices of assigning the code and the related human inefficiencies, we propose a machine learning solution that uses ELECTRA to suggest relevant five-digit purpose codes in CRS for aid activities, achieving an accuracy of 0.9575 for the top-3 recommendations. We also conduct qualitative research based on semi-structured interviews and focus group discussions with SDG experts who assess the model results and provide feedback. We discuss the policy, practical, and methodological implications of our work and highlight the potential of AI applications to improve routine tasks in the public sector and foster partnerships for achieving the SDGs.

1 Introduction and Background

Since the United Nations adopted the Sustainable Development Goals (SDGs) in 2015, various international financing sources have been discussed relative to the imperatives articulated by the SDGs. These include bilateral and multilateral aid, debt relief, and private-sector contributions. Official Development Assistance (ODA)—referring to foreign aid from OECD donor countries designed to promote socioeconomic development and welfare in developing countries [Hynes and Scott, 2013]—is considered the most direct means of implementation of the SDGs [Alsayyad, 2020]. As specifically highlighted in SDG 17 addressing global partnership, ODA plays a crucial role,[1] especially in the poorest and most vulnerable countries with limited domestic resources.[2] Tracking ODA is important for informed decision-making to ensure aid is transparently and effectively targeted to achieve the SDGs. As such, sharing high-quality and timely data–including ODA monitoring and enhancing statistical capacity to track and evaluate the progress of sustainable development (Target 17.18) by utilizing ICTs (Target 17.8)–has been recognized as the key means of successfully implementing the SDGs.

The OECD Creditor Reporting System (CRS) is the most authoritative database to monitor and evaluate aid flow. Annually, donor governments must report all ODA projects and classify them by type and purpose pursuant to the internationally agreed codification scheme detailed in the CRS. Among the codes, the five-digit ‘purpose code’ defines the characteristics and sector of development activities (see Table 1). This coding task is time-consuming and labor-intensive, given that over 250,000 new projects are launched annually worldwide. This study adopts and builds a natural language processing (NLP) model to assist in the classification and monitoring of aid activities for sustainable development by overcoming existing challenges at various levels. SDGs and ODA reporting require multi-stakeholder partnerships; thus, data sharing and reporting processes are established at multi-institutional levels. The CRS reporting process is executed at three levels: organizational, inter-organizational, and international. Within a government, development agencies collect aid data, classify the aid sector, and report to a national ODA reporting institution, which sends the data to the OECD for the final review (see figure in the appendix).

- **Organizational Level:** CRS codes are first classified by project managers in multiple organizations and undergo several verification procedures at the national ODA reporting institution. In the case of South Korea, the subject domain of this study, the Office for Government Policy Coordination (OPC), takes on this reviewing role. In summary, the proposed purpose code by the project managers is shared with the OPC. After that, the data are returned for review if modifications are made; otherwise, they are accepted.
- **Inter-organizational Level:** Following this data refinement at the organizational level, ODA reporting institutions check and revise the CRS reporting information before submitting it to the OECD Development Assistance Committee. In Korea, the OPC reviews and compiles

[1] see SDG 17.2 and SDG 17.3

[2] United Nations (Gen. Assembly), A/Res/70/1, “Transforming our world: the 2030 Agenda for Sustainable Development”

data from over 6,000 projects annually. Project data is submitted by organizations like the Korea International Cooperation Agency (KOICA) and the Export-Import Bank of Korea (KEXIM); which have separate internal review processes and are therefore assumed to be credible. However, project data from 66 other organizations needs to be reviewed thoroughly.

- **International Level:** CRS reporting at the OECD follows two rounds: the preliminary round in March and the final in July. The data received for the preliminary round from ODA reporting institutions for the interim data publication and update undergoes a month-long deliberation process to eliminate errors. The data submitted for the final round undergoes another comprehensive verification, taking over four months from July to November. The OECD gives feedback on the purpose codes several times a year, which is reviewed again before publication.

The statistics derived from these data constitute a significant social overhead capital supporting policy decisions [Holt, 2007]. Managing the CRS data and purpose codes involves important yet mundane human inspection and imposes various challenges. For example, studies have identified human inefficiencies in categorizing ODA projects and led to a call for new techniques and tools for managing CRS data [Ericsson and Mealy, 2019; Pincet *et al.*, 2019]. Our work started from this need. We interviewed the largest aid reporting agency in South Korea to understand the demands of ODA experts and the difficulties that different reporting institutions face, and identified the following:

- **Systematic Challenge:** Although a single-label classification is preferable for aggregating data and simplifying statistics, experts from multiple fields stated that the code classification could often not be delineated as desired [Schaekermann *et al.*, 2020; Yang *et al.*, 2019; Frid *et al.*, 2020; Tsoumakas and Katakis, 2007]. For instance, one ODA project may belong to multiple categories due to its multidimensional nature (e.g., a project for establishing a telecenter for farmers in a rural area could belong to both agriculture and education). Nevertheless, the current practice directs the classification into a single category.
- **Practical Challenge:** Although code allocation heavily relies on human experts, each expert has different capabilities and knowledge of the system. In addition, some institutions regularly rotate their experts, resulting in competency disparities.
- **Institutional Challenge:** There are institutions where multiple professionals collaborate to create CRS data. Although the purpose code is adequately allocated following the recommendation of the project manager, there are no clear criteria or guidelines to harmonize the conflicting opinions in the case where professionals disagree on the classification of the purpose code.

