

INTERNATIONAL MONETARY FUND

Predicting the Law

Artificial Intelligence Findings from the IMF's Central Bank Legislation Database

Khaled AlAjmi, Jose Deodoro, Ashraf Khan, Kei Moriya

WP/23/241

***IMF Working Papers* describe research in progress by the author(s) and are published to elicit comments and to encourage debate.**

The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

**2023
NOV**



WORKING PAPER

IMF Working Paper

Information Technology Department and Monetary and Capital Markets Department

Predicting the Law: Artificial Intelligence Findings from the IMF's Central Bank Legislation Database**Prepared by Khaled AlAjmi, Jose Deodoro, Ashraf Khan¹, and Kei Moriya**¹ Corresponding author

All authors contributed equally to this work

Authorized for distribution by Bachir Boukherouaa and Jihad Alwazir

November 2023

IMF Working Papers describe research in progress by the author(s) and are published to elicit comments and to encourage debate. The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

ABSTRACT:

Using the 2010, 2015, and 2020/2021 datasets of the IMF's Central Bank Legislation Database (CBLD), we explore artificial intelligence (AI) and machine learning (ML) approaches to analyzing patterns in central bank legislation. Our findings highlight that: (i) a simple Naïve Bayes algorithm can link CBLD search categories with a significant and increasing level of accuracy to specific articles and phrases in articles in laws (i.e., predict search classification); (ii) specific patterns or themes emerge across central bank legislation (most notably, on central bank governance, central bank policy and operations, and central bank stakeholders and transparency); and (iii) other AI/ML approaches yield interesting results, meriting further research.

RECOMMENDED CITATION: Al Ajmi, K., J. Deodoro, A. Khan, K. Moriya, 2023, *Predicting the Law: Artificial Intelligence Findings from the IMF's Central Bank Legislation Database*. IMF Working Paper 23/241. Washington, D.C.: International Monetary Fund.

JEL Classification Numbers:	C53, C55
Keywords:	central bank legislation; central banking; artificial intelligence; machine learning; Bayesian algorithm; Boolean algorithm; central bank governance; law and economics
Author's E-Mail Address:	kalajmi@imf.org ; jdeodoro@imf.org ; akhan4@imf.org ; kmoriya@imf.org

WORKING PAPERS

Predicting the Law

Artificial Intelligence Findings from the IMF's Central Bank Legislation Database

Prepared by Khaled AlAjmi, Jose Deodoro, Ashraf Khan, and Kei Moriya¹

¹ Production assistance provided by Julie Vaselopulos.

Contents

Glossary	4
Introduction	5
AI/ML Approaches.....	8
User Statistics	22
Conclusion.....	26
Annex I. CBLD Coding App.....	27
Annex II. Overview of CBLD Search Categories	29
References.....	30

FIGURES

1. CBLD Data Coverage 2010/2015, 2020/2021).....	6
2. Number of Included Laws (by country).....	7
3. CBLD Data Flow	9
4. Example of the Landing Page and a Coding Page of the CBLD Coding App	10
5. Multiplicity of Categories	11
6. Most Relevant Words and Coefficients for Category 1.01	13
7. Fraction of Correct Categories Against Ranking.....	14
8. Fraction of Correct Categories Against Ranking Across All Categories	14
9. Most Counted Tokenized Words	15
10. Themes of Frequently Occurring Tokenized Words in the CBLD	16
11. Word Network Graph for Common Bigrams (Albania).....	17
12. Word Network Graph for Common Bigrams (Italy).....	18
13. Word Correlation in Laws for United Kingdom and India	18

Contents (concluded)

FIGURES (concluded)

14. Proportion of Bigram Combinations on “Independ” and “Autonom” Across Central Bank Legislation	19
15. Coverage of Bigram Combinations of “Independ” and “Autonom” Across Central Bank Legislation	20
16. CBLD External Users: Number of Searches	23
17. CBLD External Users: Regional Coverage (in percent).....	23
18. CBLD Daily User Queries (2021, 2022)	24
19. CBLD Daily User Queries by Search Category (2021–2023)	25
20. CBLD Daily User Queries by Country (2021–2023)	25
21. CBLD User Queries by Country (2021–2023)	25

Glossary

AI	Artificial Intelligence
AIV	IMF Article IV
AML/CFT	Anti-Money Laundering/Countering the Financing of Terrorism
API	Application Programming Interface
BoE	Bank of England
CBI	Central Bank Independence
CBLD	IMF Central Bank Legislation Database
CBT	IMF Central Bank Transparency Code
ELA	Emergency Liquidity Assistance
EMs	Emerging Markets
ESCB	European System of Central Banks
FSAP	Financial Sector Assessment Program
FX	Foreign Exchange
IMF	International Monetary Fund
ITD	IMF Information Technology Department
LICs	Low-Income Countries
LLM	Large Language Models
LOLR	Lender of Last Resort
MCM	IMF Monetary and Capital Markets Department
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MOID	IMF Monetary Operations and Instruments Database
NBU	National Bank of Ukraine
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RBI	Reserve Bank of India
TF-IDF	Term Frequency – Inverse Document Frequency

Introduction

Artificial intelligence (AI) and machine learning (ML) offer numerous opportunities for financial sector participants. The FSB (2017) highlights that the various factors that spurred the use of fintech in general, have also led to further adoption of AI/ML in financial services. This includes the availability of more computing power, cheaper storage, parsing, and analysis of data, as well as the “rapid growth of datasets for learning and prediction owing to increased digitization and the adoption of web-based services.” Accordingly, central banks, supervisors/regulators, and market participants can deploy AI/ML tools¹ to improve their products, services, risk management, compliance, and—specifically for legislators/regulators—their development of relevant legislation and regulations.

The IMF's Central Bank Legislation Database (CBLD) is large text-based dataset that offers an interesting testing ground to analyze developments in central bank legislation worldwide. The CBLD is the most comprehensive central bank legislation database in the world.² The CBLD currently contains laws of 175 central banks and monetary unions and has 273 specific search categories that allow users to run granular queries on nearly any topic that could involve a central bank. The CBLD includes datasets from four specific update moments: 2010, 2015, 2020/2021, and 2023. Going forward, annual updates of the CBLD will take place in a prioritized manner (i.e., a prioritized selection of laws will be included in the annual update, based on, e.g., laws that have recently been fully updated, laws that include novel concepts, laws from countries that have limited data in the CBLD, etc.). CBLD data can be accessed by searching by country, or by pre-set groups of countries (notably, by region, income level, exchange rate arrangements, and membership of a monetary union). The database was opened to the public in February 2019; access requires a one-time (free) registration via the CBLD's website (<https://data.imf.org/cblld>).

Central banks are a heterogenous group of public institutions in terms of their legal framework, mandates, policies, operations, institutional structure, and size. Central banks are, in all cases, governed by a legal framework that outlines their objectives, functions, and instruments/tools. This legal framework often includes a central bank law, as well as frequently a country's constitution and other laws (such as banking, budget, Anti-Money Laundering/Countering the Financing of Terrorism (AML/CFT), (digital) currency, payment systems laws, and other financial sector laws), in as far as they contain relevant references to the central bank. The level of detail of central bank and other laws also differs—some laws can easily count hundreds of pages; other laws are one-pagers. No unified comparison can be drawn between laws that govern central banks, because legal traditions and legal frameworks differ significantly between countries.

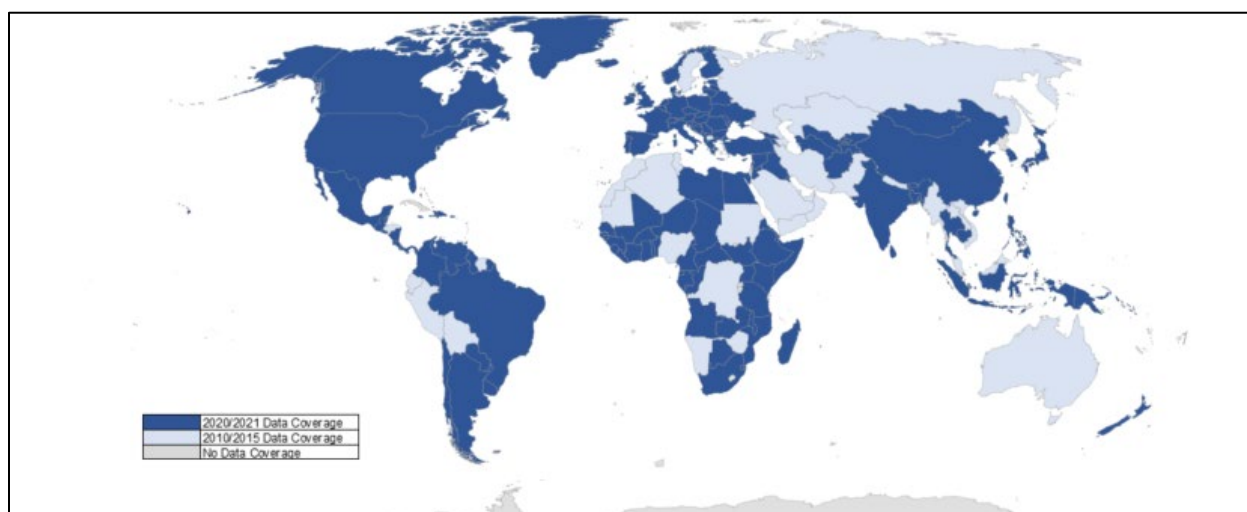
The CBLD includes data provided by central banks and monetary unions. For the updates of 2010 and 2015, a questionnaire was sent to all central bank governors and monetary unions presidents, requesting them to submit their relevant central bank legislation to the IMF CBLD team. From the 2020/2021 update onwards,

¹ AI and ML are related, but not identical concepts. AI refers to the general ability of computers to perform tasks that normally require intelligence, such as thinking, solving problems, and learning. Machine learning is a branch or a subset of AI that uses algorithms to automatically learn from data and form predictive models. Machine learning is a current application of artificial intelligence that we utilize in our day-to-day lives. In other words, machine learning is a pathway to artificial intelligence. This subcategory of AI uses algorithms to automatically learn insights and recognize patterns from data, applying that learning to make increasingly better decisions.

