Text Analysis of Project Completion Reports

César E. Montiel Olea, Leonardo R. Corral*

June 8, 2021

Abstract

Project Completion Reports (PCRs) are the main instrument through which different multilateral organizations measure the success of a project once it closes. PCRs are important for development effectiveness as they serve to understand achievements, failures, and challenges within the project cycle the can feed back into the design and execution of new projects. The aim of this paper is to introduce text analysis tools for the exploration of PCR documents. We describe and apply different text analysis tools to explore the content of a sample of PCRs. We seek to illustrate a way in which PCRs can be summarized and analyzed using innovative tools applied to a unique dataset. We believe that the methods presented in this investigation have numerous potential applications to different types of text documents routinely prepared within the Inter-American Development Bank (IDB).

JEL classifications C1; O1; O12; O22

Keywords Text Analysis, Development Effectiveness, Project Performance

*Montiel Olea: Inter-American Development Bank, cesarmo@iadb.org. Corral: Inter-American Development Bank, leonardoc@iadb.org.

1 Motivation and Background

Text analysis (often referred as text mining) is the process of transforming unstructured text in documents into normalized, structured data that is suitable for analysis. Text analysis has become a fast-growing area of empirical research. Text analysis has been recently used to study media slant in U.S. news coverage (Gentzkow and Shapiro, 2010), internal policy makers' deliberations from the Federal Open Mark Committee (FOMC) (Hansen et al., 2018), and tax changes in the U.S. from presidential speeches and reports of Congressional committees (Romer and Romer, 2010). In several international development organizations, a great amount of data is available in text documents. This is the case of Project Completion Reports (PCRs). A PCR is a self-evaluation instrument that is undertaken at the end of the execution cycle of a project. More specifically, the purpose of a PCR is to promote accountability and elicit learning lessons to replicate achievements and avoid mistakes in ongoing and future projects.

PCRs are a rich source of text data and are by nature unstructured and high-dimensional. To the best of our knowledge, text analysis of PCRs has rarely been performed in multilateral organizations. One of the few works on the subject is the one from Carneiro and Garbero (2018), in which they perform a descriptive content analysis of 72 PCRs from the International Fund for Agricultural Development (IFAD). Using an IFAD-specific conceptual framework, their work analyzes the types of achievements and challenges presented in the PCRs, classifying them into topics and sub-topics and categorizing claims about project results according to their tone, indicator level (impact, outcome or output), and type of evidence presented. Their results suggest that the majority of claims from PCRs relate to output or outcome level results and are not explicitly supported by evidence. Importantly, though, is that the authors do not make use of text automation and algorithmic categorization, but instead they rely in human coding which limits the reliability of their results as it hinges upon substantive subject-matter knowledge.[1]

In this paper we present the first quantitative analysis of the PCRs from the Inter-American Development Bank (IDB). The purpose of the paper is two-fold. First, we seek to construct a unique, structure-form dataset that allows us to organize the content of the PCRs in a manner that is suitable for quantitative analysis. Second, we illustrate a way of describing and analyzing

[1]The authors themselves acknowledge that one of the issues with descriptive content analysis is the varying interpretations and understandings of the texts arising from human coding. Despite that they use 3 different coders and two inter-reliability tests to reduce this bias, this issue could have been ameliorated by using alternative text automation tools such as machine learning.

the PCRs using different tools from text analysis. In order to do so, and to ensure consistency among sections, we use a sample of 44 PCRs in Spanish that were approved under the 2018 PCR Guidelines.

The remainder of this paper proceeds as follows. Section 2 presents the study design, describing the structure of the Bank's PCRs, their validation process, the data that we use and the methodology that we follow. In section 3 we illustrate how to perform an exploratory data analysis of the PCRs using text preprocessing tools and text analysis methods. In section 4 we use a unsupervised machine learning model to understand the content of the PCRs. We characterize the PCRs based on the underlying "topics" they cover and illustrate a way to associate the topics to metrics of project's success and effectiveness. Section 5 presents conclusions from of our analysis and possible venues of future research.

2 Study Design

2.1 Structure and score rating of PCRs

According to the 2018 IDB's PCR Guidelines, the Bank requires a PCR for each lending operation. There are different types of lending operations. In this paper we focus our attention on investment loans.[2][3] Typically, a PCR is prepared when a projects reaches operation closure (CO) and it must be prepared by the Unit that was responsible of the project execution. The PCR is evaluated against the specific development objectives stated at the moment of Board approval, and the specific objectives are measured against concrete results by the time of the project's completion date. Therefore, the evaluation of the specific objectives vis-à-vis results, mainly in the form of outcome indicators, is the basis for the project's performance score.

The PCR evaluates the project performance according to four central criteria: Relevance, Effectiveness, Efficiency, and Sustainability. Once a PCR is approved by the corresponding Country Manager, it is sent to the Office of Evaluation and Oversight (OVE) for external validation.[4] There are 6 overall validation ratings that OVE can assign to a PCR based on the

[2]A PCR is required for Investment Loans, Policy-Based Loans (PBLs), stand-alone Reimbursable Technical Cooperation (RTC), and stand-alone Investment Grants (IGR).

[3]Investment Loans are identified by an acronym composed by the two letters of the country name, followed by the L letter and then 4 digits. For instance, a project in Argentina is identified as AR-L1045, in Bolivia as BO-L1039, and so forth for the rest of the countries.

[4]The PCRs also addresses two non-central criteria: the performance of the Bank and the performance of the

assessment of the 4 core criteria: Highly Successful, Successful, Partly Successful, Partly Unsuccessful, Unsuccessful, and Highly Unsuccessful (this is an ordinal scale from the highest to the lowest score). In the following sections we will classify a project to be Successful if its overall rating is Partly Successful or above, and Unsuccessful if its overall rating is Partly Unsuccessful or below.

2.2 Data

The sample of PCRs that we use in this study comprises exclusively the documents that were prepared or validated under the 2018 PCR guidelines. This helps minimize heterogeneity that could arise from different structure and scope in the documents across years. Ninety PCRs were completed under these guidelines in the study year, however, we only make use of PCRs that were written in Spanish as it would not be very impractical to mix different languages when applying text mining tools. Hence, we are left with a sample of 44 PCRs that fall under the category of Investment Loan: 25 reached CO in 2018 and 19 reached CO in 2017. Figure 1 shows the distribution of PCRs across Sectors and CO year. As we can see, whereas in 2017 CSD has the highest proportion of PCRs, that is 32%, in 2018 IFD and SCL have the highest share, 28% respectively.[5]

2.3 Methodology

Text analysis employs a variety of tools to analyze and describe the data. The first thing to do when working with text data is to preprocess the raw data. It is extremely important that text preprocessing is clear and transparent in order to obtain meaningful findings and to build reliable statistical models. Hence, we follow standard preprocessing steps for text data including normalization, noise removal, and tokenization.[6]

In a first step, we transform the pdf documents (PCRs) into text files in order to facilitate their manipulation in a statistical software. After that we use regular expressions, that is, string-search algorithms, to extract the text of the 4 central criteria from the PCRs. Once we do so, we transform our texts into a *Corpus*, which is simply a collection of text documents containing

Borrower, however, they are not taken into account for the OVE validation.

[5]Figure 1 from Appendix B shows the number of Investment Loans PCRs by Division and CO year.

[6]The text analysis presented in this paper is performed in the statistical software R. We make use of standard libraries associated with text mining including *tm*, *tidytext*, *janeaustenr*, *tidymodels*, and *topicmodels*.

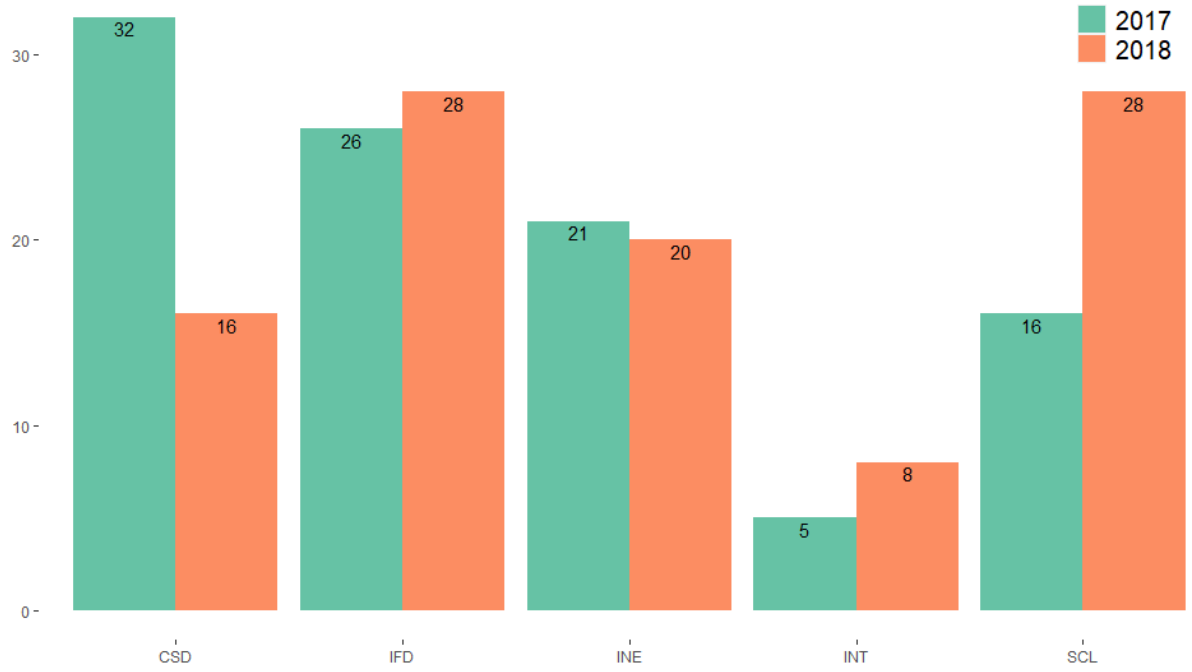


Figure 1: Percentage of PCRs by Sector and CO year

natural language. Once we have our corpus, we convert all the text to lower case, remove numbers, punctuation, special characters, and strip white spaces. After this, we remove common stopping words including non-alphabetical words, prepositions, conjunctions, as well as one to three-character words and irrelevant acronyms. Finally, we convert our remaining words to their linguistic roots through stemming.[7]

After we complete these steps, we construct a document-term matrix (DTM). A DTM is a way of representing the words in the text as a table of numbers. In this case, the rows from the DTM represent one document and its columns represent one term. Each value in the DTM contains the number of counts of a term in each document. Our DTM for the text covering the 4 core criteria is a matrix of dimension 44 x 7,456 (number of documents x number of unique words in the corpus vocabulary). The total number of words in the corpus, which includes repeated stems, is 36,815.[8] In general, DTM objects are not suitable for analysis and cannot be easily manipulated using text mining tools. Therefore, to further analyze our data, we turn

[7]Stemming is the process of eliminating affixes from words to obtain their original stem or root. For instance, the root of words such ”exportar”, ”exportaciones”, ”exportador”, ”exportación”, and ”exportadoras” will be ”export”. To get a sense of how stemming works, Table A from Appendix C shows some of the words corresponding to particular stems.

[8]Table B from Appendix C presents a short version of the DTM of our corpus.

our DTM into a data frame through tokenization in order to have a one-token-per-document-per-row.[9]

3 Illustration: Exploratory Data Analysis

3.1 Analysis of full text in 4 central criteria

In this section, we present the results of the exploratory analysis on our sample of PCRs. Before preprocessing the raw text we present a summary of the total number of words appearing in the PCRs. Then we followed the processing procedure presented in Section 2.3 to calculate word frequencies and other standard metrics for text analysis.

