World Bank - Digital Development for Africa

Aditi Gajjar | Andrew Kerr | Liam Quach | Cameron Stivers

I. Introduction

To enhance shared prosperity in developing nations and foster a more livable planet, the World Bank provides funding and technical knowledge to low and middle-income countries across over half a million projects. These projects support development programs, such as bridges, roads, and schools, to improve economic prospects and quality of life.

Snapshots of results and performance indicators from these projects are required to be available for the Board, World Bank Group management, and the public as part of the World Bank's accountability framework. However, encapsulating the World Bank's impact is challenging due to the complexity and quantity of these projects. Our solution is to build an automated system that classifies each project into distinct types and provides a detailed summary of the World Bank's activities over the past year using these classifications.

II. Data

We were provided with both ICR and ISR data, however, to simplify our process we solely examined the ISR data. This data set began with 461,556 observations containing 64,930 unique Indicator Names, however after removing observations with missing Project IDs or missing Indicator Names we were left with 421,701 observations containing 61,500 unique Indicator Names.

While removing missing Indicator Names was performed prior to implementing our classification approaches, removing missing Project IDs and additional data cleaning happens post-classification, and is further discussed in Section III, Part 2 of this report.

III. Methods

The overall process of our solution can be broken up into three steps. To begin, each project is classified into specific categories that describe what the project aims to improve. We focused on classifying projects as whether they relate to broadband connectivity or not, and this is done by analyzing their indicator names. Once each project is classified, appropriate filtering steps are taken to narrow projects down to categories of interest. Within these projects, numbers and percentages are extracted and aggregated to measure progress since each project's last reporting date.

1. Classification

For classification, two approaches, a Naïve approach and a Large Language Model (LLM) approach, were taken and compared.

a. Naïve Approach

We were provided with two sets of keywords, one focused on classifying projects as 'Digital' (526 keywords) and the other as 'Broadband' (14 keywords). With these keyword sets, we adopted a two-tiered classification strategy: initially determining whether an indicator was digital, followed by discerning whether those that are digital are related to 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 fiber-optic cables.

b. LLM Approach

i. BERT MODEL

The BERT model, short for Bidirectional Encoder Representations from Transformers, is an open-source machine learning framework designed for natural language processing (NLP). This model helps computers understand ambiguous language in text by using surrounding text to establish context. In our project, we used a pre-trained model for classifying text indicators related to broadband connectivity. Initially, the model reads a file containing the training data, which is then split into training and validation sets. The text data is tokenized using a tokenizer before loading a pre-trained BERT model for sequence classification. Sequence classification is a type of task in NLP where the goal is to assign a label or category to a sequence of text, such as a sentence, paragraph, or document. This involves analyzing the entire sequence and making predictions based on the content and context provided within the sequence. The model is then trained and validated using the Hugging Face `Trainer` class, and the trained model is saved. Once the training model is saved, it can predict classes using new input data.

ii. LLAMA MODEL

The Meta-LLaMA model, short for Meta Large Language Model Accelerator, is an open-source machine learning framework designed for natural language processing (NLP). This model aims to assist computers in understanding and generating human-like text by leveraging a large amount of pre-trained data. In our project, we used a pre-trained Meta-LLaMA, model for classifying text indicators related to broadband connectivity. Initially, the model processes a file containing the training data, which is then prepared by removing duplicates and irrelevant columns. The text data is tokenized using a specialized tokenizer before being fed into the pre-trained Meta-LLaMA model for sequence classification. Sequence classification is an NLP task that involves assigning a label or category to a sequence of text, such as a sentence, paragraph, or document, based on its content and context. The model is then used to classify the relevance of text indicators to broadband use. The results are saved in an Excel file, providing a comprehensive overview of the classifications and their associated probabilities.

2. Filtering

Classified indicators are filtered to ensure they are the most up-to-date version of their Project Id-Indicator Id combination. In other words, we ensured that there are no duplicate Project Id-Indicator Id combinations, and if there were then the instance with the most recent `Progress Date` is kept. This process brought us from 421,701 observations to 69,426 observations. Next, we only kept indicators where the `Unit of Measure` was labeled as a "Number" or "Percentage," decreasing the amount of observations to 49,084.

Filtering the indicators to observations labeled `Number` and `Percentage` did not guarantee the `Progress Value` was numerical. Additionally, some values used commas rather than decimal points to separate decimal values. After handling these situations by converting text values to numeric when possible, otherwise removing the observation, we were left with 48,965 observations.

3. Aggregation

Before we could aggregate any values, we needed to complete two steps: determine the country where each project took place and compute the progress made for each indicator.

We received (Project Id - Country Code) pairs and country population data. We merged the classified and filtered ISR data with the (Project Id - Country Code) pairs table, which increased our observations to 57,755 because some projects spanned multiple countries. However, after merging the country population data on country and year we decreased to 54,583 observations.

We calculated the Progress as the difference between the `Progress Value` and the `Baseline Value`. Lastly, for observations where the `Unit of Measure` was "Percentage," we multiplied the Progress by the population and divided by 100 to convert it to counts.

IV. Approach Comparisons

Given the absence of pre-labeled data to validate our classification methods directly, we adopted a comparative analysis to evaluate the effectiveness of our approaches. Our analysis focused on two primary metrics: the percentage of indicators classified as broadband-related and the differences in classifications among the methods.