² See also Khan (2017).

data was largely collected by a similar questionnaire (in 2020) and then supplemented with additional legislation (in 2021) by the IMF CBLD team without consulting central banks³—given that for almost all countries the relevant legislation is nowadays available online.³ All collected data is in English only—either provided by the authorities or translated by the IMF (in which case a disclaimer would be added, noting that the provided text is not an official translation). Figure 1 below shows CBLD data coverage for the 2020/2021 update, in addition to data coverage for the combined CBLD updates of 2010 and 2015. Countries that are not included either did not respond to the IMF's questionnaire, are not members of the IMF, or the IMF CBLD team was not able to include relevant legislation (either because amendments were ongoing and it would make sense to wait till those had passed, or because of lack of time).

Figure 1. CBLD Data Coverage (2010/2015, 2020/2021)



Source: IMF CBLD.

The CBLD has strong added value for low-income countries (LICs) and emerging markets (EMs). Given legal capacity constraints, limited resources, and legacy issues with central bank legal frameworks in LICs and EMs, there is a visible demand from authorities in those countries to efficiently access central bank legal frameworks of relevant peer countries, especially in the case of ongoing legislative amendment processes. The CBLD allows such access and can assist country authorities of making well-informed decisions about their current central bank legal frameworks and possible (future) amendments of relevant laws.

The 2020/2021 CBLD update significantly expanded the content covered by the database. The 2020/2021 update of the CBLD expanded the search categories to 273 (from the previous 112). The search categories now also include detailed search options for a wide host of possible central bank functions, varying from macroprudential policy to foreign exchange-related policy and reserve management, as well as aspects related to payments systems, financial integrity, consumer protection, and relatively newer topics such as fintech and climate change.⁴ The main reasons for this are the significant expansion of central bank (legal)

³ In cases where the IMF CBLD team would have questions (for instance, several countries would have different versions of the central bank law published by, e.g., the central bank, the constitutional court, and the ministry of finance), team members would reach out to the respective central banks' legal departments to ask for clarification.

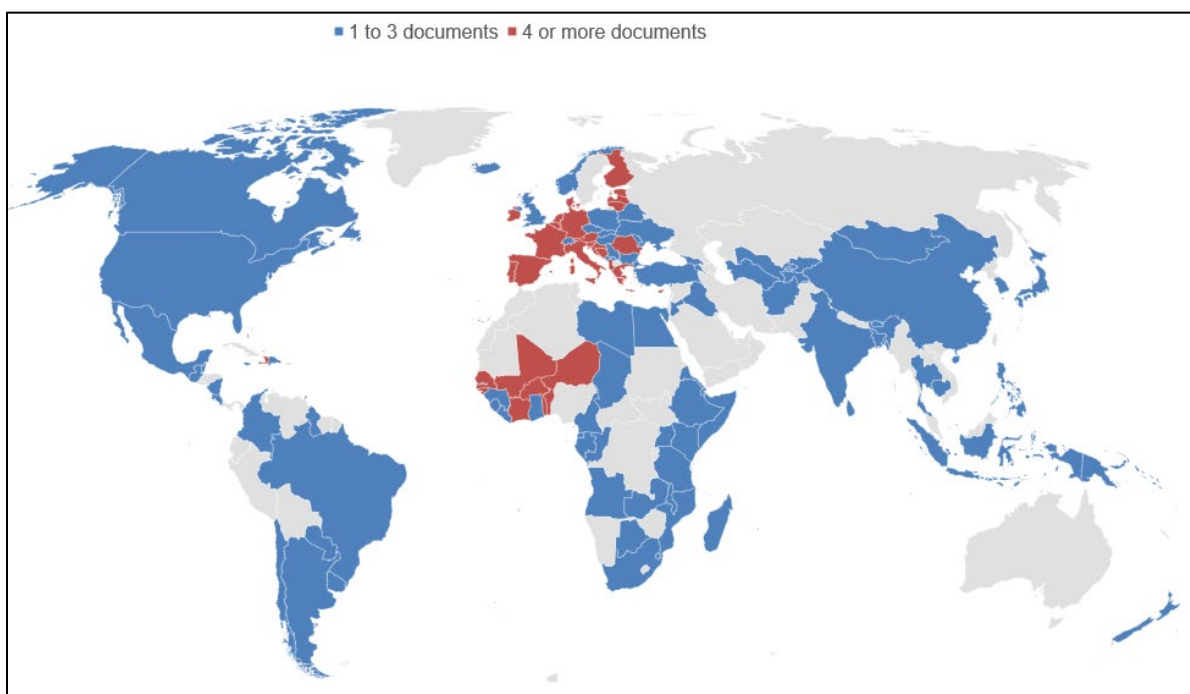
³ The only exception being Eritrea, where the central bank law is not published online.

⁴ Annex II contains a detailed overview of the CBLD's main search categories.

mandates since the Global Financial Crisis, and the (linked) development of central banks taking up an increased number of functions within a single organization.

As noted above, documents included in the CBLD are often central bank laws/charters, relevant excerpts of constitutions, as well as numerous other laws that relate to the central bank in one form or another. In those cases where a central bank is part of a monetary union (with or without national central banks), the relevant supranational legislation would be included as well (for instance, in the case of the European Monetary Union, this would include the EU Treaty and the EU Statute). As an example, Figure 2 depicts the number of documents included in the CBLD per authority. For example, for central banks of countries that are members of the European Union and for the central bank of the West African Monetary Union, the CBLD contains more than 4 separate laws per central bank. For most central banks, the number of documents is limited to 1-3. Though this is not relevant for a substantive analysis of the content of those laws, it is good to take note of this when analyzing the machine learning applications that are applied to the CBLD—see next section.

Figure 2. Number of Included Laws (by country)



Source: IMF.

The CBLD also includes other laws than central bank laws. Though most documents included in the CBLD are central bank laws, increasingly other laws are included as well, in as far as they pertain to the central bank. This includes, for instance, payments system laws, banking/supervision laws, AML/CFT laws, and resolution laws. Financial stability laws are similarly included, with a focus on micro prudential supervision, macroprudential oversight, ELA/LOLR, and resolution.

The CBLD has been designed keeping the newly introduced IMF Central Bank Transparency Code in mind. In 2020, the IMF published its Central Bank Transparency Code (CBT). The CBT⁵ is a voluntary international standard containing 47 principles (most of the principles also include sub principles, which provide more detailed guidance on the specific topics), divided into five groups: (i) transparency over central bank governance; (ii) transparency over central bank policy; (iii) transparency over central bank operations; (iv) transparency over central bank outcomes; and (v) transparency over central bank official relations. For most of the principles, three different sets of practices are listed (core, expanded, and comprehensive). The CBLD and the CBT are aligned in terms of specific search categories for central bank transparency by function. For instance, CBLD search category 2.20 relates to a central bank's general policy on institutional transparency, with various other categories relating to central bank function-specific transparency arrangements (including 5.11 on currency, 6.13 on monetary policy and operations, 7.04 on international reserves, 8.09 on foreign exchange policy, 10.08 on macroprudential policy, 11.11 on financial supervision, 12.12 on lender of last resort, 13.11 on financial integrity, 14.07 on consumer protection, 15.21 on resolution, and 16.08/16.17 on payment systems and financial market infrastructure).

Accordingly, the CBLD offers interesting opportunities for AI/ML approaches. Given the size of the dataset and the qualitative (i.e., textual) nature of the data, the CBLD offers possibilities for AI/ML approaches to help identify possible patterns. Section 2 provides an overview the various AI/ML approaches explored in the context of the CBLD. Section 3 provides additional insights from current CBLD user statistics and the interest in specific search categories. Finally, Section 4 offers concluding thoughts on how these ML approaches could benefit the policy discussions surrounding central bank legislation—including on specific topics such as central bank independence and central bank objectives.

AI/ML Approaches

Section A will describe the methodology of the CBLD Coding App, with section B providing more details on the algorithms explored for the CBLD Coding App.

The last two sections demonstrate additional ML use cases, focusing on word frequency comparison (section C) and—by means of example—sentiment analysis of the concept of “central bank independence” (section D).

A. CBLD Coding App

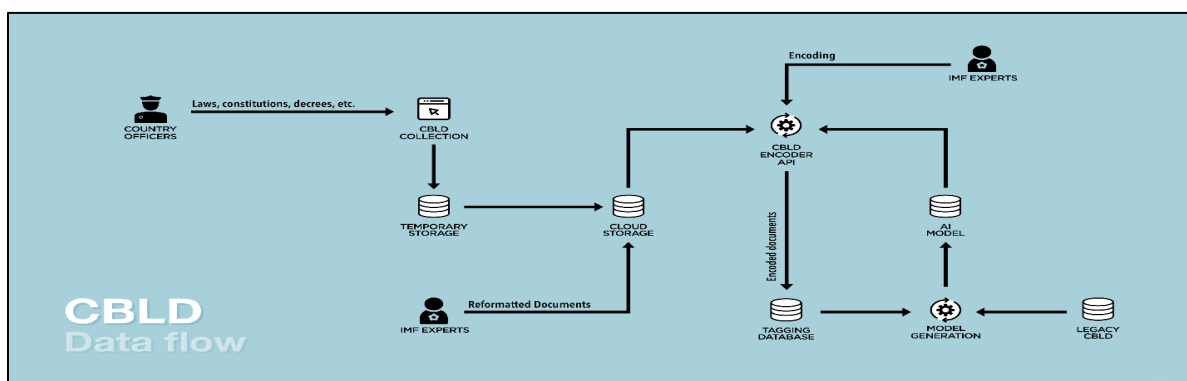
The CBLD 2020/2021 data collection allowed central banks to upload relevant files for inclusion in the CBLD, supplemented with additional laws selected by the IMF CBLD team. Between January and May of 2020, the IMF provided a public website for CBLD data collection. An email from the IMF's Monetary and Capital Markets Department (MCM) was sent to all IMF member countries' central banks and monetary unions (addressed to the Governor/President), requesting input from the respective authority. The hyperlink included in the email guided users to a dedicated IMF website, where they were presented with a registration form that collected basic information (such as user identification and e-mail address), and enabled users to subsequently upload document files containing the relevant legislation. During the remainder of 2020, the IMF CBLD team

⁵ <https://www.imf.org/external/datamapper/CBT/> and also IMF (2020).

collected additional central bank legislation (from central bank websites) to be added to the database; this mostly related to non-central bank laws, such as banking laws and AML/CFT laws.

