# World Bank: Digital Development

Aditi Gajjar, Andrew Kerr, Cameron Stivers, Liam Quach
CDSM, California Polytechnic State University, San Luis Obispo

## INTRODUCTION

To effectively increase shared prosperity in developing countries to create a more livable planet, the World Bank, the largest source of funding and knowledge for low and middle-income countries, has initiated numerous projects. Many of these projects are aimed at specific regions such as Africa, with the purpose of improving a broad arrangement of systems in these developing area. However, given the complexity of the data, it is difficult to summarize the effect of the World Bank on these regions.

Our goal is to refine the classification of these projects into digital categories by utilizing the names of indicators, specific components of projects that contain vital data and outcomes. By doing so, we can aggregate the outcomes of similar projects, standardizing their units of measurement to produce a singular, comprehensible figure representing the results.

## PURPOSE

Provide snapshots of results and performance indicators to the Board, WBG management, and the public as part of the institution's accountability framework

## OUR DATA

**ISR/ICR Progress Reports**

- Result indicator-related data for active and completed projects at the World Bank.

**Indicator Keywords**

- A list of words related to Digital topics and a list of words related to Broadband topics. Keyword lists were created and provided by the World Bank.

**Country Population**

- Dataset consisting of country codes and populations from 1960 to 2022 for all countries in the world. Population data is reported by the World Bank.

## Naïve Approach

We were provided with two sets of keywords, one focused on classifying projects as 'Digital' and the other as 'Broadband'. With these keyword sets, we adopted a two-tiered classification strategy: initially determining whether an indicator was digital, followed by discerning whether those digital are broadband connectivity.

This method allows for the use of more broad keywords when classifying an indicator as broadband since we have already confirmed that the indicator relates to a digital subject. For example, the keyword 'fiber' could reference the fiber in food, or cables such as optical fiber cables.

Additionally, our approach records what keywords were used to label an indicator as digital and/or broadband. Saving these found keywords helped us verify whether our approach works and provided insight into why an indicator is classified into a particular category.

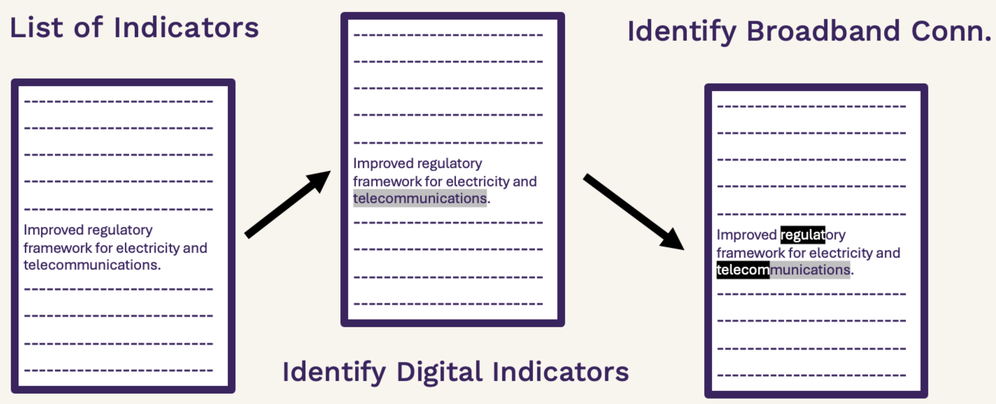


Figure 1. Visual Process of Naïve Approach

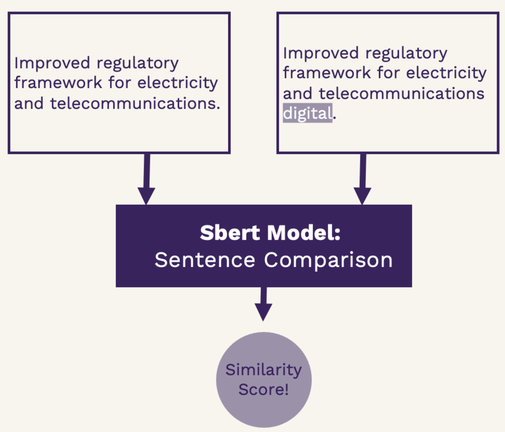


Figure 2. Visual Process of Sbert LLM Approach

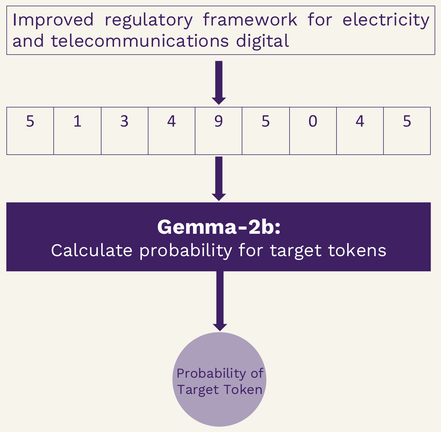


Figure 3. Visual Process of Gemma LLM Approach

## LLM MODELS

**SBert Model – Using Sentence Transformers**

- Used to calculate the semantic similarity between sentences. It does this by encoding sentences into high-dimensional vectors and then computing the cosine similarity between these vectors. This allows the model to evaluate how semantically close or different the sentences are based on their contextual meanings.
- The model is fine-tuned with a specific dataset that consists of pairs of sentences and a similarity score.

**Gemma Model – Target Token Probability**

- Tokenizing the text data, then select target tokens (Broadband or Not Broadband).
- Feed tokenized text into Gemma 2b to obtain probability estimates for each target token.
- Set thresholds based on probability values to make classification decisions and evaluate performance.

## Next Steps

- Aggregate similar projects for cohesion.
- Integrate best performing method to the World Bank internal system.
- Create a dashboard system for the World Bank to implement models and automate organization

## ACKNOWLEDGEMENTS

- Thank you to the World Bank team: Clement Gevaudan, Lucine Park, and Julia Chen, along with Elise Mary St. John.
- Thank you to our faculty advisors, Dr. Hunter Glanz and Dr. Jonathan Ventura.
- Thank you to COSAM for hosting this research conference.