3.1.1 Word totals

We begin by considering the full text of the PCRs, that is, all the words included in the 4 central criteria. First, we analyze the number of words in the PCRs and whether there are statistically significant differences in terms of the CO year and the overrall project rating given by OVE.

Table 1: Word differences in PCRs by year of CO

Closure year	n	Avg. words	sd
2017	19	10,239	3,250
2018	25	10,180	2,059

As we can see in Table 1, 2017 PCRs are on average slightly longer than 2018 PCRs, however, a test for mean group difference shows that the difference is not statistically significant (*p-value* = .941).[10]

Table 2: Word differences in PCRs by OVE rating

PCR rating	n	Avg. words	sd
Unsuccessful	17	10,655	3,009
Successful	27	9,923	2,332

[9] Tokenization is a text segmentation process in which sentences composed of strings are split into smaller pieces or tokens, these could be unigrams (single words), bigrams, trigrams, etc. We tokenize using the tidy text format from the tm package in R

[10] Figure 2 from Appendix A shows the number of words per project and CO year.

As we turn our attention to word differences with regards to project score given by OVE, we see that Unsuccessful projects have a higher average number of words than Successful projects, however, the difference is not statistically significant different from zero (*p-value* = .37).

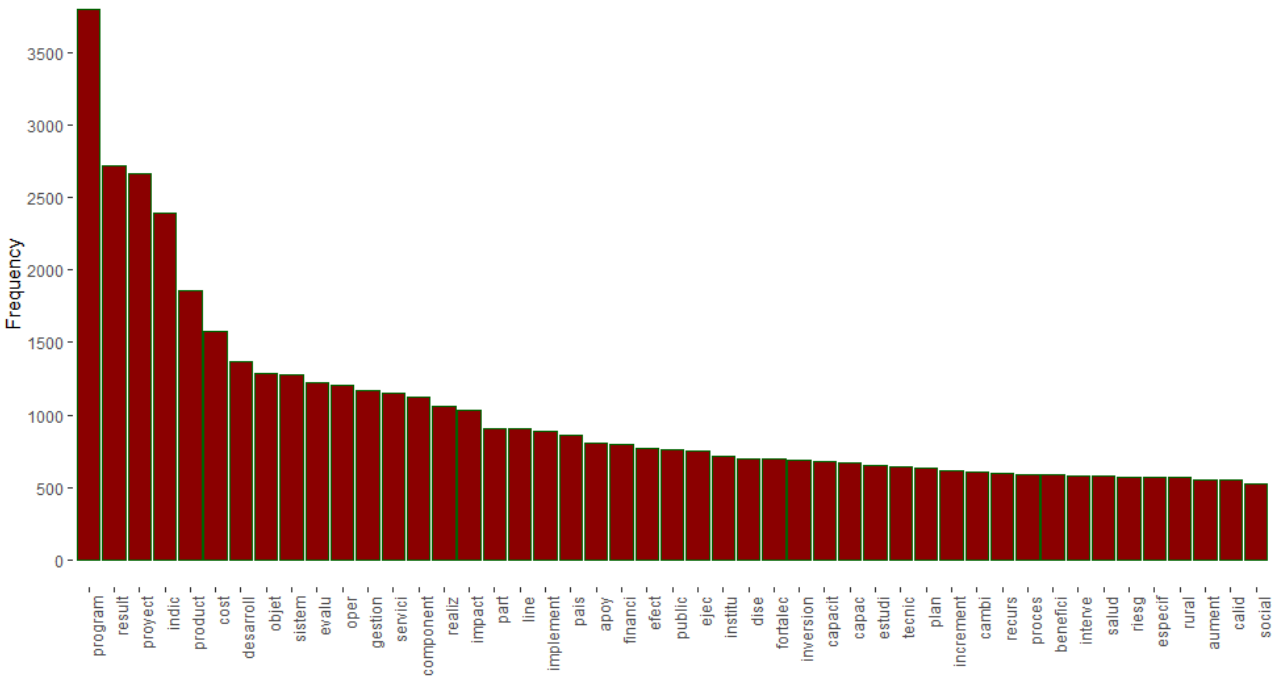


Figure 2: Frequency distribution of unique terms in the corpus. For visualization purposes, only words with a frequency equal or higher than 500 are shown.

3.1.2 Word frequency and Term Frequency-Inverse Document Frequency

A central question in text analysis regards how to understand what the documents are about. A first measure to summarize a set of documents is by counting how often a word occurs or appears in them. It is important to mention that our original DTM is 88% sparse, which means that 88% of the entries (rows) in the DTM are zero. Thus, in Figure 2 we start by showing the frequency distribution of unique terms in our sample of PCRs. As we can see the distribution is right-skewed, meaning that there are few words that have extremely high frequencies.[11] In general, it is easier to visualize the most frequent words with a word cloud using the tidy text data. Hence, Figure 3 shows the word cloud of terms for the 4 central criteria

[11] Figure 1 from Appendix D shows the word frequency distribution of the terms in the corpus with a frequency higher than 100.

in the sample of PCRs. The most popular stems are ”program”, ”result”, ”proyect”, ”indic”, and ”product”. Unsurprisingly, these are typical words in a PCR and in principle do not convey a lot of meaningful information.

Figure 3: Word cloud of terms in 4 the central criteria of the PCRs. The size of the words represents their frequency. We present the 100 most popular words.

In spite of the fact that we apply preprocessing techniques to reduce the dimension of the data and we eliminate sparse terms, we may need to further reduce dimensionality. There are two types of redundant terms: extremely common words in the corpus (which in fact can be treated as stopping words) as well as high term frequencies of certain terms due to document length. To address these issues, we quantify the words in the documents using the *term frequency-inverse document frequency* (tf-idf). The tf-idf is a transformation of the text data which measures how relevant a word is to a document in a collection of documents (Silge and Robinson, 2017). More formally, for any given term the tf-idf is made of two algorithms multiplied together:

$$tf \times idf = \frac{F_{wd}}{N_v} \times log\left(\frac{D}{D_w}\right) \tag{1}$$

Where F_{wd} is the number of times term w appears in a document, N_d is the total number of words in any given document, D is the number of documents in the corpus, and D_w is the number of documents in which term w appears. Whereas the tf algorithm measures how frequent a word is relative to the length of a document, the idf measures how unique a word is in the corpus. Note that because of the tf algorithm, longer documents with more words will no longer overshadow smaller documents with fewer words that may contain valuable information. And because of the idf algorithm too frequent words across documents with less value will be punished. As a matter of fact, words appearing in every document, such as the those displayed in bigger size in Figure 3, will have an idf and a tf-idf of zero.[12] Furthermore, extremely common words appearing in *most* of the PCRs will have a low idf, and provided that they appear many times across documents, they will have a very low (near to zero) tf-idf.

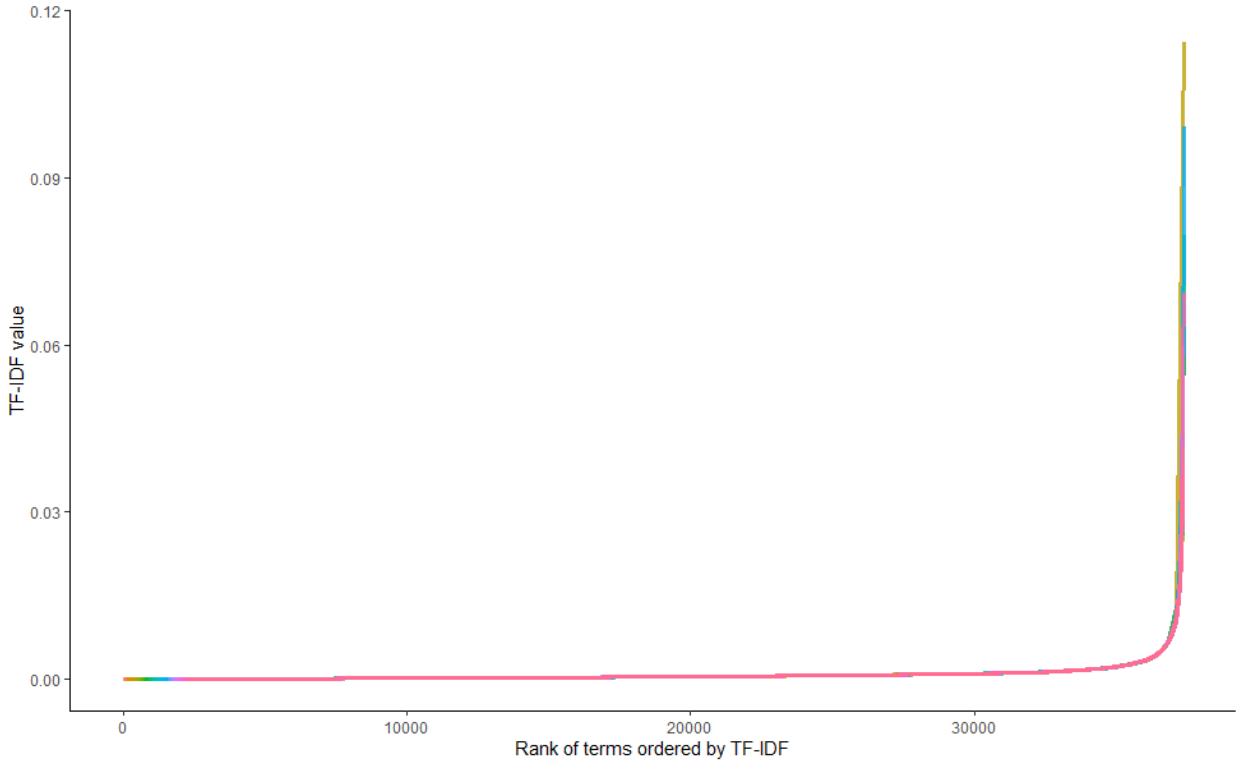


Figure 4: Rank of terms using Term Frequency-Inverse Document Frequency.

[12] If a term appears in every document, then the idf will be natural logarithm of 1 which is zero.

Figure 4 plots the tf-idf values for each word in the corpus. The curve has a long left tail indicating that most of the terms have a zero or close to zero tf-idf value. More specifically, from the 36,815 terms in the corpus, 2,464 have a tf-idf of zero and 26,859 have a tf-idf lower or equal than .001. From this Figure it is evident that we still face a high-dimensionality problem. We proceed to make further reductions in the size of the data but before doing so we use tf-idf weights to analyze which words are characteristic of different types of PCRs. As we mentioned before, a PCR can be Successful or Unsuccessful according to the external validation from OVE. Thus, in Figure 5 we present the top 10 terms with the highest tf-idf value separating by Successful and Unsuccessful PCRs.