The IMF CBLD team manually coded every line in the to-be-included central bank legislation. Following the submission phase, a team of two central bank lawyers and an IMF research officer, overseen by an IMF senior expert on central bank legislation, conducted so-called “coding” of every submitted piece of central bank legislation. The coding process consisted of annotating every article, and often specific sentences within an article, against the CBLD’s list of 273 categories. Articles and text within articles, even to the level of specific sentences, were often coded against multiple CBLD categories, allowing for very granular search queries. In the previous update rounds of the CBLD (most notably, 2010 and 2015), this coding was done on printed versions of the respective laws, and subsequently manually reviewed and entered into the system. However, with the onset of the COVID-19 pandemic, the use of paper printouts was no longer feasible, and the team worked with IMF IT experts to develop a tailored electronic approach to coding. See Figure 3 below for a more structured overview of the steps involved with including central bank legislation into the CBLD.

Figure 3. CBLD Data Flow



Source: Authors.

Building on the electronic approach, the IMF developed an in-house AI tool for faster and more efficient CBLD data coding. A prototype of the so-called “CBLD coding app” was built based on an open-source project.⁶ The app presented users with a view of a specific PDF document, enabling them to highlight and annotate sections of plain text or screenshots. It is published under a Massachusetts Institute of Technology (MIT) open-source license, which allows anyone to download, alter, and deploy the original source code as necessary, as long as the original license is available with the altered code. The prototype was successfully tested by the IMF CBLD team, and it resolved various coding issues. First, as the new application was hosted on a Microsoft Azure Cloud website, it naturally provided concurrent access for IMF staff and external legal experts and required no specialized software as it was accessible from major web browsers. Second, it required no specialized knowledge: documents were presented in the web browser in their original format, with users highlighting text and choosing the respective CBLD categories in an interface similar to any Microsoft Word or PDF document viewer. Third, the coding app merely required a simple conversion one-to-one from submitted central bank legislation documents to the PDF format, easily achievable with available software tools.

⁶ [GitHub - agentcooper/react-pdf-highlighter: Set of React components for PDF annotation.](https://github.com/agentcooper/react-pdf-highlighter)

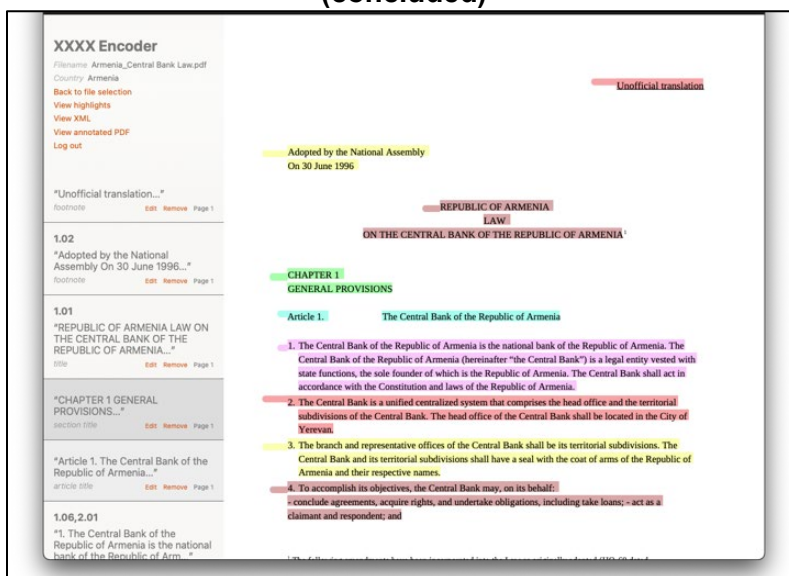
The CBLD coding app allowed IMF experts to pick specific laws to analyze and assign relevant CBLD categories to the legal text. The coding app's landing page (see Figure 4 below) provided an overview of all available files (with a single file generally being a stand-alone law), by country, in alphabetic order. After selecting a specific file, the law would be loaded along with previous annotations (if any), allowing users to then select specific text and assign one or more CBLD search categories to the text. The selected text could consist of multiple articles, one article/section, or even specific sentences or words within the article.

An AI component was added to the coding app to analyze coding actions and subsequently predict future coding actions. Custom fields were added to the coding app, allowing for the deployment of an AI/ML recommender system, as well as the recording of the AI/ML recommender's performance. The recommender used reinforced learning, as it analyzed the highlighted text and choices of CBLD categories by team members. This allowed it to subsequently present a selection of "best CBLD categories" for next coding actions. Team members would then be able to accept or reject those selected categories, thereby reinforcing or adjusting the recommender's selection. When a user would highlight a specific part of text in the document, the text would be sent to an Azure *server-less* Application Programming Interface (API), which utilized a pre-computed recommender model to compute a list of CBLD categories with the closest affinity to the submitted text. Those categories would then be suggested back to users by visual cues. Initially, categories were reordered based on the affinity scores computed by the recommender model. After a few weeks, users reported that the ever-changing category order was confusing. The interface was recoded: category lists now defaulted to numeric order, highlighting categories with the highest affinity scores in orange and adjusting the intensity proportionally to each category score. The AI/ML recommender system is based on a Naïve Bayes model, trained with the previous (2015) dataset of the CBLD—see subsection B. The model was loaded into the API at the beginning of the coding exercise. Once coding was underway, annotations with text and the relevant CBLD categories were periodically re-computed by the recommender model and reloaded in the API.

Figure 4. Example of the Landing Page and a Coding Page of the CBLD Coding App

Country	Document
Armenia	Armenia_Central Bank Law.pdf
Armenia	Armenia_Constitution_CBA chapter.pdf
Australia	2022 CBLD_Reserve Bank of Australia 1959 Amendments through 2020.pdf
Austria	Austria Central Bank Act_1984_june_2018.pdf
Bahamas	Bahamas_Central Bank of The Bahamas Act, 2020.pdf
Bahrain	2022 CBLD_THE CENTRAL BANK OF SAHRAIN AND FINANCIAL INSTITUTIONS LAW ENGLISH.pdf
Bahrain	2022 CBLD_CBB-Law Amendment-2015.pdf
Bahrain	2022 CBLD_Central Bank of Bahrain Legislation Amendment 2.pdf
Bahrain	2022 CBLD_Central Bank of Bahrain Legislation Amendment 3.pdf
Barbados	Barbados_Central Bank of Barbados Act 2018-20.pdf
Barbados	Barbados_CBB Amendment Act 2018-38.pdf
Barbados	Barbados_Central Bank of Barbados Act CAP323C (corrected copy).pdf
Belarus	Belarus_Banking Code of Belarus.pdf
Belarus	Belarus_statute of the National Bank of the Republic of Belarus.pdf
Belgium	Belgium_Loi organique - EN 01.2020 (clean).pdf
Belize	Belize_Central Bank of Belize Act_revised-edition-2011.pdf
Belize	Belize_Act No. 8 of 2018 Central Bank of Belize Amendment.pdf
Belize	Belize_Act No. 28 of 2017 Central Bank of Belize.pdf
Belize	Belize_Central Bank of Belize Act No. 19 of 2016 revised edition 2011.pdf
Bhutan	Bhutan_The Financial Services Act of Bhutan 2011.pdf
Bhutan	Bhutan_FOREIGN EXCHANGE RULES AND REGULATIONS 2018.pdf
Bhutan	Bhutan_FOREIGN EXCHANGE OPERATIONAL GUIDELINES 2018.pdf
Bhutan	Bhutan_Prudential Regulations 2017.pdf
Bhutan	Bhutan_Corporate Governance Rules and Regulations 2018.pdf
Bhutan	Bhutan_Rules and Regulations for Central Registry for Secured Transactions in Bhutan 2019.pdf
Bhutan	Bhutan_RMA_Act_2010.pdf
Bhutan	Bhutan_Anti-Money Laundering Act 2018.pdf
Bolivia	Bolivia_Political Constitution of the State_Articulos 326 300 CPE Politica monetaria y financiera_translated_this.pdf
Botswana	Botswana_BANKING AMENDMENT ACT.pdf
Botswana	Botswana_Finacial Intelligence Act.pdf
Brazil	Brazil_LEI N 4.595, DE 31 DE DEZEMBRO DE 1964_translated.pdf
Bulgaria	Bulgaria_BG - Law on the Bulgarian National Bank (as amended till February 2020).pdf
Chile	Chile_Law No 20.789 - Financial Stability Council (extract) (in force).pdf

Figure 4. Example of the Landing Page and a Coding Page of the CBLD Coding App (concluded)

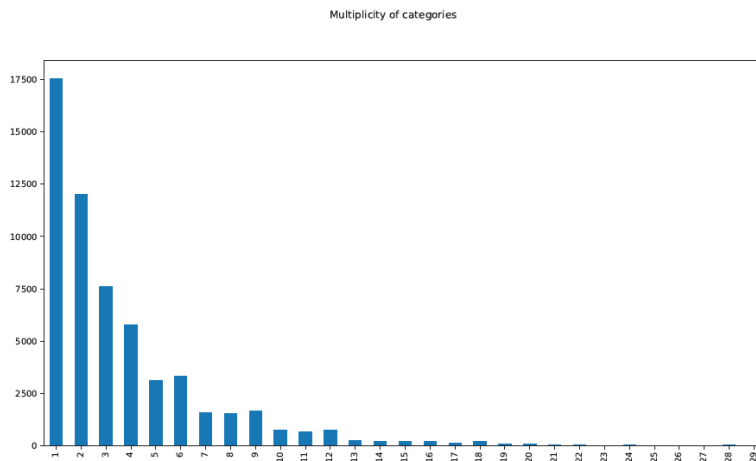


Source: Authors.

B. ML Algorithm

This subsection will explore the AI algorithm used for the coding app, as well as identify its increased efficiency as more data was added. The applied algorithms were selected based on their enhanced application on datasets consisting of short amounts of text, categorized into the pre-defined CBLD search categories (i.e., the average length of CBLD text—such as an article—is 97 words, with a median of 51 words). The ML approach’s goal was to take this text and predict the correct category out of the available 273 CBLD categories. As noted above, the same text could be categorized as belonging to multiple categories. A histogram of the multiplicity of categories for each text is shown in Figure 5 (which highlights that each text most commonly has one category assigned to it, though there are also texts that have up to six categories assigned to it).

Figure 5. Multiplicity of Categories



Source: IMF calculations.

