


Fiscal data in text: Information extraction from audit reports using Natural Language Processing

Alejandro Beltran* 

The Alan Turing Institute, London, United Kingdom

*Corresponding author. E-mail: abeltran@turing.ac.uk

Received: 31 January 2022; **Revised:** 13 November 2022; **Accepted:** 03 February 2023



Key words: auditing; corruption; natural language processing; subnational governments; text-as-data

Abstract

Supreme audit institutions (SAIs) are touted as an integral component to anticorruption efforts in developing nations. SAIs review governmental budgets and report fiscal discrepancies in publicly available audit reports. These documents contain valuable information on budgetary discrepancies, missing resources, or may even report fraud and corruption. Existing research on anticorruption efforts relies on information published by national-level SAIs while mostly ignoring audits from subnational SAIs because their information is not published in accessible formats. I collect publicly available audit reports published by a subnational SAI in Mexico, the Auditoria Superior del Estado de Sinaloa, and build a pipeline for extracting the monetary value of discrepancies detected in municipal budgets. I systematically convert scanned documents into machine-readable text using optical character recognition, and I then train a classification model to identify paragraphs with relevant information. From the relevant paragraphs, I extract the monetary values of budgetary discrepancies by developing a named entity recognizer that automates the identification of this information. In this paper, I explain the steps for building the pipeline and detail the procedures for replicating it in different contexts. The resulting dataset contains the official amounts of discrepancies in municipal budgets for the state of Sinaloa. This information is useful to anticorruption policymakers because it quantifies discrepancies in municipal spending potentially motivating reforms that mitigate misappropriation. Although I focus on a single state in Mexico, this method can be extended to any context where audit reports are publicly available.

Policy Significance Statement

Annual audits by supreme audit institutions produce important information on the health and accuracy of governmental budgets. These reports include the monetary value of discrepancies, missing funds, and corrupt actions. This paper offers a strategy for collecting that information from historical audit reports and creating a database on budgetary discrepancies. It uses machine learning and natural language processing to accelerate and scale the collection of data to thousands of paragraphs. The granularity of the budgetary data obtained through this approach is useful to reformers and policymakers who require detailed data on municipal finances. This approach can also be applied to other countries that publish audit reports in PDF documents across different languages and contexts.

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

1. Introduction

Fiscal decentralization has awarded subnational governments (SNGs) with more spending responsibilities and resources for collecting revenue at the local level. SNGs in Latin America control 26% of all public spending (Radics et al., 2022), this is a significant proportion that highlights the importance of accounting SNG budgets and of understanding their fiscal practices. More localized spending implies more opportunities for malfeasance (Rose-Ackerman and Palifka, 2016); therefore, keeping track of SNG spending is an integral component in the fight against corruption. Across the developing world, survey respondents note that corruption at the local level is a salient issue that directly affects them in their daily lives (Pring and Vrushi, 2019). To tackle malfeasance and improve the financial health of local governments, the international community and multilateral funding organizations agree that effective oversight through the accounting of SNG budgets is important to confirm these are complying with their approved budgets and to identify errors or variations (OECD, 2016; Jeppesen, 2018). Supreme audit institutions (SAIs) are the organization responsible for the oversight of government spending and may implement audits across all levels of government. As the budgets of SNGs have grown though, national level SAIs have struggled to continuously implement effective audits. They are often forced to manage their workload by implementing audits at random, at alternating time periods, or for only a sample of the total budget (Santiso, 2007). This implies that a large portion of budgetary items are not reviewed by the national SAI, but one approach that helps alleviate pressure is the delegation of full audits to subnational SAIs in federal countries. The subnational SAIs often share the legal structure and tools of a national SAI but are only responsible for auditing governments within their defined jurisdiction and can audit the complete scope of local government finances.

SAIs, whether at the national or subnational level, have access to the receipts of items purchased, of bank transfers made, and of other information that they use to evaluate the accuracy of spending. These organizations have the tools and information necessary to identify evidence of misappropriation or of corrupt operations. After performing an audit, the SAI publishes their findings in an audit report that summarizes the procedures followed and provides government officials with recommendations on how to correct fiscal deficiencies. Audit reports can contain data on the value of missing funds, on types of errors or omissions, or details on the compliance of the audited party. In general, audits produce information that local governments can use to improve their accounts and the application of an audit should serve as a deterrent for individuals with corrupt intentions. The discrepancies identified through an audit do not imply outright corruption, but on occasion, an audit may explicitly state that a particular bank operation was fraudulent, these findings are often turned over to a financial crimes prosecutor who uses the information from the audit to prepare an indictment (Gutmann and Voigt, 2020). Regardless, audits report a total monetary value for the discrepancies identified that is useful for assessing compliance, performance, and as a proxy for potential corruption. These rich data are severely underutilized in the existing literature and are rarely available to reformers and civil society groups. Although national-level SAIs attempt to publish this information in a reasonable format,¹ national audits only reflect a partial picture of subnational spending. The reports generated by subnational SAIs contain more contextualized and granular information of what local governments are spending on. Unfortunately, in contexts where many subnational SAIs exist it is difficult for these to publish a centralized database on their findings, especially given the heterogeneity in each institution's procedures and mechanisms for reporting findings. In this paper, I try to breach that gap by introducing a methodology for systematically extracting data from published audit reports using innovations in machine learning and natural language processing that can be adapted to the unique audits of each subnational SAI.