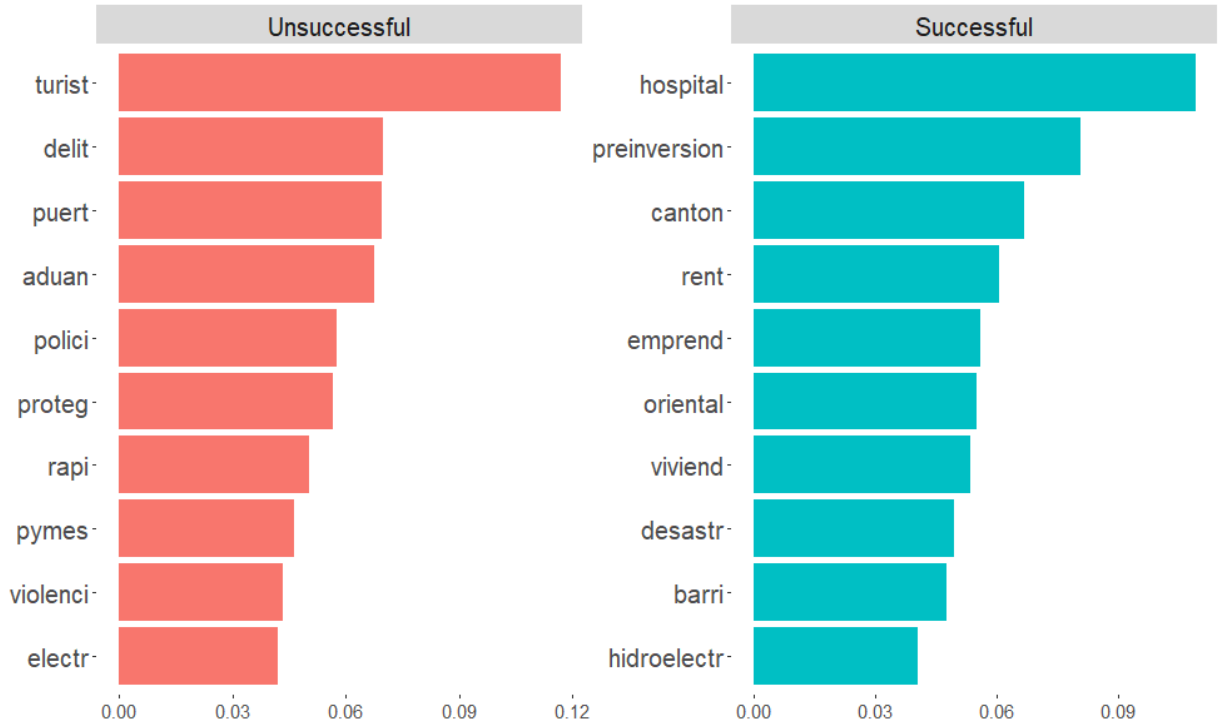


Figure 5: Top 10 terms with the highest tf-idf. The left-hand side plot shows the terms in Unsuccessful projects, while the right-hand side the terms in Successful projects.

The first thing to note from Figure 5 is that most of the stems are characteristic of PCRs from specific sectors or divisions. For example, within Unsuccessful projects we find terms such as ”turist” from RND or ”delit” from IFD, whereas in Successful projects we find terms such as ”hospital” from SCL or ”hidroelectr” from ENE. In principle, this could suggest that projects that contain these specific-sector words should be link to Successful or Unsuccessful projects. This, however, is misleading as some of these terms appear only in very few PCRs. For example,

the term ”turist” appears in 8 documents a total of 314 times, although it appears 299 times in a single Unsuccessful PCR (a tourism project in Bolivia). Something similar happens with the term ”canton”, which appears 156 times in 3 Successful PCRs, albeit 152 times in just one of them (an agricultural project supporting cantons in Ecuador). With the if-idf formula these terms will automatically have a higher weight in spite of the fact that they are driven by a single document. This individual-document bias is also characteristic of a text corpus, and we need to correct it in order to better characterize our sample of Successful and Unsuccessful PCRs.

3.1.3 Vocabulary adjustment

In the previous section, we applied the tf-idf formula to our corpus and ended up with a corpus mostly conformed of terms with a very low tf-idf weight. Furthermore, we saw that because of the uniqueness of our corpus, we also have terms with a high tf-idf in spite of the fact that they might not be very informative. More precisely, we have terms that seem relevant at first glance, as they are present in very few documents, but whose frequency in the corpus is driven by specific documents. Therefore, in this section we apply different rules to remove redundant terms from the corpus, thus, reducing the size of the vocabulary.[13]

First, we would like to remove generic terms from the corpus (i.e. words that are typical of a loan proposal and thus of a PCR). One way to accomplish this is by dropping the terms with a tf-idf of zero. After doing so, we eliminate 56 unique stems and a total of 2,464 terms from the corpus. Figure 6 shows an updated version of the word frequency distribution of the corpus after removing the terms with a tf-idf of zero.[14] Clearly, the distribution looks flatter and, although we still have terms that have a high frequency, now the highest frequency of a term in the corpus is of 885 (this correspond to the term ”implement”). A detailed inspection of this Figure (showed in Figure 1 from Appendix E) shows that there are generic terms such as ”public”, ”fortalec”, ”inversion”, ”capacit”, and ”estudi” as well as sector words such as ”salud”, and ”rural”, that are still present in the corpus. In general, we would like to remove the remaining

[13] Removing terms from the word frequency distribution is a common practice in text mining. Not only does it avoid having too many dimensions in the data but it also prevents overfitting statistical text models. This however is a little but subjective for there are not specific rules on how to drop meaningful terms. For instance Hansen et al. (2018) established a threshold to drop terms that fall below a certain ranking. Furthermore, based on their own assessment, they remove terms that appear in two or fewer FOMC statements.

[14] There are 4,031 terms that appear in one document but for visualization purposes the corresponding bar is not shown in the Figure.

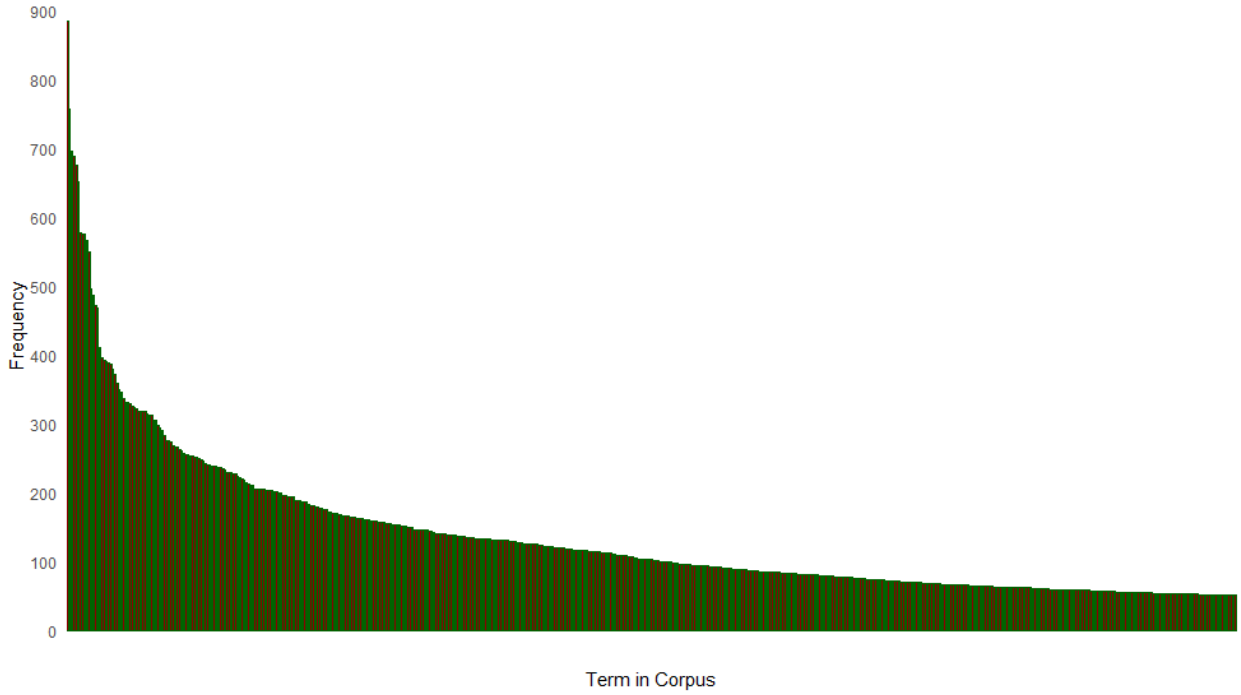


Figure 6: Frequency distribution of unique terms after removing terms with a tf-idf of zero.

noise from the word frequency distribution while getting rid of the noise caused by terms with a high tf-idf weight (but whose frequency is driven by one or two documents). To get a better sense of which terms might be the ones that we need to remove from the corpus at this stage, Figure 7 plots the number of times that each term appear in each document from the corpus. For example, there are 946 terms that appear in two documents, 452 that appear in three documents, 13 terms that appear in 42 documents, 15 terms that appear in 43 documents, etc.

A detailed examination of the terms from the right tail of Figure 7 reveals that the words that are present in 37 or more documents (85% of all documents) are still likely meaningless terms that do not add too much information to the corpus (e.g. ”particip”, ”cumplimient”, ”elegibilid”, and ”public”).[15] Thus, after a thorough inspection we decide to drop the terms that appear in 4 documents or less, in 37 documents or more as well as the terms with a single-document frequency equal or higher than 125.[16]

[15]Some other terms are ”implement”, ”calid”, ”gobiern”, ”ambiental”, ”relev”, ”vertical”, ”aline”, ”presupuest”, ”crecimient”, ”propuest”, ”document”, ”inversion”, and ”riesg”.

[16]Only 15 unique terms has a single-document frequency equal or higher than 125.

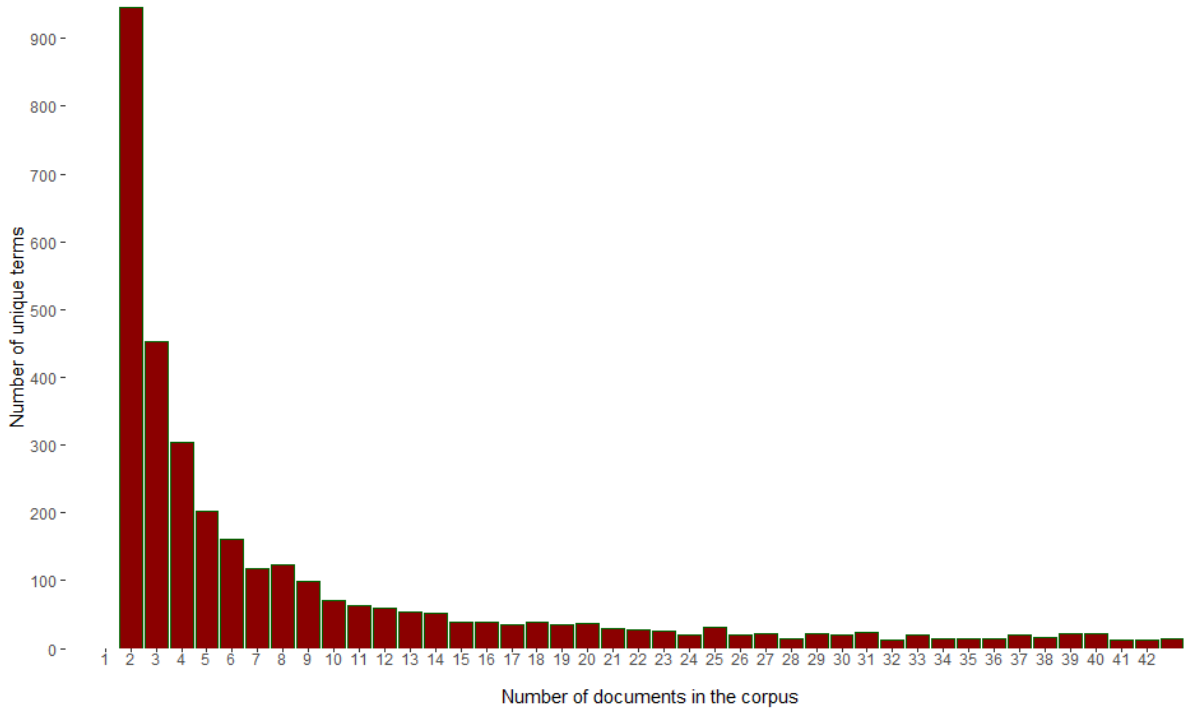


Figure 7: Repetition of unique terms across documents in the corpus

Table 3 synthesizes the evolution of our text data in each proprocessing step starting from the raw data.

