TECHNICAL NOTE Nº IDB-TN-02679

# Crossroads in a Fog: Navigating Latin America's Development Challenges with Text Analytics

Leopoldo Avellán
Steve Brito

Inter-American Development Bank
Office of Strategic Planning and Development Effectiveness
Country Department Central America, Haiti, Mexico, Panama and the Dominican Republic
December 2023

# Crossroads in a Fog: Navigating Latin America's Development Challenges with Text Analytics

Leopoldo Avellán
Steve Brito

Inter-American Development Bank
Office of Strategic Planning and Development Effectiveness
Country Department Central America, Haiti, Mexico, Panama and the Dominican Republic
December 2023

Cataloging-in-Publication data provided by the Inter-
American Development Bank
Felipe Herrera Library

Avellán, Leopoldo.

Crossroads in a fog: navigating Latin America's development challenges with text analytics / Leopoldo Avellán, Steve Brito.

p. cm. — (IDB Technical Note ; 2679)

Includes bibliographical references.

1. Economic development-Latin America. 2. Economic development-Caribbean Area. 3. Sustainable development-Latin America. 4. Sustainable development-Caribbean Area. 5. Text data mining-Latin America. 6. Text data mining-Caribbean Area. I. Brito, Stev. II. Inter-American Development Bank. Office of Strategic Planning and Development Effectiveness. III. Inter-American Development Bank. Country Department Central America, Haiti, Mexico, Panama and the Dominican Republic. IV. Title. V. Series.

IDB-TN-2679

http://www.iadb.org



# Crossroads in a Fog: Navigating Latin America's Development Challenges with Text Analytics

Leopoldo Avellán and Steve Brito*

December 2023

**Abstract**

Latin America and the Caribbean are facing challenging times due to a combination of worsening development gaps and limited fiscal space to address them. Furthermore, the region is contending with an unfavorable external environment. Issues such as rising poverty, climate change, inadequate infrastructure, and low-quality education and health services, among others, require immediate attention. Deciding how to prioritize efforts to address these development gaps is challenging due to their complexity and urgency, and setting priorities becomes even more difficult when resources are limited. Therefore, it is crucial to have tools that help policymakers prioritize current development challenges to guide the allocation of financial support from international financial institutions and other development partners.

This paper contributes to this discussion by using Natural Language Processing (NLP) to identify the most critical development areas. It applies these techniques to detailed periodic country analysis reports (Country Development Challenges, CDCs) prepared by country economists at the Inter-American Development Bank (IDB) from 2015 to 2021. The study reveals that despite the perception that new development challenges have become more critical lately, the region continues to struggle with the same challenges from the past, particularly those related to the government's institutional capacity, fiscal policy, education, productivity and firms, infrastructure, and poverty.

**JEL Codes:** D83, F53, O19, O54, Z13
**Keywords:** Development Challenges, Text Analytics, Natural Language Processing, Latin American and the Caribbean.

* Leopoldo Avellan (leopoldoa@iadb.org) is Economics Principal Specialist at the Office of Strategic Planning and Development Effectiveness (SPD) and Steve Brito (stevebr@iadb.org) is Economic Specialist at the Country Office at Dominican Republic, both at the Inter-American Development Bank. The authors thank Francesca Castellani, Daniel Jiménez, Arnoldo López, Lucía Martín, and participants at the Half-Baked Lunch Seminar at the Office of Strategic Planning and Development Effectiveness (SPD) for insightful comments during the preparation of this Technical Note. The opinions are exclusively the authors' and do not represent those of the Inter-American Development Bank or its Board of Directors.

## 1. Introduction

Latin America and the Caribbean (LAC) faces a grim post-pandemic outlook. Recently, the United Nations Economic Commission for Latin America and the Caribbean (ECLAC) stated that the region was going through a stagnation phase worse than the one experienced in the 1980s, a period known as “the lost decade.”

Although the COVID-19 pandemic hit hard, vaccination programs made residents hopeful that a rebound was around the corner. But then the war in Ukraine broke out, further complicating matters by spiking energy, food, and fertilizer prices while central banks around the world increased interest rates to curb inflation. This backdrop is a heavy drag on the region’s development prospects.

At this pivotal moment, Latin American policymakers stand at a crossroads in a fog. The pressing need to act decisively and reduce development gaps is clear. Yet, with the region lagging in so many areas, deciding what to prioritize and where to concentrate scarce resources can feel overwhelming. The myriad paths before policymakers represent the numerous challenges and potential solutions, but the fog of complexity and uncertainty makes it hard to discern the best way forward, particularly because different policy objectives might imply different priorities, some of them competing. For example, inadequate port infrastructure may affect competitiveness and economic growth. Poor project planning or corruption may delay investment in critical infrastructure projects, including ports. Some development gaps may have a more severe impact in the long term than in the short term. Likewise, the consequences of the recent drop in school attendance in the region will not make their full mark in development prospects until minors turn into adults and look for a job in the next decade without having the required skills. Some other development challenges may have disproportionate effects on specific segments of the population. For example, unreliable internet connectivity is worse in rural areas and for poor households. In fact, there are so many competing priorities that choosing where to concentrate and allocate resources is a daunting task because every issue is critical.

However, trying to navigate every path at once is not feasible, especially when public funds are limited. Prioritization becomes essential in guiding discussions on the most critical areas of focus. Over the years, various policymakers and scholars have recognized this, presenting frameworks (Hausman et al. 2005) and empirical models (Izquierdo et al. 2016 and Borensztein et al. 2014) that shed light through the fog, pointing out the key gaps that, once addressed, can unleash substantial economic growth.

This paper contributes to this discussion by introducing a new approach to uncover the potential critical sectors in the region. It analyzes the Country Development Challenges (CDCs) reports with novel text analysis techniques. Teams of experts produce these reports every five years to review the constraints that are holding back development in each country. They do not explicitly rank the areas of intervention, but the reports can provide some clues to learn what the critical barriers are.

In this regard, the only assumption made in this analysis is that those development gaps that are relatively more intense are covered at greater length in these reports. This assumption is reasonable, as one would expect that thematic experts will discuss in more detail issues that are more severe to their assigned countries. To this end, the text is first pretreated to eliminate repeated and common words so that the main keywords remain in the sample. Then a classification algorithm is applied to sort and identify the topics present in the text. Next, the themes that received more attention in each CDC report are computed and ranked by calculating their share in the report. Finally, the results are aggregated at the regional level to obtain a snapshot of the whole region.

Ranking the resulting topics by their share in the text, the analysis finds that *"Institutional Capacity of the Government"* is the top development challenge. It records the largest share of text out of 14 possible topics. This finding is consistent with the vast literature that highlights the role of institutions as the main element behind long-term growth (Acemoglu et al. 2005; Acemoglu and Robinson 2008) and is also aligned with literature that points at institutions as the culprits for LAC's disappointing economic performance (Edwards, Esquivel, and Márquez 2007). Furthermore, *"Institutional Constraints"* dominates as the most critical topic at the country level, as these challenges are the top priority in 20 out of 26 IDB borrowing member countries. Furthermore, in 5 out of the 6 remaining countries, it is in the top three. Only in Venezuela does this topic rank below the top five. Next in the regional intensity ranking are the topics of *"Fiscal Policy,"* and *"Education and Job Training,"* followed by *"Productivity and Firms," "Infrastructure,"* and *"Poverty and Social Protection."* Additionally, results suggest that the structural challenges facing the region during the last decade remain binding, within the subset of countries for which a CDC report was prepared after the COVID-19 pandemic hit.

The application of Natural Language Processing and text analysis in economics and social sciences has been evolving. For example, Gentzkow and Shapiro (2010) study what drives media slant in US newspapers; Romer and Romer (2010) identify tax changes from the US presidential speeches and Congressional reports; Hansen, McMahon, and Prat (2018) calculate central bank transparency from the deliberations of the US Federal Open Market Committee (FOMC); and Biasi and Ma (2022) estimate the diffusion of frontier knowledge analyzing the syllabi of higher education courses.

The rest of the discussion is organized as follows: Section 2 assembles the data set (Corpus), explores its main descriptive statistics, and conducts some preliminary analysis based on frequencies of words occurring together as sequences of two words (s) and three words (trigrams). Section 3 discusses the topic model and estimates the main topics present in the CDC reports. Section 4 discusses the results, and Section 5 concludes.

## 2. An Overview of Sample and Text Processing

### 2.1 Main Descriptive Statistics

The sample includes 39 CDC reports available for 26 borrowing countries in LAC.[1] Table 1 shows their production timeline between 2015 and 2021. Every IDB borrowing member country has at least one CDC in the sample, and 13 countries have two. Twenty-six reports were written in Spanish and 13 in English.[2] To perform the same linguistic analysis, the reports in English were translated into Spanish. The focus of the text analysis is the sectoral section of CDC reports.[3]

**Table 1. CDC Reports Timeline**

| Country | CDC Years of Publication | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| Argentina | | C | | | C | | |
| The Bahamas | | | C | | | | |
| Barbados | | | | C | | | |
| Belize | | | | | | C | |
| Bolivia | C | | | C | | | |
| Brazil | C | | | C | | | |
| Chile | | | | C | | | |
| Colombia | C | | | | U | | |
| Costa Rica | | | | C | | | |
| Dominican Republic | | C | | | | U | |
| Ecuador | | | | C | | U | |
| El Salvador | | | | | C | | |
| Guatemala | | C | | | U | | |
| Guyana | | C | | | U | | |
| Haiti | | | C | | | | |
| Honduras | | | | C | | | |
| Jamaica | C | | | | | | U |
| Mexico | | | | | C | | |
| Nicaragua | | | C | | | | |
| Panama | C | | | | C | | |
| Paraguay | | | C | | | | |
| Peru | | | C | | | U | |
| Suriname | | | C | | | U | |
| Trinidad and Tobago | | C | | | | U | |
| Uruguay | | | | | C | | |
| Venezuela | | | | | | C | |

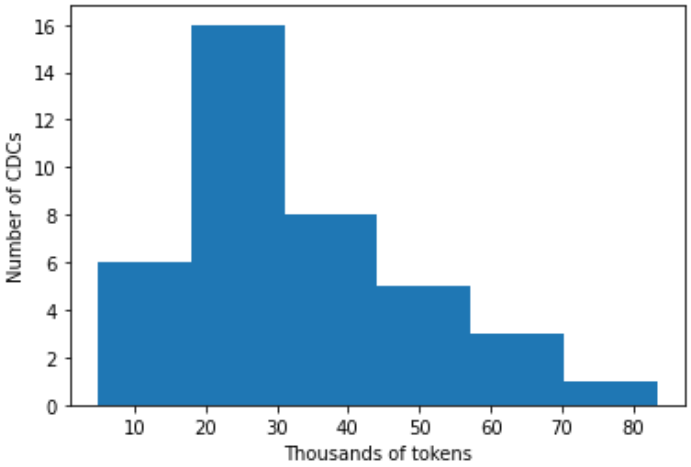
*Source:* CDC Reports at the Inter-American Development Bank.
*Note:* C corresponds to a complete CDC report. U corresponds to a CDC report update.