For this task a Naïve Bayes classifier was chosen, implemented by the scikit-learn library⁷ in the Python programming language. The Naïve Bayes algorithm⁸ is a powerful and scalable supervised classifier often used in document classification and spam email filtering. The classifier models the probability of a class variable y being dependent on features x_1 through x_n and makes the simplifying assumption that each x_i is independent. This simplification, together with Bayes' theorem, leads to a model where the probability of class y is proportional to the product of each $P(x_i|y)$ which is the probability of a given word being included in that class. The classification algorithm calculates each of these $P(x_i|y)$ which allows inspection of which words are important for each category, providing transparency of the model. A Naïve Bayes model can be estimated in a short (linear) time with very limited data compared to other ML algorithms, and the explainability of the model is useful for detailed analysis.

The algorithm is well-suited for the CBLD coding task, as each text may correspond to multiple categories, with correct categories therefore not being unique. As the probability of each category for a given text is given by the sum of the log of probabilities $P(x_i|y)$, it is possible to calculate the probability of each category and evaluate where the correct category ranks. This is unlike other classification algorithms, such as decision trees, where the only answer provided would be the predicted category and no further information would be available on the rank of the correct category within the model.

Multiple Naïve Bayes algorithms were tested. In some cases, a stemmer (Porter or Snowball stemmer provided by the Python Natural Language Toolkit (NLTK) package⁹) in combination with the MultinomialNB or BernoulliNB classifiers from the scikit-learn Python package. The MultinomialNB classifier uses the frequency of words to calculate the log probability $\ln [P(x_i|y_j)]$ for each word x_i in class y_j . These are referred to as weights or coefficients, and the probability of a given class y_j for a given combination of x_i is simply the sum of these terms. Note that due to this feature it is possible to calculate the *ranking* of the correct category. This helps as in general it is difficult to predict the correct category as the best one out of 273 categories, especially in the case where the “correct” category can be multiple categories. However, being able to calculate the ranking allows insights into whether the correct category appears in the top N, which will be useful for a suggestion system. Additionally, a further technique to unweight frequently appearing words is Term Frequency-Inverse Document Frequency (TF-IDF), which was applied to some models.

The final set of models consisted of:

- 1) Porter/Snowball stemmer combined with raw counts/TF-IDF counts and modeled as MultinomialNB; and
- 2) counts/tf-idf counts modeled by MultinomialNB/BernoulliNB

leading to eight distinct Naïve Bayes models. In all cases, the data was split into a training and test dataset, and after training the model using the training dataset, estimates of model performance were done independently on the testing set so that all results are “out of sample.”

⁷ Pedregosa e.a., 2011, *Scikit-Learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830.

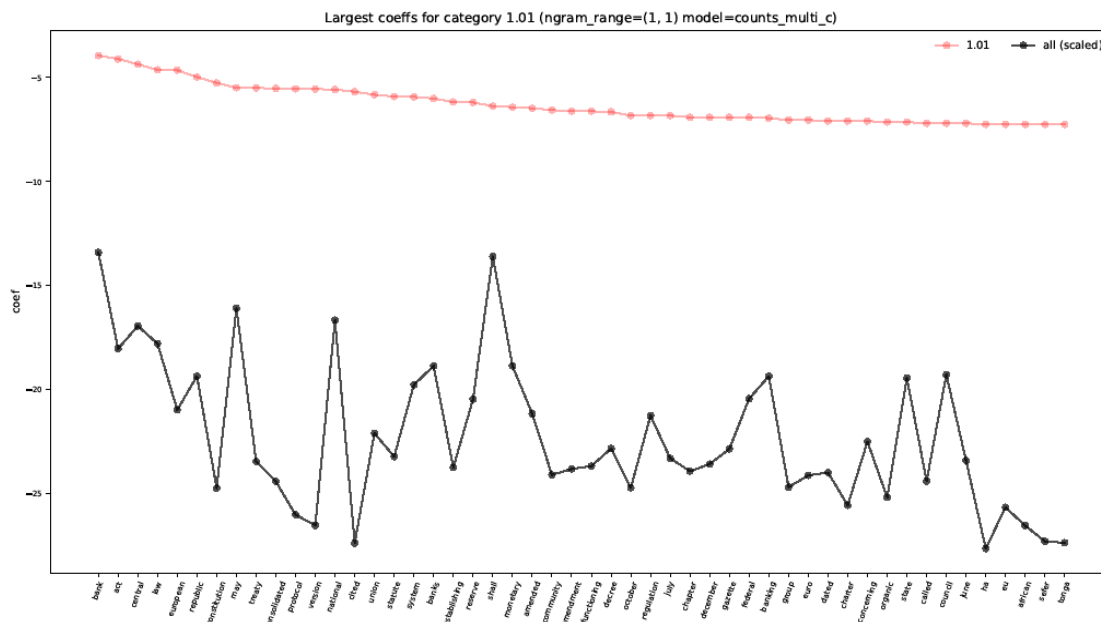
⁸ For further information on the Naïve Bayes algorithm, see A. McCallum, Nigam, K., 1998, *AAAI-98 workshop on learning for text categorization*. Vol. 752. Other algorithms, such as random forest, were also tested but improvements in performance were not found.

⁹ S. Bird, Klein, E., e.a., 2009, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc. <https://scikit-learn.org/stable/>.

Figure 6 shows the most relevant coefficients for the model created for category 1.01. Words are sorted in order of weights for category 1.01 as the red dots and line, and for reference the value of each coefficient across all categories is shown as the black line, scaled down by 100. These charts assist in deducing which words are important factors when categorizing text to different categories.

Figure 6. Most Relevant Words and Coefficients for Category 1.01

(Note: The probability of the text being classified in this category is proportional to the sum of coefficients shown)

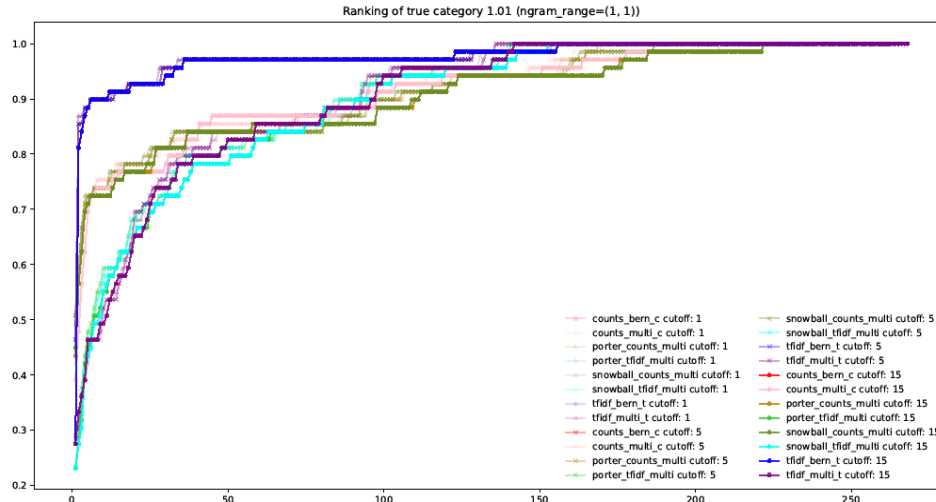


Source: Authors.

By using these coefficients, it is possible to calculate the *ranking of the correct category (or in some cases categories) of each text*. This is useful for a recommender system if the correct category appears within the top five to ten candidates for users to select. The ranking of the correct category is computed, i.e., when the probability of all classes is calculated for a given text, and the rank the correct category shows up in, a number between one and 272. If multiple categories have the same rank as the correct rank, the correct rank is taken as the minimum rank. Then, for each category the fraction of times the correct category was chosen against this ranking is calculated. This is shown in figure 6 where an example of one category (1.01) is shown. It demonstrates that with the best model, the probability of having the correct category within the top five candidates is more than 90 percent, indicating that there is a good chance that the best category can be suggested if five candidates are shown.

Figure 7. Fraction of Correct Categories Against Ranking

(Note: For this category, the probability of identifying it correctly as the best category is around 25 percent, but with the best model the probability that it is in the top five is around 90 percent)

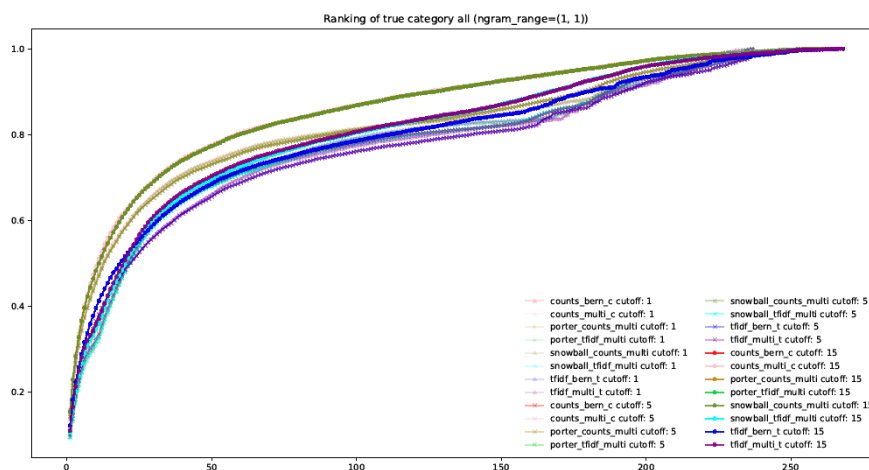


Source: Authors.

Figure 7 shows this ranking probability against all categories, where we see that overall, the chances of suggesting the correct category within the top 10 candidates is about 50 percent. The overall best model is the simple model of using word counts and a multinomial Naïve Bayes model, followed by a model using the Snowball stemmer together with counts of words in a multinomial Naïve Bayes model. In general, as can be seen by comparing Figure 6 and Figure 7, different categories have different models that perform best, so it would be possible to select the best available model for each category separately based on empirical performance.

Figure 8. Fraction of Correct Categories Against Ranking Across All Categories

(Note: With the best overall model, the chances of suggesting the correct category within the top 10 candidates is about 50 percent)



Source: Authors.