Table 3: Data Reduction in Each Preprocessing Step

	(1) Raw text	(2) Cleaning corpus	(3) TF-IDF adjustment	(4) Trim by Document
Total words	77,558	36,815	34,351	20,807
Unique words	17,255	7,456	7,400	1,530

We start with a high-dimensional raw data composed of 77,558 words corresponding to 17,255 unique stems or words. We went through several steps to clean the data and end up with 1,530 unique stems in the corpus. Moreover, due to our preprocessing work we manage to reduce the size of the vocabulary to a little less than 10% of its original size. One important thing to highlight is that the most substantial reduction on the size of the data was in the transition from step (3) to step (4), that is, when we remove the words appearing in fewer than 4 documents, in more than 37 documents, and with a frequency in a single document equal or

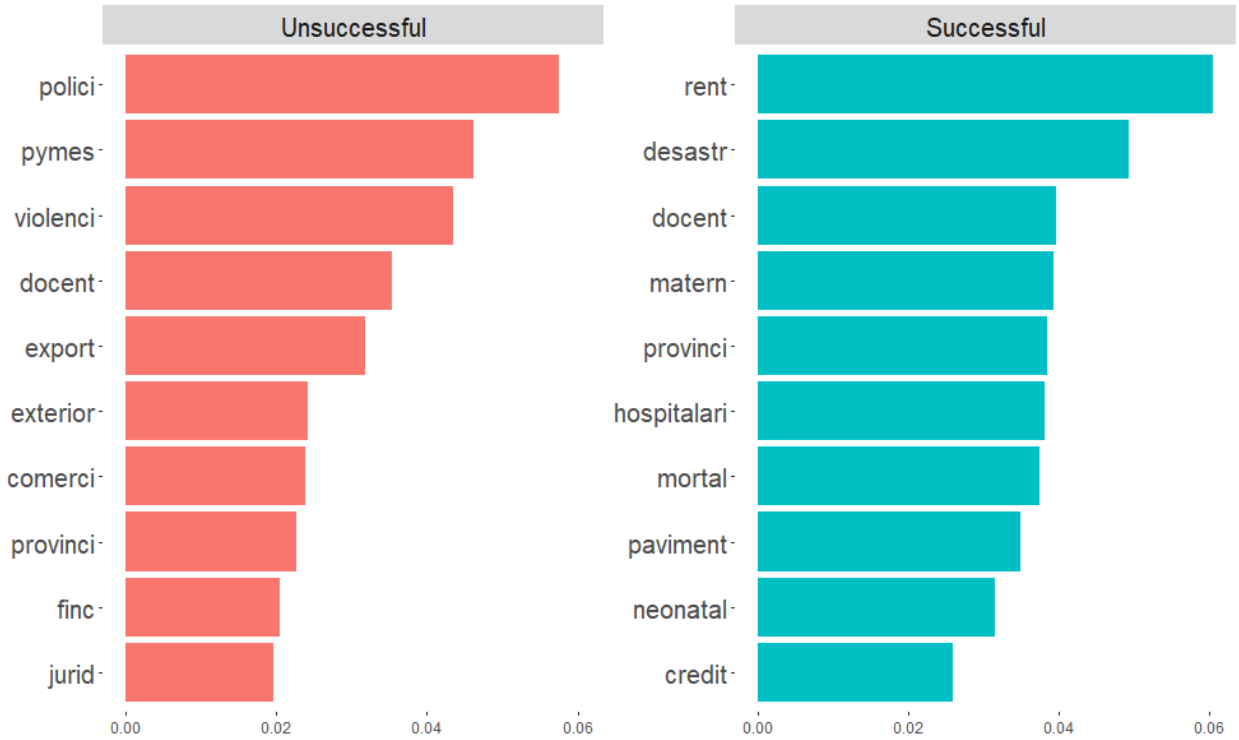


Figure 8: Top 10 terms with the highest tf-idf removing words appearing in fewer than 4 documents, more than 40 documents as well as those with a single-document appearance equal or higher than 125. The left-hand side plot shows the terms in Unsuccessful projects, while the right-hand side the terms in Successful projects.

higher than 125. What this signifies is that a great bulk of the words were concentrated in either few documents or a great number of documents. It is clear that these words were producing noise in the data and it was necessary to remove them from the corpus.

Once our corpus has been completely clean, in Figure 8 we plot again the top 10 words with the highest tf-idf score. We see notable difference with respect to Figure 5. Note that now the words with the highest tf-idf weight in Unsuccessful projects are ”polici”, ”pymes”, ”violenci”, and ”docent”, among others, whereas words such as ”turist” and ”delit” no longer appear. On the other hand, the most relevant words in Successful projects are now ”rent”, ”desastr”, and ”docent”, whereas ”hospital”, ”preinversion” and ”canton” no longer appear. We believe that this division of unigrams is very informative and illustrates a way to analyze the most relevant words in different types of PCRs.

4 Topic Modeling

In text analysis, we often want to characterize each document based on the "topics" that stand out across documents. To illustrate this idea, we use a popular machine learning tool for topic modeling known as Latent Dirichlet Allocation (LDA). LDA is a probabilistic model which reduces the dimension of the data by grouping words into topics on the basis of repeated co-occurrence across documents. Unlike supervised machine learning models, LDA is a unsupervised learning algorithm in which we do not have a defined task to perform (for example, predicting an outcome), instead, the model divides the documents based on the proportion of space covering to a variety of hidden "topics". We can think of LDA as a sort of clustering model in which we seek to find clusters of words and assign or classify them into topics.[17]

LDA was initially developed by Blei et al. (2003) to study discrete text data. LDA has been widely used in many research areas. For instance, it has been used in political science to gain insights from legislative speeches (Quinn et al., 2010) or in economics to examine the effect of transparency on the deliberation of the FOMC transcripts (Hansen et al., 2018). Here, we use LDA to explore and understand the topics that appear across PCRs. We begin by providing a brief description and a general intuition of the model based on Hansen et al. (2018) and Ke et al. (2019). After that we present the results of the model applied to our sample of PCRs.

4.1 LDA model

In LDA we begin with a corpus C which is composed of a collection of D documents indexed by an integer $d \in \{1, ..., D\}$. A document is a sequence of N_d words denoted by $\mathbf{w} \in \{w_{1,1}, ..., w_{d,n}\}$ which belong to a vocabulary V of unique unique terms or tokens. The algorithm works by making the assumption that there are K latent topics, each of which is a distribution $\beta_k \in \triangle^{V-1}$ over the V tokens in the corpus vocabulary. Here, we can imagine a process in which we assign weights to the words, then we clustered together those words that reflect a similar theme, and with that we construct a set of topics. Intuitively, we can think about the corpus of documents and the number of topics as two key inputs of the model.

For each document, LDA also generates a predictive distribution over the K topics, $\theta_d \in$

[17] LDA is method for unsupervised learning that can classify documents, but it should not be confused with machine learning classifiers such as Supporting Vector Machines or Ensemble Methods for those are supervised machine learning models.

$\triangle^{K-1}$. Informally, we can say that each document is characterized by the share of space that it assigns to each topic. For example, suppose that there are two topics in our topic space ($K = 2$): topic 1 which is "Technology" and topic 2 which is "Education". This means that in a two-topic model with only 2 PCRs, one PCR might be composed 60% about topic 1 (Technology) and 40% about topic 2 (Education). The other PCR might be composed 20% about topic 1 and 80% about topic 2. The percentage of each topic represents the probability of a document to belong to a certain topic. The topics B and the topic compositions θ_d determine the mixture model for each word in document d.

Therefore, in LDA documents are characterized as a mixture of latent topics, and each topic has a probability distribution over words. The generative process for a word $w_{d,n}$ is as follows:

1. Choose a topic k: $z_{d_n} \sim Categorical(K, \theta_d)$[18]
2. Choose a term of V from topic z_{d_n}: $w_{d_n} \sim Categorical(V, \beta_{zd,n})$

To put it succinctly, each word in a document is generated by sampling a topic from the topic-distribution associated with the document, and then sampling a term from the term-distribution associated with the topic. The probability that a term t appears in document d is given by $p_{dv} \equiv \sum_{k=1}^{K} \beta_{t,k}\theta_{k,d}$. And the overall likelihood is given by two products: $\prod_{d=1}^{D}\prod_{t=1}^{V} p_{td}^{n_{td}}$, where n_{td} is the number of times term t appears in document d.

We can view LDA as a dimensionality-reduction method which finds the right assignment of a topic to every word in the corpus. Moreover, given a document-word frequency matrix (which in our case will be of size 44 x 1,625) and a number of topics as inputs, LDA yields two outputs: 1) per-topic-per-word probabilities, called matrix β (*beta*), in which each row represents a topic and each column a word: a value in row i and column j will represent how likely topic k contains a word; and 2) per-topic-per-document probabilities, called matrix γ (*gamma*), in which for each topic we will get a probability for a document belonging to that topic.

Because the essence of the LDA model relies in the co-occurrence of words that are put together to create a topic, it is key that the corpus contains a large amount of word co-occurrence information and the topic model has the ability to correctly capture the amount of the word co-occurrence (Chen and Kao, 2015). What this signifies is that topic models will perform very

[18] As Ke et al. (2019) state this is tantamount to having a multinomial distribution where the number of trials is implicitly assumed to be equal to 1.

poorly in short text corpus. Thus, the LDA model will improve its accuracy to discover high quality topics when a greater number of documents is provided. Intuitively, if we were to create 40 topics with our 44 PCRs, there will not be enough words and documents for the LDA model to learn the distribution of words across topics and of topics across documents.

4.2 Topic Coherence and Comparability

One important question in LDA is how to choose the right number of topics. We follow different methods —including a pure empirical, computationally-intensive approach as well as a more intuitive approach which estimates a goodness-of-fit of LDA models for different number of topics—and find out that an ideal number of topics for our data would be 12. Appendix F shows the different approaches that we follow to find the optimal number of topics in the data.

4.2.1 Words in Topics

LDA does not specifically tells us what the topics are about, thus, we need to analyze the most representative words in each of the topics to get a sense of their meaning. As we mentioned before, the first interesting output from LDA is the probability that each word is generated from each topic. In this case the vector $\hat{\beta}_{\mathbf{k}}$ provides an estimate of the relationship between each word in the corpus and each topic. In general, topics have few words with relatively high probability of appearing in each of them and a greater number of words with relatively low probability.[19] Overall, the intuition behind the words contained in the topics is that they might form natural grouping of words, hence, creating natural labels for each topic (although nothing guarantees that this will be true).

Figure 9 presents a heatmap of the terms that most distinguish each of the 12 estimated topics. The dark the color of a term the more likely it will appear in that topic. Probabilities of the words associated with a topic can be vanishingly small, indicating that the interpretation of the topic is less clear. At first glance, some of the words within the topics are related to each other, suggesting that indeed some topics may be characteristic of certain types of PCRs. We analyze the most likely words appearing in each topic in order to produce rough descriptive labels for each of them. Before doing so there are two caveats that are warranted. The first one

[19] One feature of LDA is that there could be co-occurrence of words across topics.