To address these challenges, we aim to support public officials in managing the ODA data and help them more efficiently allocate the codes. Our work contributes to SDG 17's call to form "Partnerships for the goals," which prioritizes the means of implementation and revitalizes the global partnership for sustainable development, particularly by assisting the essential financial ODA processes for all other SDGs [Alsayyad and Nawar, 2021]. We investigate the use of language models in managing CRS data. We created a multi-class classification model to recommend the top three (Top-3) codes, which model has a reasonable accuracy of 95.75 percent. The top five (Top-5) suggestions increase the model's accuracy by 1.3 percent. This enhancement implies that the model can reduce human labor and increase the cost-effectiveness of the classification job. Moreover, our model allows for the assignment of multiple classes to ODA projects. Our contributions are summarized as follows:

- **Empirical Contribution**: Despite calls for a better data management system in the SDG sector, implementing new systems has been slow. Our work serves as a feasibility test for applying machine learning to the SDGs.
- **Methodological Contribution**: We present a model for the OECD's CRS data. Our model adaptively identifies the top-k suggestions instead of generating a single suggestion, which has multidimensional characteristics.
- **Data Contribution**: Prediction results, codes, and collected data are released to facilitate SDG research.[3]
- **Survey Contribution**: We interviewed stakeholders involved in reporting the CRS data to the OECD in South Korea. The interview identified key challenges officials face in managing the reporting system and the potential for adopting a machine learning-based approach.

2 Related Work

2.1 AI Applications in the Public Sector

AI-driven models are rarely used in the public sector. This is partly due to the public administration's inherent slowness in adopting new technologies. However, the risks associated with computerized decision-making in public sector domains, such as accountability, responsibility, and transparency, are a more significant factor. Notwithstanding, AI technologies have great potential for enhancing routine tasks in the public sector. According to Deloitte, automating federal work in the United States can save up to 1.2 billion federal hours annually, representing savings of up to $41 billion. The report mentions additional benefits, such as reducing backlogs, improving the accuracy of projections, and sifting through millions of documents in real-time to identify the most pertinent content [Eggers *et al.*, 2017]. However, it remains unclear whether the application of AI in the public sphere provides more benefits than potential risks and complications. AI models are effective at both routine and non-routine tasks, including product classification [Zahavy *et al.*, 2018], outlier detection [Han *et al.*, 2021], and content recommendation [Covington *et al.*, 2016]. Likewise, AI can assist the public sector by relieving, dividing, replacing, and augmenting the work of public officials. Notwithstanding, AI must undergo a comprehensive review of its legal and technical

[3] https://github.com/elias-lee/kaist-crs-cs

Project Title	Training AI data labeling workforce to create non-contact jobs through decent ICT education programs in Ghana
Long Description	To improve ICT Literacy of Ghanaian Youth through IT S/W practical education in various fields
Purpose Code	11320

Table 1: CRS Data Sample

implications [Wirtz *et al.*, 2019], societal and ethical values [Dignum, 2017; Brown *et al.*, 2019], and a diverse array of challenges [Sun and Medaglia, 2019; Kawakami *et al.*, 2022] prior to implementation. For this reason, scholars have studied ways to improve performance and minimize its side effects by combining human expertise and machine learning models [Wang *et al.*, 2018; Kamar *et al.*, 2012]. Newer models emphasize the collaboration between humans and machines by referring to these systems as *AI-advised human decision-making* or *human-in-the-loop* [Bansal *et al.*, 2019; Gross *et al.*, 2019]. One successful and practical application of an AI-advised human decision-making model in the public sector is classifying product codes [Lee *et al.*, 2021] and detecting frauds [Kim *et al.*, 2022; Park *et al.*, 2022] in customs.

2.2 ML Applications in ODA and UN SDGs

Amidst the recognition of the benefits AI can bring to SDG [Porciello *et al.*, 2020; Vinuesa *et al.*, 2020], there is no exact precedent in applying ML techniques to categorize CRS data by purpose codes. There have been a number of applications of ML techniques to international development data to achieve other objectives. Boosted Regression Trees model has been applied to an SDG dataset to identify the most synergetic ways SDGs can be achieved and allocate resources more efficiently [Asadikia *et al.*, 2021]. Text mining and LASSO regression were utilized in a case study on the development project data of the International Fund for Agricultural Development to uncover food system dynamics across development projects [Garbero *et al.*, 2021].