Individual audit reports vary in length, where in some cases an SAI may publish a short summary of their findings, in others it may extend their reporting throughout hundreds of pages. Systematically extracting information from an ever increasing number of pages over hundreds of audits is a significant challenge. When producing these documents, an SAI's audit procedures allow them to use the same

¹ Mexico's national SAI produces a database of audits, see <http://www.asfdatos.gob.mx/>

document format across reports leading to these replicating the same boilerplate content that describes legal procedures, laws they adhere to, and other redundant information. Although an individual could spend an afternoon reading an audit report to identify relevant information, this process quickly becomes cumbersome as the number of reports increases. Thus, at a larger scale there is a “needle in the haystack” problem where important data exists, but it is buried within hundreds of irrelevant paragraphs. I break this problem into two separate challenges: first, filtering out text that is irrelevant; second, identifying the specific items of interest and extracting the relevant information from the text. I address these challenges by building a text classification model that can discern between relevant and non-relevant content that filters out boilerplate language and other noise from the text, and training a named entity recognizer that identifies key concepts in the text and extracts the monetary values associated with them. I use both tools to systematically extract the total monetary value of discrepancies reported in the audits of a subnational SAI.

For this paper, I focus on audits published by a single subnational SAI in Mexico, the *Auditoria Superior del Estado de Sinaloa* (Supreme Audit Institution of the State of Sinaloa, here on referred to as ASE). Data on corruption are scarce throughout Mexico and often rely on anecdotal evidence or surveys asking people about their experience with corrupt officials. Hard measures that contain the value of missing funds are rare. The *Auditoria Superior de la Federacion* (Federal Supreme Audit Institution, ASF) publishes information on discrepancies identified in municipal audits, but the institution is limited to the review of funds transferred from the central government. The existing literature has exclusively focused on audits performed by national SAIs while mostly ignoring those applied by subnational institutions. The subnational-level SAIs review the entire scope of municipal budgets and therefore have access to more detailed information. Mexico is an excellent starting point for the analysis of subnational SAI audits because all 32 states have implemented a version of the ASE. It opens the possibility of comparing performance across SAIs or observing their varying effectiveness in enforcing horizontal accountability. The main limitation though is a lack of accessible data, both in terms of data in usable formats and of the actually published audit reports. Across Mexico, the ASEs vary in how much information they publish through their online web portals, although some may publish the entire collection of audits applied in the past 20 years, none publish spreadsheets or databases of an audit’s results. The ASE of Sinaloa is a great starting point because they apply biannual audits to each of their municipalities and publish all their historical audits online.² Although Sinaloa is not representative of the rest of the country, it does present an interesting case study that offers a snapshot of what data may be available in other state’s ASE documents. From the audits in Sinaloa, I extract the monetary sum of *pliegos de observaciones* or budgetary discrepancies. On their own, pliegos do not imply corruption but are evidence of discrepancies that merit correction by municipal officials and can lead to an investigation or sanction. These data are then combined with official spending data to evaluate the percentage of municipal budgets that are associated with fiscal irregularities. I report an example of a potential analysis made possible by these data, through two ordinary least squares (OLS) regressions I find that discrepancies are not conditional on a municipality’s size, and that a previous year’s value of discrepancy may lead to a larger discrepancy in the following year’s audit.

In the following section, I present a brief review of the literature that uses audit data and of text-as-data in political science research. Then I summarize the process of systematically collecting audit reports and the procedures I use to break them down into machine-readable text. After, I explain the process of building training data for a binary classification model and then describe how the named entity recognizer was created. I end this paper with a discussion on how the data extracted through this process are useful for policymakers and researchers in identifying and tracking poor fiscal management across time.

2. Literature Review

SAIs exist across the developing world. Santiso (2007) summarizes the different types of audit institutions that vary based on the legal tradition of each country. Each SAI has a varying degree of autonomy and

² See <https://www.ase-sinaloa.gob.mx/>

independence that influences the extent of their audits and the mechanisms used to evaluate public spending (Blume and Voigt, 2007; Gustavson and Sundström, 2018; Beltran Aguirre, 2021). The International Organization of Supreme Audit Institutions (INTOSAI) offer a set of guidelines that a country SAI must follow to guarantee their independence so that these have the ability to implement effective audits (INTOSAI, 2007; Otbo, 2009). Although SAIs may be organized differently, they all apply audits that evaluate a government's implementation of their budget in a given fiscal year. A rapidly growing literature uses data from audits in their empirical analysis.

2.1. Research using audit data

Ferraz and Finan (2008) are the first to use detailed audit data in linking these with the results of elections. They find that corrupt mayors exposed in an SAI audit are less likely to be reelected in Brazil. Audit data from Brazil are also used in Ferraz et al. (2012) to study the effects of corruption on education. Multiple studies have utilized the rich audit data from Brazil to study the effects of corruption on accountability, reelection, government performance, and public sentiment (Melo et al., 2009; Pereira et al., 2009; Pavão, 2018; Boas et al., 2019). Brazil stands out as a frequent case study because their SAI randomly implement audits, and their reports are publicly accessible with readily available data. Outside of Brazil, Dunning et al. (2019b) find that across six countries the effects of audit information on elections vary by context (Dunning et al., 2019a). In South Africa, Berliner and Wehner (2021) find that municipal audits can lead to an electoral response when these report poor performance.