To conduct this analysis, we executed the Naïve approach and the BERT model across the entire ISR dataset. Due to its computational intensity, we limited the Llama model to a random sample of 200 unique indicator names.

1. Comparative Analysis

The comparative results are summarized in Table 1, which outlines the total number of differing responses among the classification methods. The table also details the breakdown of classifications where one model classified an indicator as broadband-related while another did not.

Table 1: Classification Discrepancies

	Total Differing Responses	Broadband by Model X, Not by Y	Not Broadband by Model X, Broadband by Y
Naïve vs. BERT	14,892 (3.53%)	14,252 (95.7%)	640 (4.30%)
Naïve vs. Llama	195 (97.5%)	0 (0.00%)	5 (100%)
BERT vs. Llama	188 (94.0%)	0 (0.00%)	12 (100%)

Note: Naïve and BERT were run on the entire ISR dataset (421,701 observations), while Llama was run on a random sample (200 observations).

2. Findings

Overall, the Naïve approach classified 0.89% of observations as broadband-related while the BERT and Llama models classified 4.13% and 99.50% as broadband-related respectively.

a. Naïve Approach and BERT Model

The Naïve approach and BERT models showed the lowest amount of disagreement in classification, as evidenced by the low number of discrepancies. The 14,252 observations classified as broadband-related by the Naïve approach, but not by the BERT approach, highlight the Naïve approach's tendency to over classify indicators as broadband-related. Meanwhile, the 640 observations classified as broadband-related by the BERT model, but not by the Naïve approach, suggests BERT has better contextual understanding in certain cases.

b. LLAMA Model

The Llama model classified nearly all sampled observations as broadband-related, indicating its tendency to over classify. This over-classification indicates that while the Llama model has an advanced contextual understanding, it may also be too sensitive, interpreting nearly all nuances as relevant to broadband.

V. Discussion

Between our two classification methods, we saw several strengths and weaknesses. The Naïve approach, by design, is straightforward and easy to implement. The simplicity of the approach also makes it easy to explain: If a word related to the category is present in the indicator, the indicator will be labeled as that category. Therefore, its performance is entirely dependent on the comprehensiveness of the lists of keywords used to define categories. Because the World Bank had been classifying indicators and projects manually before, the lists that they provided us helped capture a large percentage of all digital and broadband connectivity indicators using this method. Where the Naïve approach falls short is in its inability to capture the context of the indicators, and the broadness of keywords can lead to misclassification. An example of this is in the indicator “Freight cost for automobile per wagon.” The algorithm recognizes the word “mobile” in this indicator and believes it to be both a digital indicator and a broadband connectivity indicator, when in reality it is used to describe a vehicle or transportation.

Across our LLM approaches, we realized that these methods do not provide any reasoning behind their classification decisions, creating a lack of explainability. While we cannot explain why these models make each individual classification decision, we are able to assist it in capturing the context of the indicator name if it is fine tuned and trained on relevant textual data. For example, in the automobile indicator, the BERT model was able to recognize “automobile” as related to vehicles instead of digital content. In a model that we will discuss in Next Steps, we saw much improved results because it was pre-trained on data relevant to technology, which is helpful for labeling digital and broadband connectivity indicators, but it may not perform well with other categories because of the context of the training data.

VI. Next Steps

First, we propose a hybrid approach that leverages both the Naïve and LLM methods, combining BERT and TF-IDF for robust classification. By integrating TF-IDF vectorization with a pre-trained BERT model, we can capture both the contextual and keyword-specific features of the indicators. The process would involve using TF-IDF to generate feature vectors that highlight key terms in the indicators, which can then be fed into the BERT model for context-specific classification. This hybrid approach would balance the strengths of both methods: the interpretability and simplicity of the Naïve approach and the contextual understanding provided by BERT.

Also, we suggest refining our filtering and aggregation methods to ensure consistency and accuracy across both Naïve and LLM-derived outputs. Currently, these methods have only been applied to the Naïve approach's results. However, the same processes can be adapted for LLM outputs without requiring significant changes. By standardizing the filtering steps, such as removing duplicates and ensuring the most recent progress dates are retained, we can maintain data integrity and accuracy. Additionally, further enhancements in handling numeric values and ensuring accurate aggregation of progress metrics will be crucial. This improvement will allow us to produce reliable and consistent summaries of the World Bank's project impacts, regardless of the initial classification method used.

Next, once we receive pre-labeled data with accurate indicator classifications from the team at the World Bank, we can greatly enhance our model training and evaluation processes. With this high-quality labeled dataset, we will be able to train Naïve and LLM models more effectively and evaluate their performance using essential metrics such as accuracy, precision, recall, and F1-score. These metrics will provide a comprehensive understanding of each model's strengths and weaknesses, allowing us to refine our approaches for optimal performance. This collaborative effort will ensure that our classification models are both accurate and reliable, ultimately enhancing the World Bank's ability to track and report on its digital development initiatives.

To further enhance our classification capabilities, we have introduced the BART-large-mnli model for zero-shot classification of indicators. This model allows us to classify indicators without the need for extensive labeled training data. By providing a list of labels, the BART-large-mnli model can accurately classify each indicator based on its contextual understanding. Initial tests using this model have shown promising results, with improved classification accuracy and the ability to handle a wide range of indicator descriptions. The implementation of the BART-large-mnli model represents a significant advancement in our approach, offering robust and reliable classification outcomes that will enhance the World Bank’s ability to track and report on its digital development initiatives.