[1] Nine CDC reports are updates (presented in Table 1 as U=update). The difference between a complete CDC report and a CDC report update is that the last one keeps the same development challenges and evaluates how they changed since the complete report was elaborated. On the other hand, in a complete CDC report, the development challenges are re-evaluated.
[2] Countries with their CDC reports originally written in English are The Bahamas, Barbados, Belize, Brazil, Chile, Guyana, Jamaica, Suriname, and Trinidad and Tobago.
[3] Chapter 1 includes the socioeconomic context of the country. Chapter 2 of a complete CDC report identifies the most binding constraints to inclusive and sustainable growth and development. For the CDC report updates of Colombia, the Dominican Republic, Ecuador, Guatemala, Guyana, Jamaica, Peru, Suriname, and Trinidad and Tobago, chapter 2 includes the usual contents of chapter 3. Chapters 3 to 5 were used for Venezuela's CDC report.

The next step after collecting the sectoral chapters of the CDC reports is to convert them into a Corpus,[4] which is a collection of text transformed into a data set that allows the use of text analytics techniques. Tokenization is an important step for modeling text data and a common task in NLP. Tokenization is performed on the Corpus to obtain tokens. Tokens can be either words, characters, numbers, or n-grams (a contiguous sequence of *n* items or words).[5]

Figure 1 shows the histogram and descriptive statistics of the Corpus after tokenization of the selected CDC reports chapters. On average, CDC reports have 32,200 tokens, ranging from a minimum of 4,900 tokens for Guatemala in 2019 to a maximum of 83,300 tokens for Argentina in 2016. However, many of these tokens are stop words: a set of commonly used words in any language, such as *I, you, the, who, having, doing, between, because, though, from, further, some, any,* and so on.[6] In addition to removing stop words, digits, punctuation, special characters, and blank spaces are removed.

**Figure 1. Histogram and Descriptive Statistics of the Text Analysis for the CDC Reports**




| **Descriptive Statistics** (Thousands of tokens) | |
|---|---|
| CDC Reports | 39 |
| Mean | 32.2 |
| Standard Deviation | 16.5 |
| Minimum | 4.9 |
| Percentile 25% | 21.4 |
| Percentile 50% | 28.5 |
| Percentile 75% | 39.1 |
| Maximum | 83.3 |

*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

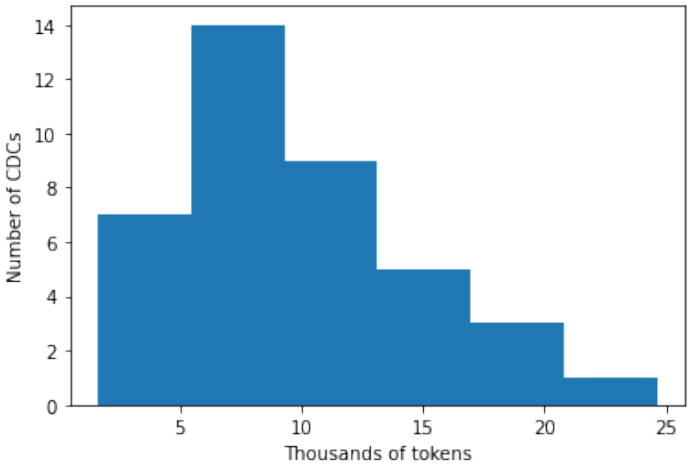
The next step for text processing is lemmatization,[7] which is the procedure of grouping together words with the same root or in their primitive form, to analyze them as a single item. For example, the word *women* has *woman* as lemma, the word *children* has *child* as lemma, and the word *policies* has *policy* as lemma. In this way, words that have the same or similar meaning can be grouped under the same term. Another advantage of lemmatizing the documents is the possibility of identifying linguistic features, such as the types of words, which could be nouns, adjectives, verbs, adverbs, pronouns, prepositions, and so on. Verbs such as *increase* or

[4] The programming language Python and the open-source web application Jupiter Notebooks are used to process and analyze the text of the CDC reports. These applications' advantages are the access to many open-source libraries designed for Natural Language Processing and text analysis.
[5] The most conventional approach of forming tokens is based on blank spaces. For example, the tokenization of the sentence "Reduce poverty and inequality" results in four tokens: *Reduce, poverty, and inequality.*
[6] The Spanish-language list of stop words of the SpaCy library in Python contains 551 words.
[7] Python library SpaCi for lemmatization: https://spacy.io/usage/linguistic-features.

*reduce* or prepositions like *below* or *above* provide little information about the development challenges in the region. Thus, only words that are nouns or adjectives are kept in the database[8] to identify the development challenges in the reports.

Figure 2 presents the histogram and descriptive statistics for the number of tokens in the complete set of CDC reports after applying the text processing techniques explained above. Now the reports have on average 9,800 tokens, a 69.5 percent reduction in the average of words in the sample. The minimum number of tokens (1,600) is found in the 2019 report for Guatemala and the maximum (24,600) is found in the 2016 report for Brazil.

**Figure 2. Histogram and Descriptive Statistics of the CDC Reports after Text Processing**




| Descriptive Statistics (Thousands of tokens) | |
|---|---|
| CDC Reports | 39 |
| Mean | 9.8 |
| Standard Deviation | 5.0 |
| Minimum | 1.6 |
| Percentile 25% | 6.7 |
| Percentile 50% | 8.5 |
| Percentile 75% | 11.6 |
| Maximum | 24.6 |

*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

### 2.2 Document-Term Matrix and Term Frequency-Inverse Document Frequency

The next step of the analysis is setting up the Document-Term Matrix (DTM). This is a matrix in which each CDC report is summarized in a row, while each word is represented in a column. To build the matrix, the DTM generates a vector count of the set of all words in all CDC reports, and it counts how many times these words appear in each document. Only the words that are in a report are counted and the others are recorded as zero. For example, if the Corpus has the following three documents with their corresponding text:

*Document 1 = Policies to promote education*
*Document 2 = Access to water*
*Document 3 = Promote access to housing*

The Document-Term Matrix would be:

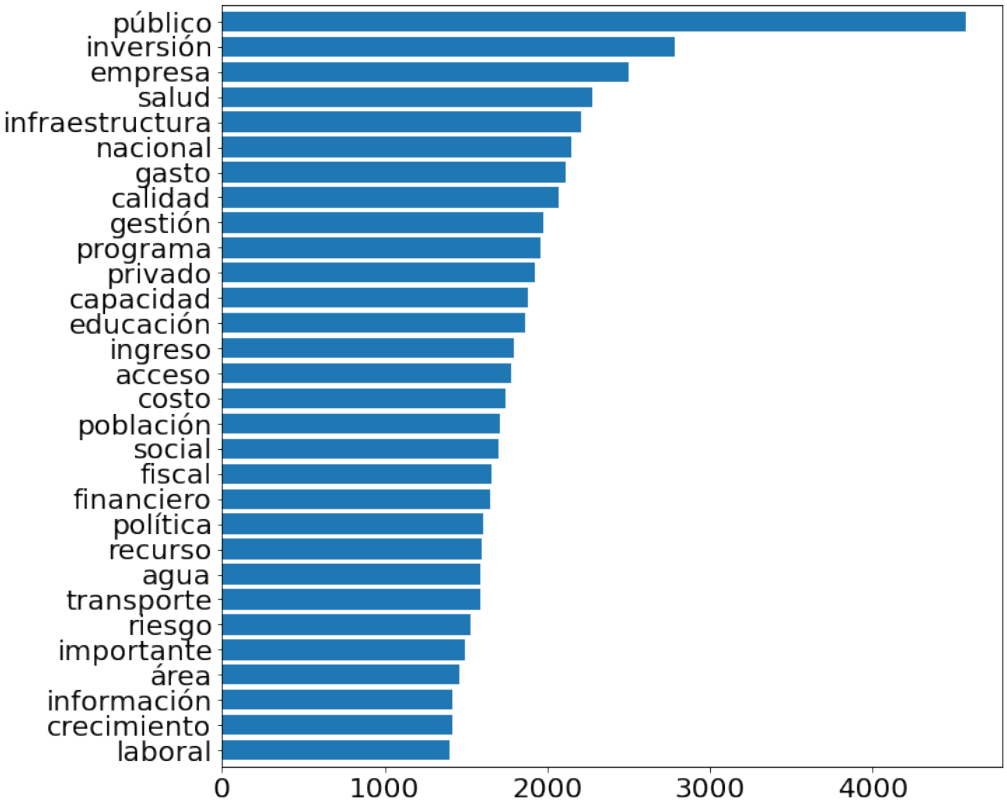
[8] Country names and nationalities are also dropped from the database. As expected, these words appeared many times in the corresponding CDC report of each country, but they do not add value to the analysis.

| | policies | to | promote | education | access | water | housing |
|---|---|---|---|---|---|---|---|
| **Document 1** | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| **Document 2** | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| **Document 3** | 0 | 1 | 1 | 0 | 1 | 0 | 1 |

The number of unique words in the Corpus vocabulary in the sample is 14,361. The size of the DTM is 39 rows (corresponding to the 39 CDC reports) by 14,361 columns. Sparsity and density are terms used to describe the percentage of cells in the DTM that are not populated and are populated, respectively. Consequently, the sum of the sparsity and density should equal 100 percent. The sparsity of the calculated matrix is 74 percent (cells populated with zero values), and the density is 26 percent (cells populated with non-zero values).

Based on the DTM, figure 3 presents the top frequency of the 30 most repeated words of the Corpus of the CDC reports. Only a few words have high frequency. For example, the word *public* appears 4,571 times in the CDC reports and the word *investment* appears more than 2,779 times.

**Figure 3. Top 30 Words of the CDC Reports**
(Number of times appearing in the Corpus)




*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

A word cloud is a helpful visualization that shows word frequency in a text by displaying words in sizes proportional to the times they appear in the Corpus. Figure 4 shows the 100 words with highest frequency in the CDC reports, which together with the previous figure, give some preliminary insights about the development challenges of the region. The most repeated words are: *public, investment, company, health, infrastructure, spending, management, private, education, and so on.*

**Figure 4. Word Cloud of the CDC Reports**
(the size of the words represents their frequency)


*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

A bag of words is defined as a row vector of frequencies that represents a document, disregarding grammar and the order of the words. With the tokens obtained after cleaning and processing the documents, the Document-Term Matrix contains 39 bags of words corresponding to the text of each CDC report. Even though it is possible to analyze the most repeated words in the Corpus and in each bag of words with the word count, this has the disadvantage of not considering the heterogeneity in the size of each CDC report. As figures 1 and 2 show, there are important differences in the size of CDC reports. Additionally, looking at the frequency count could be misleading in efforts to analyze the importance of each word within each CDC bag of words with respect to the others. For example, the acronym IDB appears many times in each document, but the term is common and repetitive in all CDC documents.