A few cases exist where ML has been applied specifically to the CRS database. One example is the three-tiered approach used to overcome the classification challenges of development projects with Science, Technology, and Innovation (STI) components [Ericsson and Mealy, 2019]. Ericsson suggested reinforcing the STI ODA identification method by utilizing NLP algorithms to retrieve projects that were previously classified under a different category. The most similar work to our paper was done by Pincet in 2019 [Pincet *et al.*, 2019]. Pincet utilized approximately 200 reports to train XGBoost and generate the Random Forest model for classifying each CRS project to a specific SDG goal. The accuracy of the model varied by country, with an average of 87%.

3 Dataset

Our work utilized the CRS database, initially established by the OECD and the World Bank in 1967, to provide the participants with regular data on indebtedness and capital flows.[4] Every year, around 250,000 aid project records are submitted and stored in the database, entailing nearly 50 fields of information on development activities. The data has several classification variables, including fund provider and recipient; channel of delivery (entity implementing the activity); sectors (the primary sector(s) targeted by the activity); financial instruments (grants, loans); and project type. We present the full list of features in the appendix.

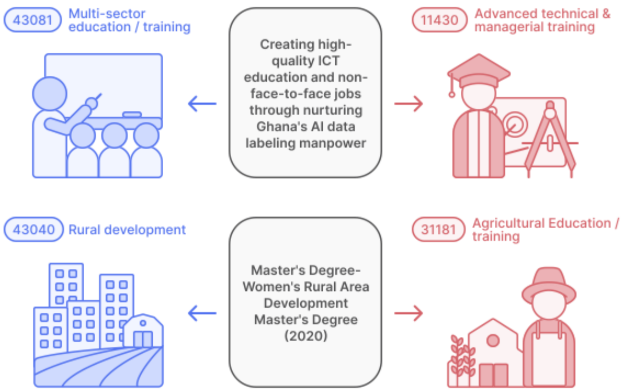
43081
Multi-sector education / training
Creating high-quality ICT education and non-face-to-face jobs through nurturing Ghana's AI data labeling manpower
11430
Advanced technical & managerial training
43040
Rural development
Master's Degree-Women's Rural Area Development Master's Degree (2020)
31181
Agricultural Education / training

Figure 1: Composition of the purpose code, with the purpose code 43081 for multisector education/training. The first two (43) and three digits (430) of the purpose code represent the project sector and category, and the last two (81) represent the detailed aid type.

Purpose Code. The data defines this as the five-digit code mainly used for classification, the first three digits refer to the aid category (or sector), and the last two digits represent the detailed aid function, as shown in Figure 1.[5] There are 234 unique purpose codes, yet their utilization is uneven.

Long Description. The single feature utilized to develop the AI model in our research is the *long description* field, a text description of the reported project within 4,000 words. This paper excludes non-textual features because our research focuses solely on using NLP to address the problem. The *long description* field is employed to classify the sector (or the purpose code) of each ODA project. One limitation we faced in our study was the high degree of inconsistency in the quality of descriptions provided by each country, which led us to limit our analysis to Korean data.

3.1 Korean CRS Dataset

There are numerous reasons why Korea offers a good setting for testing AI modeling on aid projects. These include data homogeneity, cleanliness, and compatibility among entities submitting reports. CRS data become more consistent when looking at a single country or organization, improving the quality of the analysis. [Pincet *et al.*, 2019] showed that the purpose code accuracy ranges from 0.87 to 0.96 depending on the country. Korea's CRS data are evaluated as complete and

[4] Technical Guide to Terms and Data in the Creditor Reporting System (CRS) Aid Activities Database: http://bit.ly/3Y53seI

[5] For a complete description of the purpose codes and sectors, please visit the OECD website at https://tinyurl.com/2vn3zmem.

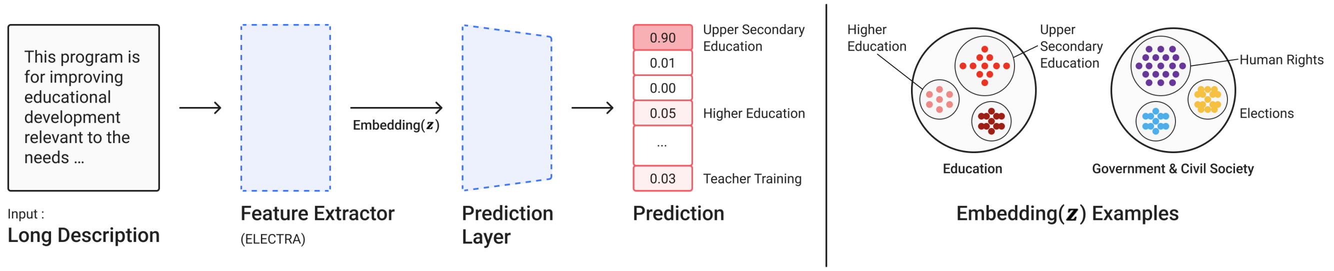

Figure 2: Workflow of the proposed framework. Left: The long descriptions are converted into embeddings by feature extractor and the embeddings are used to predict the labels (purpose codes). Right: Each input sample forms clusters in the embedding space.