Recent literature has also utilized audit data published by the national-level SAI in Mexico. Chong et al. (2015) use these audit reports to run an experiment exploring whether voters punish corrupt mayors. Arias et al. (2019) build on this and implement an experiment in Mexico to identify how reporting on audit discrepancies to voters affects their support of candidates. Larreguy et al. (2020) also use national-level audits to explore the effects of corruption on elections and find that information that is widely disseminated through radio and television can hurt the electoral returns of the accused party. In all three cases, the national-level SAI data are readily available on the ASFs online platform for reviewing specific cases. The national-level SAI in Mexico is limited to audits of intergovernmental transfers from the central government to SNGs, but given Mexico has over 2,500 SNGs it is impossible for this institution to audit each and every one of them, instead only auditing a sample of SNGs. In 2021, for example, the ASF of Mexico audited 249 municipalities of 2,471 (Auditoría Superior de la Federación, 2022), leaving more than 90% of local governments outside of that year's analysis. As noted above, state SAIs are responsible for performing full audits of local governments and thus have significantly better geographic coverage. Each ASE is responsible for disseminating the results of their audits to the general public, but none publish a searchable database of their findings. Instead, their findings are in the annual publications of their audits, with the data and details stored within the text of each document. This information is neither pre-processed nor readily available for use in research which explains why the existing literature has largely ignored it as a source of data. To overcome this limitation, I use natural language processing to extract fiscal data from the text.

2.2. Text-as-data in political science

Recent applications of text-as-data approaches exploit the ever-increasing sources of online archived news for research in the social sciences. Grimmer and Stewart (2013) utilize text-as-data to understand what actors are involved in achieving policy change and what topics are of interest to voters. Applications of machine learning in generating data are a growing approach in political science (Wilkerson and Casas, 2017). These tools are also being popularized in public administration literature which accelerates the data collection process (Hollibaugh, 2019). In conflict research, text-as-data is also used to track drug trafficking organizations (Osorio and Beltran, 2020) and the presence of insurgents (Osorio et al., 2020). In this paper, I leverage natural language processing tools to systematically extract information from audit reports.

3. Audit Reports

3.1. Case study: ASE-Sinaloa

The case study for this paper are the audits produced by the subnational SAI in the state of Sinaloa, Mexico. The ASE-Sinaloa is an excellent pilot because their audit reports are similar and consistent throughout the time period studied which facilitates the systematic extraction of information. Their website publishes historical audit reports from 2002 through the present. Over each time period, the contents and structure of the reports vary given updates in audit procedures and requirements, but the vocabulary and items audited remain the same. For example, in 2002 the audit report simply consists of a recommendation of whether the municipalities finances were in order that excluded any detailed information about specific accounts. By 2010, the content of an audit report became much more detailed, specifying the amounts audited and the location of different funds. Starting in 2012 and through 2016, these audit reports began specifying the total value of *pliegos de observaciones* or fiscal discrepancies identified through an audit of a given municipality. Each document would summarize the discrepancies and publish their total monetary value. Pliegos do not explicitly imply corruption, but they do signal errors, omissions, or potential misappropriation that may merit a further investigation. The values of pliegos are not reported in any official budgetary databases compiled at any government level in Mexico, this information only exists in the audit report that identified them. If a researcher were interested in studying what percentage of municipal funds was tied to discrepancies, they would have to individually review each audit to determine that amount and then combine this information with official budgetary data.

3.2. Collecting audit documents

Each state in Mexico has its own separate ASE that inspects the finances of the state and municipal governments. All 32 ASEs publish their audits through their online portals, but only a few publish all their historical audits. I reviewed the website of each ASE and identified the temporal coverage of the audits they do publish. Table 1 breaks down the years covered and the number of state ASEs within each category of coverage. Of the 32, only 9 publish more than 10 years of audits while the majority had less than 5 years published on their website. I had initially placed an access to information request to obtain the audits from ASEs that did not publish their entire historical collection online through Mexico's National Institute for Access to Information (INAI in Spanish), but these were temporarily put on hold as the world entered various lockdowns caused by COVID-19. Given I did not have complete coverage of all ASEs, I decided to focus on Sinaloa because their information was readily available but also because of my particular familiarity with the contents of their audits and the budgets of municipalities in this state.³ The ASE-Sinaloa website publishes all of their audits in scanned PDF files that are simple to download. As the first case study, it offers an opportunity for identifying the procedures and content review required to extend extraction of fiscal data to audits produced in other states that also publish a collection of their historical audits.

The collection of audit documents was downloaded from the ASE-Sinaloa website through a web crawler using *Scrapy* in Python. The crawler searched for all PDF documents throughout its website and downloaded them, this process collected a total of 3,759 documents. Many of the documents contained

Table 1. Number of ASEs that publish historical audits

Years of audits	Num. of ASEs
+10 years	9
5–10 years	9
1–5 years	14
Total	32

³ See Verdugo Lopez and Beltran Aguirre, 2017

legal information on the organization and for specific audits, others contained appendixes and supporting materials. I subset the actual audit reports published for each municipality on a biannual basis from 2008 through 2016. This reduces our collection to 321 total audit reports that cover Sinaloa's 18 municipalities over a 9 year period. Unfortunately, each audit report is a scanned copy of the official document that contains signatures, stamps, and alterations to the document made during its progression through the state legislature as it was reviewed and approved. Contrary to traditional PDF documents where a user can search through the text, scanned images are much more difficult to process because the items on a page are saved as pictures rather than words. I overcome this limitation by systematically pre-processing each document using optical character recognition (OCR) software `textextract` and `tesseract` in Python to convert each PDF into a machine-readable plain text file.

The documents are written in Spanish and the OCR software struggled to recognize various accented characters. For example, the character *ó* was frequently confused with a *6* which highlights some of the existing limitations with using OCR outside of the English language. To correct for these errors, I identified a list of commonly misspelled words across the documents and generated a dictionary of their correct spelling, and then I programmed a function that read through each document and replaced misspelled words, this helped standardize the errors produced by the OCR and reduced the amount of noise in the text.