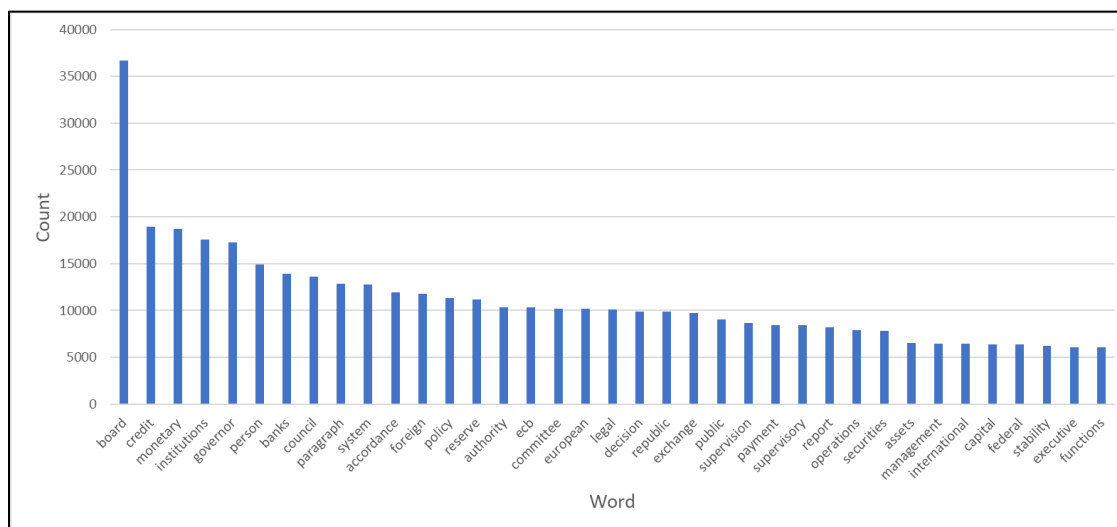
Figure 8 shows the ranking of the correct category against the iteration of data, with results using more training data shown with opaque markers. Across models the efficiency of predicting the correct category goes up with iteration. This suggests that while further dramatic predicting capabilities may require much more data, most gains to be had for this data are already available through the data that has already been categorized.

Ultimately, the predicting capabilities of the coding app would allow a form of “predicting the law.” With its current efficiency of predicting the relevant (273) CBLD search categories for specific articles and text within those articles, the coding app’s algorithm will be able to facilitate efficient coding of future legislation. More importantly, the IMF CBLD Team will be able to use the algorithm to predict which components (based on the CBLD search categories) are most frequently occurring, and in which groups of countries (using the CBLD’s selection groups based on region, income level, and exchange rate regime), as well as which components are increasing in frequency, as well as the connection between those components. The predictive value therefore lies in the algorithm identifying patterns between legal text and the CBLD search categories, and connections between those search categories in legal texts, beyond what central bank lawyers could identify outright based on human text analysis.

C. Word Frequency Comparison

The option of word frequency comparison provided additional interesting insights into the themes emerging from the laws included in the CBLD. To be able to analyze the frequency of words occurring in the entire CBLD dataset, every word in all CBLD 2020/2021 documents were tokenized (i.e., all laws that were coded using the CBLD coding app), amounting to a total number of words of 6,746,772. After filtering out of numbers, special characters, and stop words, the total number of words was reduced to 2,763,901. Out of these, the highest-ranking words were examined further—see Figure 9 below. With commonly used words such as “bank,” “central,” “legal,” “act,” or “financial,” it is not a surprise that the word “board” was most often occurring, both in its reference to central banks as well as commercial banks.

Figure 9. Most Counted Tokenized Words

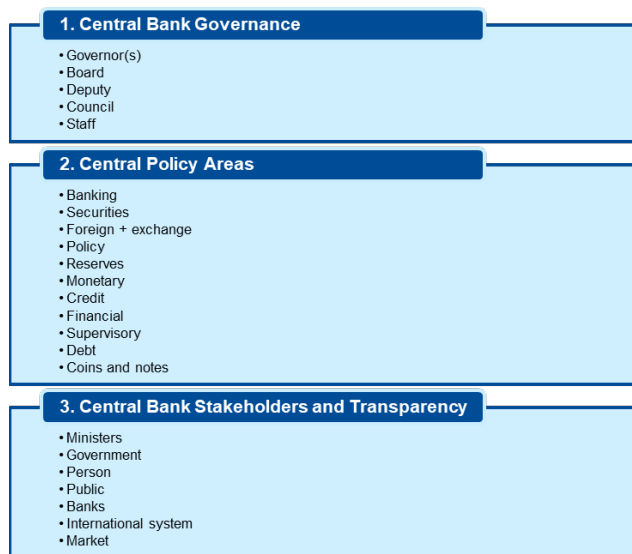


Source: Authors.

Further analysis of tokenized words can be done using word clouds. In the context of this paper, a “token” is defined as any given word in the law documents and tokenization is “the process of breaking down sentences by isolating tokens.” N-grams are “an ordered sequence of either characters or words of length N” using sequences of words (Bengfort, e.a., 2018). Both tokenization and the construction of N-grams are programmed in the R programming language. Frequently occurring stop words (such as “and,” “but,” and “the”) have been omitted from the analysis. The same holds for other frequently occurring words that are not stop words, but that likely not add much value to the analysis. This includes words such as “article,” “section,” “national,” “appointed,” “conditions,” “issued,” “pursuant,” “amount,” and “provisions.”

The main themes emerging from the word clouds relate to central bank governance, policy, and stakeholders and transparency. Based on the various N-grams, excluding frequently occurring stop words, a tentative, careful first overview of groupings (which merits further research) can be identified in terms of: (i) governance (governor, governors, deputy, council, staff); (ii) policy areas (banking, securities, foreign + exchange, policy, reserves, monetary, credit, financial, supervisory, debt, coins and notes); and (iii) stakeholders and transparency (ministers, government, person, public, banks, international system, market). See Figure 10 below. Central bank legislation clearly assigns significant text to issues relating to decision-making and the internal organization of the central bank, as well as to the central bank’s policy areas. Interestingly, stakeholder management, and, accordingly transparency and accountability, highlight that central banks do not operate in a vacuum, but legally are made aware of the responsibilities they carry vis-à-vis the larger financial sector community or society as a whole.

Figure 10. Themes of Frequently Occurring Tokenized Words in the CBLD



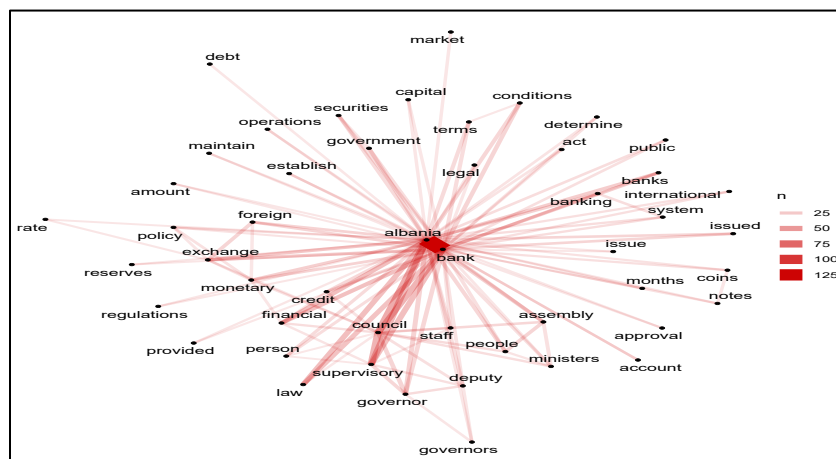
Source: Authors.

As an example, the figures below show common bigrams for Albania (Figure 11, based on the Bank of Albania Law) and Italy (Figure 12, based on Banca d’Italia Act). Here too, the bigrams are based on the limited tokenized word set (i.e., where stop words and errors have been filtered out). Both countries were randomly selected as examples for this paper after running visual representations on a larger number of countries.

For Albania, the frequency of connections between “Albania” and “bank” are clear and are similar to most central bank laws. Here too, groupings can be identified as noted above in terms of: (i) central bank governance; (ii) policy areas; and (iii) stakeholders and transparency. See Figure 11 below. The connections between “Albania” and “bank”, and “council” and “supervisory” seem to be strongest. This could imply a simple connection stemming from references to “Bank of Albania” and “Supervisory Council,” which is the central bank’s body in charge of oversight over management. However, any assumptions would require further analysis—the word network graphs should be used as such: to identify patterns that could be explored in more detail.

In the case of Italy, the European context plays a much stronger role. As with Albania, the bigram for the law of the Banca d’Italia (Figure 12) shows similar patterns in terms of: (i) governance (board, commission, directors, governor, deputy, senior, regents, meetings, head, office, management, bodies); (ii) policy areas (monetary, currency—though no reference to supervision); and (iii) stakeholders and transparency (shareholders, parliament, auditors, institutions). However, it is also clear that some of the law’s strongest connections appear to focus on Eurozone-specific references: European System of Central Banks (ESCB), governing + council, statute, treaties, euro, and possibly commission). The fact that the Banca d’Italia is one of the few central banks with shareholders is clearly visualized through the very strong connection between shareholders, meeting, directors, board, proposal, and governing and council.

Figure 11. Word Network Graph for Common Bigrams (Albania)

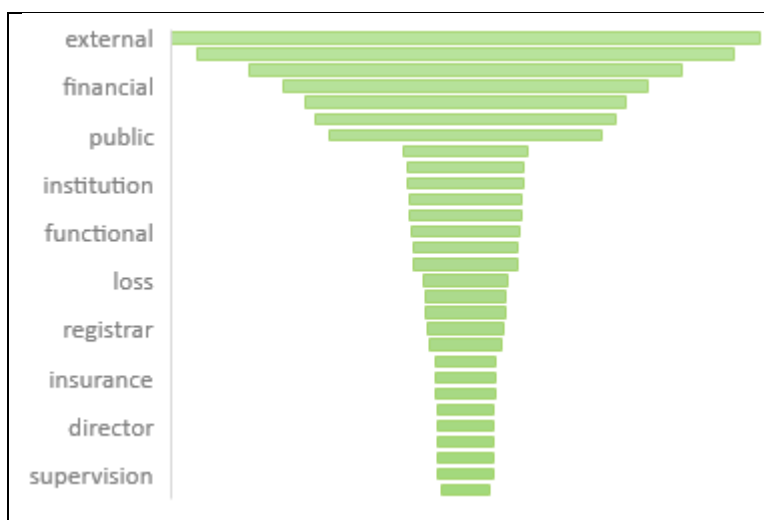


Source: Authors.