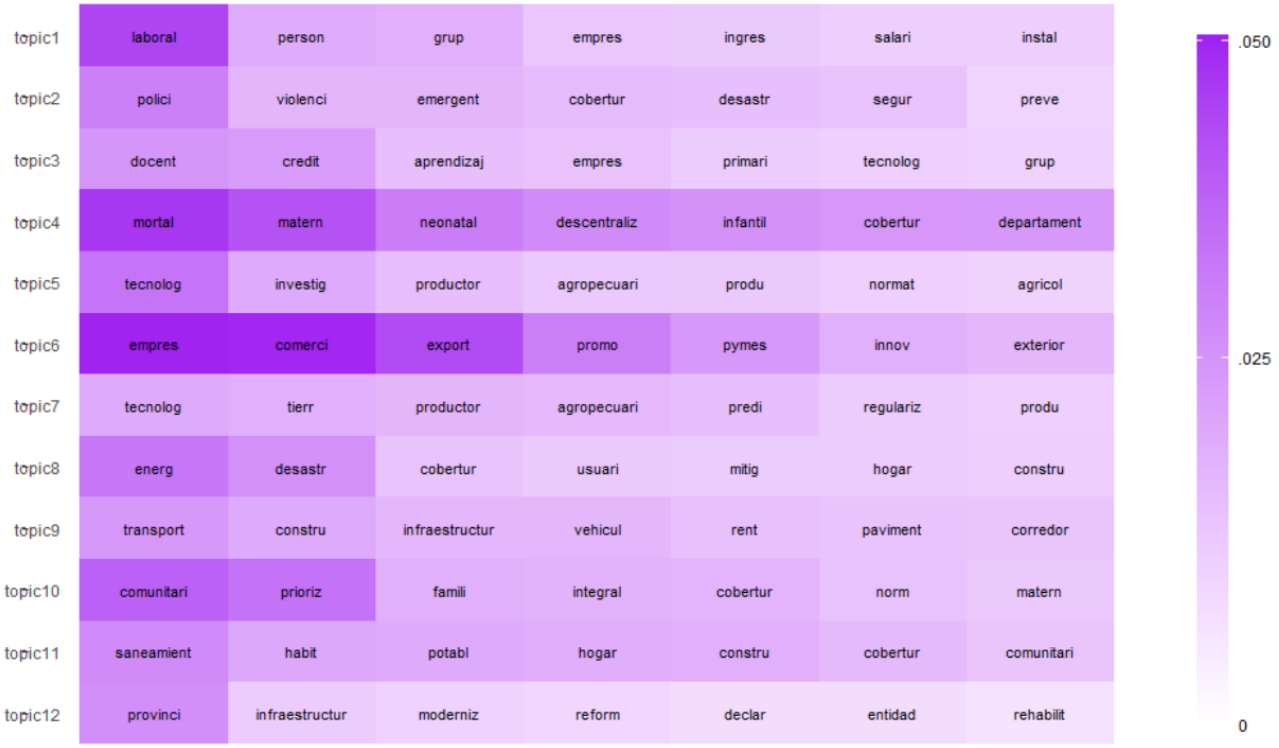


Figure 9: Heatmap with the 7 terms with the highest probability of appearing in each of the 12 topics.

is that the order of the topics does not matter. In other words, topic 1 is no more important than topic 3, and topic 5 is no more relevant than topic 12.[20] The second is that the interpretation of the labels do not follow any formal or statistical analysis. Moreover, we construct the underlying themes or labels for each topic based on the connection between the keywords from Figure 8 as our knowledge working with PCRs. Therefore, we recognize that the labels could not be exhaustive and our interpretations subjective as they could differ from one person to another. However, the popularity of the LDA model relies precisely in its ability to yield easy-to-interpret topics that can be used for further analysis.

In topic number 1 we find words such as ”laboral”, ”person”, ”ingres”, ”salari”, and ”ofert”, suggesting that this topic is related to labor and unemployment. In topic number 2 we find words such as ”polici”, ”violenci”, and ”preve”, suggesting that this topic might be related to citizen security services and the development of strategies to prevent violence. However, the terms ”emergent”, ”cobertur”, and ”desastr” might also indicate that this topic is related to nat-

[20] The labels were ordered according the the default procedure given by R.

ural disasters management. In this case, we label this topic as a combination of two possible underlying themes, ”Citizen Security” and ”Natural Disaster Management”, although we recognize that this might not be very practical, especially when we proceed to classify our PCRs into specific topics. In topic number 3, there is a mix of interesting words. On the one hand, the stems ”credit”, ”empres”, and ”tecnologi” suggest that this topic might be related to small and medium-sized enterprises (SME). On the other hand, the terms ”docent”, ”aprendizaj”, ”primari”, and ”tecnologi” suggest that the topic might also be related to improving learning in education through the use of technology. Hence, we also label this topic as a combination of these two themes. In topic number 4, we have a list of words all of which are connected to addressing neonatal and maternal mortality issues. In topic number 5, we find the word ”tecnolog” (which also appears in topic number 3) along with words such as ”investig”, ”productor”, ”agropecuari, and ”produ”. It seems that this topic is related to innovation in agricultural and livestock practices. In topic number 6 the most relevant words are unequivocally related to entrepreneurship and commerce issues. Moreover, the words ”empres”, ”comerci” and ”export” are marked in dark purple confirming the labelling of the topic. In topic number 7, we find a few words that are shared with topic number 4 including ”tecnolog”, ”productor”, and ”agropecuari”. Unlike that topic, however, we also find word such as ”predi” and ”regulariz”. Taken together, it seems that this topic is related to land tenure regularization issues. In topic number 8, we find terms such as ”energ”, ”desastr”, ”cobertur” and ”usuari”, which at first glance suggest that the topic could referred to issues on increasing electricity coverage. In topic number 9, all of the terms are plainly related to road safety projects. The terms in topic number 10 seem a little bit generic because the two most popular terms are ”comunitari”, and ”prioriz”. When we consider the terms ”famili”, ”integral”, and ”cobertur” the topic seem to become more meaningful. We believe that together the connection of the terms from this topic is related to projects that seek to improve integral community services in targeted areas. In topic number 11, we find terms such as ”saneamient”, ”hogar”, and ”potabl” suggesting that this topic is related to water and sanitation issues. Finally, in topic number 12 we find terms such as ”provinci”, ”infraestructur”, ”moderniz”, ”reform”, and ”rehabilit”. Thus, this topic is dominated by issues on rehabilitating infrastructure or modernizing public services. In table 4 we present the labels that we assign to each of the topics. We only show the keywords that we deem most relevant to label each topic.

Table 4: Topic labels for the 12 topic-model

Topic label	Keywords
Labor and Unemployment	*laboral, person, grup, empres, ingres, salari, instal*
Citizen Security / Natural Disaster Management	*polici, violenci, segur, preve*
Improve Learning in Education / SME	*docent, credit, aprendizaj, empre, primari, tecnolog*
Neonatal and Maternal Mortality	*mortal, matern, neonatal, infantil, cobertur*
Agricultural and Livestock Innovation	*tecnolog, investig, productor, agropecuari, produ, normat, agricol*
Entrepreneurship and Commerce	*empres, comerci, export, promo, pymes, innov*
Land Tenure Regularization	*tecnolog, tierr, productor, agropecuari, predi, regulariz, produ*
Increasing Electricity Coverage	*energ, cobertur, usuari, mitig, hogar, constru*
Road Safety	*transport, constru, infraestructur, vehicul, rent, paviment, corredor*
Community Services	*comunitari, prioriz, famili, integral, cobertur, norm*
Water and Sanitation Services	*sanamient, habit, potabl, hogar, contru, cobertur, comunitari*
Rehabilitation-Modernization of Public Services	*provinci, infraestructur, moderniz, reform, rehabil*

4.2.2 Document Classification

In principle, based on the different connections from the keywords showed in table 4, our LDA model does a good job at describing the topics. However, we would like to know to what extent our interpretation of the topics is correct. One advantage of the LDA model is that it automatizes the process of classifying our documents into topics.[21] Thus, we can also examine per-document-per-topic probabilities to check how well our topics are associated with each PCR. To do so, we analyze the matrix γ (*gamma*). Each of the values from this matrix is an estimated proportion of words from a document that are generated from a topic. For instance, the model estimates that each word in PCR number AR-L1045 has a .0006% probability of coming from topic 3, a .0176% probability of coming from topic 6, and a .975% probability of coming from topic 5. From table 4 we know that topic 3 is labeled as "Improve Learning in Education / SME", topic 5 is labeled as "Neonatal and Maternal Mortality", whereas topic 6 is labeled as "Agricultural and Livestock Innovation". In this case, it seems that document AR-L1075 is weakly associated with topic 3 and 6, but strongly associated with topic 5. An inspection of the nature of this PCR reveals that this project sought to improve the productivity, environmental sustainability and socioeconomic equity of the agricultural and livestock sector in Argentina, putting emphasis in strengthening the agricultural innovation system and increasing

[21] In a very practical sense, another method in which we may classify our PCRs would be to read them. The problem with approach is that it will it take a considerable amount of time and will also require subject-expert matter from the reader. This signifies that a different reader may classify the content of the same PCR differently.

the capacity and generation of new technologies. Hence, our LDA model did a good job at classifying document AR-L1075 with its corresponding topic. We use another example to see how well our LDA model worked at classifying documents into topics. The model estimates that each word in PCR number ES-L1050 has a .0286% probability of coming from topic 2, a .0449% probability of coming from topic 7, and a .926% probability of coming from topic 9. Recall that topic 2 is labeled as "Citizen Security / Natural Disaster Management", topic 7 is labeled as "Land Tenure Regularization", and topic 9 is labeled as "Road Safety". After reading the content of this PCR we find that the project sought to improve the conditions of public transportation between the provinces of Soyapango and Downtown from El Salvador. Again, as with PCR number AR-L1045, the LDA model classified PCR number ES-L1050 into its right topic, "Road Safety".

Figure 10: Distribution of documents across the 12 topic-model

For each document, we take its maximum probability from matrix γ (*gamma*) in order to classify it into one of the possible 12 topics from the LDA model. Figure 9 shows the distribution

of documents across the labeled topics.[22] The numbers inside each segment of the circle indicate the number of PCRs that were assigned to a particular topic. In general, the distribution of documents in the topics is more or less balanced. For instance, there are 3 PCRs assigned to the topic labeled as "Labor and Unemployment", 5 PCRs assigned to the topic of "Citizen Security and Violence", 4 PCRs assigned to the topic of "Improve Learning in Education / SME", and so forth for the rest of the topics. We find that 31 PCRs (70% percent of all documents) are more than 90 percent likely to be from a single topic. Although we lose some information by treating the maximum γ (*gamma*) probability for each document as an indicator of the topic into which a PCR should be classified, our LDA model made a decent job describing the documents assigned to each (substantive) topic.

We decide to use the probability threshold of greater or equal to .9 as a proxy for "right" assignment. Based on this rule, we have that 70% of our documents were classified correctly whereas 30% were misclassified. To put it differently, 13 out of 44 documents seem to have been assigned to the wrong topic. For instance, the model assigns PCR number CO-L1132 to topic 3, "Improve Learning in Education". However, this project was intended to support the government of Colombia to strengthen the competitiveness of the productive sector by financing investment, productive reconversion, and business and export development. Thus, in this manner, this PCR should have been assigned into topic 6, "Entrepreneurship and Commerce" and not into topic 3. Another salient case is PCR number PE-L1087. According to the LDA model this model was assigned to topic 5, "Agricultural and Livestock Innovation". A careful reading of this PCR tells that the project's aim was to increase the efficiency of Peru's public sector by i) modernizing different processes and instruments related to financial execution, ii) strengthening the institutional capacity of the Ministry of Economy and Finance, iii) and improving the decision-making processes on the allocation of public resources. Clearly, this project is not related to agricultural and livestock services, and in principle it could have been assigned to topic 12 "Rehabilitation-Modernization of Public Services". Moreover, it is possible that very few proportion of words from this document were generated from any of the twelve topics.