uniform, which can be attributed to the two domestic deliberation cycles required before submission to the OECD. Korea is also known for digitized governance and open data, ranking among the top five countries in digital adoption across three dimensions: people, government, and business. According to the 2016 Digital Adoption Index; it was ranked first for governmental adoption worldwide.[6] Adding to this, Korea was also recently ranked top in the OECD's digital government index [OECD, 2020]. In 2021, the Korean government declared the third strategic plan for international development cooperation to strengthen monitoring and evaluation systems in aid and digitize the aid management process.[7] These efforts culminated in 2022, with Korea ranking third among reporting countries in the Data Transparency Index [Publish What You Fund, 2022]. These metrics strongly evince that Korean data is a reliable source for machine learning tasks, especially for CRS data. This work has also benefited from close collaboration with experts in the fields of sustainable development and domestic data access in Korea. The Korean government has been reporting approximately 6,000 new aid projects annually since 2010. While the most frequent codes appear over 4,000 times, codes like General Budget Support-Related Aid, Forest Industries, and Alternative Agricultural Development appear only a handful of times showing a high concentration on particular SDG goals. The high degree of asymmetry in the purpose code frequency can hinder prediction performance for less frequent categories.

Pre-processing. We pre-processed according to the following steps: (1) cleaning text descriptions that were empty, duplicate, or fewer than 25 characters; (2) aggregating purpose codes relative to sectors (focusing on the first three digits); and (3) eliminating sectors with fewer than 100 observations. After the pre-processing, among the original 63,977 observations, 60,540 remained.

4 Proposed Method

Problem Statement. CRS purpose code classification can be specified as a general text-to-code classification task. The text input sequence with length n, $\mathbf{x} = [x_1, ..., x_n]$ is the long description of ODA projects and the one-hot label $\mathbf{y} \in \mathbb{R}^c$ where x_i is word tokens and c is the number of class labels. The task is to train the classifier C, which produces the prediction result for the given $\mathbf{x}$; $\hat{\mathbf{y}} = C(\mathbf{x})$, where $\hat{y} \in \mathbb{R}^c$ refers to the probabilities for the output labels.

Proposed Framework. The process is as follows: The classifier is split into two components, feature extractor f and prediction layer g (i.e., $C = g \circ f$). For given input $\mathbf{x}$, the feature extractor f is used to extract the feature representation $\mathbf{z} \in \mathbb{R}^d$ from $\mathbf{x}$ (i.e. $\mathbf{z} = f(\mathbf{x})$), where d is the embedding dimension. After that, the prediction layer g is used to produce the softmax output probability (i.e., $\hat{\mathbf{y}} = g(\mathbf{z})$). Cross-entropy loss is used for training. Feature extractors can be any model and we experiment with TextCNN, BiLSTM, BERT, and ELECTRA. Figure 2 presents the proposed method.

Training Details. One of the advantages of algorithmic classification is its ability to pool candidates flexibly. We, therefore, tested two scenarios in which CRS experts are assumed to receive the Top-3 or Top-5 purpose code recommendations from the algorithm. If any of these suggestions contain the correct classification, the expert can review them and decide among this pool (or start searching from the proposed set). Based on these scenarios, we computed the macro-F1 and AUC scores for the Top-k recommendations. Table 2 compares the model accuracy, macro-F1, and AUC statistics. We reported the follow-up results using the ELECTRA model. For the evaluation of top-k prediction, we calculated the accuracy and the F1 metrics.

5 Evaluation

We explain the setup of our proposed method and the baseline models used for comparison.

5.1 Experimental Set-up

The data were divided into training and testing sets at a ratio of 3:1 with no overlap, leaving 45,405 samples of ODA project descriptions for training and 15,135 samples for testing. The data were split up regardless of the year entries were submitted, so the purpose code distribution is the same for training and testing sets. We used several backbone networks for the feature extractor. Figure 2 shows that the prediction pipeline has a prediction layer to produce the prediction. We utilize contextualized sentence embeddings from language models instead of traditional embedding approaches such as Word2Vec [Mikolov *et al.*, 2013]. The choice of architecture

[6] World Bank, Digital Adoption Index, http://bit.ly/3Yaui5c
[7] ODA Korea, https://bit.ly/3X6bxOT

for the feature extractor can be any model. Here, we used two representative models: BERT and ELECTRA.

- **Random Forest**: This baseline is used to assess other non-textual input features.
- **TextCNN**: This model uses a convolution layer to represent the input text [Kim, 2014].
- **BiLSTM**: The recurrent neural network, especially the Long Short Term Memory (LSTM) architecture, exhibited better performance on various NLP tasks by solving the long-term dependency problem [Graves and Schmidhuber, 2005].
- **BERT** (Bidirectional Encoder Representations from Transformers): BERT [Devlin *et al.*, 2019] is known for good performance in various downstream tasks due to its masked-language modeling that generalizes well.
- **ELECTRA** (Efficiently Learning an Encoder that Classifies Token Replacements Accurately): This model is another cutting-edge language model [Clark *et al.*, 2020] that is known to be light in training. In contrast to BERT, this model corrupts the input by replacing some tokens with plausible alternatives sampled from a small generator network.