Another example of applying n-grams is in the context of specific topics of relevance to the central banking community. In addition to comparing laws from different countries, using an n-gram approach to search for, for instance, the combinations of words with the roots “independ” and “autonom” could provide helpful insights for policymakers and authorities as well. In this case, words that relate to the concept of central bank independence (or autonomy) are captured in the bigrams listed below in Figure 15 and Figure 16. The added value of a bigram approach in this case (over, for instance, a simple word search), is that the algorithm analyzes the frequency of connection of those words with independence/autonomy, even if these words are placed elsewhere in a sentence or in the article text itself.

Figure 14 gives a visual impression of the most frequently occurring bigrams in all legislation in the CBLD for concepts related to central bank independence. The correlation between independence and words such as “act,” “external,” “functions,” “financial,” “audit,” and “entities” is clearly visible—which could allow for interpretation (or at the least a starting point for more detailed research) on how central bank independence ties into aspects such as the financial position of the central bank, (internal or external) audit, and the relation of the central bank vis-à-vis other entities. References to “assessment,” “legal,” and so on could provide the starting point for further legal analysis.

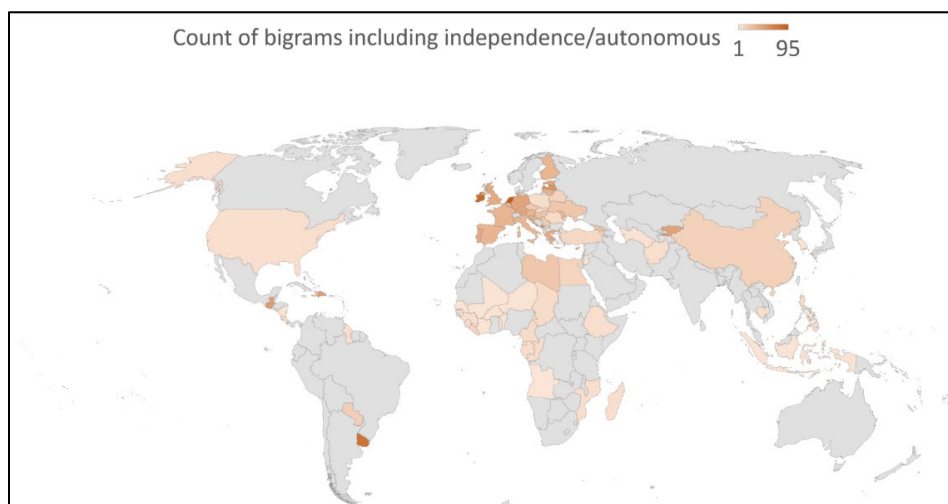
Figure 14. Proportions of Bigram Combinations on “Independ” and “Autonom” Across Central Bank Legislation



Source: Authors.

Figure 15 gives a visual impression of countries where the most frequently occurring independence/autonomy bigrams in all CBLD legislation occur. Central bank legislation in the European region, the US, China, and a significant number of countries in Africa contain the highest numbers of bigrams relating to any correlation between independence/autonomy and other words.

Figure 15. Coverage of Bigram Combinations on “Independ” and “Autonom” Across Central Bank Legislation



Source: Authors.

Counts: From 1–95 bigrams including independence/autonomy (including root words). Shows countries where all the included law documents have one or more bigram combinations.

The field of Natural Language Processing (NLP) is rapidly evolving and will offer even more opportunities for legal text analysis. Some aspects of NLP have been explored in this research, including text extraction and classification, implementation of AI/ML in the prediction of law text categorization in accordance with the CBLD's categories, and text mining to inspect correlations among different laws. Development of Large Language Models (LLMs) was more recently empowered by the OpenAI's Chatbot (Zhou et al., 2023). In the context of the current research, OpenAI's Chatbot was asked the following:

Summarize similarities and differences between the Bank of England Act and the Reserve Bank of India Act

And the response was:



The Bank of England Act and the Reserve Bank of India Act are the central banking laws of England and India respectively.

Similarities:

1. Both laws establish the central bank as the primary monetary authority in their respective countries.
2. Both laws give the central bank the responsibility to regulate the money supply and maintain monetary stability.
3. Both laws give the central bank the authority to issue currency and act as a lender of last resort.

Differences:

1. The Bank of England is owned by the UK government, while the Reserve Bank of India is owned by the Indian government with partial shareholdings by private banks.
2. The Bank of England has a dual mandate of monetary stability and support for the UK's economic policy, while the Reserve Bank of India has a singular mandate of price stability.
3. The Bank of England is a part of the European System of Central Banks and operates within the framework of the European Central Bank, while the Reserve Bank of India operates independently.

In summary, while the Bank of England Act and the Reserve Bank of India Act have some similarities in terms of central banking functions, they also have distinct differences in ownership, mandates, and relationships with other institutions.

The impact of this development in the field on NLP will assist with further analysis of central bank legislation. In addition to the ML options explored in this paper, the IMF CBLD Team will be able to use OpenAI to supplement findings and draw on other data sources (including academic literature and non-central bank legislation). Of course, as with all AI, potential ethical aspects such as social stereotyping and unfair discrimination, false or misleading information, outdated information, and the propriety nature of OpenAI's ChatGPT, will need to be taken into account as well (Zhou e.a., 2023, and Boomarito and Katz, 2022).

D. Sentiment Analysis of “Independence”

Another detailed example of applying AI/ML to the CBLD dataset relates to sentiment analysis. Building on the abovementioned example of central bank independence/autonomy, sentiment analysis is another AI/ML approach that could help analyze how these concepts are reflected in central bank legislation. The IMF CBLD team examined all currently coded references in CBLD data (through pulling data on CBLD categories 2.13 through 2.17—relating to central bank independence and its various forms), and manually coded these as positive or negative towards central bank independence (a positive connotation is assumed if the respective reference is intended to strengthen central bank independence; a negative connotation is assumed if the respective reference limits central bank independence). This list of positive and negative connotations¹⁰ was subsequently used to train an algorithm to identify words associated with positive or negative influences on central bank independence. Subsequently, the algorithm analyzed the entire CBLD dataset (i.e., not only the sections from the CBLD that relate to central bank independence) to assess whether an entire law (which could be a central bank law, or any other law included in the CBLD) was more positively or negatively inclined towards the concept of central bank independence. Initial studies show that a Naïve Bayes approach will not be able to capture the intricacies of whether a sentence leans towards independence or not, as the simple algorithm cannot account for connections and meanings between words. However, with much more complex models such as the aforementioned ChatGPT, further automatic assignment of sentiment and meaningful analysis of the contents will likely be possible.

User Statistics

Lastly, user statistics provide helpful insights into the demand for CBLD data. Given that users need to register before being able to access the CBLD, the IMF can keep track of the number of queries, the categories and countries that are being searched for, and the domain name of the users. These insights are helpful to determine overall demand for CBLD data, but also to identify rudimentary patterns in CBLD searches.

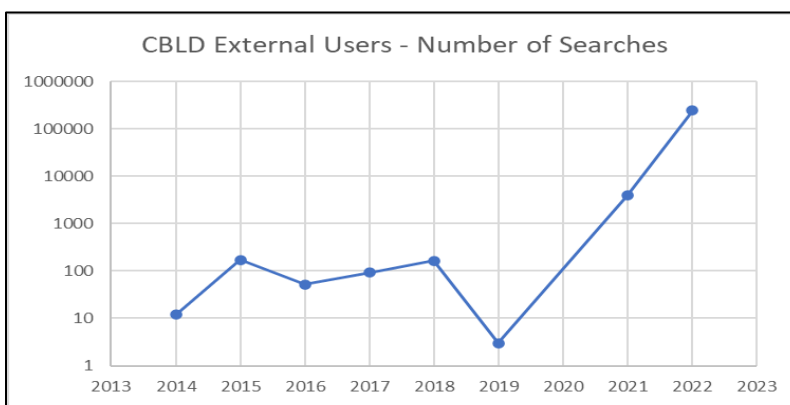
The number of users of the CBLD has increased tremendously. Since the opening of the CBLD to the public in February 2019, as Figure 16, Figure 17, and Figure 18 demonstrate, the number of individual queries run in the CBLD has grown from 323 on average per day in 2021 (between September 7 and December 7), to 5,441 on average per day.¹¹ Leaving out the days in 2022 where queries ranged above 10,000 per day—which were likely due to IMF researchers working on intensive data analysis—the average number of daily queries would still be 932 in 2022.

¹⁰ The total list consisted of several hundred sentences pre-selected on the basis of the relevant CBLD search categories. And an additional 100 sentences (not based on the CBLD search categories) were similarly coded; another 100 “fuzzy” (i.e., not clearly related to a negative or positive connotation without further analysis) were also added.

¹¹ Note that each data point represents single day access. This includes searches, browsing, and queries, i.e., the sum of all actions for that specific day (note that one user might perform multiple actions).

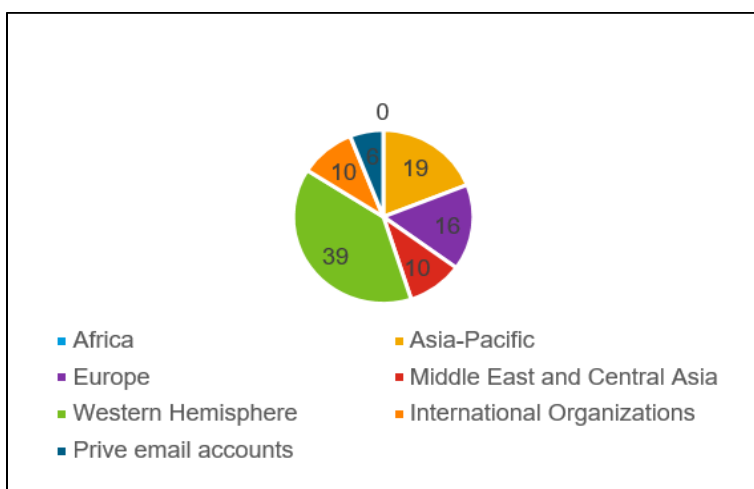
There could be various reasons for the increase in CBLD users. For instance: (i) central banks worldwide are facing more pressure on their policies and operations and are looking for data on how other central bank legal frameworks provide details for a host of issues, including on newer topics such as fintech and climate change; and (ii) as explained in more detail above, the CBLD’s search categories were expanded, and the user interface was significantly improved to allow for easier data searches across topics and countries. Figure 17 shows that many of these user queries came from countries (which includes public institutions such as central banks and ministries of finance, as well as research institutions) in the Western hemisphere, as well as the Middle East, Europe, and Asia-Pacific. Additionally, many CBLD queries came from users using private email accounts (e.g., Gmail, Yahoo)—it is difficult to identify if these users represent any specific organization, though experience has shown that staff from central banks especially in the larger African region often use private email accounts as well. Selected commercial organizations have been using the CBLD as well, though this is a relatively small number compared to the users from various public institutions.

