Reversa.ai Legal Publication Data Analysis

Andrew Darwin, Tony Lomeli, Ethan Zheng, J.T. Reilly

Project Overview

Goal:

- Analyze a structured dataset of legal publications extracted from Spanish official bulletins.
- Identify patterns across legal domains (ramas jurídicas) and industrial divisions (divisiones).
- Prepare a cleaner dataset to support clear, insightful visualizations and interpretable trends in legal data.

Tools:

- Pandas for data manipulation
- NumPy for numerical operations
- Seaborn/Matplotlib for visualization
- Scikit-learn for the machine learning model
- Google Colab + Google Drive for cloud execution and storage

Dataset Description

- Extracted from the Boletín Oficial del Estado (BOE) — the official gazette of Spain that publishes legal, administrative, and regulatory texts.
- Provided in CSV format as UCLABOE.csv, located in a shared Google Drive folder.

Column	Description
titulo	Full legal title of the document.
fecha_publicacion	Date the document was published. This column is further decomposed into: year, month, and day to allow for temporal analysis.
num_paginas	Number of pages in the document. Used as a proxy for document length or complexity.
divisiones[0-13]	Up to 14 columns indicating the economic/industry sectors affected by the document.
ramas_juridicas[0-14]	Up to 15 columns indicating the legal domains associated with the document.
rango	Type of norm

Data Cleaning & Preprocessing

Date Conversion:

- ‘Fecha_publicacion’ converted from string to datetime using ‘pd.to_datetime()’

Handling Missing Pages:

- ‘Num_paginas’ had approximately 10% missing values (NA or equal to 0), which were imputed using the median value for corresponding type of law category (rango)
    - Rango only category with no missingness, low variability in page numbers

Restructuring Categorical Data:

- The dataset originally had fields like ‘divisiones[0],’ ‘divisiones[1],’ ..., ‘divisiones[13]’ to include multiple labels for a single publication
- Aggregated fields into a single column

Dealing with Missing Data

Issue:

- Many entries had missing or unhelpful values in ‘divisiones’ and ‘ramas_juridicas’ (Would state ‘No identificado’ or it was empty).

Solution:

- Created a custom function that was used to quantify this issue.
- Extracted keywords from ‘titulo’(title of the publication) to infer probable divisions or legal branches.
- For example: if a title includes “laboral”, it might belong to the “Labor Law” category.
- This was a rule-based NLP approximation; it wasn’t the most perfect approach, but it adds structure.

Imputation results and relevance

Results

Field	Missing Before	Missing After	Reduction (%)
divisiones	1127 rows	950 rows	**15.7% ↓**
ramas_juridicas	1174 rows	982 rows	**16.4% ↓**

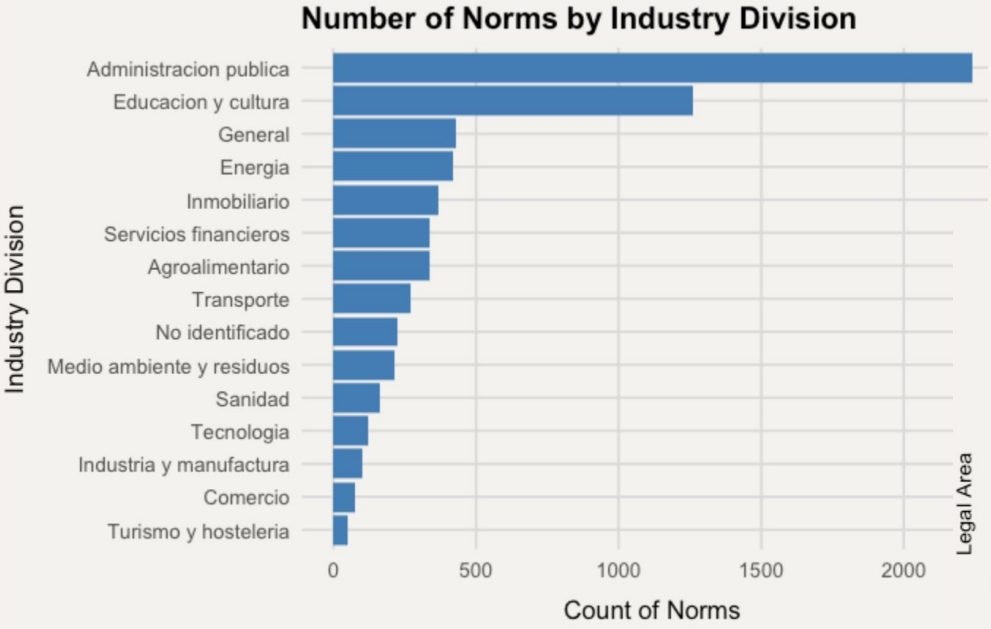
These reductions improved the **label coverage** of the dataset, enhancing the reliability of downstream analyses.