For BiLSTM and CNN models, ReLU activation was used in the prediction layer. The embedding sizes of all models were set to 300, and the hidden dimension was set to 64. The Adam optimizer was used with a learning rate of 0.001. We used Intel(R) Xeon(R) provided by Google Colab at 2.30GHz and Nvidia K80/T4 16GB for experiments. The batch size was set to 512 training took 60 seconds per epoch for CNN and 23 seconds per epoch for BiLSTM. We used five estimators (trees) for the random forest. Language models were trained for ten epochs and evaluation was based on the best-seen accuracy. We used a two-layer perceptron with ReLU activation for prediction in the BERT and ELECTRA models. The hidden dimension was set to 100. The Adam optimizer for BERT and ELECTRA was set with a learning rate of 3e-5. We used an Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz and five A100-PCIE-40GB for BERT and ELECTRA experiments. The batch size was set to 8 and training time took 15 minutes per epoch. The average inference time per sample was less than 0.1 seconds. Codes for reproducibility are available anonymously at https://tinyurl.com/bddhshme

5.2 Model Performance

Table 2 confirms that language models (BERT and ELECTRA) outperformed other baselines. Random forest performed poorly compared to models that used text features. ELECTRA and BERT outperformed BiLSTM, likely because LSTM models are better calibrated for small datasets. ELECTRA and BERT produced comparable results, with the former slightly outperforming the latter. ELECTRA is known as a relatively light model in training and may have practical advantages in terms of a lower memory requirement and faster fine-tuning and prediction times. The Top-k recommendation scenarios were tested for the best-performing model, ELECTRA. For Top-3 suggestions, the model reported a peak performance of 95.75 for accuracy and 94.44 for macro-F1. Top-5 suggestions further increase the accuracy. High accuracy and an F1 score will be critical for practical considerations in ODA management. The goal is to assist humans in their tasks and reduce the search time needed for purpose code assignments. However, we adopted the top 3 suggestions as our main model since the increase in accuracy was marginal.

Model/Eval	Accuracy	F1	AUC
Random Forest	74.71	63.40	90.61
TextCNN	87.53	84.51	99.63
BiLSTM	82.04	72.80	99.13
BERT	89.96	88.02	99.37
ELECTRA	89.71	87.40	99.44
ELECTRA (Top-3)	95.75	94.44	-
ELECTRA (Top-5)	97.05	95.74	-

*First five rows are accuracies based on Top-1 suggestion

Table 2: Evaluation of the proposed model. AUC values are not available for Top-3 and Top-5 ELECTRA models

5.3 Accuracy by Sector

We tested the accuracy of each sector in the CRS data; the purpose codes were divided into 26 sectors using the first two digits of the purpose codes as in Figure 1. Among them, we could retrieve the accuracy of only 20 sectors due to missing data. The highest accuracy was shown for aid projects the Communication, Health and Education with 98.07%, 97.38%, and 97.36%. The lowest accuracy was observed in Banking & Financial Services, Population policies/programs and reproductive health, and Business & Other Services with 67.27%, 80.00%, and 82.52%, respectively.

5.4 Multi-category Classification

The current human labeling system only considers a single label due to high maintenance costs, despite the fact that ODA projects typically aim for multiple SDG objectives by nature. Algorithmic assignment assigns multiple labels easily. We thus examine how the aid sector distribution would change with multi-classification. We consider the idea of a *weighted suggestion*, a combination of Top-3 predictions that assigns weights of 80%, 15%, and 5%, respectively, to the Top-1, Top-2, and Top-3 predictions. The weighted proportion by sector is calculated by dividing the weighted count by the total count. The weighted percentage appears in the column titled "Top-3(%)" in Table 3.

We make three observations. First, the most popular sector, Education, showed a higher proportion than the original distribution (33.9 vs. 33.07). This may imply training bias due to the imbalance in training data distribution. It may also indicate that most aid projects include an educational element. Second, the Communication sector exhibited a reduced proportion (from 9.26 to 7.83). This sector also recorded the highest prediction performance, implying that misclassification is less likely to occur, resulting in a lower weight than in the human-labeled distribution. When we counted the Top-1 prediction result, the proportion was 9.25, close to the human-labeled distribution. Lastly, other sectors showed fewer differences in proportion.