To deal with these issues, the approach of Term Frequency - Inverse Document Frequency (TF-IDF) is implemented, which quantifies how relevant a word is to a document, while considering its importance in other documents in the Corpus. TF–IDF is the product of two measurements: Term Frequency and Inverse Document Frequency:

$$\boldsymbol{TF - IDF(w, d, D) = TF\ (w, d) * IDF(w, D)},$$

where ***w*** is a given word in a document, ***d*** is a given document, and ***D*** is the collection of all documents.

There are several ways to determine the value of both terms in this formula, but the method used in this study is based on the Scikit-Learn, a free software machine learning library written in Python programing language. The Term Frequency is calculated as:

$$\boldsymbol{TF(w,d) = log[1 + f(w,d)]},$$

where $\boldsymbol{f(w,d)}$ is the frequency of word ***w*** in document ***d***. The second term, the Inverse Document Frequency is calculated as:

$$\boldsymbol{IDF(w,D) = log\left(\frac{N}{f(w,D)}\right)},$$

where *N* is the total number of documents in the dataset, and $\boldsymbol{f(w,D)}$ is the number of documents that contain the word ***w*** over all the documents ***D***.

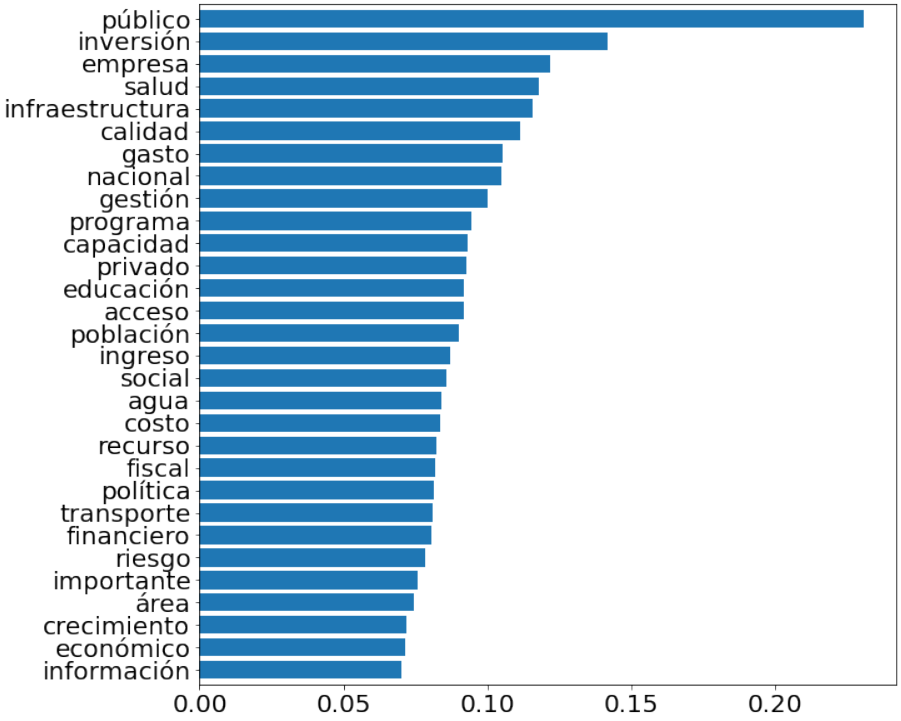
The TF–IDF value of a word ***w*** in a specific document ***d*** will increase with the frequency of the word in the document but will decrease with the appearance of the word in other documents. Therefore, words with a higher TF value in a document are more frequent and words with a higher IDF are more unique in the Corpus. The combination of these two terms allows for the identification of how important and distinctive a word is in a group of documents. For example, the word *country* has a high frequency in each CDC report (higher TF value), but this result is penalized because the word appears in all the documents (lower IDF value) indicating that the word is not central in the text analysis to the development challenges discussed in the CDC reports.

Figure 5 presents the top 30 words with the highest average TF–IDF value. The average is calculated with the TF–IDF values of the corresponding word for the 39 CDC reports. Like figure 3, the most important words are: *public, investment, firm, health, infrastructure, quality,* and so on. Even though the highlighted words are similar to those in figure 3, the TF–IDF results make sure that the analysis is taking into consideration the heterogeneity of the size of the CDC reports and the relevance of each word with respect to the Corpus of documents.

Figure 6 displays the word cloud for the top 100 words with the highest TF–IDF average values. Words like *public, fiscal, program, quality, management, institutional, capacity, information, and policy* have high frequency, and they might capture the need for institutional strengthening in the region. Words like *investment, infrastructure, transport, private, productivity, firm, and growth* might summarize the lack of productivity growth and private sector involvement in boosting income. Words like *health, education, women, income, water,* and *climate* indicate the importance of social, gender, and environmental issues for the region.

**Figure 5. Top 30 Words of the CDC Reports**

(Average TF–IDF value)




*Source:* Authors calculations.

*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports. TF–IDF refers to Term Frequency - Inverse Document Frequency.

**Figure 6. Word Cloud of the CDC Reports**

(The size of the words represents the average TF–IDF value)


*Source:* Authors calculations.

*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports. TF–IDF refers to Term Frequency - Inverse Document Frequency.

### 2.3 Bigrams and Trigrams

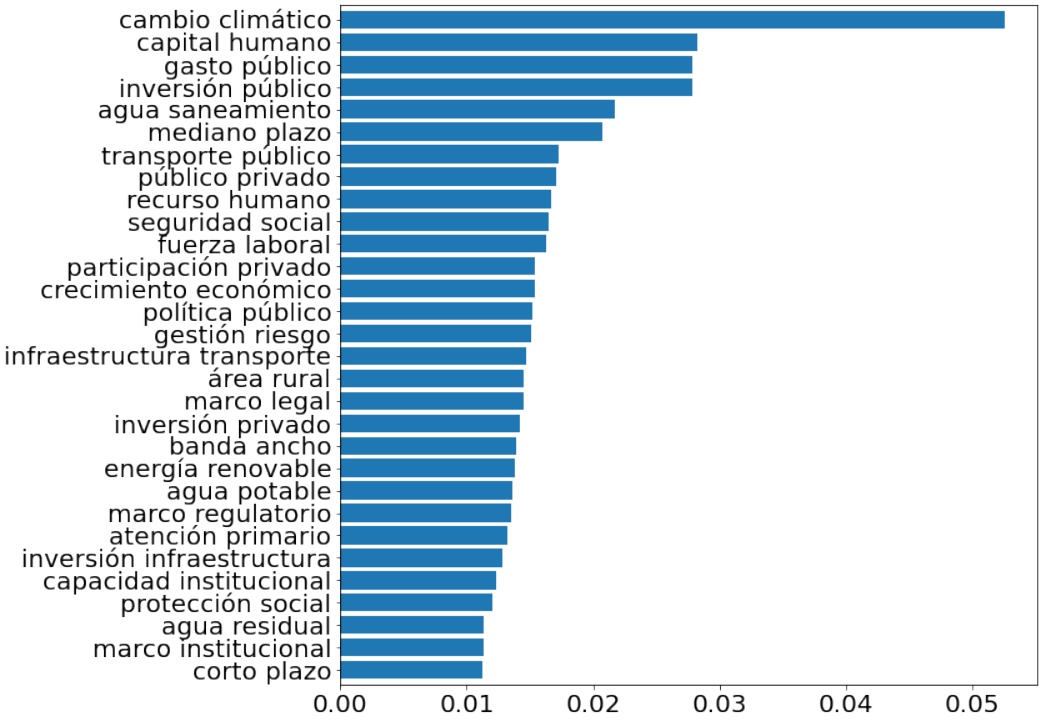
The previous section presents some of the most frequent words describing the development challenges in LAC. However, it could be challenging to connect some of the words without the corresponding context. For example, one of the most frequent words in the previous analysis is *investment*, but this word could be related to *public investment* or *private investment*. Therefore, the use of n-grams can help uncover the most frequent topics in the region with more precision. An n-gram is a sequence of words in a document that is defined as the neighboring sequences of tokens in a text. The sequence of two words is called bigrams and the sequence of three words is called trigrams. For example, consider the following n-gram:

*Document 4 = Climate change is a worldwide challenge*

Based on the words of Document 4, there are five bigrams in the text: *climate change*, *change is*, *is a*, *a worldwide*, and *worldwide challenge*. In the case of trigrams, there are four phrases in this document: *climate change is*, *change is a*, *is a worldwide*, and *a worldwide challenge*.

**Figure 7. Top 30 Bigrams of the CDC Reports**

(Average TF-IDF value)




Source: Authors calculations.

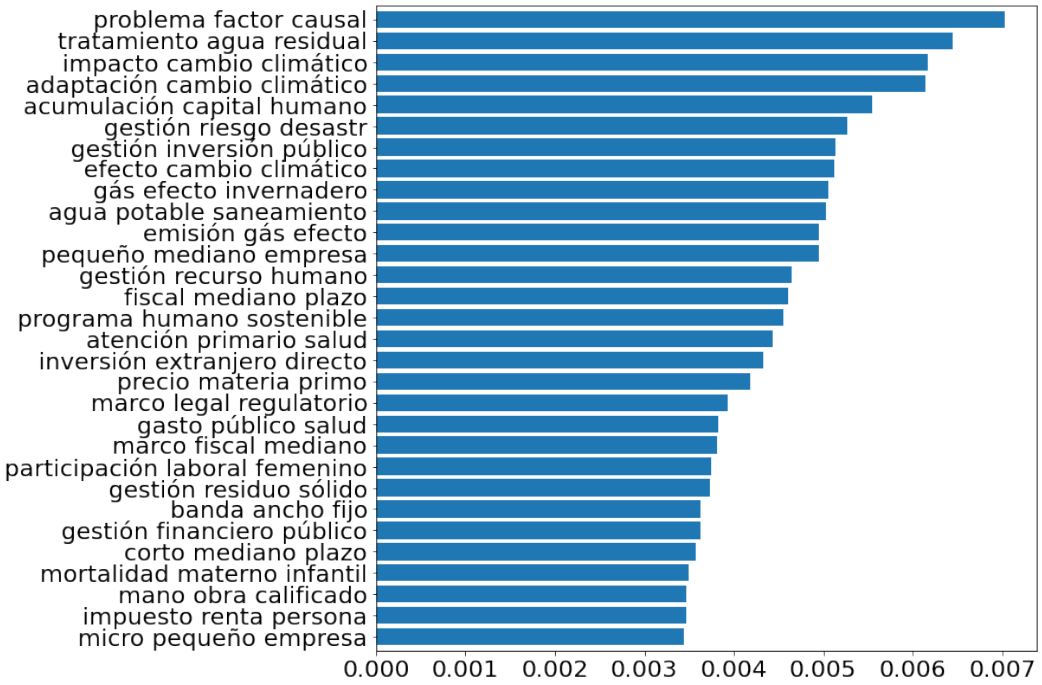
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports. TF–IDF refers to Term Frequency - Inverse Document Frequency.