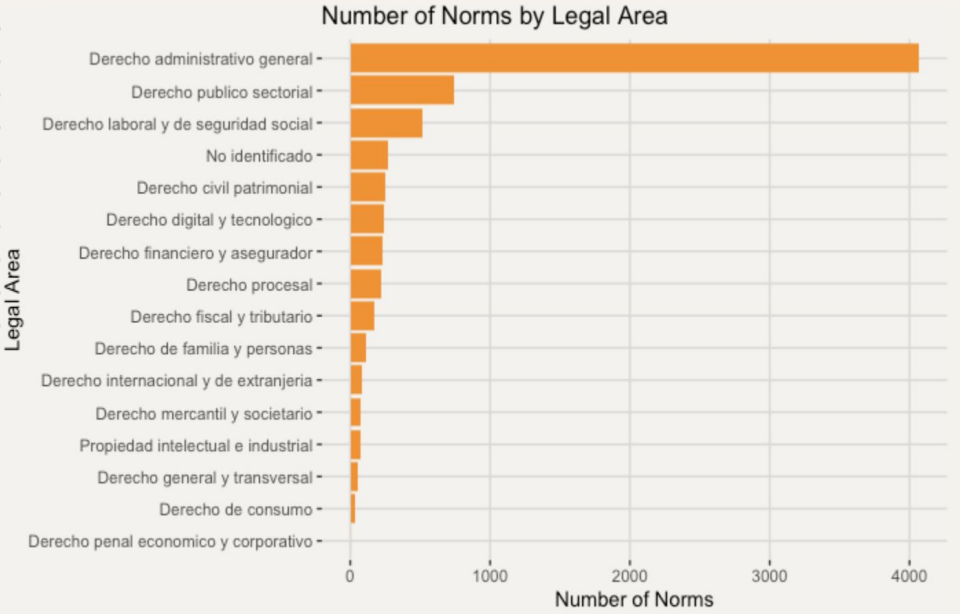
This also increases the **statistical power** and **interpretive depth** of our analysis.

- Remaining missing values were labeled as “No identificado” for uniformity with preexisting data