Sector Name	Count (%)	Budget	Accuracy	F1	Top-3 (%)
Education	20,026 (33.07)	2,204	97.36	97.97	33.99
Government & Civil Society	7,279 (12.02)	898	90.41	90.16	11.94
Communications	5,607 (9.26)	540	**98.07**	**98.25**	7.83
Agriculture, Forestry, Fishing	5,110 (8.44)	974	96.76	96.25	8.37
Health	5,011 (8.27)	1,994	97.38	96.54	8.81
Other Multi sector	3,113 (5.14)	597	94.31	93.63	4.98
Other Social Infrastructure & Services	2,719 (4.49)	312	89.15	91.27	4.51
Industry, Mining, Construction	2,187 (3.61)	169	94.16	93.07	3.39
Transport & Storage	2,149 (3.54)	**2,353**	95.61	96.63	3.25
Water Supply & Sanitation	1,307 (2.15)	1,253	95.12	95.12	2.15
Trade Policies & Regulations	1,168 (1.92)	91	92.35	82.00	2.79
General Environment Protection	923 (1.52)	166	85.58	88.37	1.39
Energy	862 (1.42)	923	94.41	97.12	1.46
Unallocated / Unspecified	813 (1.34)	465	88.05	85.30	1.32
Emergency Response	812 (1.34)	522	92.75	94.81	1.24
Disaster Prevention & Preparedness	421 (0.69)	83	88.59	90.99	0.70
Business & Other Services	402 (0.66)	33	82.52	81.33	0.67
Banking & Financial Services	248 (0.40)	14	67.27	74.00	0.34
Population Policies/Programmes	211 (0.34)	97	80.00	75.21	0.52
Administrative Costs of Donors	172 (0.28)	498	82.97	87.64	0.24

Table 3: Evaluation by sectors. Test accuracy, f1 score, and weighted count proportion by sectors. (proportion), and budget (total sum budget) are calculated from the full dataset (Total budget values in 1 million USD).

5.5 Main Takeaways

ODA activities are skewed in most countries, with governments focusing disproportionately on a subset of sectors. This limits us from contemplating the accuracy of every sector. Instead, we grouped purpose codes by their popularity.

Frequency Matters. Frequent purpose codes showed better prediction accuracy than infrequent ones. Figure 3 shows that although the change is small, there was an upper trend in the average prediction accuracy where there was an increase in the frequency of the purpose codes, implying that more training data will lead to better prediction accuracy. The stability of the prediction also increased, with a smaller standard deviation as popularity increased. The results indicate that purpose codes that appear 1,000 or more times in the data could be accurately classified based on their long text descriptions. This means human experts can rely on machine learning for frequent labels, leaving more time for less frequent items.

Budget Matters. Another evaluation criterion can be whether the algorithm performs well for high-budget sectors. Note that the frequency of projects is unrelated to their budget size (e.g., large construction and facility developments cost substantially more yet occur less frequently). Our analysis also showed that higher-budget sectors also benefited from accurate classification results. The top three sectors in terms of budget Transport & Storage (ranked top), Education (ranked second), and Health (ranked third) showed accuracies of 95.61%, 97.36%, and 97.38% in Table 3.

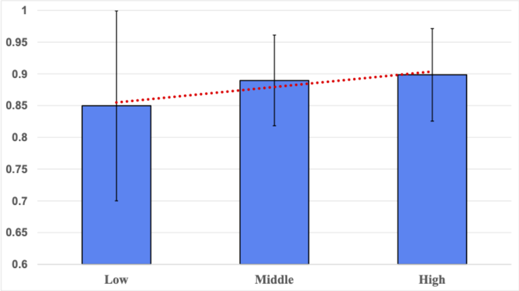

Figure 3: Prediction accuracy by popularity. The y-axis shows the average accuracy of each aid project (grouped by popularity), and the x-axis shows the group popularity. High: purpose codes above 1000 counts, Middle: below 1000 and above 300, Low: Below 300. The line in the middle of the bars represents standard errors.

6 Qualitative Evaluation

We executed an exploratory pilot interview to identify the challenges experts face in the reporting cycle and their demands for an AI model. Then, following the prototype design, two additional focus group interviews were conducted to assess our model. The selection of participants was informed by the key stakeholders' mapping in the field of ODA reporting mechanisms of the Korean government [Kvale and Brinkmann, 2009]. The four government agencies—OPC, KOICA, KEXIM, and Statistics Korea—represent the Korean government entities responsible for policy formulation, coordination, and implementation of ODA projects. Six data experts from the organizations listed above were selected for the interviews based on the close relevance of their job responsibilities to ODA statistics.

The major feedback on the model was in three parts: 1) necessity, 2) applicability, and 3) potential. First, the necessity of the AI model can vary depending on one's perspective. For example, individual project managers can forgo the time required to assign purpose codes from scratch. Personnel responsible for compiling all the individual projects that each ministry or institution oversees will save time and review them with greater precision. In addition to saving time, accuracy may improve, which will further benefit the OECD statistical officers who oversee global projects.

Second, the interviews indicated two major application areas for machine learning: 1) the domestic allocation and review procedure and 2) the international review procedure. As stated by the expert from Statistics Korea below, project managers have diverse understandings and knowledge of the purpose codes, and the model suggestion can be of great value in screening out candidates at the initial stage. For development agencies in charge of a large number of projects, the suggestion model can be used to evaluate each project. The OECD Office of Statistics, which examines numerous projects annually, can use the ML model on a larger scale to identify misclassified purpose codes.

