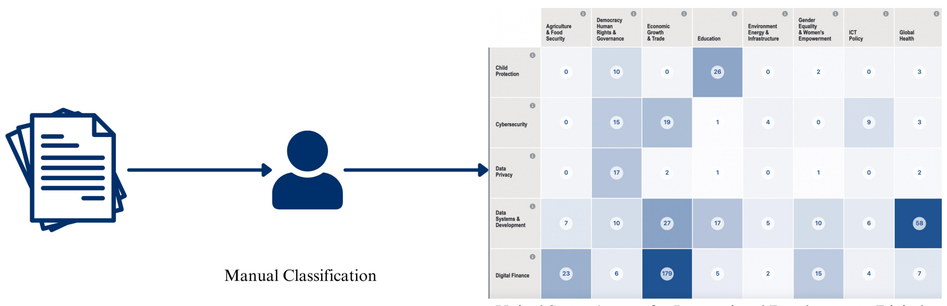
A ONE vs REST APPROACH FOR MULTI-CLASS DOCUMENT CLASSIFICATION

INTRODUCTION

Automated document classification is a trending topic in Natural Language Processing due to extensive growth in digital databases. However, a model that fits well for a specific classification task might perform weakly for another dataset due to differences in data. Thus, training several algorithms and evaluating performances is necessary to optimise the results.

PROBLEM

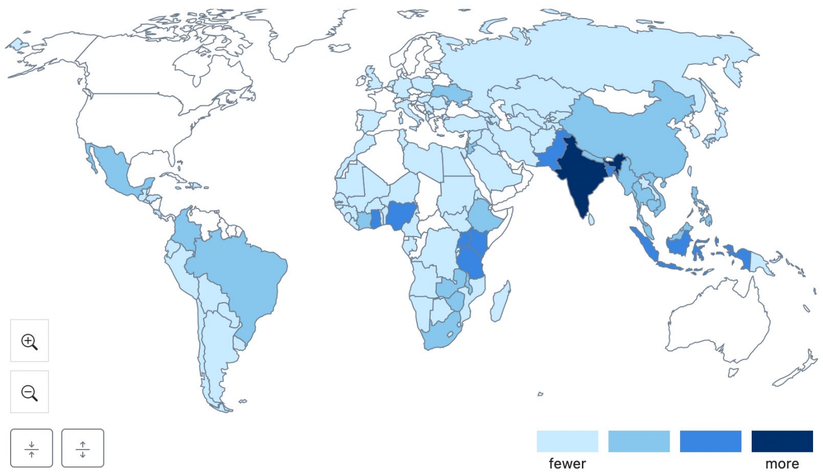
Propose an efficient and reliable approach to classify the development reports into intervention areas specified by USAID.



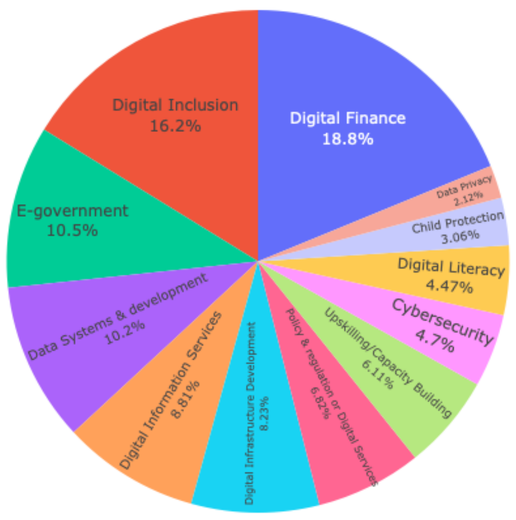
United States Agency for International Development - Digital Ecosystem Evidence Map

ABOUT USAID DEEM

A searchable database that stores worldwide digital development evidence, making it a world map for digital initiatives.



It includes **851 documents** classified under **12 pre-defined intervention areas**:



Word clouds for intervention areas show similarities and dissimilarities among classes.
.

OBJECTIVES

1. To evaluate approaches in **handling class imbalanced data** in emerging fields.
2. To develop new models using **ML and DL algorithms** to classify digital reports.
3. To compare the performance of different models in **multi-class classification** of digital reports.

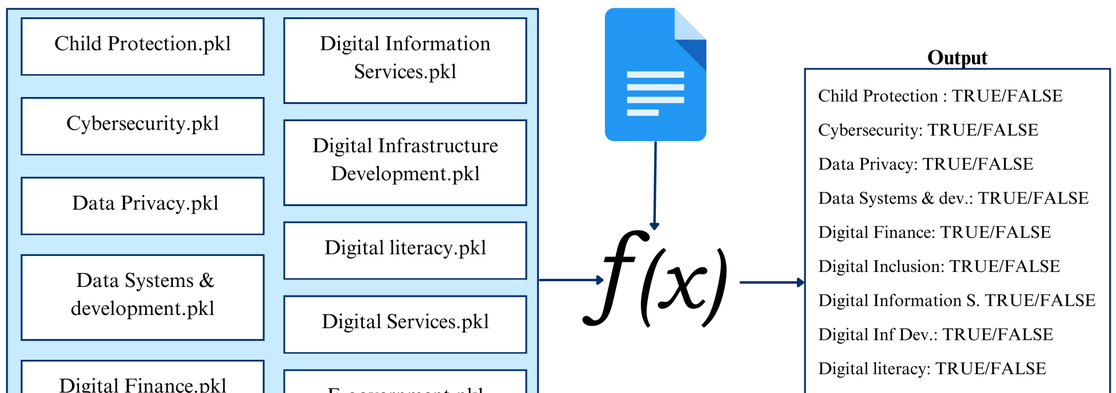
METHODOLOGY

- Extracting reports and preparing the dataset
- Preprocessing: Treat duplicates, Lowercasing, Remove URLs, Remove HTML tags, Remove emails, Remove stopwords, Remove punctuations and other characters, and lemmatisation.
- Train-test split: **Train set**: 615 reports; **Test set**: 185 reports
- Handling class-imbalanced nature: Random Oversampling
- Single model for multiclass classification: **Deep Learning algorithms:** CNN, LSTM, BERT; **Machine Learning algorithms:** Decision Trees, k-NN, SVM, AdaBoost, SGD, Naïve Bayes, Logistic Regression
- One vs Rest Approach (OvR)
- Measuring model performances: **F1 score** = $\frac{2 \times precision \times recall}{precision + recall}$

COMPARISION OF RESULTS

Intervention Area	Single Model (Logistic Regression)	OvR
Child Protection	0.78	0.86 (SGD)
Cybersecurity	0.67	0.76 (SVM)
Digital Inclusion	0.52	0.47 (Naïve Bayes)
Digital Finance	0.82	0.8 (Logistic Regression)
Data Privacy	0	0.67 (SGD)
Digital Infrastructure	0.35	0.32 (Naïve Bayes)
Digital Literacy	0	0.35 (SGD)

PROPOSED SOLUTION: ONE VS REST APPROACH



CONCLUSIONS

1. OvR strategy provides an understanding of complexities linked to class-imbalanced datasets.
2. OvR approach lays the groundwork to expand into multi-label classification.
3. The amount of data is not the sole factor affecting the performance; features like similarity within classes and dissimilarity among classes are crucial.
4. ML algorithms outperform DL algorithms due to the limited availability of data points.
5. BERT-based transformer showed promising performance despite limited data points.
6. The recommended algorithm is logistic regression if one seeks an approach that trains a single ML model for multiclass classification.

Name : A.M.S. Keshan, S.C. Liyanage, B.C. Mawitagama, R.A.U.A Ranaweera
Department : Department of Decision Sciences