It is worth noting that by randomly reading some of the documents and verifying their assignments to the topics we were able to discover some nuances in our classification. For instance, our LDA model estimated that PCR number BO-L1039 was misclassified into topic 1,

[22] Table A from Appendix G shows assignment of each PCR into its corresponding topic.

"Labor and Employment", as the words from this topic had only a .53% probability of coming from topic 1. However, after reading this document, we find that this PCR is indeed related to labor and employment issues as its objective was to increase the tourism expenditure and the employment generated within this sector in Bolivia.

As we mentioned before, a topic model would render more coherent topics when we have a higher amount of word co-occurrence in the documents. This is particularly important if we seek to obtain a reliable way to link the topics, and thereby their associated PCRs, to measurements of project success and effectiveness. For instance, we would like to know which topics are associated with more Successful PCRs (or with less Unsuccessful PCRs) according to the external validation made by OVE. In spite that we may have few documents to explore this association, one easy way to do is by calculating the percentage of Successful and Unsuccessful PCRs within each topic.

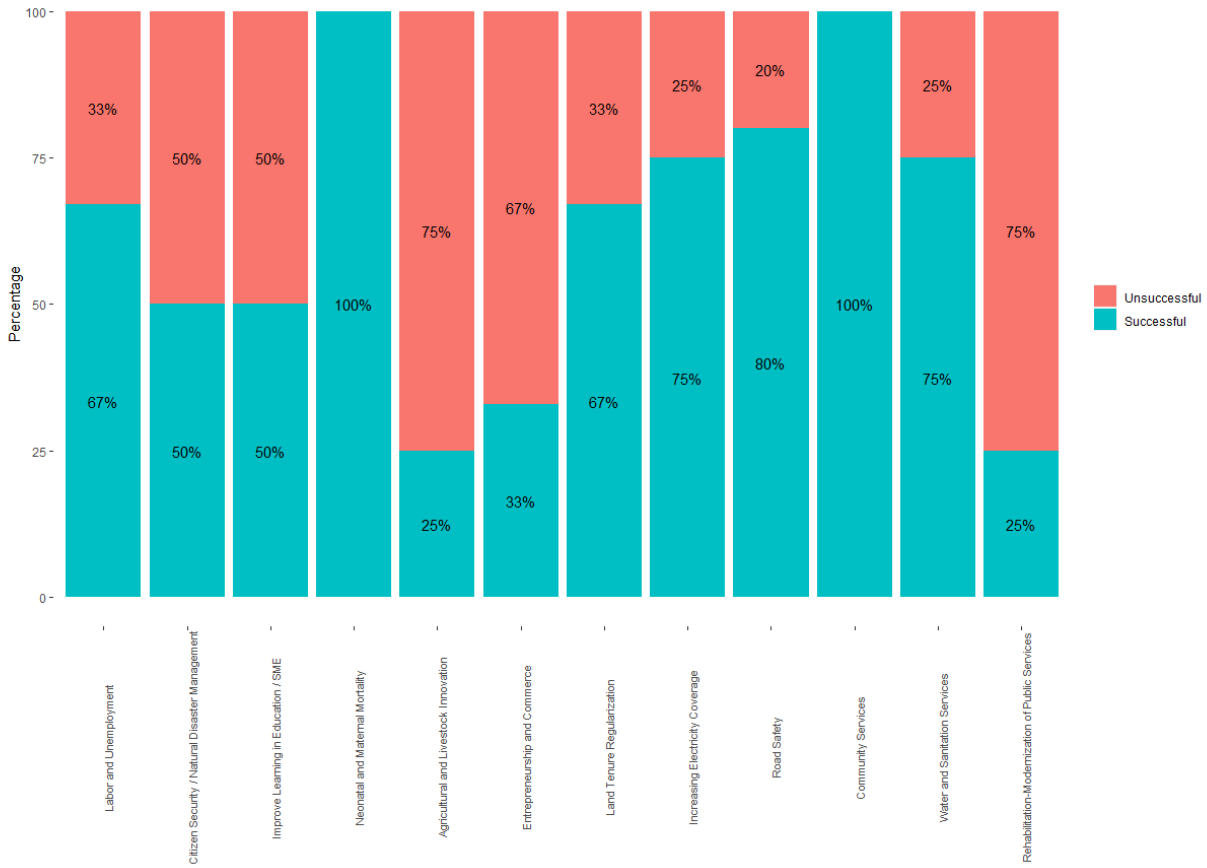


Figure 11: Percentage of projects with a Successful and Unsuccessful overall OVE score by topic.

As we can gleaned from Figure 11, topics 4 ("Neonatal and Maternal Mortality") and 10 ("Community Services") are only associated with Successful PCRs. In Topic 9 ("Road Safety") 80% of the PCRs were Successful which suggests that this topic might be more associated with Successful PCRs than with Unsuccessful ones. On the other hand, topics 5 ("Agricultural and Livestock Innovation") and 12 ("Rehabilitation-Modernization of Public Service") are more associated with Unsuccessful PCRs than with Successful ones. Obviously, this classification might not be too relevant for those topics with few PCRs associated (and actually it would make more sense if we have more documents in each category), but it is shown here just for the sake of illustrating a way to link the topics to measurements of project success and effectiveness. Furthermore, we know that the classification of the PCRs into topics was not completely accurate, thus, the association between topics and project success will also improve as we add more documents into our LDA model.

5 Concluding Remarks

In this paper we present a text analysis of Project Completion Reports from the IDB. We construct a unique, structure-form dataset that allows us to organize the content of a sample of PCRs in a manner that is suitable for quantitative analysis. To summarize and analyze the content of the PCRs, we introduce and make use of typical concepts from text analysis and natural language processing such as corpus, string-search algorithms, stemming, dictionary of terms, a document-term matrix, and tokenization. After that we provide an exploratory data analysis of the PCR documents through different text analysis statistics including word counts, word clouds, and term-frequency inverse-document frequency. Using these tools we were able to identify the most relevant terms in Successful and Unsuccessful PCRs based on the externally validated ratings given by the Office of Evaluation and Oversight (OVE). Despite that we only made use of unigrams, our exercise can easily be extended to analyze more complex terms, including bigrams or trigrams. Moreover, instead on focusing on overall ratings from OVE, the analysis could also be disaggregated to analyze specific sections from a PCR, to compare different departments and divisions or regions and countries. However, the reliability of a more disaggregated analysis is dependent on having a greater number of PCRs available. Our paper also highlighted the value of unsupervised machine learning tools to analyze text data. We il-

lustrate a way to understand the content of the PCRs by finding their underlying ”topics”. Using a probabilistic topic modeling, called Latent Dirichlet Allocation, we are able to reduce the dimensions of our text data to find meaningful word groupings or clusters in the data. Moreover, the LDA model allowed us to classify each one of our PCR into a topic. Insofar we acquire more documents and generate a richer corpus, we will be able to eventually associate the topics with successful and unsuccessful types of projects, as well as with other measurements of project effectiveness. The use of the tools described in this paper will help in providing another avenue for exploring characteristics of successful projects. This type of analysis can be complementary to more traditional analysis such as that presented in Álvarez et al. (2021). We are also positive that methods presented in this investigation have numerous potential applications to other text documents routinely prepared by the IDB.

References

Álvarez, C., Corral, L., Cuesta, A., Martínez, J., Montiel, C., and Yepez, C. (2021). Project completion report analysis: Factors behind project success and effectiveness.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781.

Carneiro, B. and Garbero, A. (2018). Supporting impact with evidence: A content analysis of project completion reports. *The Journal of Development Studies*, 54(8):1426–1449.

Chen, G.-B. and Kao, H.-Y. (2015). Word co-occurrence augmented topic model in short text. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 2, December 2015-Special Issue on Selected Papers from ROCLING XXVII*.

Deveaud, R., SanJuan, E., and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Ke, S., Olea, J. L. M., and Nesbit, J. (2019). A robust machine learning algorithm for text analysis. Technical report, Working paper.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Romer, C. D. and Romer, D. H. (2010). The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. *American Economic Review*, 100(3):763–801.

Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach*. " O'Reilly Media, Inc.".

Appendix A.

To convert the pdf files to text files we cut off the first five pages of every PCR. The reason to do so is because the first pages of every pdf contain text that is not relevant for our analysis. This text refers to official communications between IDB Departments informing the approval of the PCR, the official cover of the PCR with the original project members and the PCR team members, the index of the PCR, the electronic links contain in the document, the acronyms used in the document, and basic information about the project including date of approval by the Board, date of closure, date of eligibility, loan amount, and information about the disbursements. Once we convert all the pdf files into text files, we use regular expressions to trim the content of each PCR. For instance, when we analyze the four core criteria we need to cut off unnecessary sections from the PCR, including the introductory section (which every PCR typically has), the text from the non-central criteria and the section of Findings and Recommendations. To do so, we take advantage of the fact that in general the title of each section within the PCR starts with an upper case letter, or in some cases the whole title of a section appears in upper case letters. For instance, a PCR has the title ”CRITERIOS CENTRALES” or ”Criterios Centrales” to identify the text belonging to the 4 central criteria. Hence, we use string-matching (heuristic based-keyword) to identify the first appearance of the word ”CRITERIOS” or ”Criterios” in the text. Once our heuristic recognizes this word, we delete the text that precedes this appearance. We follow the same procedure to cut off other unnecessary sections. Given that the section for the non-central criteria has the title ”CRITERIOS NO CENTRALES” or ”Criterios no Centrales”, we use another heuristic to find (starting from the end of the text), the word ”CRITERIOS” or ”Criterios”. Again, once the algorithm recognizes this word, we delete the the text that follows this appearance. Finally, to remove the stopping words and stem the words into their root we use the *tm* package in *R*.