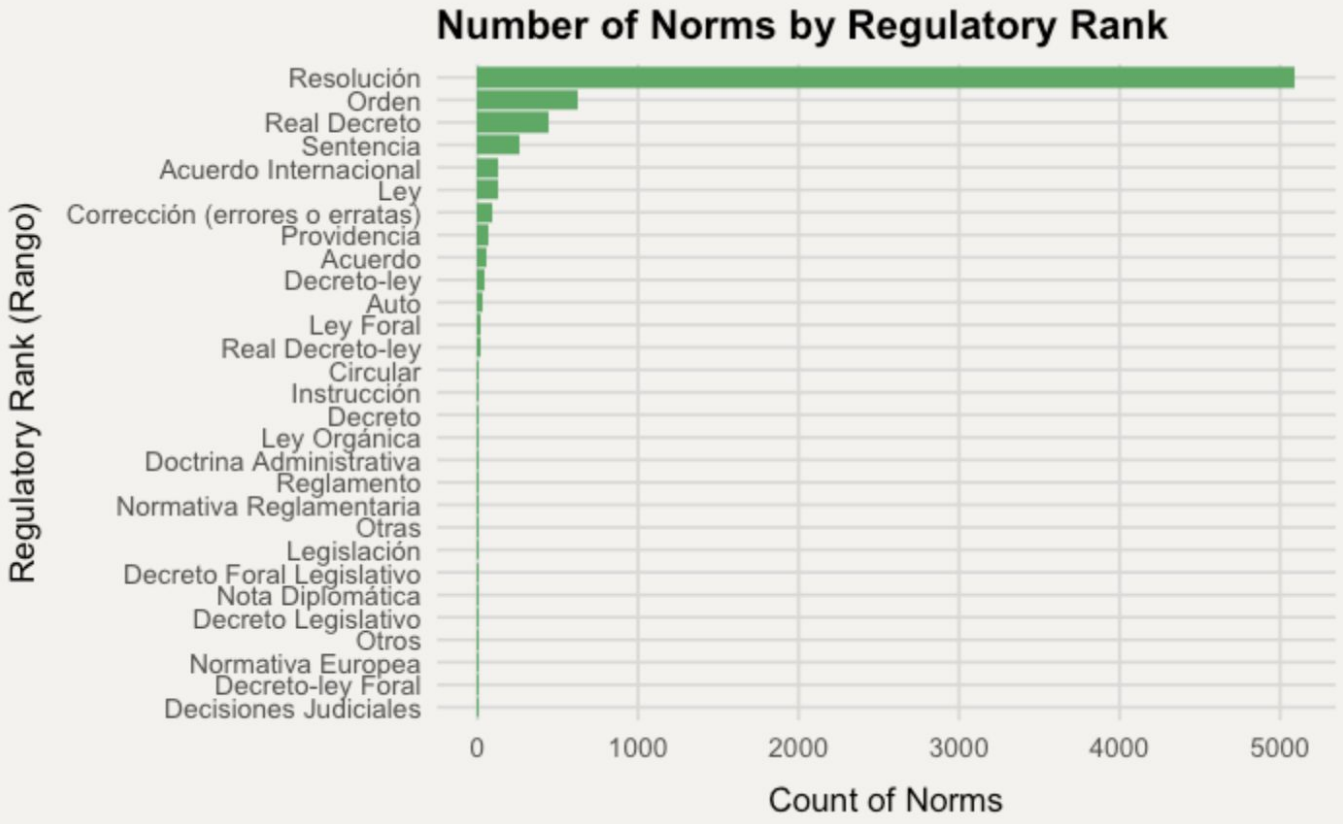
Visualizations

Norms by Industry, Legal Area, and Law Type

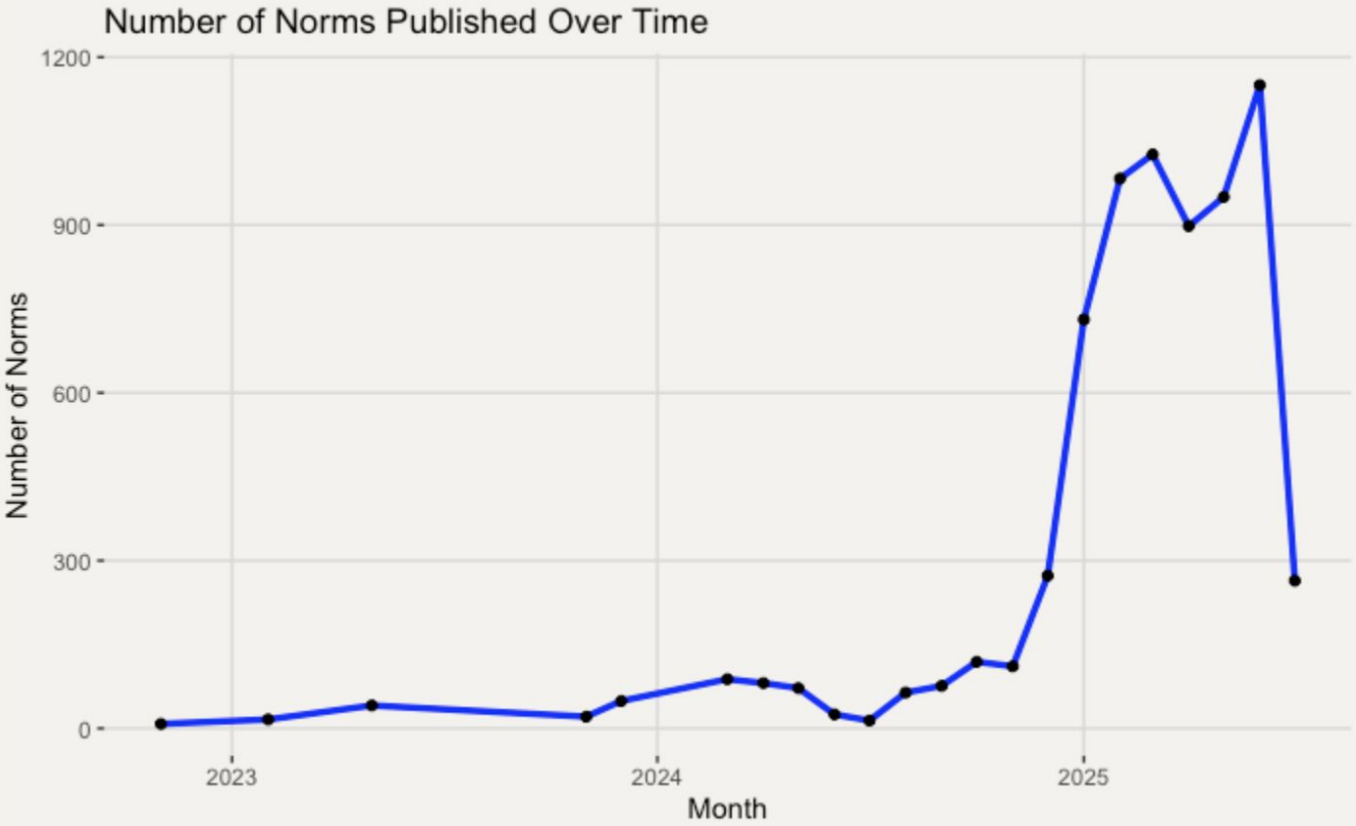




Norms by Industry, Legal Area, and Law Type