Figure 16. CBLD External Users: Number of Searches

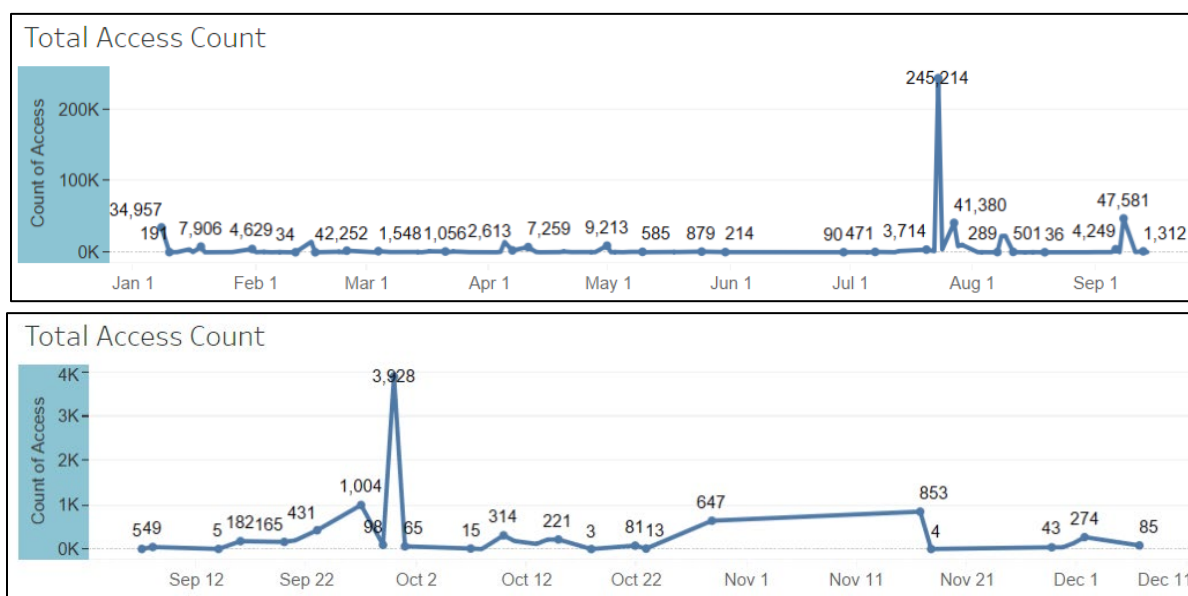


Source: IMF.

Figure 17. CBLD External Users: Regional Coverage (in percent)



Source: IMF (CBLD, 2022–2023 data).

Figure 18. CBLD Daily User Queries (2021, 2022)

Source: IMF.

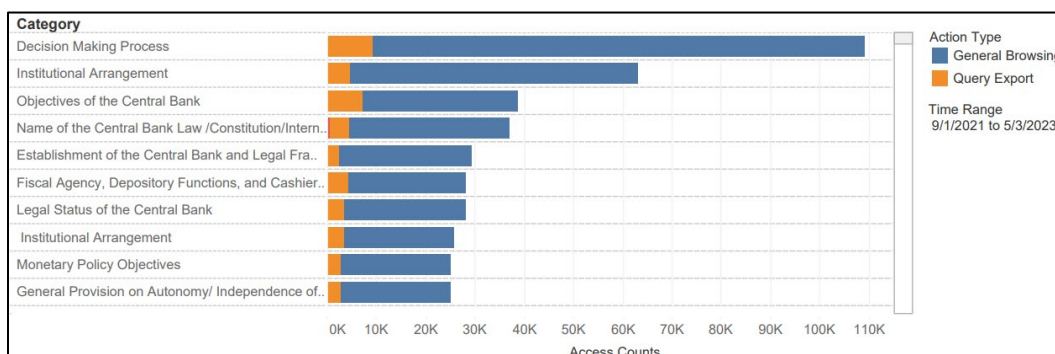
Insights into user statistics can help to refine the CBLD. Identifying search patterns in user queries (e.g., by category/topic) could help to identify CBLD search categories that are less relevant, and search categories that might be missing. Details on users can also help identify key users, who might even be of help in the future if the IMF would consider crowdsourcing of the coding work to selected groups of users (see above).

Most CBLD users search for governance, policy, and accountability-related topics. In line with our earlier findings based on n-grams, CBLD user data¹² shows that most users run queries with the search categories that relate to governance (e.g., decision-making process, institutional arrangement), policy (objectives and functions), and accountability (accountability framework and reporting obligations), in addition to specific legal categories (e.g., legal status and name of the central bank law). See Figure 19.

Additionally, in terms of countries, most CBLD users search for Hungary, CAMU, EU, US, and various other countries. As Figure 20 and Figure 21 show, the countries/monetary unions most searched for include Hungary (40,635 searches), followed by Czech Republic, Lithuania, Poland, Belgium, Slovenia, the Central African Monetary Union, and the European Monetary Union. Similarly, Argentina, the United States, and India score high as well. Countries that only have one search in the period of 2021–2023 are Seychelles, Panama, Suriname, Maldives, Morocco, UAE, Lao, Barbados, Venezuela, Antigua and Barbuda, Vietnam, Mozambique, Comoros, Eritrea, and Eswatini.

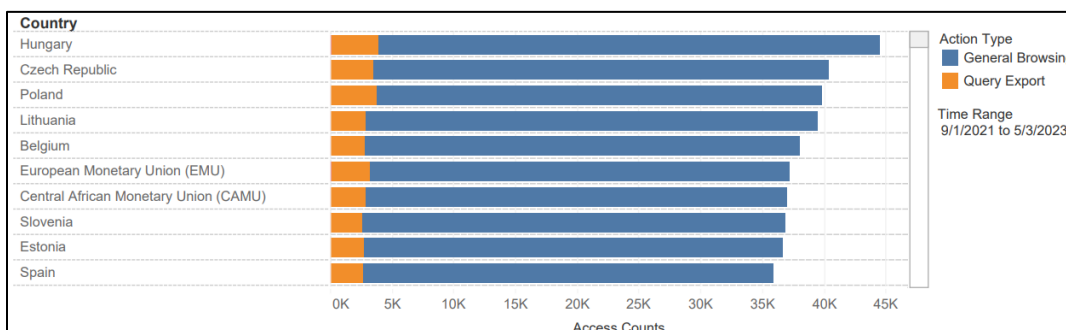
¹² CBLD user data from September 1, 2021–May 3, 2023.

Figure 19. CBLD Daily User Queries by Search Category (2021–2023)



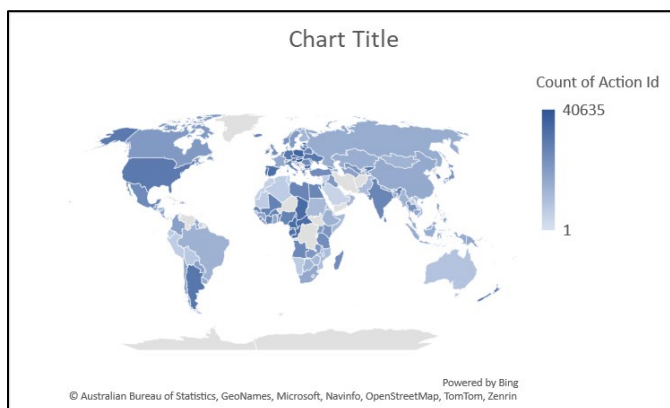
Source: IMF.

Figure 20. CBLD Daily User Queries by Country (2021–2023)



Source: IMF.

Figure 21. CBLD User Queries by Country (2021–2023)



Source: IMF.

Though technically not an AI/ML approach, user statistics’ analysis is helpful to use jointly with the various AI/ML approaches noted above. For instance, user demand for specific search topics or categories could serve as input for detailed n-gram analysis; and the order in which users look for data belonging to specific CBLD search categories could provide info on which search categories are more closely related, which could help to finetune to coding app’s predictions.

Conclusion

The IMF's Central Bank Legislation Database is a critical tool for central banks. It offers insights into central banking developments, in the first place in terms of legislation and regulation of course. These insights can be of help for national legislators and central banks themselves to identify possible patterns, topics, and modalities that feed into policy discussions or even legislative amendments.

However, central bank and other laws relating to the central bank are not always the most accessible documents for central banks and other policy makers. Legal documents are often phrased (regardless of the drafting language) using more complex terminology and syntax, and with a certain structure in mind to guide the users/recipients of the law to specific (im)possibilities. This makes it difficult for central banks and other policy makers to identify more general patterns or linkages between the various concepts and stipulations laid down in the respective law.

The various Artificial Intelligence and Machine Learning approaches explored by the IMF for central bank legislation, provide important tools for understanding, analyzing, and identifying patterns. The IMF's in-house developed coding app was initially intended to serve as a tool for IMF central bank lawyers to analyze and subsequently incorporate central bank legislation into the CBLD, and to make this process more reliable and cost-efficient. However, the (increasing) predictive value of the coding app also allows identification of patterns and interlinkages throughout central bank legislation. This, in its own, provides insights into the large body of global central bank legislation that could be used to understand developments in central bank legislation that central banks and policy makers were not able to identify before.