Using TF–IDF, it is possible to find the most important bigram and trigrams in the texts. In this case, the TF–IDF value would be calculated for phrases instead of single words. Figure 7 shows the top 30 bigrams based on the average TF–IDF value over the 39 CDC reports. The bigrams present a clearer picture of the region's main development challenges. The most frequent bigram in the text is *climate change*. Also, *human capital* is now a more frequent phrase than *public investment*. In addition, key phrases such as *public spending, institutional capacity, legal framework, regulatory framework, social security, labor force, economic growth, private investment, renewable energy, water sanitation,* and *social security* appear as potential pressing challenges in LAC.

In the case of trigrams, figure 8 presents the top 30 phrases based on the average TF–IDF value over the 39 CDC reports. The figure also gives some insights about the intensity of challenges in the region. For example, some trigrams with a high average TF–IDF value are *small medium enterprises* and *foreign direct investment*, which could be related to the bottlenecks in the business climate to overcome LAC's development challenges. Once again, climate change is a frequent topic for the region and appears in different contexts such as: *climate change adaptation, climate change impact,* and *climate change effect*. Additionally, with trigrams it is possible to identify vital topics related to climate change such as: *greenhouse gas effect, gas effect emission, disaster risk management,* and *disaster risk reduction*. Moreover, additional phrases appear, such as: *sustainable human development, water sanitation service, public transport system, public finance management, female labor participation, primary health attention, human capital development, reproductive sexual health,* and *infant maternal mortality*, among others.

**Figure 8. Top 30 Trigrams of the CDC Reports**

(Average TF–IDF value)




*Source:* Authors calculations.

*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports. TF–IDF refers to Term Frequency - Inverse Document Frequency.

## 3. Topic Modeling: Latent Dirichlet Allocation

The previous sections identify the most common words and phrases in the texts, but do not explain how they are interconnected. Although the analysis of words, bigrams, and trigrams provides information about development challenges based on frequencies of key words, this information is not yet organized, and it cannot provide a comprehensive overview based on the most challenging topics. Another widespread NLP procedure, topic modeling, helps systematically identify which topics are discussed in a series of documents. It allows the topics that are common among documents to be grouped based on their content and helps yield conclusions about the most prominent subjects in the text. The clusters of words associated with a topic emerge from the documents themselves and not from an external criterion imposed by the reader. This section implements topic modeling using the Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003), which is an unsupervised machine learning statistical algorithm.

The LDA model groups words into topics based on their probability of repeated occurrences across documents. A topic is a set of words or phrases that, taken together, suggest a common theme. This type of model is commonly used in the classification of news, by grouping articles into topics based on the importance of the most relevant words in each of them. Thus, news articles in which the most frequent words are *president, government, election, campaign,* and *vote* could be associated with the implicit topic *"politics."* Importantly, the LDA model does not provide a label or name for the topic. It just clusters text elements sharing similar words. It is up to the researcher to label the topic.

Methodologically, the LDA model depends on two essential assumptions. First, topics should be found by searching for clusters of words that commonly appear together in the documents. Second, documents with similar topics use similar words. Therefore, based on these assumptions, topics could be interpreted as probability distributions of words, and documents as probability distributions of some latent topics. Consequently, the LDA model characterizes the documents as combinations of topics through the attribution of certain probabilities of the words in the text.

First, it is necessary to choose a fixed number of topics before implementing the LDA model, based on an empirical criterion. Next, the algorithm goes through each document, and randomly assigns each word to one of the topics. Then it iterates over every word in every document, grouping words with higher probability of appearing together in the documents. For every document ***d*** and for each topic ***t***, the algorithm calculates:

***p (topic t | document d) = number of words of topic t as the share of total words in document d.***

Then, the algorithm calculates:

***p (word w | topic t) = number of assignments of topic t as the share of all documents that come from the word w.***

And then, it reassigns *w* to a new topic, where the algorithm chooses topic *t* with probability:

***p (topic t | document d) * p (word w | topic t),***

which is essentially the probability that topic ***t*** generated word ***w***. Then, after repeating the previous steps many times, the LDA model eventually converges to a solution where the words stop changing topics.[9] The LDA model outcome is a set of probabilities for the words in every document that allows the mixture of topics in each document to be identified. It is also possible to identify the most common words for every topic.

The themes in the CDC reports usually are written following a structure that in some sense mirrors the IDB's sector divisions (such as education, health, transport, water and sanitation, and so on). A priori this structure could give the impression that it is then straightforward to capture the weight of each theme/division to determine its share in total words/sentences in the document. But development problems are more complex and certainly do not follow the Bank's division structure. For example, consider this paragraph from one of the CDC reports:

> *Researchers found that better access to water and sanitation in schools tends to raise attendance rates (particularly for girls) and children's abilities to learn. Good health and nutrition are essential prerequisites for effective learning. Healthier children perform better in school just as healthier workers are more productive. It is problematic to make a diagnosis of the service quality of operators throughout the country due to the very little information that is produced and disseminated systematically and regularly.*

The text analyzes a country's access to water and sanitation as a broader theme, but in the same section it mentions the connection with health and education as well as the weak institutional capacity of the government on the matter (an interconnected and cross-cutting topic). A mechanical count of words for each section in the CDC would not capture these linkages; hence, it is key to disentangle this mixture of topics systematically.

The LDA topic model is implemented at the sentence level in CDC reports to capture the interconnection of topics within each document. The model estimates a probability distribution over topics to identify whether a sentence is addressing, for instance, climate change or the development of the private sector, rather than framing the entire content of a document. Table 2 presents the number of sentences and words in each CDC report after cleaning, vectorizing, and lemmatizing the documents. The total number of sentences is 38,637 and the total word count is 377,944, which gives an average of 10 words per sentence. Then the DTM has 38,637 rows (sentences) and 13,908 columns (unique words in the Corpus).

[9] For a detailed description of the LDA model, see Blei, Ng, and Jordan (2003).

**Table 2. Sentences by CDC Report**

| Country and year of report | Number of sentences | Number of words | Average number of words by sentence |
|---|---|---|---|
| Argentina, 2016 | 2,343 | 24,418 | 10 |
| Argentina, 2019 | 1,202 | 10,929 | 9 |
| Bahamas, 2017 | 847 | 7,965 | 9 |
| Belize, 2020 | 1,826 | 18,000 | 10 |
| Bolivia, 2015 | 530 | 5,226 | 10 |
| Bolivia, 2019 | 931 | 8,562 | 9 |
| Brazil, 2015 | 572 | 5,168 | 9 |
| Brazil, 2018 | 1,587 | 14,623 | 9 |
| Barbados, 2018 | 1,086 | 9,555 | 9 |
| Chile, 2018 | 1,171 | 10,486 | 9 |
| Colombia, 2015 | 835 | 7,648 | 9 |
| Colombia, 2019 | 656 | 5,745 | 9 |
| Costa Rica, 2018 | 1,648 | 15,659 | 10 |
| Dominican Republic, 2016 | 687 | 8,076 | 12 |
| Dominican Republic, 2020 | 681 | 7,351 | 11 |
| Ecuador, 2018 | 940 | 10,506 | 11 |
| Ecuador, 2020 | 616 | 5,043 | 8 |
| Guatemala, 2016 | 652 | 6,292 | 10 |
| Guatemala, 2019 | 168 | 1,636 | 10 |
| Guyana, 2016 | 2,032 | 18,642 | 9 |
| Guyana, 2019 | 703 | 6,704 | 10 |
| Honduras, 2018 | 766 | 6,597 | 9 |
| Haiti, 2017 | 955 | 9,485 | 10 |
| Jamaica, 2015 | 843 | 8,220 | 10 |
| Jamaica, 2021 | 1,068 | 10,690 | 10 |
| Mexico, 2019 | 1,098 | 10,537 | 10 |
| Nicaragua, 2017 | 1,407 | 13,891 | 10 |
| Panama, 2015 | 204 | 1,621 | 8 |
| Panama, 2019 | 901 | 8,330 | 9 |
| Paraguay, 2017 | 973 | 11,170 | 11 |
| Peru, 2017 | 738 | 7,566 | 10 |
| Peru, 2020 | 539 | 7,077 | 13 |
| El Salvador, 2019 | 1,209 | 11,826 | 10 |
| Suriname, 2017 | 1,313 | 13,988 | 11 |
| Suriname, 2020 | 858 | 8,433 | 10 |
| Trinidad and Tobago, 2016 | 1,797 | 18,285 | 10 |
| Trinidad and Tobago, 2020 | 388 | 3,652 | 9 |
| Uruguay, 2019 | 1,473 | 15,259 | 10 |
| Venezuela, 2020 | 394 | 3,083 | 8 |
| **Total** | **38,637** | **377,944** | **10** |

*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

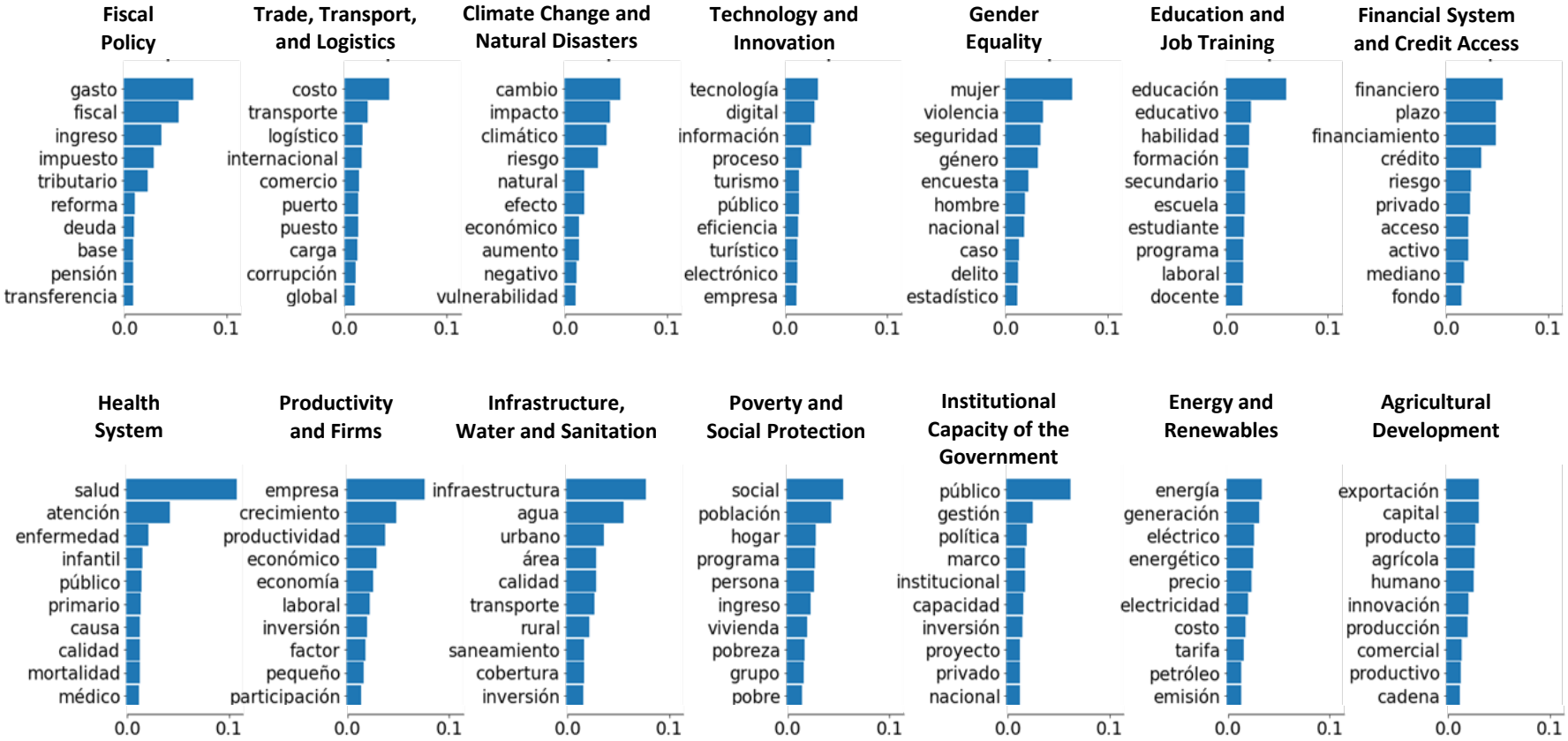
As mentioned, an important LDA parameter is choosing the number of topics. To determine it, this study maximized the topic coherence of the LDA model using different values for the