Third, the model's potential can be summarized as improving the quality of domestic and international ODA statistics. ML can improve the credibility and coherence of the assigned purpose codes, as ML suggestions minimize human subjectivity and present a standardized framework for code allocation. It can also contribute to better accuracy by reducing the room for human errors and improving efficiency through time savings from task automation.

The donor countries and the OECD extensively utilize purpose codes to determine sectoral financial flows, devise national strategies, and produce analysis. Since the CRS purpose code information is the only data source that divides development financing into different sectors, numerous other development institutions and scholars use this information. Therefore, it is crucial to assign accurate purpose codes, as they serve as primary data for much other research.

7 Conclusion

7.1 Discussion

We discuss the main findings of this research and outline future directions for applying AI models in the public sector.

First, we demonstrated the feasibility of a machine learning model to assist in the multi-classification of purpose codes, thereby overcoming an existing systematic challenge. Through testing AI models over actual data, we discovered that while most purpose codes fit almost perfectly into one classification, many projects can still appear across multiple purpose codes. In this sense, we hypothesized assigning multiple codes to each project (as a weighted assignment of the Top-3 suggestions). Such a method can comprehensively encompass the sectors that each ODA project targets.

Second, the model can improve classification accuracy by overcoming practical challenges. Practical challenges happen for two main reasons: resource availability and sociopolitically constructed incentives. Our model helps solve the resource problem by recommending candidate sectors. Interviews with aid institutions indicated that data managers' workload is high; they annually oversee 1,500 new reports that need to be reviewed by a small team of sectoral and statistics experts. As in many public sectors, aid offices are understaffed, and this is just one of many tasks they perform. Having an AI model that suggests initial sectors can decrease the amount of work required of humans and increase accuracy. Considering that ODA classification is not limited to Korea and that approximately 250,000 new projects are reported annually, this can substantially reduce costs at the OECD level. Our model also addresses the sociopolitical incentive-one example is reactivity. Professionals are likely to be reactive unless they have a reason to be proactive. In this regard, having a model that recommends preliminary sectors based on similarities to previous descriptions can reduce external political influence.

Third, the AI model can help solve institutional problems by bridging the gap among individuals with diverse perspectives. Cultural challenges arise when there is disagreement among stakeholders, and there need to be clear criteria for resolving different perspectives. The preliminary purpose codes an AI model recommends can serve as a starting point for consensus. To empirically test how our model resolves disagreements, we monitored four ODA projects during the deliberation phase and tested whether the model correctly classified the purpose code. Our AI model correctly classified all four ODA projects, demonstrating that it can assist in resolving disputes regarding project classifications.

7.2 Limitations and Future Directions

We conclude with a discussion of future improvements that can be made to this research and to the classification of CRS data using the AI model. First, we recommend improving the transparency of the model by highlighting the most important words that led to each classification. This aggregation will allow professionals to revise the AI model's recommendations faster. Also, this study was limited to classifying Korean ODA projects with consideration of data quality (see discussion in Section 2). Regardless, an AI model will be more valuable if it can be expanded to all data-reporting countries reporting the data. A multilingual version of the classifier can be developed to expand to additional nations. Another problem is that the training data are not directly comparable because they come from different source organizations and are not described in the same manner. Therefore, one can consider standardizing the data elaboration process for internationally comparable data. Setting an appropriate description length could be one such solution.

We restricted our analysis to text features for classifying the purpose codes (other features totaling 91 are listed in the Appendix). One feature that can be considered in the future is the input *year* information. This is because ODA projects are subject to the government's agenda and priorities, which can change from year to year. Accommodating features like the submitted year of the project entry can improve the model's accuracy by controlling for heterogeneous elements.

Acknowledgments

This work was supported 80 percent by the National Research Foundation of Korea (2022S1A5A8052217), 10 percent by the Institute for Basic Science (IBS-R029-C2), and 10 percent by the IITP grant funded by the Korea government (MSIT) (No. 2019-0-01842).

Contribution Statement

JL and HS contributed equally as first authors and analyzed the main data. DL participated in data collection and focus group interviews. All authors participated in data interpretation and paper writing. MC and KRP conceptualized this study and are co-corresponding authors.

References

[Alsayyad and Nawar, 2021] Amina Said Alsayyad and Abdel-Hameed Hamdy Nawar. The nexus of oda and the sdgs: a scoping review of performance and statistical methodology. *International Policy Centre for Inclusive Growth*, pages 72–76, 2021.

[Alsayyad, 2020] Amina Said Alsayyad. Linkages between official development assistance and the sustainable development goals: a scoping review. *International Policy Centre for Inclusive Growth*, page 446, 2020.

[Asadikia *et al.*, 2021] Atie Asadikia, Abbas Rajabifard, and Mohsen Kalantari. Systematic prioritisation of sdgs: Machine learning approach. *World Development*, 140:105269, 2021.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *proc. of the AAAI Conference on Artificial Intelligence*, pages 2429—-2437, 2019.

[Brown *et al.*, 2019] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *proc of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[Clark *et al.*, 2020] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *proc. of the International Conference on Learning Representations (ICLR)*, 2020.