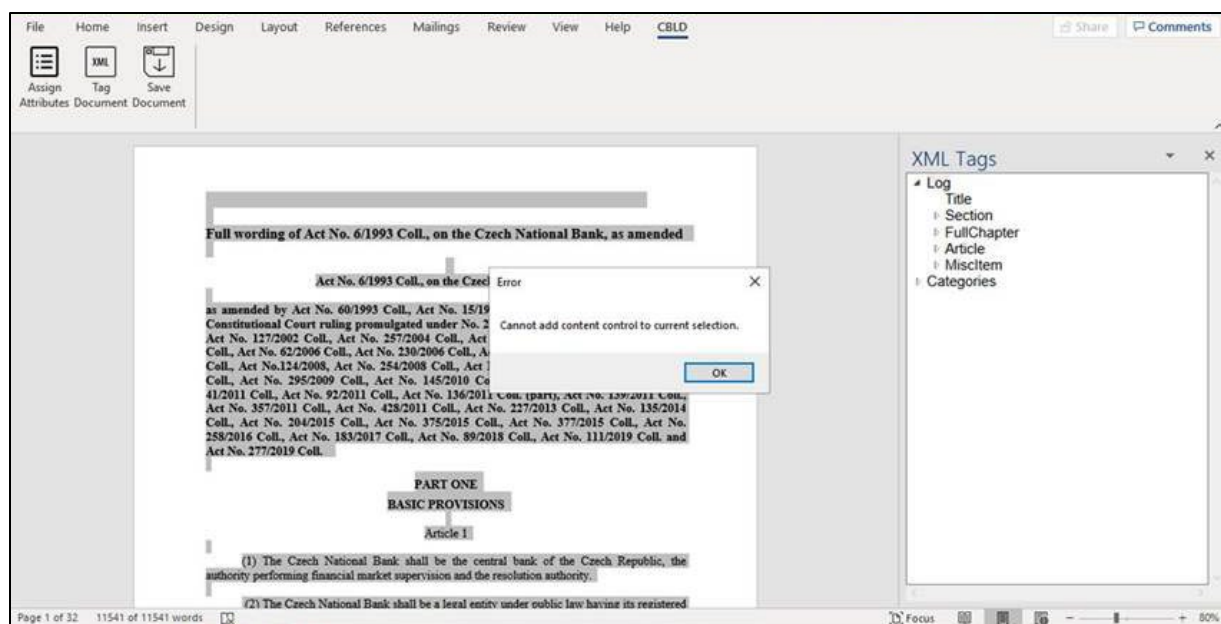
Through these various AI/ML approaches, a careful, first approach can be explored to identify key themes in central bank legislation. By means of tokenizing individual words used in the entire CBLD dataset, the main themes of: (i) central bank governance; (ii) central bank policy; and (iii) central bank stakeholders and transparency can be identified. New topics within those themes (for instance, fintech or climate change as part of central bank policy) are not clearly emerging. This is understandable as amendments to legislation take time, and legislation therefore is, by definition, behind the curve.

The CBLD AI/ML approach is only a first step. Going forward, the CBLD's annual update by the IMF will provide the opportunity to further enhance the CBLD coding app, apply other AI/ML approaches, and use the outcomes to identify additional and more detailed patterns in central bank legislation. For instance, jurisdictions that had a colonial past or particularly strong trading ties might have very strong legal similarities because the legal frameworks were directly copied at one point in time. The CBLD AI/ML approach could be used to identify those patterns, deviations over time, and highlight subsequent legal developments—which could be of help for countries looking to amend their central bank legislation and either avoiding legal pitfalls or further enhancing their legal framework. It would be helpful for the IMF to work together with selected academic partners and country authorities to get a better understanding of the AI/ML approaches' outcomes, and finetune these further to get input that can help countries understand and improve their own central bank legislation. Ultimately, this will also benefit access to the CBLD's data by external users, through more enhanced search options.

Annex I. CBLD Coding App

The process of coding central bank legislation prior to the 2020/2021 cycle was somewhat convoluted. The process required IMF central bank lawyers to use paper and pen to mark codes on the legislation submitted by the country authorities, and, subsequently, an IMF research assistant to manually generate output using a tool hosted in Microsoft Word, in a specific domain language (XML)—see Figure 1 below. This could only be done one piece of legislation at a time, making the process time-consuming and prone to possible human error.

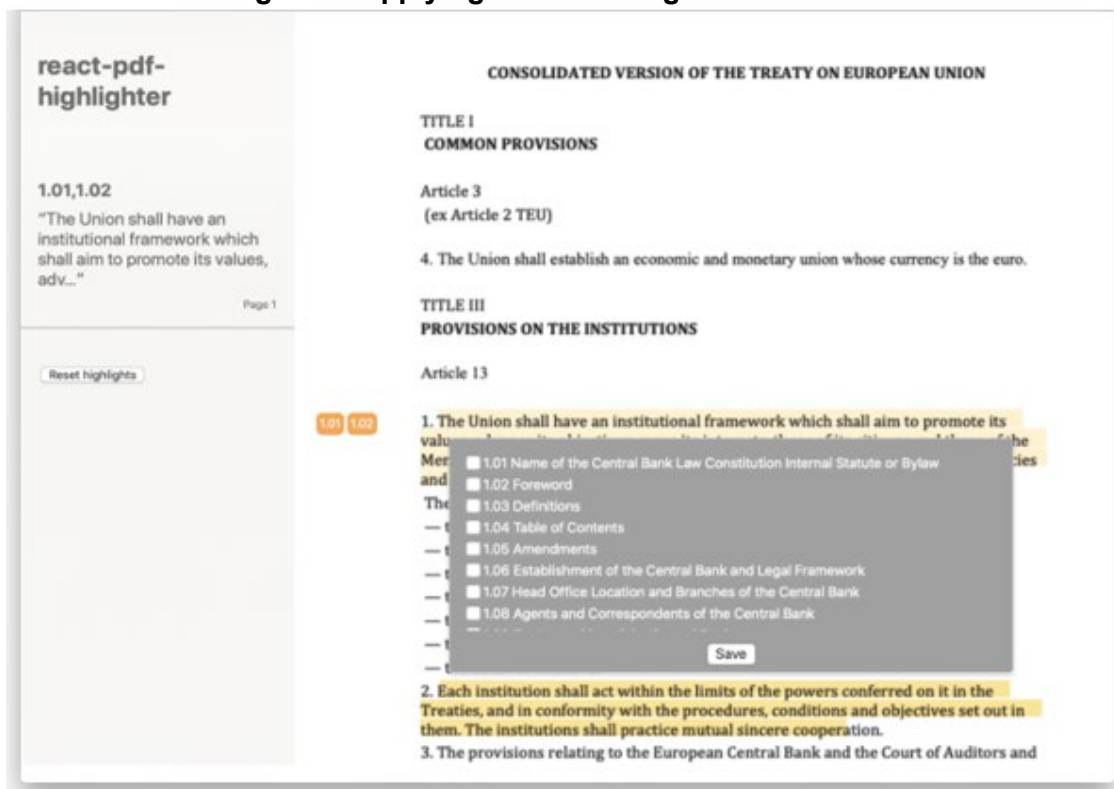
Figure 1. Original Coding Application



Source: Authors.

The newly developed AI coding app allowed for easier, faster, and more systematic coding of central bank legislation. As described in section 3A, the newly developed coding app allowed the IMF CBLD Team to electronically access all uploaded legislation (whether provided by country authorities or collected by the team from central bank websites and other publicly accessible sources), and subsequently code the any text within the specific document against one or more of the 273 CBLD search categories. Multiple team members could access documents at the same time (and even work in the same document), with coding being saved automatically, with suggestions for possible categories provided by the coding app's AI.

Figure 2. Applying Search Categories to Selected Text



Source: Authors.

After selecting a specific document, the IMF's coding staff were presented with a detailed view of the specific document. As Figure 2 above shows, the coding staff could then freely navigate the contents of the selected document (using mouse and keyboard) and highlight any portion of text by clicking the left mouse button and dragging it until the end of the desired text. This could be the entire article or section, or specific sentences or words within an article. Once highlighted, the coding staff was presented with the list of codes, and able to select multiple codes using mouse. Each highlight action was routed via the server, which would then run the ML algorithm to calculate the best-matched codes to that specific text. A list of scores would then be sent back to the interface, and to the coding staff using the app. In the earlier versions of the coding app, codes were reordered according to the score given by the algorithm. However, user experiences indicated that the changing order of the app's suggestions were confusing, and the mechanism was reviewed. In later versions, scores were color-coded. Scores were normalized from zero to one, with items in orange added in proportion to the score (i.e., items with a zero score were listed in white; the item with the highest score in orange).

The integration of legislation within the coding app was successful, but not without hurdles. Automated scripts were used to convert laws received in the collection phase. Subsequently, new laws received via e-mail were added to the application via the same process. However, the team detected that the conversion process was unable to handle the formatting of a small set of documents. While team members could see their coding, the underlying code was corrupted by invisible spaces and line breaks. These defects affected the quality of CBLD search results (for instance, words were inexplicably broken off, or rendered unreadable through the inclusion of additional characters). The team had to manually reformat those text sections and reload those into the application. Accordingly, with every annual update of the CBLD it is likely that the coding app, including its user interface and the ML algorithm, as well as the underlying repository, will be refined further.

Annex II. Overview of CBLD Search Categories

The CBLD includes 273 search categories with the following main topics, allowing users to perform detailed searches within those main topics:

1. Background information
2. Legal Status, Objectives, Functions, Powers
3. Organization & Administration
4. Financial Provisions
5. Monetary Unit, Banknotes, Coins
6. Monetary Policy and Operations
7. International Reserves
8. Foreign Exchange Policy and Operations
9. Relations between the Central Bank and the State
10. Macroprudential Policies
11. Financial Regulation and Supervision
12. Lender of Last Resort
13. Financial Integrity
14. Consumer Protection
15. Resolution & Crisis Management
16. Payment Systems and FMIs
17. Accounts, Financial Statements, Audits, Publication
18. Miscellaneous

References

- Bengfort, B., R. Bilbro, T., 2018, *Applied text analysis with python: Enabling language-aware data products with machine learning*. Sebastopol, CA: O'Reilly Media, Inc.
- Bommarito II, M., D.M. Katz, 2022, *GPT Takes the Bar Exam*. arXiv preprint arXiv:2212.14402. Ithaca, NY: Cornell University.
- FSB, 2017, *Artificial intelligence and machine learning in financial services – Market development and financial stability implications*, FSB paper, November 1, 2017. Basel: Financial Stability Board.
- IMF, 2020, *Central Bank Transparency Code*. IMF Board Paper. Washington, D.C.: International Monetary Fund.
- Khan, A., 2017, *Central Bank Legislation in the Aftermath of the Global Financial Crisis*. IMF Working Paper 17/101. Washington D.C.: International Monetary Fund.
- Silge, J., D. Robinson, 2017, *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media, Inc.
- Zhou, T. Y., Y. Huang, C. Chen, e.a., 2023, *Exploring AI Ethics of ChatGPT: A Diagnostic Analysis*. arXiv preprint arXiv:2301.12867. Ithaca, NY: Cornell University.



PUBLICATIONS

Predicting the Law: Artificial Intelligence Findings from the IMF's Central Bank Legislation Database
Working Paper No. WP/2023/241