number of topics over the Corpus and the dictionary of unique words.[10] Topic coherence measures a single topic by the degree of semantic similarity between high scoring words in the topic. The results show that the optimal number of topics is 14. More details on this analysis are provided in appendix A.

After applying the LDA model to the Corpus of sentences, 10 words with the highest probability of belonging to every latent topic emerge. They are presented in figure 9. In the *"Fiscal Policy"* topic, the word with the highest probability of representing it is *spending*, followed by *fiscal, income, tax, tributary, reform, debt,* and so on. For the *"Energy and Renewables"* topic, the top word is *energy*, followed by *generation, electrical, energetic, price, electricity*, and so on. These words provide elements to identify latent topics. As previously explained, the labeling of the topics is up to the researcher using as reference the top words of each topic. Importantly, some words may appear in different topics because they are presented in different contexts across the documents. For example, the word *risk* is important in the topic "*Climate Change and Natural Disasters*," but also for the topic "*Financial System and Credit Access.*" This is an advantage of topic modeling. It considers the context of the most relevant words rather than just counting them mechanically.

[10] The library Scikit-Learn from Python is used to calculate the LDA topic modeling.

**Figure 9. Top 10 Words in the Distribution of Words by Topic in CDC Reports**
(Probability of belonging to every topic)




*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

Now that it is possible to estimate the development challenges topics presented in the CDC reports, the next step is to associate the main topic with each sentence. The LDA model estimates the probability distribution that a sentence belongs to any topic. Consider, for example, this sentence taken from one of the CDC reports:

> *"The low academic performance in standardized exams is not due to the universalization of basic education."*

It has an 87 percent probability of belonging to the topic labeled as *"Education and Job Training."* Take another example:

> *"Thus, climate models indicate on average an increase in extremes of high temperatures and rainfall in most of the country."*

It has an 88 percent probability of belonging to the topic *"Climate Change and Natural Disasters."* Then, with these probability estimates, it is straightforward to map each sentence to the topic with the highest probability.

## 4. Estimation Results

Table 3 displays the heatmap for the share of sentences allocated to the 14 estimated topics. The allocation rule linked sentences to topics using their highest probability of belonging to each topic. The exercise is performed over the latest available CDC report for each country.[11] Simple and weighted averages for the region appear at the bottom of the table. This study assesses which are the key challenges in each country by assuming that tougher challenges receive greater attention in the CDC report through a more extensive discussion, which results in a larger proportional share of the text.

The estimates show that five key topics constitute almost half of the sentences in the reports. Among these topics, "*Institutional Capacity of the Government*" emerges as the most critical challenge in the region; it constitutes the largest share (17 percent) of sentences in the CDC reports. Notably, these result remains consistent regardless of the weighting criteria used for the regional averages. When analyzing the results weighted by real GDP, it becomes evident that challenges associated with "*Fiscal Policy,*" "*Education and Job Training,*" "*Productivity and Firms,*" "*Infrastructure, Water and Sanitation,*" and "*Poverty and Social Protection*" continue to be important areas holding back progress in the region.

Institutional constraints also stand out at the country level; they are the top priority in 20 out of 26 IDB borrowing member countries. Furthermore, they are in the top three priorities in 5 out of the 6 remaining countries.

[11] For nine countries (Colombia, Dominican Republic, Ecuador, Guatemala, Guyana, Jamaica, Peru, Suriname, and Trinidad and Tobago), their latest CDC report is an update, and it complements the main challenges found in the previous full version of the CDC (see updates indicated in table 1). For these cases the results of the sentences include both the update and the previous full version of the report. In the other cases, the latest CDC corresponds to a new and complete version of their development challenges report.

## Table 3. Heatmap of the Distribution of Topics by Country in CDC Reports

(Percentage of sentences with the highest probability of belonging to each topic by country)

| Country | Fiscal Policy | Trade, Transport, and Logistics | Climate Change and Natural Disasters | Technology and Innovation | Gender Equality | Education and Job Training | Financial System and Credit Access | Health System | Productivity and Firms | Infrastructure, Water, and Sanitation | Poverty and Social Protection | Institutional Capacity of the Government | Energy and Renewables | Agricultural Development | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Argentina | 10.5 | 4.3 | 3.1 | 8.3 | 3.0 | 9.7 | 7.7 | 3.6 | 8.1 | 7.9 | 8.2 | 13.3 | 6.2 | 6.2 | 100 |
| The Bahamas | 8.4 | 6.4 | 8.9 | 4.3 | 9.8 | 11.0 | 5.3 | 6.3 | 11.6 | 3.0 | 6.8 | 14.4 | 1.9 | 2.1 | 100 |
| Belize | 10.9 | 7.5 | 7.3 | 13.3 | 3.3 | 7.0 | 5.3 | 3.4 | 6.1 | 5.6 | 6.5 | 11.9 | 5.3 | 6.5 | 100 |
| Bolivia | 9.5 | 7.0 | 4.7 | 3.4 | 2.7 | 5.5 | 4.5 | 4.4 | 15.1 | 9.9 | 7.4 | 12.9 | 4.6 | 8.4 | 100 |
| Brazil | 9.6 | 7.3 | 3.3 | 9.2 | 2.1 | 5.2 | 8.6 | 2.3 | 6.3 | 6.7 | 3.6 | 23.6 | 6.7 | 5.5 | 100 |
| Barbados | 14.6 | 6.8 | 7.6 | 7.0 | 3.8 | 10.4 | 3.4 | 5.0 | 7.2 | 5.4 | 7.6 | 11.6 | 5.9 | 3.6 | 100 |
| Chile | 3.6 | 4.4 | 8.4 | 9.0 | 3.3 | 10.8 | 5.0 | 6.1 | 9.2 | 8.8 | 8.0 | 14.3 | 2.6 | 6.4 | 100 |
| Colombia | 15.8 | 5.2 | 3.2 | 4.2 | 4.5 | 7.2 | 4.8 | 3.6 | 7.3 | 10.8 | 13.7 | 12.7 | 3.2 | 3.8 | 100 |
| Costa Rica | 12.3 | 4.2 | 7.5 | 7.1 | 3.3 | 12.0 | 3.7 | 3.4 | 5.2 | 11.2 | 4.0 | 15.8 | 5.6 | 4.8 | 100 |
| Dominican Republic | 12.9 | 5.9 | 4.5 | 5.2 | 3.8 | 9.4 | 4.8 | 4.9 | 5.7 | 5.2 | 9.9 | 17.1 | 4.5 | 6.3 | 100 |
| Ecuador | 17.8 | 4.0 | 7.4 | 3.6 | 2.1 | 6.4 | 3.6 | 4.0 | 11.2 | 5.5 | 11.1 | 13.2 | 4.0 | 6.2 | 100 |
| Guatemala | 10.0 | 4.3 | 3.8 | 3.2 | 3.7 | 13.8 | 2.4 | 9.6 | 5.1 | 7.8 | 6.5 | 22.3 | 2.1 | 5.5 | 100 |
| Guyana | 6.8 | 6.9 | 5.8 | 4.2 | 7.2 | 7.6 | 5.6 | 5.3 | 8.2 | 4.7 | 5.7 | 20.1 | 5.3 | 6.6 | 100 |
| Honduras | 3.8 | 7.2 | 6.1 | 7.3 | 6.1 | 7.0 | 6.8 | 6.0 | 7.0 | 7.7 | 7.8 | 18.3 | 3.7 | 5.1 | 100 |
| Haiti | 2.6 | 5.9 | 7.0 | 4.5 | 7.2 | 6.2 | 3.1 | 4.4 | 8.7 | 6.7 | 6.8 | 19.6 | 6.2 | 11.1 | 100 |
| Jamaica | 8.1 | 4.8 | 7.4 | 7.8 | 9.3 | 9.4 | 4.6 | 7.3 | 7.8 | 4.1 | 6.1 | 13.3 | 5.0 | 5.1 | 100 |
| Mexico | 7.7 | 2.7 | 5.1 | 9.5 | 3.2 | 8.6 | 7.7 | 4.3 | 4.9 | 11.4 | 7.9 | 17.8 | 6.1 | 3.2 | 100 |
| Nicaragua | 6.5 | 4.5 | 4.0 | 5.3 | 2.1 | 10.7 | 9.7 | 8.7 | 7.5 | 7.4 | 7.2 | 11.4 | 6.4 | 8.5 | 100 |
| Panama | 5.5 | 5.0 | 4.7 | 10.2 | 2.3 | 11.1 | 2.6 | 6.0 | 5.7 | 9.7 | 6.8 | 21.5 | 2.9 | 6.1 | 100 |
| Paraguay | 6.3 | 4.6 | 4.5 | 4.4 | 2.9 | 4.3 | 4.5 | 3.5 | 4.8 | 13.1 | 7.7 | 26.1 | 6.0 | 7.3 | 100 |
| Peru | 7.6 | 4.5 | 4.3 | 4.3 | 3.3 | 8.0 | 3.6 | 3.8 | 11.8 | 7.8 | 9.0 | 23.7 | 1.9 | 6.4 | 100 |
| El Salvador | 6.7 | 5.5 | 6.7 | 8.1 | 7.3 | 7.4 | 5.5 | 3.5 | 6.0 | 7.0 | 9.2 | 15.6 | 4.7 | 6.6 | 100 |
| Suriname | 5.3 | 6.3 | 7.3 | 4.7 | 3.0 | 9.4 | 4.6 | 4.7 | 7.1 | 6.7 | 4.0 | 24.4 | 5.5 | 7.1 | 100 |
| Trinidad and Tobago | 7.5 | 5.8 | 5.1 | 3.5 | 7.6 | 8.2 | 6.0 | 5.8 | 6.7 | 6.3 | 6.7 | 21.2 | 5.6 | 4.0 | 100 |
| Uruguay | 6.3 | 3.6 | 4.8 | 9.3 | 3.9 | 10.9 | 5.4 | 4.3 | 6.9 | 7.2 | 8.1 | 15.3 | 4.8 | 9.2 | 100 |
| Venezuela | 10.2 | 6.3 | 9.4 | 2.3 | 3.6 | 4.6 | 1.3 | 6.1 | 8.1 | 13.7 | 13.7 | 8.1 | 9.1 | 3.6 | 100 |
| **Simple average LAC** | **8.7** | **5.4** | **5.8** | **6.3** | **4.4** | **8.5** | **5.0** | **5.0** | **7.7** | **7.7** | **7.7** | **16.9** | **4.8** | **6.0** | 100 |
| **Sentences weighted average LAC** | **8.8** | **5.5** | **5.8** | **6.4** | **4.5** | **8.6** | **5.2** | **4.9** | **7.6** | **7.3** | **7.3** | **17.3** | **4.9** | **6.0** | 100 |
| **Population weighted average LAC** | **9.5** | **5.3** | **4.6** | **7.5** | **3.1** | **7.2** | **6.6** | **3.8** | **7.0** | **8.7** | **7.3** | **18.7** | **5.6** | **5.2** | 100 |
| **Real GDP weighted average LAC** | **9.4** | **5.1** | **4.5** | **8.0** | **3.0** | **7.6** | **6.9** | **3.8** | **6.8** | **8.8** | **7.2** | **18.6** | **5.6** | **5.0** | 100 |