[Covington *et al.*, 2016] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *proc. of the ACM Conference on Recommender Systems*, pages 191–198, 2016.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[Dignum, 2017] Virginia Dignum. Responsible autonomy. In *proc. of the 26th International Joint Conference on Artificial Intelligence*, pages 4698–4704, 2017.

[Eggers *et al.*, 2017] William D Eggers, David Schatsky, and Peter Viechnicki. *AI-augmented Government*. Deloitte University Press, London, UK, 2017.

[Ericsson and Mealy, 2019] Fredrik Ericsson and Sam Mealy. Connecting official development assistance and science technology and innovation for inclusive development: Measurement challenges from a development assistance committee perspective. *OECD Development Co-operation Working Papers*, (58), 2019.

[Frid *et al.*, 2020] Emma Frid, Celso Gomes, and Zeyu Jin. Music creation by example. In *proc. of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[Garbero *et al.*, 2021] Alessandra Garbero, Bia Carneiro, and Giuliano Resce. Harnessing the power of machine learning analytics to understand food systems dynamics across development projects. *Technological Forecasting and Social Change*, 172:121012, 2021.

[Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[Gross *et al.*, 2019] Tom Gross, Kori Inkpen, Brian Y Lim, and Michael Veale. The human(s) in the loop—bringing ai and hci together. In *proc of the IFIP Conference on Human-Computer Interaction*, pages 731—-734, 2019.

[Han *et al.*, 2021] Sungwon Han, Hyeonho Song, Seungeon Lee, Sungwon Park, and Meeyoung Cha. Elsa: Energy-based learning for semi-supervised anomaly detection. In *proc of the British Machine Vision Conference (BMVC)*, 2021.

[Holt, 2007] D Tim Holt. The official statistics olympic challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1):1–8, 2007.

[Hynes and Scott, 2013] William Hynes and Simon Scott. The evolution of official development assistance. *OECD Development Co-operation Working Papers*, (12), 2013.

[Kamar *et al.*, 2012] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *proc. of the International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, pages 467—-474, 2012.

[Kawakami *et al.*, 2022] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-ai partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *proc. of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[Kim *et al.*, 2022] Sundong Kim, Tung-Duong Mai, Sungwon Han, Sungwon Park, Thi Nguyen Duc Khanh, Jaechan So, Karandeep Singh, and Meeyoung Cha. Active Learning for Human-in-the-Loop Customs Inspection. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *proc of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[Kvale and Brinkmann, 2009] Steinar Kvale and Svend Brinkmann. *Interviews: Learning the craft of qualitative research interviewing*. Sage, 2009.

[Lee *et al.*, 2021] Eunji Lee, Sundong Kim, Sihyun Kim, Sungwon Park, Meeyoung Cha, Soyeon Jung, Suyoung Yang, Yeonsoo Choi, Sungdae Ji, Minsoo Song, and Heeja Kim. Classification of goods using text descriptions with sentences retrieval. *arXiv preprint arXiv:2111.01663*, 2021.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[OECD, 2020] OECD. Digital government index: 2019 results. *OECD Public Governance Policy Papers*, (03), 2020.

[Park *et al.*, 2022] Sungwon Park, Sundong Kim, and Meeyoung Cha. Knowledge Sharing via Domain Adaptation in Customs Fraud Detection. In *proc. of the AAAI Conference on Artificial Intelligence*, pages 12062–12070, 2022.

[Pincet *et al.*, 2019] Arnaud Pincet, Shu Okabe, and Martin Pawelczyk. Linking aid to sustainable development goals: A machine learning approach. *OECD Development Cooperation Working Papers*, (52), 2019.

[Porciello *et al.*, 2020] Jaron Porciello, Maryia Ivanina, Maidul Islam, Stefan Einarson, and Haym Hirsh. Accelerating evidence-informed decision-making for the sustainable development goals using machine learning. 2(10):559–565, 2020.

[Publish What You Fund, 2022] Publish What You Fund. *Aid Transparency Index 2022*. 2022.

[Schaekermann *et al.*, 2020] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. Ambiguity-aware ai assistants for medical data analysis. In *proc. of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[Sun and Medaglia, 2019] Tara Qian Sun and Rony Medaglia. Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2):368–383, 2019.

[Tsoumakas and Katakis, 2007] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[Vinuesa *et al.*, 2020] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):233, 2020.

[Wang *et al.*, 2018] Cancan Wang, Rony Medaglia, and Lei Zheng. Towards a typology of adaptive governance in the digital government context: The role of decision-making and accountability. *Government Information Quarterly*, 35(2):306–322, 2018.

[Wirtz *et al.*, 2019] Bernd W Wirtz, Jan C Weyerer, and Carolin Geyer. Artificial intelligence and the public sector — applications and challenges. *International Journal of Public Administration*, 42(7):596–615, 2019.

[Yang *et al.*, 2019] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *proc. of the CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[Zahavy *et al.*, 2018] Tom Zahavy, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor. Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. In *proc. of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.