The audit reports all follow the same structure, they use the same language, and all contain the content of interest. They also use the same boilerplate language to specify the legal standing of their audits that repeat a description of the procedures used in applying the audit, and of other organizational information that is redundant across all the documents. [Figure 1](#) shows a sample boilerplate paragraph identified in the audit reports. This is their opening paragraph specifying that article 53 of the state's Constitution establishes the legal obligation for the ASE to review the public accounts of the state government and of the municipalities.

Buried within the boilerplate content and legal terminology is the actual result of the audit, specifically the value of discrepancies identified in the audit. Given the abundance of noise in these documents, it is necessary to filter out the non-relevant text from each document. In order to filter the irrelevant chunks of text, I split each document into paragraphs so that I can distinguish between sections of text that contain important information related to the results of the audit and those that simply contain descriptions of the legal requirements they followed, details on how and when the audit was applied, and other information not relevant to my pre-defined classification criteria. To each paragraph, I apply a cleaning function that eliminates white space, rogue characters, and other errors introduced through the OCR conversion. This produces over 28,000 paragraphs across the 321 documents, the high volume of paragraphs further highlights the challenge of reviewing each document.

3.3. *Text classification*

After having split each document into paragraphs, it is now possible to create a text classification algorithm that can read the contents of each paragraph and determine the likelihood of it being relevant or non-relevant. Reviewing all of the 28,000 paragraphs would be a time-intensive task that might produce inconsistent results based on the reviewer's attention, fatigue, or disinterest with the topic as the number of paragraphs read increases. Instead, I use supervised machine learning to train a text

I. El artículo 53 de la Constitución Política del Estado, establece que para el cumplimiento de la obligación de revisar las Cuentas Públicas del Gobierno del Estado y de los Municipios, esta representación social se apoyará en la Auditoría Superior del Estado, que es el órgano técnico de fiscalización general en la entidad.

Figure 1. Boilerplate paragraph example.

classification model that can categorize the population of paragraphs based on the structure, contents, and similarities shared with a sample of paragraphs reviewed by a human. The supervised approach requires that the individual reviewer follows consistent classification criteria throughout the sample of articles read, and to review a sample of paragraphs that contain a wide range of potential paragraphs. This method allows a model to make inferences about all the documents based on the contents of a sample reviewed in detail and produces more consistent results.

From the 28,000 documents, I extracted a random sample of 1,300 paragraphs to use as the training data for the text classification model. I relied on the text annotation software *Prodigy* in Python that offers a friendly user interface where the reviewer can read the text and assign a label to each paragraph. In *Prodigy*, I read the individual paragraphs and labeled them as relevant or non-relevant based on the following criteria: if a paragraph mentioned a type of discrepancy, the approval of a specific line item, the monetary value of a discrepancy, or any other mentions of specific results of the audit, then these were classified as relevant, all else was classified as non-relevant since they are not of interest to the research objective of this paper. This differs from a string matching method where the researcher can search within the text for keywords or paragraphs that contain numbers. In such an approach, one would have to supply the classification algorithm with a robust and comprehensive set of rules to match on that will miss variations in spelling or formatting not covered in the rules. A text classification model uses machine learning to learn what variations of word combinations, variations in spelling, and particularities in paragraph formatting to determine the probability of a paragraph classified to either label. The advantage of it is that the model will learn what features are important without the user explicitly stating them, this minimal input is then used to generate out-of-sample predictions on paragraphs it has not seen. Given the OCR process introduced a lot of noise into the text, a machine learning text classifier adapts better to the variation in the spelling and structure of words rather than relying on a finite list of terms to match on.

The main limitation of machine learning is that it requires a large corpus and a diverse set of training data for it to learn the variety of features within each label. A typical text classification model would require thousands of observations for it to produce reliable results, but *Prodigy* contains a set of internal tools that accelerate the production of an accurate model. For this project, I used the “model-in-the-loop” feature that learns which paragraphs are most likely to be informative to a classification model. It focuses on examples that are the most ambiguous while avoiding the reclassification of items that it already understands. For example, all of the audits started with the same paragraph specifying the legal standing of the audit that contained no relevant information for the project, after the third time reading this paragraph the internal model understood that this content was very likely to be classified as not relevant and thus stopped suggesting it in the training data and default classifying all of the same examples as not relevant. It also understood that paragraphs that made any mention of monetary values were very likely to be classified as relevant and stopped suggesting them. Instead, the interface provided me with the paragraphs it was the least certain about, specifically the most ambiguous and fringe cases contained in the sample that it would learn the most from. This allowed me to focus on paragraphs that contained unique information or examples that were less common across the documents and reduced the classification time from weeks to a few hours. The result is 800 paragraphs labeled as non-relevant and 482 as relevant. The annotated paragraphs serve as training data for a logistic regression classification model that produces a combined F1 score of 0.91 for both labels.

I use the logistic regression model to generate predictions on all of the 28,000 paragraphs. The model reviews the contents of each paragraph and estimates the probable label for each based on the features it learned from the training data. The audits from the ASE-Sinaloa were very similar across time and this made the implementation of a classification model fairly simple. In general, the model performed well on the boilerplate content but struggled with paragraphs that contained overlapping information of boilerplate and relevant data. Fortunately, these were most often classified as false positives rather than false negatives which implies the classification over-identified relevant paragraphs rather than over-excluding them from the final output. The text classification model when implemented on the full population of documents allowed me to reduce the 28,000 paragraphs across 321 audits to 6,517 paragraphs that contained relevant information.