Appendix B

Figure 1. Distribution of PCRs by Division and CO year

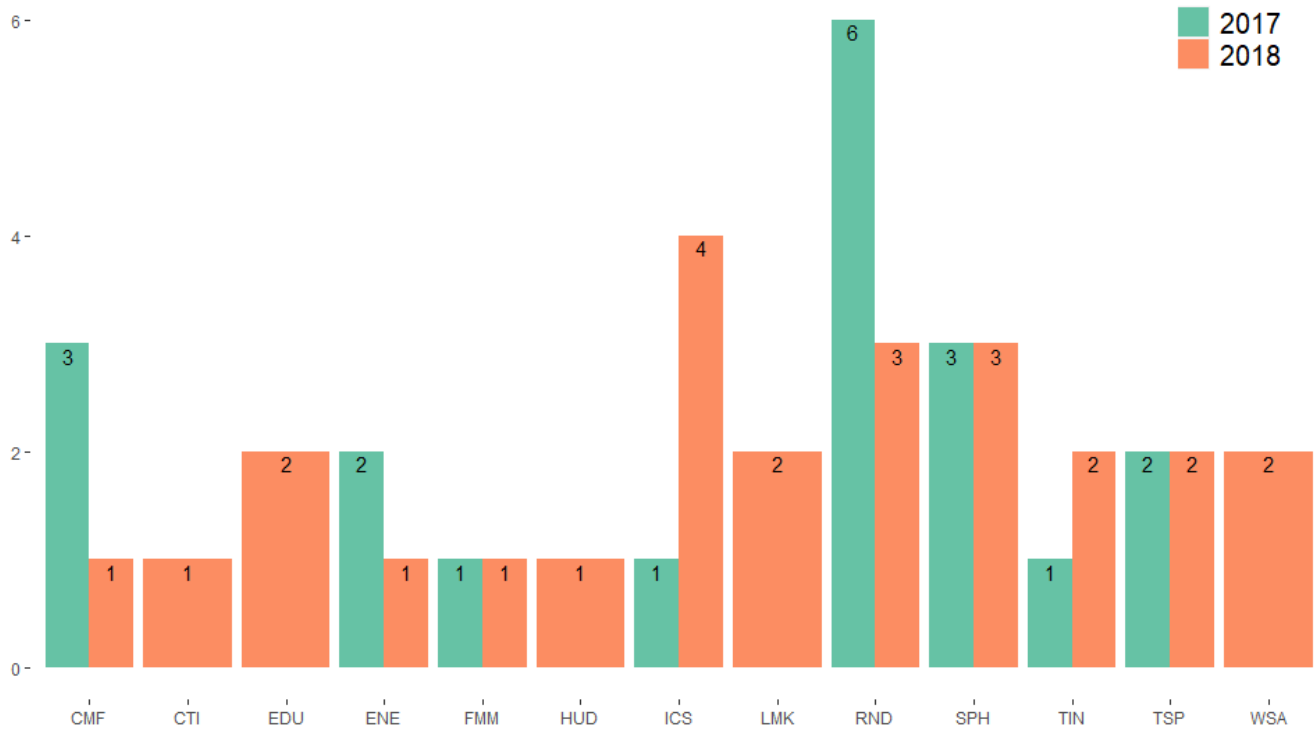
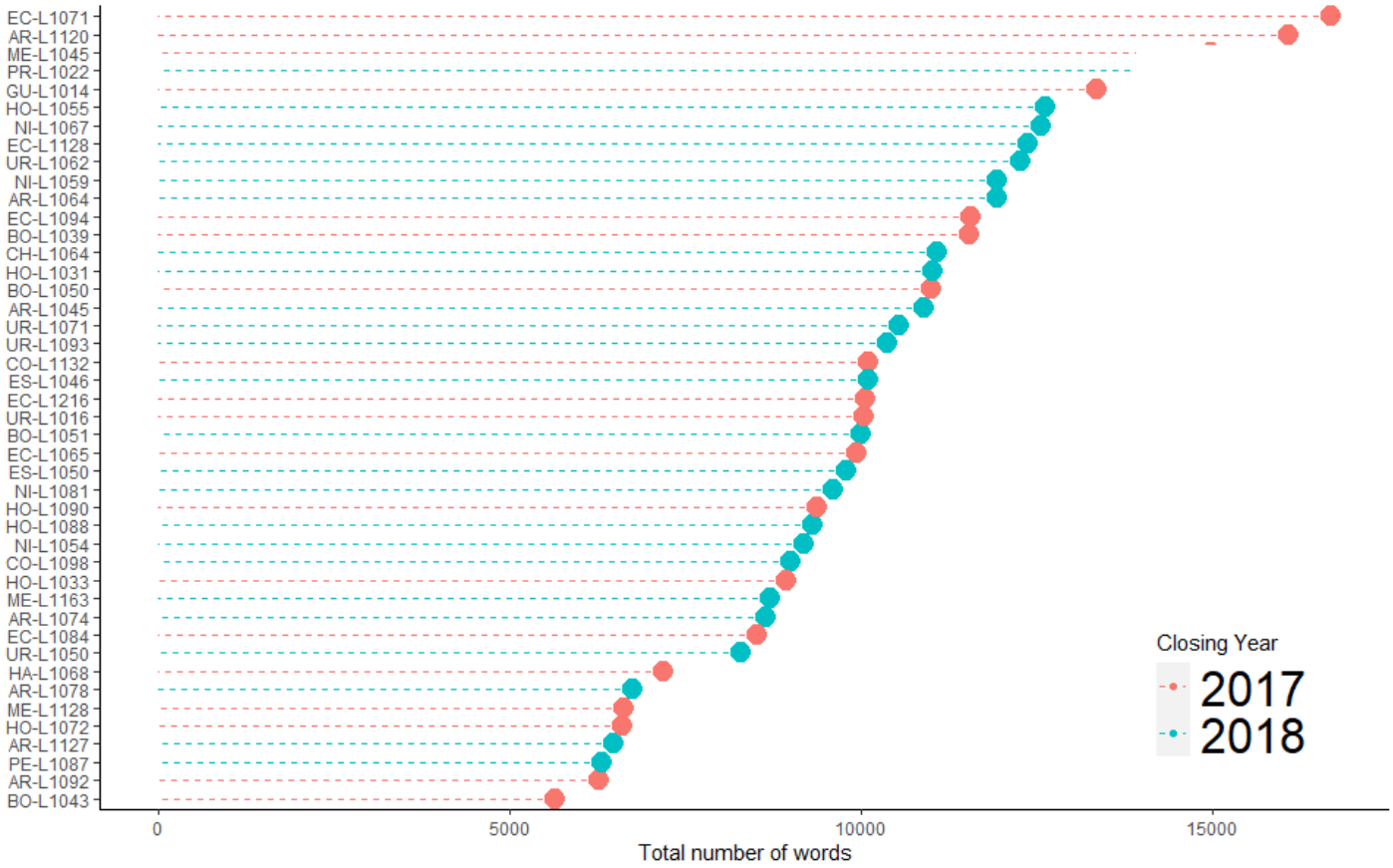


Figure 2. Total words in each PCR by CO



Appendix C

Table A. Example of words associated with a stem

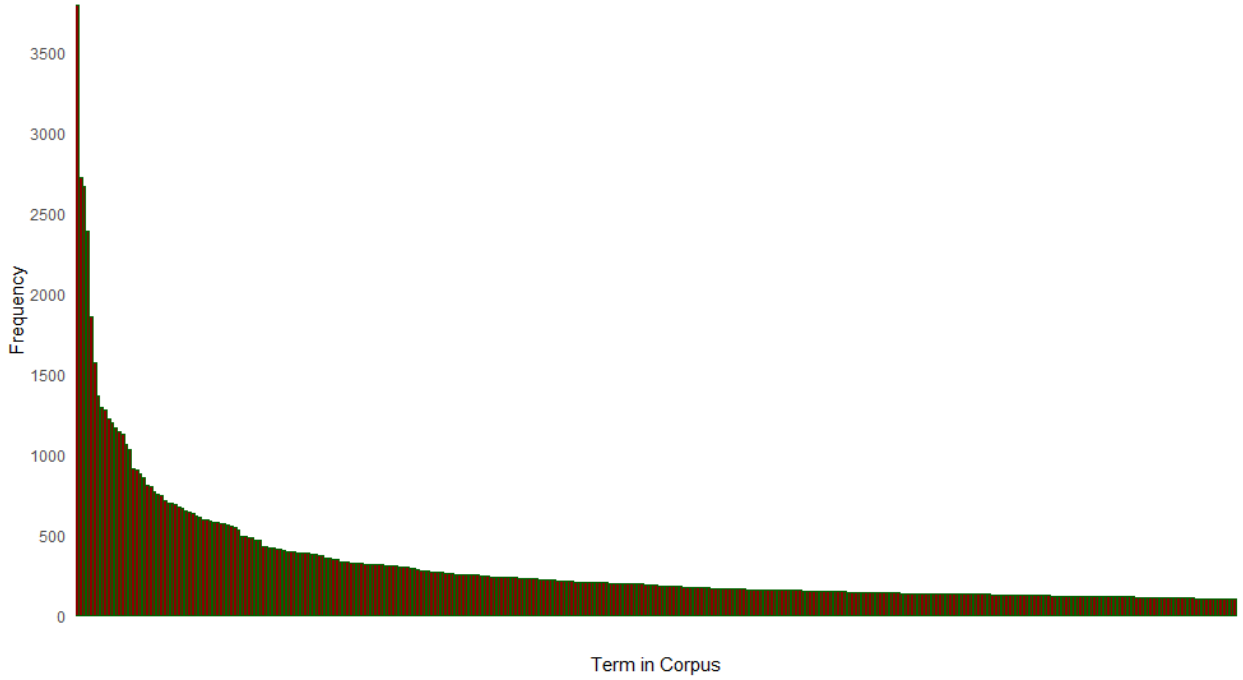
Stem	Associated words
agropecuari	agropecuaria, agropecuario, agropecuarias, agropecuarios
cancel	cancelar, cancelaron, cancelación, cancelados, canceladas
ciudadan	ciudadana, ciudadanos, ciudadano
clas	clase, clases, clasificar
comerci	comercio, comercial, comercializar, comerciales
comunitari	comunitaria, comunitario, comunitarias, comunitarios
constru	construcción, construidas, construir, construidos, constructores
credit	crediticia, crediticias, crediticios, crédito
descentraliz	descentralizada, descentralizado, descentrilizados, descentralización
empres	empresas, empresarios, empresarial, empresariales
energ	energía, energías, energéticos, energética
export	exportar, exportaciones, exportador, exportación, exportadoras
famili	familia, familias, familiar, familiares
habit	habitan, habitantes, habitaciones
innov	innovación, innovadores, innovador, innovadoras
investig	investiga, investigación, investigaciones, investigadores
polici	policia, policial, policiamiento, policing
produ	productividad, producción, productos, productivas, productores
promo	promoción, promover, promueven, promotores
reform	reforma, reformas, reformarze
rehabilit	rehabilitación, rehabilitaron, rehabilitar, rehabilitada
tecnolog	tecnologías, tecnología, tecnológicos
transport	transporte, transportitas, transportadas, transportación
victim	victimas, victimización, victimarios
viaj	viaje, viajes, viajan, viajo

Table B. Exemplary DTM for 4 random documents and 5 terms

PCR	desarroll	indic	mejor	product	result
AR-L1064	52	110	30	83	76
AR-L1120	55	51	84	151	78
EC-L1071	57	57	27	9	88
GU-L1014	30	54	22	173	75

Appendix D

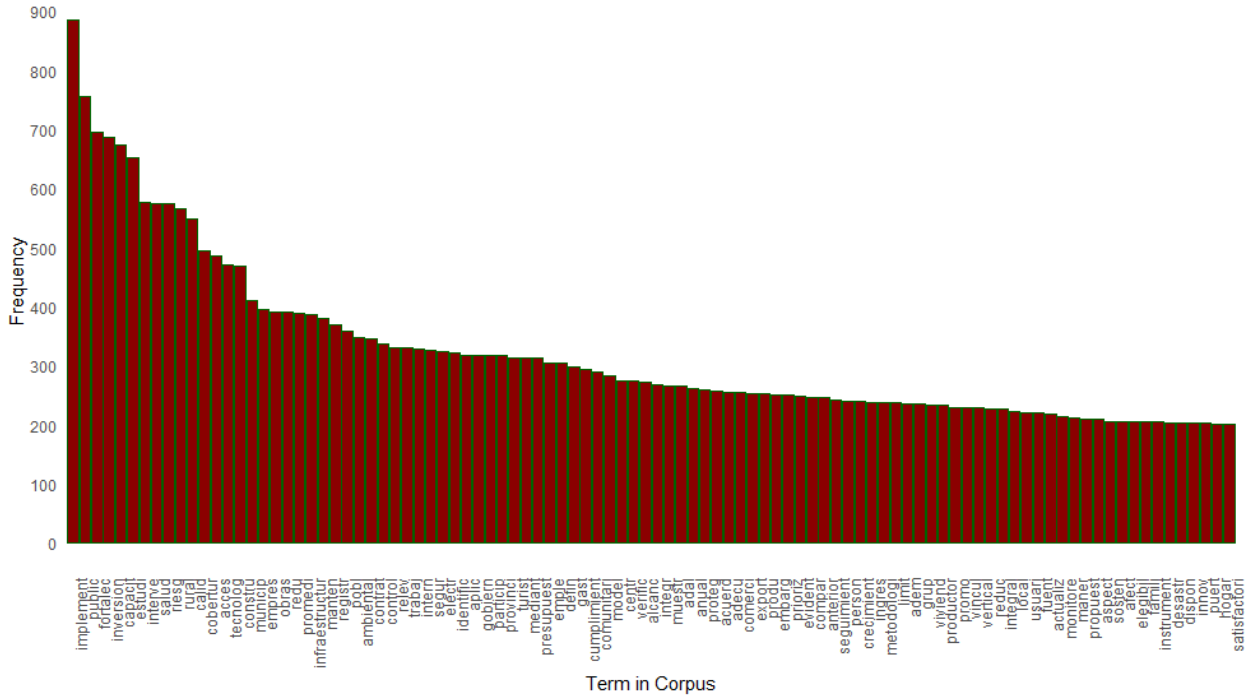
Figure 1. Word Frequency Distribution of the Corpus



We show the word frequency distribution of all the terms in the Corpus. We do not show the label for each term as it would be impractical to visualize. Also, we only show those terms with a frequency higher than 100. As in a typical word frequency distribution, the distribution is skewed to the right, meaning that the distribution has few terms that have extremely high frequencies while most terms have very low frequencies.