*Source:* Authors calculations.

*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports. The heatmap colors are based on the topic values for each country and the averages for Latin America and the Caribbean. Population and real GDP data used to calculate the weighted averages come from the World Bank's World Development Indicators, and the year of reference corresponds to 2019. Real GDP is gross domestic product in constant 2017 international dollars using purchasing power parity (PPP) rates. The data come from World Bank International Comparison Program, except for the value for Venezuela, which is calculated using information from the International Monetary Fund (IMF) World Economic Outlook (WEO) 2021 database.

### 4.1 The cross-cutting nature of institutional capacity constraints

Any public policy program or project could be affected by capacity constraints. So, it is informative to disentangle where these constraints are more binding. Consider, for example, the following text from a CDC report:

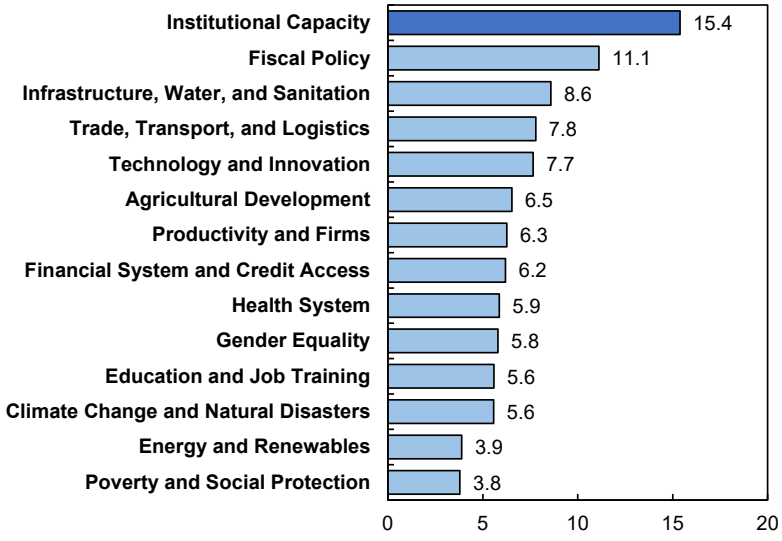
*"Building a More Effective Government: recommended measures to build a more effective government are organized into five policy areas: (1) improve the efficiency of public spending and build public sector capacity; (2) strengthen the framework of fiscal federalism; (3) rationalize spending and improve the quality of public spending; (4) use e-government solutions to promote transparency, accountability, and efficiency; (5) promote equal opportunities."*

It has a 52 percent probability of belonging to the topic *"Institutional Capacity of the Government,"* but the second most likely topic is *"Fiscal Policy,"* with a 24 percent probability. Consider another example:

*"Likewise, it introduced new tools for civil service reforms through performance evaluation and the identification of labor competencies."*

It has a 62 percent probability of belonging to the topic *"Institutional Capacity of the Government,"* and its second most likely topic is *"Education and Job Training,"* with a 29 percent probability.

**Figure 10. Second Most Probable Topic for Sentences Classified as Institutional Capacity of the Government**
(percentage of sentences in the topic in CDC reports)




Source: Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

Based on the LDA model probability distribution estimates, it is possible to assess the sector where institutional constraints are more likely to bind. This information allows the areas where deficient institutional capacities might be more pronounced to be identified. So, the analysis further associated the 6,037 sentences that were originally linked to *Institutional Capacity* with the topic with the second highest probability. Figure 10 displays the results including the topic of *"Institutional Capacity"* for those sentences where it was not possible to assign a second most likely topic (15.4 percent of sentences) because the second highest probability is too low[12]. The figure depicts the cross-cutting nature of the topic, ranging from 11.1 percent of the sentences related to *"Fiscal Policy"* to 3.8 percent of them related to *"Poverty and Social Protection."* It suggests that institutional constraints play a comparatively larger role in fiscal policy and in infrastructure.

### 4.2 The stability of development challenges over time

The time elapsed between two editions of the CDC report for a country is between two to six years (see table 1), with an average time of 3.5 years. Half of the 26 countries have at least two CDC reports in the sample. To analyze whether the identified development priorities remain constant over time, it is useful to compare the similarity of the CDC reports for the subset of countries with two documents in the sample. One of the most popular techniques to compare documents and assess how close they are is the cosine similarity (Manning, Raghavan, and Schütze 2008). It is defined as the geometric representation of the scalar product of the vectorial representation of two documents given by the following expression:

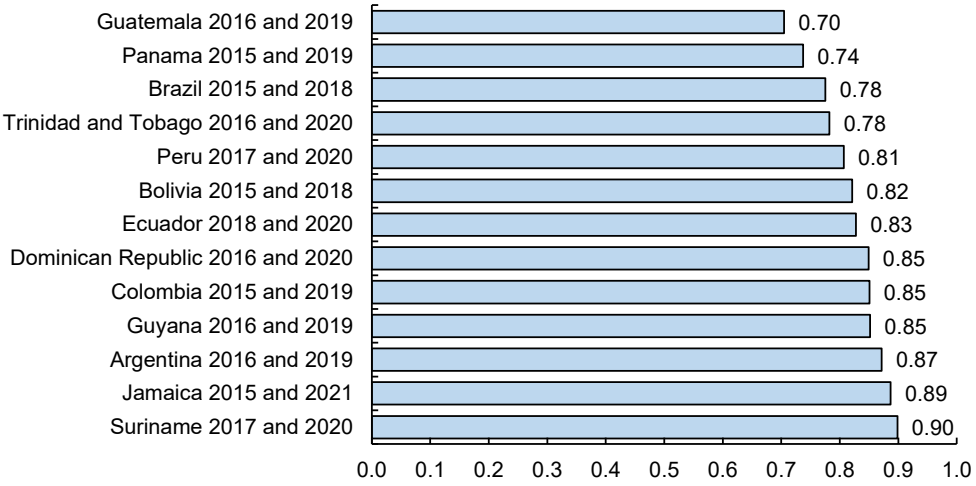
$$Cosine\ similarity = \frac{A * B}{\|A\| * \|B\|} = \cos(A, B),$$

where *A*B* is the inner product of vectors *A* and *B*, $\|A\|$ is the Euclidean norm of vector *A*, and $\|B\|$ is the Euclidean norm for vector *B*. The measure computes the cosine of the angle between the two vectors. A cosine value of 0 means that the two vectors are orthogonal to each other (related at a 90-degree angle), and thus the documents are not similar. The closer the value is to 1, the smaller the angle between the two vectors and hence the greater the similarity of the documents. The vectorial representation of the documents is computed as the TF–IDF matrix presented in section 2, which captures how relevant every word in a document is, while considering their importance in the other document.

Figure 11 shows the cosine similarity score for the 13 countries with 2 CDC reports available. The resulting cosine similarity scores are high in all cases (Han, Kamber, and Pei 2012), ranging from 0.7 for Guatemala to 0.9 for Suriname. The figure presents evidence suggesting that the diagnosed development challenges have not changed much over time for the countries in the sample. This result is somehow expected, given that the structural challenges that countries in LAC are facing are unlikely to change within a decade.

[12] A threshold of 10 percent was set to determine that the probability of the second most likely topic was too low to make an assignment.

**Figure 11. CDC Reports Cosine Similarity of LAC Countries**
(A higher score represents higher similarity between CDC reports)

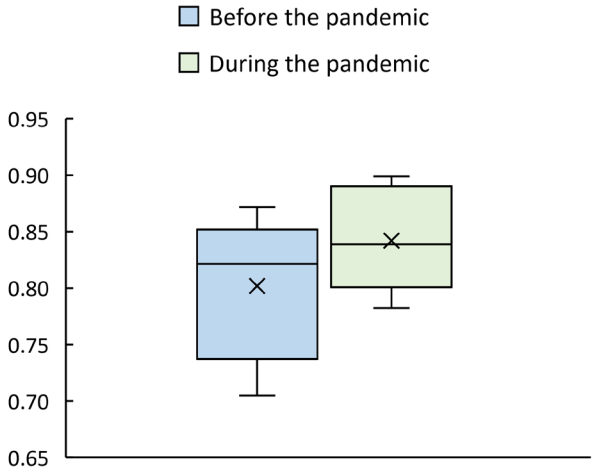



*Source:* Authors calculations.
*Note:* CDC refers to the Inter-American Development Bank (IDB) Country Development Challenges reports.

The COVID-19 pandemic provides a useful setting to study the stability of development challenges, as some could argue that the pandemic has changed the sense of urgency of some challenges, with some development gaps now more pressing than before. To test this hypothesis, the subsample is further divided between countries with two CDC reports written before the pandemic and countries with the last CDC written during the pandemic. Out of the 13 countries with two CDC reports in the subsample, 7 have CDC reports written before the pandemic, and 6 have the latest CDC report written during the pandemic (2020 and 2021) (see table 1). If the pandemic changed development challenges drastically, then the similarities of CDC reports in the first group should be larger than in the second group, because for any country the post-pandemic CDC reports should be different (less similar) than the pre-pandemic one.