3.4. Information extraction

Although the text classification process filtered out thousands of paragraphs and over a million non-relevant words, the product is still a collection of 6,517 paragraphs that would be challenging to manually review. The audits may include information on specific line items of a municipalities spending, references to purchases procured, and a disaggregated explanation of each discrepancy identified. It may also contain information on items that were not purchased, excess in tax collections or fees, and other information that contains a monetary value. This information is important, but for the purposes of this paper, I focus on a single item that summarizes the findings of the audit to showcase the usefulness of these documents and to present a method for extracting information that can be extended to the rest of the data. My objective is to identify the total value of all discrepancies reported in the audit and extracting that information from the text.

To extract the monetary value of the sum of discrepancies, I implement an additional tool from natural language processing: named entity recognition (NER). The text classifier assigned labels to an entire paragraph, a NER model reviews each word and the contextual information surrounding it to classify entities such as locations, nouns, specific terms, and monetary values. The entities I am searching for are unique to this context and thus require a NER trained from scratch. I again rely on Prodigy to build a NER model that can identify the specific entities that I am searching for. The first are mentions of *pliegos de observaciones* which translates to a “folder of observations.” Pliegos are all of the discrepancies identified in the audit and are generally followed by the monetary value of their sum. This is the actual amount of monies tied to the discrepancies, together they are the specific information I am interested in extracting.

Using the Prodigy interface, I annotate a random sample of 150 paragraphs from the 6,517 classified as relevant to create the training data required by the NER model. Figure 2 presents an example of the annotation interface along with a paragraph I annotated with the labels of interest. The paragraph states that after comparing the approved budget with the actual amount spent, there is a variation that totals a specific amount. Simply searching for monetary values would return a long list of amounts without any contextual information, that is why I have trained the NER model to identify three specific items: monetary values, a mention of a discrepancy, or a mention of a budgetary item recommended for congressional approval. Prodigy specifies the label assigned to each annotated entity, in this example

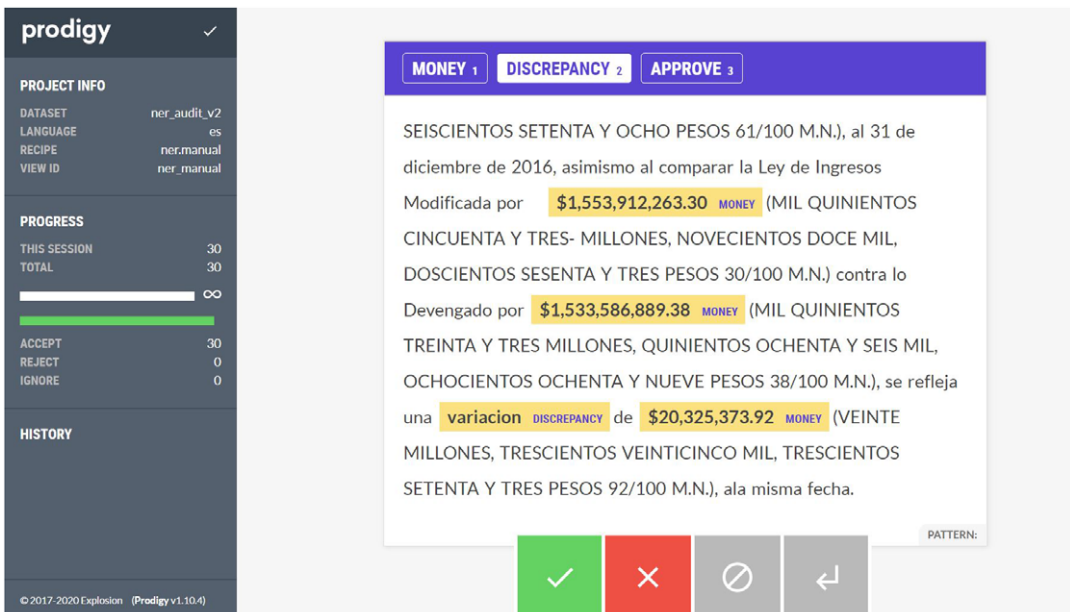


Figure 2. NER annotation in Prodigy.

next to the highlighted amounts the label of “MONEY” appears, and next to the word *variacion* the label “DISCREPANCY” is tagged. Although this example is specific to the items I am particularly interested in, it could easily be extended to the identification of other key terms and values across the documents, it would require the user to annotate additional paragraphs that contain the information the user is searching for. This opens the possibility of identifying payments to companies or individuals, the procurement procedures followed, the amount of unspent sums, excess taxes collected, or other items of interest to researchers and policymakers. The user could elect to build a NER model from scratch or to simply update the model trained here with the additional observations, reducing the time and effort required to extract information significantly.

The 150 annotated paragraphs are then used to train the NER model using `Spacy` in Python. The NER model returned a 0.93 F1 score which suggests it is performing well at classifying the relevant entities. I reviewed a subset of individual paragraphs and confirmed that the model was identifying words associated with a discrepancy and monetary values; therefore, I used this model on the 6,517 paragraphs. The NER model read each paragraph and identified the entities of interest based on the information it learned in the training step.

The NER model returns a list of the entities and their tag for each paragraph, specifically, it found mentions of pliegos followed by their monetary values and extracted that information. [Figure 3](#) demonstrates how the NER model highlights and tags each entity it identified. In this paragraph, the document is summarizing that the total value of discrepancies found in the audit amounted to \$2,886,965.33 MXN. From the list of entities identified by the NER model, I filter for mentions of pliegos followed by their monetary values to produce a dataset with 125 observations where each row represents the total value of discrepancies reported in the audit of a municipality for a particular year.