Appendix E

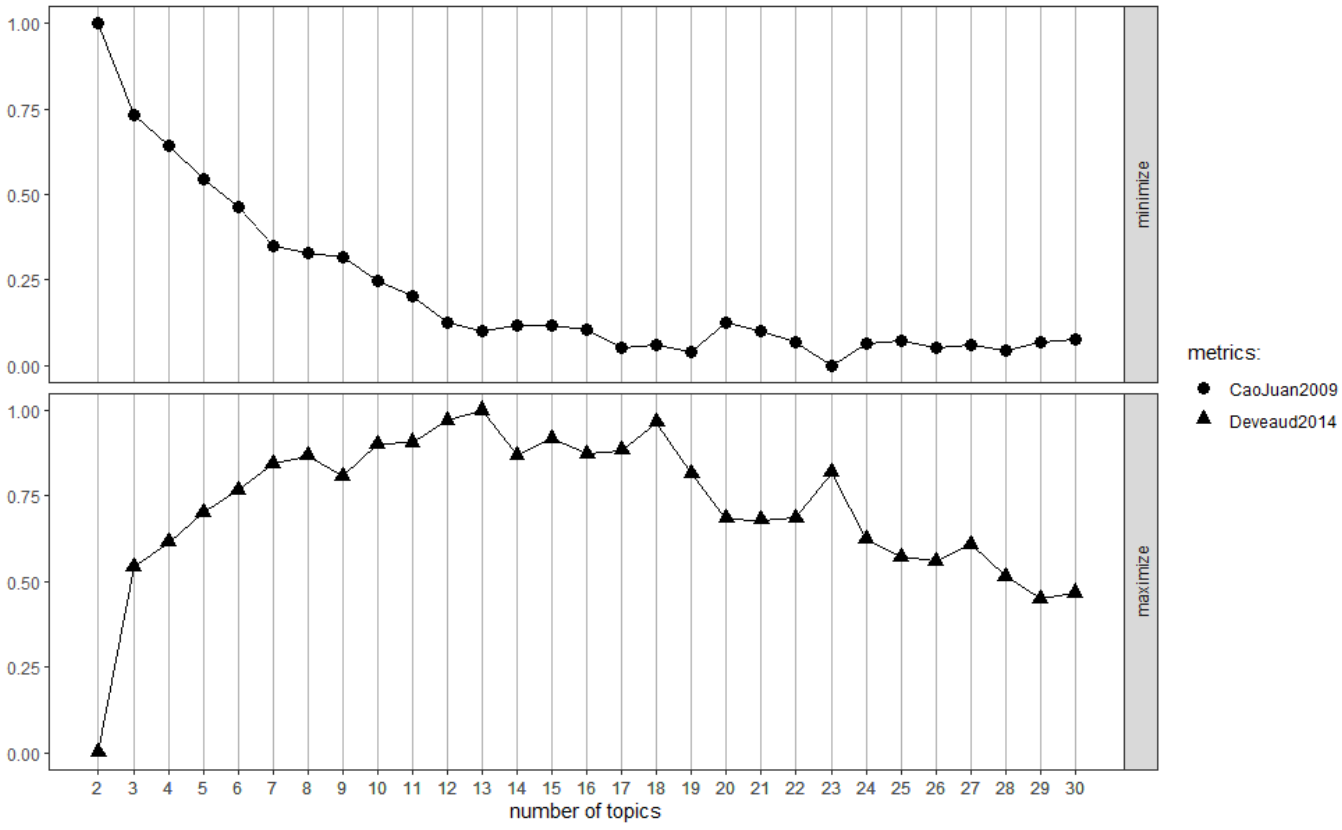
Figure 1. Word Frequency Distribution of the Corpus after removing terms with a tf-idf of zero



Appendix F

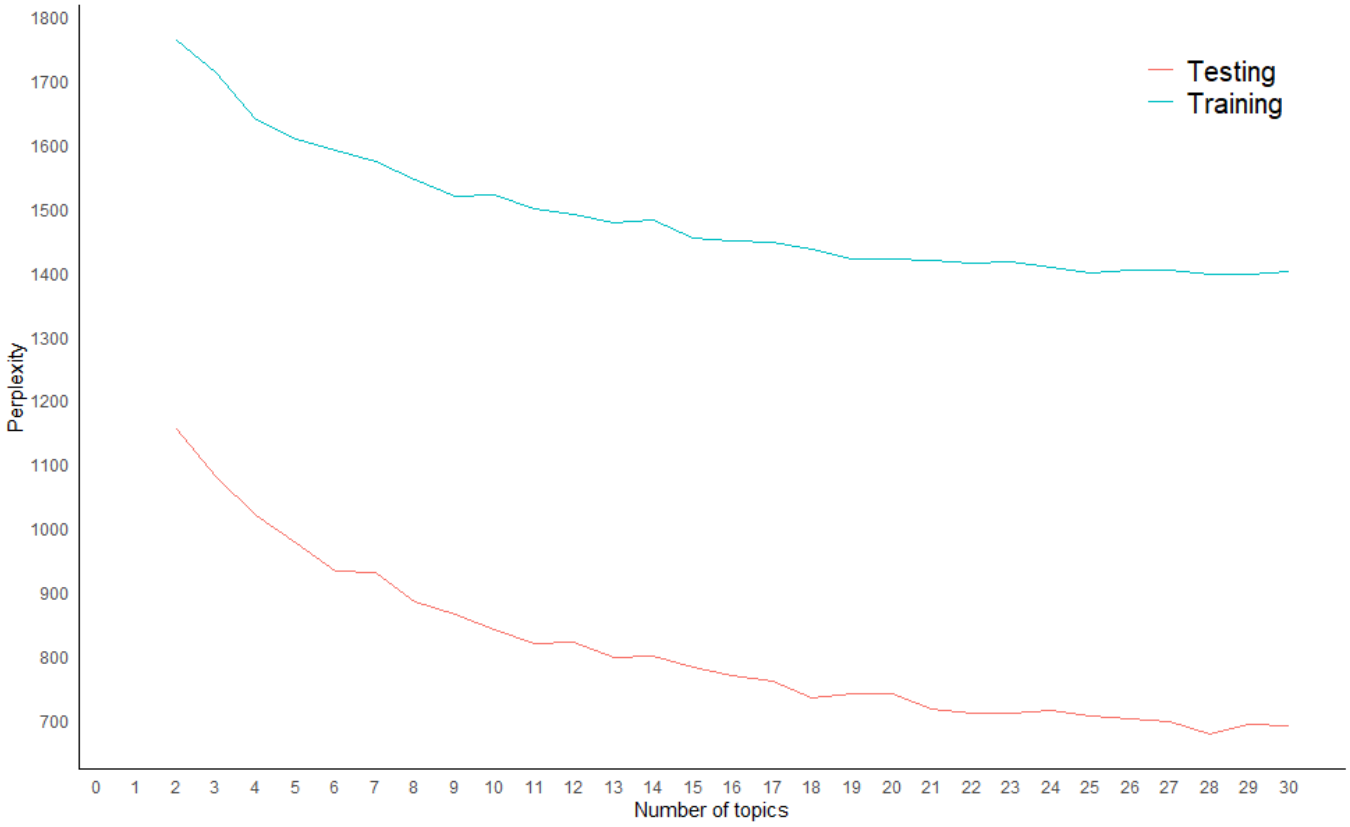
We first use the library *ldatuning* from the R software to choose the optimal number of topics for the LDA model. This method is computationally-intensive and and requires parallel computing. We use two metrics available in this package: *CaoJuan2009* and *Deveaud2014*. Both of these two methods find extrema. The *CaoJuan2009* adaptively select the best LDA model based on the density (Cao et al., 2009), whereas the *Deveaud2014* estimates the log-likelihood of the data for different number of topics using across-topic divergence (Deveaud et al., 2014). Given that this approach is extremely time-consuming we only estimate a limited number of possible topics, between 2 and 30. Based on Figure 1, it seems that across-topic divergence is maximized 13 while density is minimized at 23.

Figure 1. Two metrics to choose the optimal number of topics in LDA



We also follow Hansen et al. (2018) and use an heuristic approach referred as ”perplexity”. Perplexity is a measure of how well a probability model predicts a sample. We calculate the perplexity in a held-out set of documents. To implement this approach we need to split our data into training and testing. We use 70% in the training data and 30% for the testing data. We use 100 iterations and again we estimate the perplexity for a total number of topics between 2 and 30. In general, perplexity should always decrease as the number of topics increases. A general rule to choose the optimal number of topics is to observe the moment when the goodness-of-fit of the model no longer improves. In this case, we see that the perplexity value in the testing data declines very slightly after reaching 13 topics.

Figure 2. Evaluating the optimal number of topics using perplexity



Based on figure 1 and 2, we choose 12 topics for our corpus. We also bear a mind that we only have a limited number of documents for the corpus, 44, thus, we believe that using a greater number of topics might be inefficient in terms of topic coherence.

Appendix C

Table A. Assignment of PCRs to each topic

Topic	PCRs assigned to topic
Labor and Unemployment	BO-L1039, BO-L1051, CH-L1064
Citizen Security / Natural Disaster Management	AR-L1074, AR-L1127, EC-L1216, UR-L1062
Improve Learning in Education / SME	CO-L1132, HA-L1068, UR-L1050, UR-L1093
Neonatal and Maternal Mortality	HO-L1072, HO-L1090, ME-L1128
Agricultural and Livestock Innovation	CO-L1098, ME-L1045, PE-L1087
Entrepreneurship and Commerce	AR-L1078, AR-L1092, UR-L1071
Land Tenure Regularization	EC-L1071, NI-L1067, UR-L1016
Increasing Electricity Coverage	BO-L1043, BO-L1050, EC-L1128, HO-L1031
Road Safety	AR-L1045, EC-L1065, ES-L1050
Community Services	NI-L1054, NI-L1059, NI-L1081
Water and Sanitation Services	ES-L1046, GU-L1014, HO-L1088, PR-L1022
Rehabilitation-Modernization of Public Services	AR-L1120, EC-L1084, EC-L1094, HO-L1055