Figure 12 shows the distribution for the cosine similarity for both groups (the blue box corresponds to the 7 countries whose CDC reports were written before the pandemic, and the green box corresponds to the 6 countries whose second CDC reports were written during the pandemic). The countries that could have had their development challenges modified by the COVID-19 pandemic in their second CDC report have higher similarity scores than those whose reports that were written before the pandemic, and differences between groups are not significant. This result suggests that during the sample period, the pandemic did not introduce new challenges and did not drastically change the diagnosed developing challenges previously identified in sample countries' CDC reports. We interpret this result not as an indication that certain gaps have not worsened but rather as a sign that no new challenges have surfaced. In fact, results reflect the structural nature of development challenges and their relative stability over short periods of time.

**Figure 12. Distribution of CDC Reports Cosine Similarities and COVID-19 Pandemic**
(A higher score represents higher similarity between CDC reports)




*Source:* Authors calculations.
*Note:* The blue box corresponds to the seven countries with their Country Development Challenge (CDC) reports written before the pandemic, and the green box corresponds to the six countries with their second CDC reports written during the pandemic.

## 5. Remarks

The aim of this paper is to assess the most pressing development challenges in the region using the Country Development Challenges (CDC) reports written by country economists from the Inter-American Development Bank. A Latent Dirichlet Allocation (LDA) topic model was applied to identify 14 development topics present in the CDC reports. To rank development challenges by their intensity, it is assumed that reports would have more extensive discussions about the challenges perceived as more binding. The results suggest that *Institutional Capacity of the Government* is the most pressing challenge, accounting for about 17 percent of all sentences in the CDC reports, on average. Other crucial challenges include *Fiscal Policy*, *Education and Job Training*, *Infrastructure*, *Productivity and Firms,* and *Poverty and Social Protection*. These six challenges account on average for 60 percent of the text discussing development challenges in the CDC reports.

The exercise also finds preliminary evidence in the sample period that despite the perception that the region may be facing new challenges, this is not the case given that CDC reports are very "similar," even after the COVID-19 outbreak. This is somehow expected due to the structural nature of development challenges and their relative stability during short periods of time.

## 6. References


Acemoglu, D., S. Johnson, J. Robinson. 2005. Chapter 6 Institutions as a Fundamental Cause of Long-Run Growth, Editor(s): Philippe Aghion, Steven N. Durlauf, Handbook of Economic Growth, Elsevier, Volume 1, Part A, 2005, Pages 385-472.

Acemoglu, D., and J. Robinson. 2008. "The Role of Institutions in Growth and Development. Commission on Growth and Development." Commission on Growth and Development Working Paper No. 10, World Bank, Washington, DC. https://openknowledge.worldbank.org/handle/10986/28045.

Biasi, B., and S. Ma. 2022. "The Education-Innovation Gap." NBER Working Paper 29853, National Bureau of Economic Research, Cambridge, MA.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (January): 993–1022.

Borensztein, E., S. Miller, G. Sánchez, and P. Valenzuela. 2014. "Development Diagnostics for the Southern Cone." Working Paper IDB-WP-516, Inter-American Development Bank, Washington, DC.

Cormier, Ben, and Mark Manger. 2022. "Power, Ideas, and World Bank Conditionality." Review of International Organizations 17 (3): 397–425.

Edwards, S., G. Esquivel, and G. Márquez, eds. 2007. *The Decline of Latin American Economies: Growth, Institutions, and Crises*. University of Chicago Press for the National Bureau of Economic Research. https://EconPapers.repec.org/RePEc:nbr:nberbk:edwa04-1.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword Assisted Topic Models." Working Paper, arXiv:2004.05964.

Gentzkow, M. and J. M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." Econometrica 78 (1): 35–71.

Han, J., M. Kamber, and J. Pei. 2012. "Getting to Know Your Data." Chapter 2 In *Data Mining Concepts and Techniques, third edition.* The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 39–82.

Hansen, S., M. McMahon, and A. Prat. 2018. "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach." *Quarterly Journal of Economics* 133 (2): 801–70.

Hausmann, R., D. Rodrik y A. Velasco, (2005). "Growth Diagnostics". Cambridge, MA: Kennedy School of Government Faculty, Harvard University.

Izquierdo, A., J. Llopis, U. Muratori, and J. J. Ruiz Gomez. 2016. "In Search of Larger Per Capita Incomes: How to Prioritize across Productivity Determinants?" IDB Working Paper IDB-WP-680, Inter-American Development Bank, Washington DC.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press.

Robinson, J. A., D. Acemoglu, and S. Johnson. 2005. “Institutions as a Fundamental Cause of Long-Run Growth.” Chapter 6 in *Handbook of Economic Growth* 1A, 386–472. Elsevier.

Röder, M., A. Both, and A. Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures.” WSDM 2015, Proceedings of the 8th ACM International Conference on Web Search and Data Mining.

Röder, Michael & Both, Andreas & Hinneburg, Alexander. (2015). Exploring the Space of Topic Coherence Measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 399-408. 10.1145/2684822.2685324.

Romer, C. D., and D. H. Romer. 2010. “The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks.” *American Economic Review* 100 (3): 763–801.

Syed, S., and M. Spruit. 2017. “Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation.” 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*,* 165–174. IEEE. doi: 10.1109/DSAA.2017.61.

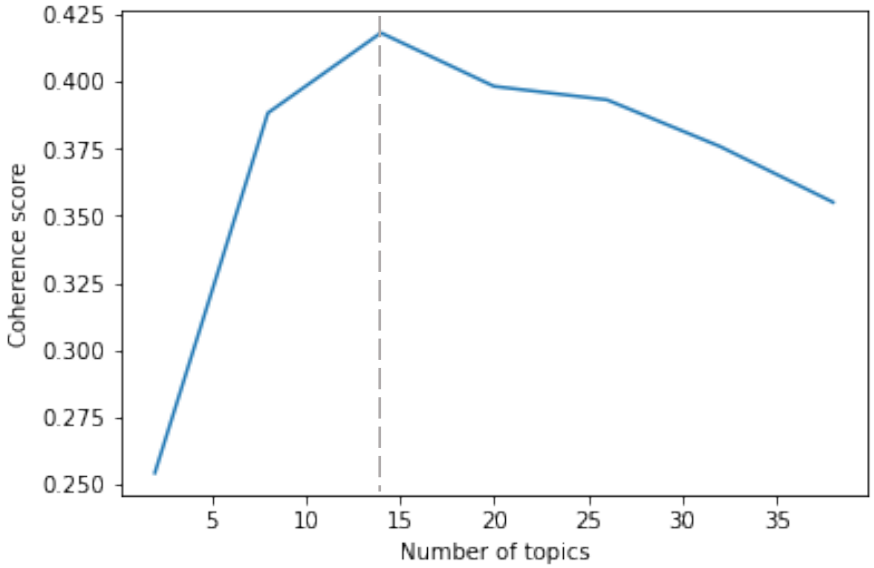
## Appendices

### Appendix A: Optimal Number of Topics

One methodology to choose the optimal number of topics is analyzing the "topic coherence" of the Latent Dirichlet Allocation (LDA) model using different values as the number of topics over the Corpus and the dictionary of unique words. Research by Röder et al. (2015) presents a framework to construct topic coherence measures in the context of text analytics and topic modeling. Topic coherence measures a single topic by the degree of semantic similarity between high scoring words in the topic. A set of statements is coherent if the statements support one another. Consequently, the approach for finding the optimal number of topics is building various LDA models with different number of topics and choosing the one that provides the highest coherence value. Studies by Syed and Spruit (2017) and Mifrah and Benlahmar (2020) provide implementations of topic coherence for the LDA model.

The library Genism[13] in Python is used to calculate the coherence scores for LDA models with different values for the number of topics. The coherence measure used is the Coherence Value method (Röder, et al. 2015). Figure A.1 shows the results of the coherence scores for the LDA models, with values ranging from 2 to 40 as the number of topics. The highest level of coherence for the IDB Country Development Challenges (CDC) reports is reached when the number of topics in the LDA model is 14.

**Figure A.1. Coherence Scores with Different Number of Topics**
*(Higher score represents higher coherence)*




*Source:* Authors calculations.

[13] The coherence pipeline of the library is presented in this link: https://radimrehurek.com/gensim/models/coherencemodel.html

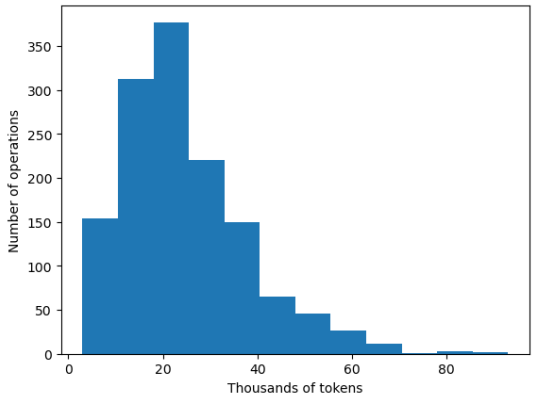
## Appendix B. Topic Classification of Sovereign Guaranteed Loans at the Inter-American Development Bank

This appendix extends the LDA model to build a keyword-assisted topic model (KeyATM) to classify sovereign guaranteed (SG) loans approved by the Inter-American Development Bank (IDB) from 2010 to 2022. This model requires the input of meaningful keywords before identifying the relevant topics (Eshima, Imai, and Sasaki 2020). It has the advantage that the resulting topics are easier to interpret, and its performance is satisfactory compared to the performance using human coders.[14]

This exercise then focuses on the loans' objectives and uses Natural Language Processing (NLP) to classify them to determine the areas where IDB SG Loan Approvals are allocated. The sample has 1,368 loans with their objectives, approved from 2010 to 2022.[15] The data set was tokenized and lemmatized, and further prepared to retain nouns, pronouns, or adjectives. Figure B.1 presents the histogram and descriptive statistics of the number of words for the loans' objectives after processing.

**Figure B.1. Histogram and Descriptive Statistics of Objectives of IDB Sovereign Guaranteed (SG) Loans after Text Processing**

**Panel A. Histogram of words**




**Panel B. Descriptive Statistics (Number of words)**

| Descriptive Statistics | |
|---|---|
| Number of loans | 1,368 |
| Mean words | 24.5 |
| Standard Deviation | 13.3 |
| Minimum | 3.0 |
| Percentile 25% | 15.0 |
| Percentile 50% | 22.0 |
| Percentile 75% | 31.0 |
| Maximum | 93.0 |

*Source:* Authors' calculations.

To implement the KeyATM, it is necessary to set the number of topics and keywords that characterize each topic. The number of topics and their respective keywords are obtained from the topic model previously estimated with the CDC reports. These reports provide a good reference to sort the IDB SG operations by the objectives of because they are exhaustive in the sense that their discussion of development challenges is comprehensive; hence, they provide a rich set of words to characterize topics in IDB loan documents. Figure 9 presents the resulting

[14] The Key-ATM model has been implemented to analyze how the World Bank's economic research and policy priorities influence its policy loan conditions (Cormier and Manger 2022).