The use of natural language processing and machine learning facilitates the extraction of information contained deep within the contents of a scanned audit report. I converted 321 scanned documents into machine-readable text and then split these documents into over 28,000 paragraphs. I train a text classification model to differentiate between paragraphs that contain relevant information from those that do not, resulting in 6,517 potentially relevant paragraphs. I extract the total monetary value of all discrepancies reported in an audit using NER and the final product is a dataset with 125 observations that can be used for additional analysis. This pipeline and the models developed can be applied to other cases with minimal user input. For audits produced by other state ASEs, the user would need to update the classification model and the NER model on a random sample of paragraphs extracted from those documents so that the model would recognize their particular nuances. This process can be scaled to a much larger collection of audits, or it can be applied to more recent audits published by the ASE-Sinaloa to continue extracting data on discrepancies. This approach can be extended to countries that do not publish their audits in Spanish, the user can follow the pipeline of splitting documents into paragraphs, training a classifier, and then a NER model to identify the specific information they are in search of.

4. Municipal Discrepancies

After extracting the fiscal information of interest, the final product is a dataset of the sum of discrepancies identified in audits each year from 2008 through 2016. The NER model returned the pliegos and their total

Asimismo, se determinó que los `pliegos de observaciones DISCREPANCY`, los cuales se toman como referente para medir el impacto en la fiscalización, ascienden a `$2,886,965.33 MONEY` (DOS MILLONES, OCHOCIENTOS OCHENTA Y SEIS MIL, NOVECIENTOS SESENTA Y CINCO PESOS 33/100 M.N.).

Figure 3. NER identification in paragraphs.

amount. [Table 2](#) presents descriptive statistics for each municipality audited. A few municipalities reported pliegos across both audits implemented in a year, these observations were aggregated to obtain the total year pliegos and reduced the number of observations from 125 to 69. Across this time period, the total annual value of discrepancies averaged \$19,009,985 MXN. On their own, these numbers offer insights into the variation of official discrepancies reported in the audits. This information is not available in official fiscal data sources and is rarely seen outside of local newspapers.

The information extracted offers important insights into municipal budgets in Sinaloa, but these data can be combined with external sources for additional analysis. To demonstrate a potential use for it, I combine each audit with the municipal-year official budgetary data published by Mexico's National Statistics Institute, INEGI (2020). On average, discrepancies are 4% of total expenditures and in one case over 26% of the municipality's annual budget. This highlights the importance of tracking this information over time given it is a significant portion of the budget under dispute that is currently not included in any of the official sources of budgetary data in Mexico.

In addition, I run a simple OLS regression to analyze the relationship between total spending and discrepancies while controlling for municipal size using population data from INEGI and incorporating year fixed effects. Values are logged to normalize exponential differences between observations. The results in [Table 3](#) suggest that there is no direct effect of total spending on the value of discrepancies. This implies that both big and small municipalities tend to have discrepancies and that their total value is not determined by the size of their expenditures.

If these discrepancies were associated with honest accounting errors, their intensity should decrease over time given municipal officials have the opportunity to learn and correct these issues. This would lead to discrepancies decreasing when regressed on their lagged value as the municipal treasury learns from the previous year's mistakes. [Table 4](#) suggests the opposite. Discrepancies in year $t-1$ have a positive effect on year t 's total value of discrepancy. [Figure 4](#) plots the marginal effects of the logged value of the previous year's discrepancy on the municipality's year t discrepancy. This implies that discrepancies are resulting in more errors over time which cannot be attributed to simple mistakes in accounting practices.

Table 2. Total value of annual discrepancies in MXN\$

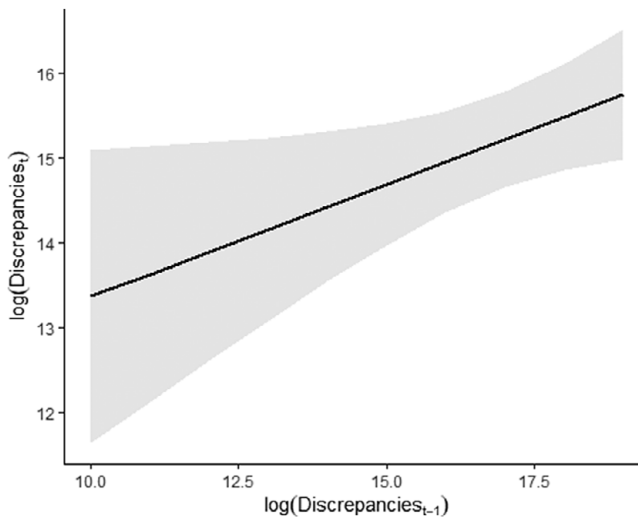
	Municipality	Mean	SD	Min	Max	n
1	Ahome	\$13,918,442	\$10,856,383	\$1,745,475	\$27,344,663	4
2	Angostura	\$2,037,750	\$1,504,349	\$448,301	\$3,432,313	4
3	Badiraguato	\$6,274,011	\$6,372,507	\$258,816	\$12,089,494	4
4	Choix	\$6,084,444	\$5,157,988	\$50,000	\$11,334,923	4
5	Concordia	\$334,133	\$244,055	\$144,745	\$609,554	3
6	Cosalá	\$3,587,062	\$1,833,190	\$1,259,580	\$5,392,141	4
7	Culiacán	\$108,374,933	\$89,151,626	\$20,194,785	\$213,316,351	4
8	El Fuerte	\$52,287,331	\$76,405,662	\$3,661,505	\$166,178,601	4
9	Elota	\$4,061,174	\$5,003,001	\$244,934	\$9,725,258	3
10	Escuinapa	\$7,050,214	\$5,921,600	\$2,076,693	\$15,098,320	4
11	Guasave	\$27,312,917	\$24,032,144	\$2,356,051	\$50,298,971	3
12	Mazatlán	\$23,865,049	\$23,354,743	\$10,474,282	\$58,784,920	4
13	Mocorito	\$9,060,324	\$1,789,617	\$7,052,375	\$11,408,735	4
14	Navolato	\$34,688,409	\$19,498,896	\$15,204,060	\$56,566,432	4
15	Rosario	\$8,218,689	\$4,352,175	\$2,492,043	\$13,038,519	4
16	Salvador Alvarado	\$6,919,343	\$4,979,834	\$1,512,400	\$12,256,169	4
17	San Ignacio	\$4,433,866	\$4,152,006	\$1,135,654	\$10,373,814	4
18	Sinaloa	\$17,341,203	\$21,231,402	\$3,899,338	\$48,979,744	4
19	All municipalities	\$19,009,985	\$37,049,461	\$50,000	\$213,316,351	69

Table 3. OLS Model 1: DV is total discrepancies per municipality

	Estimate	Std. error	<i>t</i> value	Pr(> <i>t</i>)
(Intercept)	−1.8476	9.3066	−0.20	0.8433
log(Total Spending)	0.9803	1.0177	0.96	0.3393
log(Population)	−0.0413	0.9261	−0.04	0.9646

Table 4. OLS Model 2: DV is Discrepancies_{*t*}

	Estimate	Std. Error	<i>t</i> value	Pr(> <i>t</i>)
(Intercept)	2.8687	9.5272	0.30	0.7649
log(Discrepancies _{<i>t</i>−1})	0.2646	0.1224	2.16	0.0366
log(Total Spending)	0.2703	1.0218	0.26	0.7927
log(Population)	0.4136	0.9190	0.45	0.6551

**Figure 4.** Marginal effects of Model 2.

Although discrepancies are not direct evidence of misappropriation, their increase over time may be attributed to a lack of meaningful consequences to errors found in audits that provide corrupt individuals with the opportunity for malfeasance. Of the 18 municipalities audited over the 8 years, not a single mayor was prosecuted nor did they face any meaningful legal consequences for any of their pliegos. The ASE-Sinaloa although effective at identifying these issues may simply lack the institutional support to enforce accountability on the municipalities within its jurisdiction. The results merit a broader analysis on the discretionary tools ASEs have for preventing corrupting and the mechanisms these have for recuperating monies identified in discrepancies.

5. Conclusion

In this paper, I present a pipeline for extracting information from audit reports using natural language processing. I introduce various data science techniques to automate the download of thousands of documents and to convert them into a machine-readable format. Using Prodigy, I annotated a subset of

paragraphs extracted from the audits to train a text classification model that can discern between relevant and non-relevant content. Then, I annotated the contents of relevant paragraphs to train a NER model that can extract discrepancies and their monetary values. The end product is a dataset of 125 discrepancies across 18 municipalities throughout 8 years of audits for the state of Sinaloa in northwestern Mexico.

The data extracted from the audits are not explicit evidence of corruption, and the audits themselves note that the value of discrepancies is subject to correcting on behalf of the municipality, but the list of discrepancies may merit a legal response. In countries where corruption is rarely prosecuted, hard measures on the amounts misappropriated are rare and often rely on rough approximations. The monetary value of discrepancies is a more accurate assessment of potential strategies of malfeasance that is corroborated by the organization created to provide oversight for government finances, the SAI. SAIs are specifically responsible for tracking misappropriation and sanctioning it, the rich data they produce are often locked away in scanned documents. Reviewing and identifying their information are time-intensive tasks that would require a monumental effort to manually extract.

Ideally, SAIs would make the results of their audit reports publicly accessible in a format where the typical citizen can review and understand it. Each ASE in Mexico varies in terms of the information they share, but none of the 32 institutions make their information downloadable in a spreadsheet or database format that would facilitate quantitative analysis. A discussion on why they do not share information in an open format is outside the scope of this paper.

The data extracted from each audit could be used to assess fiscal discipline and evaluate whether local governments are correcting for previous mistakes. The results presented here suggest this is not the case as observed in the positive effect that previous year's discrepancies have on the next year's audits. It can be combined with other fiscal data for a broader analysis on municipal finance and it could also be used to evaluate the performance of the ASE-Sinaloa.

Expanding this approach to multiple states is a feasible task. It would require updating the existing classification and NER models on paragraph samples extracted from audits produced in other states. Although that was the original objective of this paper, because of delays introduced by COVID-19 in the response of access to information requests I instead focused on a single case study with audits publicly available. The systematic processing of audit reports to extract relevant data from them may be a task best suited to the audit institutions themselves. The implementation of this approach signals to these organizations that it is possible to identify this information at a relatively low cost using open software that can be implemented from any computer terminal. The audits could also be used to compare whether municipal expenditures match those reported in official data sources, thus assessing the reliability of that data. The NER model could be trained to identify mentions of sanctions to particular individuals or other detailed information that I did not extract here. The audits contain a plethora of data currently stored in unusable formats, but this paper is an important first step to making the information accessible.

Acknowledgments. An earlier version of this paper was presented at the Data Analytics for Anticorruption in Public Administration Symposium organized by the World Bank in 2021. The author thanks the two anonymous reviewers for their excellent feedback. I appreciate the wonderful comments from Edella Schlager, Javier Osorio, and Alex Braithwaite on an early draft of the project. All of the views and opinions expressed are explicitly my own and do not represent the views of those who I have consulted.