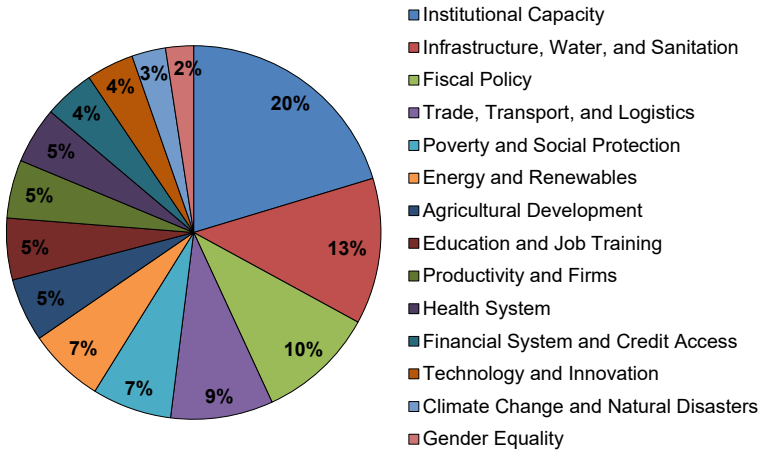
[15] Loans objectives are recorded in IDB systems. In 21 operations, because the information in the IDB system was either absent or incomplete, it was necessary to refer directly to the loan documents themselves.

topic-identifying keywords. As discussed, these keywords correspond to the resulting top 10 words with the highest probability of belonging to each of the 14 previously estimated topics.

Now it is possible to map the SG loans into the previously identified topics using their objectives and keywords from the LDA model. The KeyATM calculates the probability that the objectives correspond to each topic. For example, the objective in one of the SG loans, "The general objective of this project is to contribute to reducing morbidity and mortality from COVID-19 and to mitigate the other indirect health effects of the pandemic" has a 94 percent probability of belonging to the topic "*Health System*." For another SG loan, "The overall objective of the project is to strengthen the supply of electricity in the northern area of the National Interconnected System and improve reliability." This objective has an 85 percent probability of belonging to the topic "*Energy and Renewables*." Each loan is then mapped to a topic, based on maximizing the likelihood that its objectives belong to it.

Figure B.2 presents the resulting distribution of loans for the sample. The most common topic in the SG loans is "*Institutional Capacity*," with 20 percent of the loans referring to this theme. The following frequent topics are: "*Infrastructure, Water, and Sanitation*," "*Fiscal Policy*," and "*Trade, Transport, and Logistics*." Three topics that are linked to the reduction of poverty and inequality in the region ("*Poverty and Social Protection*," "*Education and Job Training*," and "*Health System*") represent 17 percent of the IDB SG loans. Three other topics directly linked to the development and strengthening of the private sector ("*Productivity and Firms*," "*Financial System and Credit Access*," and "*Technology and Innovation*") represent 14 percent of the SG loans from 2010 to 2022.

**Figure B.2. Distribution of 1,368 IDB Sovereign Guaranteed (SG) Loans Assigned to the Topic with the Highest Corresponding Probability, 2010–2022**



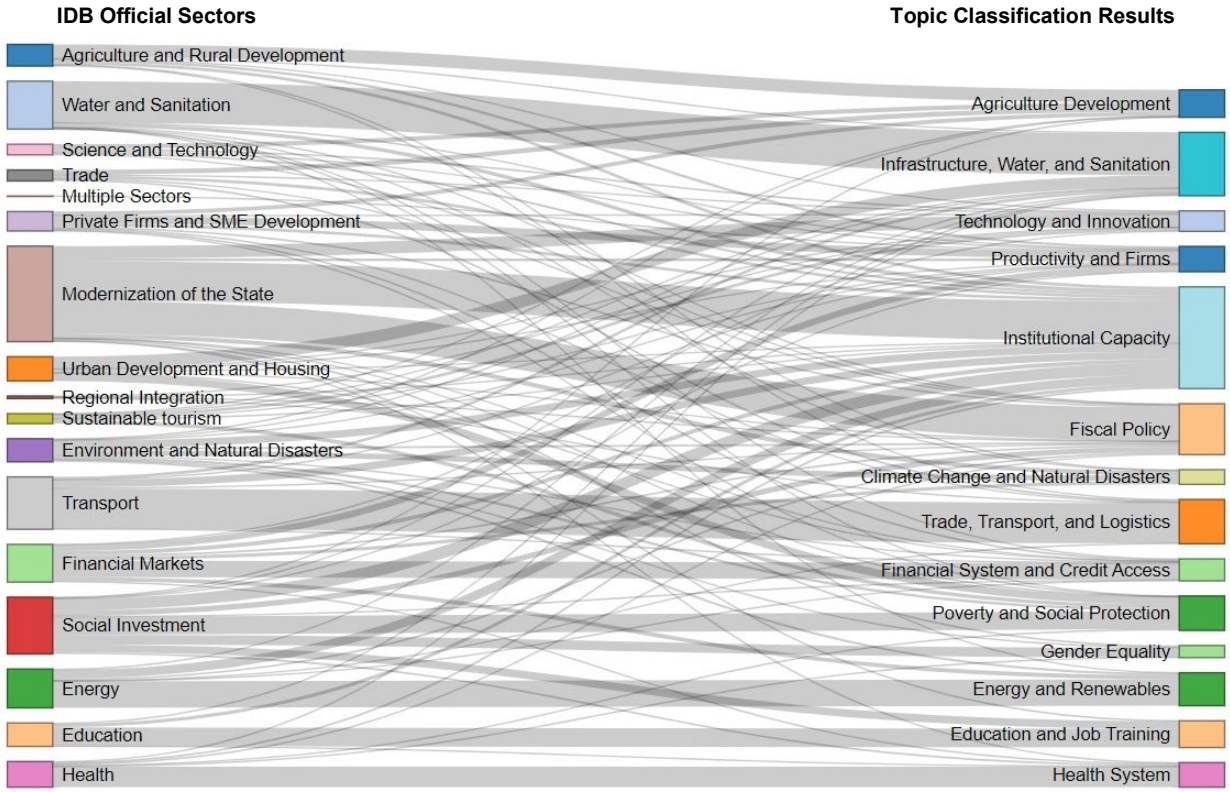

*Source:* Authors' calculations.

Figure B.3 maps the 1,368 SG loans officially classified by sectors into topics, using a Sankey Diagram. The diagram shows the differences between the official IDB sector classification[16] and

[16] Official sector categories are defined by IDB internal guidelines (OA-123).

the classification obtained with the KeyATM. It plots how the official classification of loans by IDB sectors maps into the estimated topics. In the diagram, the width of the arrows is proportional to the number of loans related to the sectors and topics. The diagram shows that in addition to the strong logical connection between the IDB official sector and its corresponding estimated topic, some projects aim at topics beyond their originating sector. For example, the objectives of operations classified as "Modernization of the State" (the brown column on the left) are mainly aimed at improving "*Institutional Capacity*," "*Fiscal Policy*," and "*Technology and Innovation*." Loan projects classified as "Social Investment" (the red column on the left), in addition to their natural connection to the "*Poverty and Social Protection*" topic (dark green column on the right), also aim to improve "*Institutional Capacity*," "*Gender Equality*," and "*Education and Job Training*." Moreover, the "*Institutional Capacity"* topic (the light blue column on the right) is strongly connected to SG loans in sectors across the IDB, which is consistent with the purpose of many operations and the cross-cutting nature of the topic.[17]

**Figure B.3. Official Sectors of the 1,368 IDB Sovereign Guaranteed (SG) Loans Mapped to the Resulting Topics, 2010–2022**

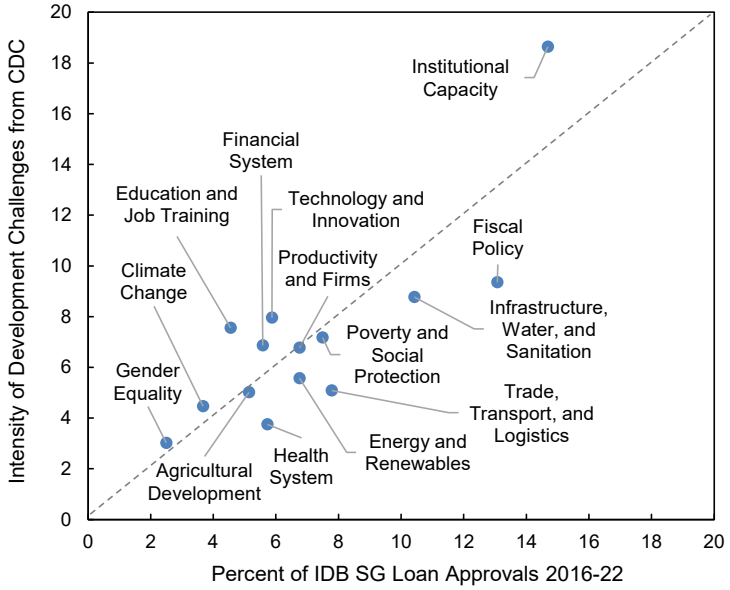



*Source:* Authors' calculations.
*Note:* The left side of the figure represents the 1,368 sovereign guaranteed loans approved by the IDB in the sample according to their IDB official sector classification. The right side depicts the disaggregation of the objectives according to the topic classification model.

[17] This topic classification exercise should not be mistaken with the mainstreaming of the IDB Group's cross-cutting strategic priorities reported in its Corporate Results Framework. While the topic classification of SG loans is performed at the objectives level, mainstreaming of strategic priorities is measured at the expected results level.

Based the results obtained so far, it is possible to compare how volumes in loan approvals at the regional level are related to the intensity of development challenges in recent years. For this exercise, the sample of the SG loans is adjusted to only include operations approved from 2016 to 2022 because CDC reports started in 2016. Figure B.4 plots the intensity of development gaps, assessed from CDC reports, against the percentage of IDB SG lending related to each topic. It shows that on average, IDB SG lending is allocated to areas where development gaps appear to be more intense based on the analysis of CDC reports. The possibility of reverse causality, implying that CDC reports might cover areas more extensively where there is increased lending activity, is lessened because the main authors of the CDC reports do not engage in loan origination efforts.

**Figure B.4. LAC Development Challenges and IDB Sovereign Guaranteed (SG) Loan Approvals, 2016–2022**




*Source:* Authors' calculations.
*Note:* A topic model was used to classify the sentences of the Country Development Challenges (CDC) reports for 26 Latin American and Caribbean (LAC) countries by development challenges. The intensity of the development challenges was computed as the share of sentences in CDC reports with the highest probability of belonging to each topic. To calculate the weighted averages over topic shares for the region, real GDP (2019) from the World Bank, World Development Indicators denominated in constant 2017 international dollars using purchasing power parity rates was used. The estimated model was applied to the objectives of the IDB SG loans approved (2016–2022) to classify them into the established topics.