Funding Statement. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing Interests. The author declares no competing interests exist.

Author Contributions. Writing—original draft: A.B.

Data Availability Statement. The data that support the findings of this study are openly available at <https://github.com/AlejandroBeltranA/Sinaloa-Audits>. The source files for each audit are publicly available at <https://www.ase-sinaloa.gob.mx/>.

References

Arias E, Larreguy H, Marshall J and Querubín P (2019) When does information increase electoral accountability? Lessons from a field experiment in Mexico. In *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press, pp. 315–344.

- Auditoría Superior de la Federación** (2022) *Sistema Público de Consulta de Auditorías*. ASF Datos. <http://www.asfdatos.gob.mx/>
- Beltran Aguirre A** (2021) *Accounting for Corruption: Evaluating State Audit Agencies in Mexico*. PhD thesis, The University of Arizona.
- Belliner D and Wehner J** (2021) Audits for accountability: Evidence from municipal by-elections in South Africa. *The Journal of Politics* 84(3), 1581–1594.
- Blume L and Voigt S** (2007) Supreme audit institutions: supremely superfluous? A cross country assessment. International Centre for Economic Research, Working Paper (No.3).
- Boas TC, Hidalgo FD and Melo MA** (2019) Norms versus action: Why voters fail to sanction malfeasance in Brazil. *American Journal of Political Science* 63(2), 385–400.
- Chong A, De La O AL, Karlan D and Wantchekon L** (2015) Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification. *Journal of Politics* 77(1), 55–71.
- Dunning T, Grossman G, Humphreys M, Hyde SD, McIntosh C and Nellis G** (2019a) *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.
- Dunning T, Grossman G, Humphreys M, Hyde SD, McIntosh C, Nellis G, Adida CL, Arias E, Bicalho C and Boas TC** (2019b) Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances* 5(7), eaaw2612.
- Ferraz C and Finan F** (2008) Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes. *Quarterly Journal of Economics* 123(2), 703–745.
- Ferraz C, Finan F and Moreira DB** (2012) Corrupting learning: Evidence from missing federal education funds in Brazil. *Journal of Public Economics* 96(9), 712–726.
- Grimmer J and Stewart BM** (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3), 267–297.
- Gustavson M and Sundström A** (2018) Organizing the audit society: Does good auditing generate less public sector corruption? *Administration & Society* 50(10), 1508–1532.
- Gutmann J and Voigt S** (2020) The independence of prosecutors and government accountability. *Supreme Court Economic Review* 27(1), 1–19.
- Hollibaugh GE** (2019) The use of text as data methods in public administration: A review and an application to agency priorities. *Journal of Public Administration Research and Theory* 29(3), 474–490.
- INEGI** (2020) Finanzas públicas estatales y municipales.
- INTOSAI** (2007). Mexico Declaration on SAI Independence. Technical report, INTOSAI, Mexico City.
- Jeppesen KK** (2018) The role of auditing in the fight against corruption. *The British Accounting Review* 51, 100798.
- Larreguy H, Marshall J and Snyder JM** (2020) Publicising malfeasance: When the local media structure facilitates electoral accountability in Mexico. *The Economic Journal* 130(631), 2291–2327.
- Melo MA, Pereira C and Figueiredo CM** (2009) Political and institutional checks on corruption: Explaining the performance of Brazilian audit institutions. *Comparative Political Studies* 42(9), 1217–1244.
- OECD** (2016). *Supreme audit institutions and good governance: Oversight, insight and foresight*. In *OECD Public Governance Reviews*. Paris: OECD Publishing.
- Osorio J and Beltran A** (2020) Enhancing the detection of criminal organizations in Mexico using ML and NLP. *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Glasgow, UK. IEEE.
- Osorio J, Reyes A, Beltran A and Ahmadzai A** (2020) Supervised event coding from text written in Arabic: Introducing Hadath. In *Proceedings of the Workshop on Automated Extraction of Socio-Political Events from News 2020*, pp. 49–56, Marseille, France. European Language Resources Association (ELRA).
- Otbo H** (2009) SAI Independence: A founding principle of INTOSAI. *International Journal of Government Auditing* 36(4), 1.
- Pavão N** (2018) Corruption as the only option: The limits to electoral accountability. *The Journal of Politics* 80(3), 996–1010.
- Pereira C, Melo MA and Figueiredo CM** (2009) The corruption-enhancing role of re-election incentives? Counterintuitive evidence from Brazil's audit reports. *Political Research Quarterly* 62(4), 731–744.
- Pring C and Vrushni J** (2019) *Global Corruption Barometer Latin America & the Caribbean 2019: Citizens' Views and Experiences of Corruption*. Transparency International.
- Radics A, Vázquez F, Pérez Benítez N and Ruelas I** (2022) Panorama de las relaciones fiscales entre niveles de gobierno de países de América Latina y el Caribe. *Banco Interamericano de Desarrollo*.
- Rose-Ackerman S and Palifka BJ** (2016) *Corruption and Government: Causes, Consequences, and Reform: Second Edition*. New York, NY: Cambridge University Press.
- Santiso C** (2007) Eyes wide shut? The politics of autonomous audit agencies in emerging economies (May 31, 2007).
- Verdugo López M and Beltrán Aguirre, A. D. J.** (2017). *Ciudades, desarrollo urbano y autonomía financiera: dilemas para la gobernanza local en México*. Culiacán, MX: Universidad Autónoma de Sinaloa, Juan Pablos Editor.
- Wilkerson J and Casas A** (2017) Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science* 20, 529–544.