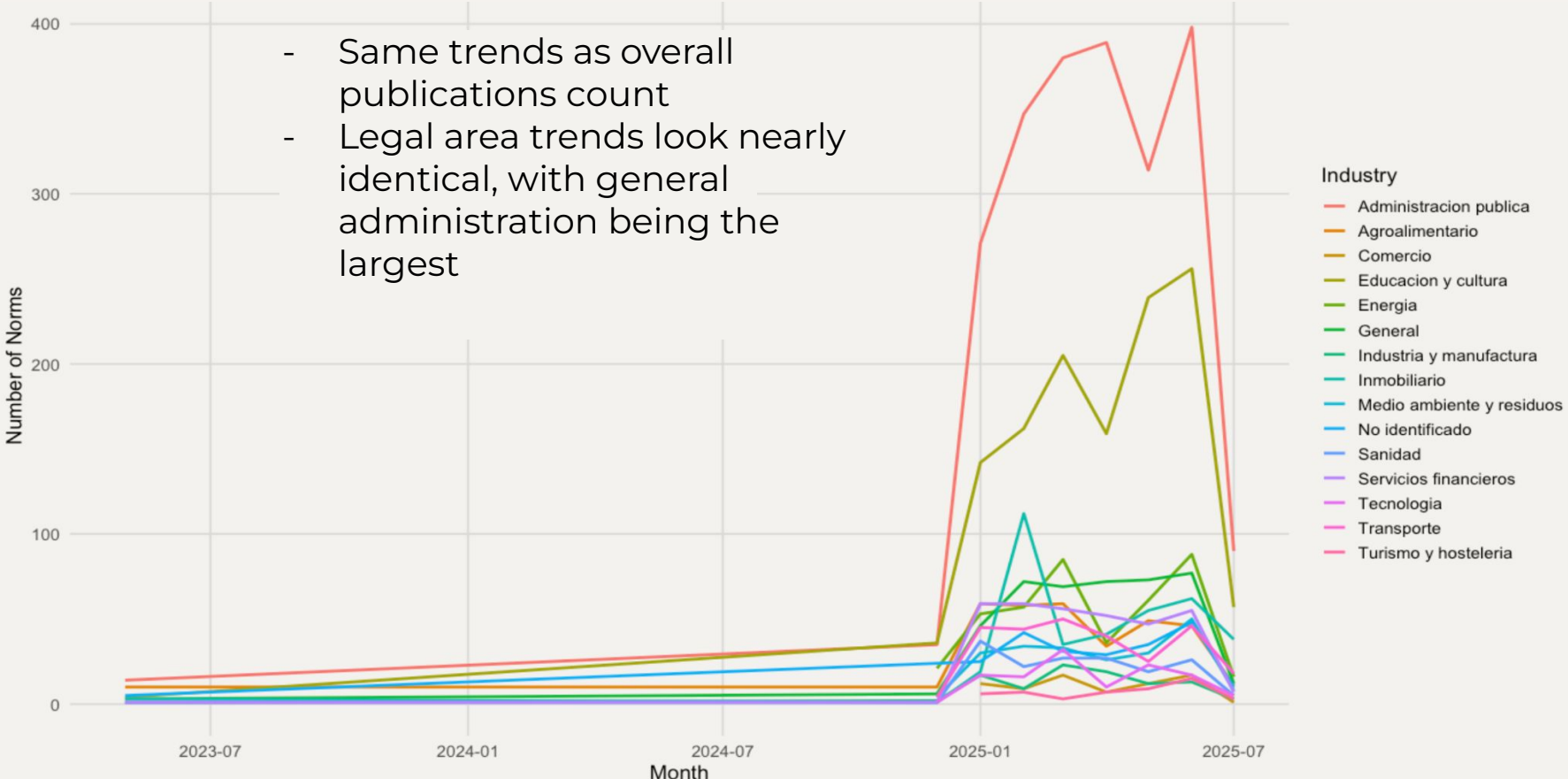
Publications over time



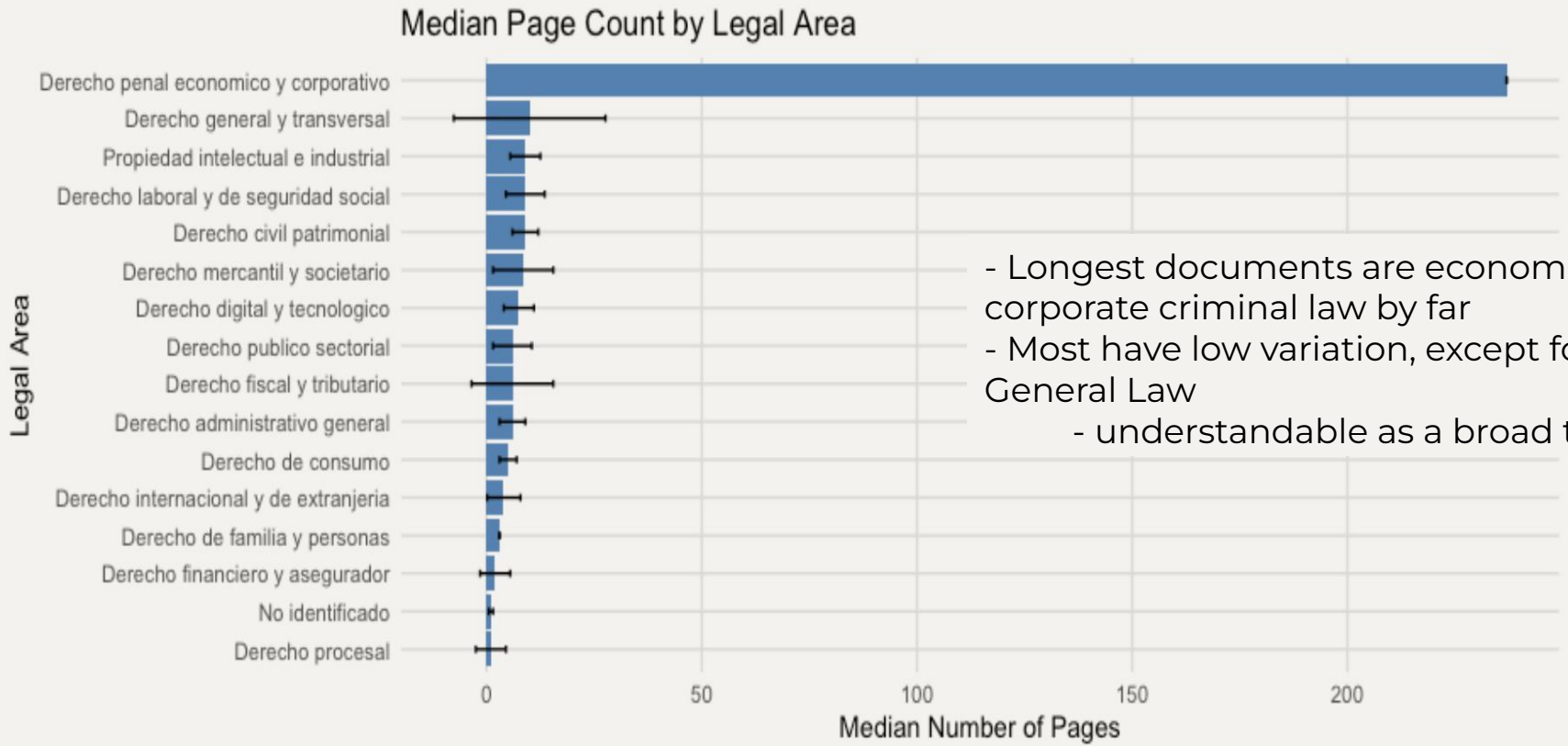
- Massive growth in number of publications in 2025
- More publications towards start/middle of years
- On downward trend

Publications over time by Industry

- Same trends as overall publications count
- Legal area trends look nearly identical, with general administration being the largest

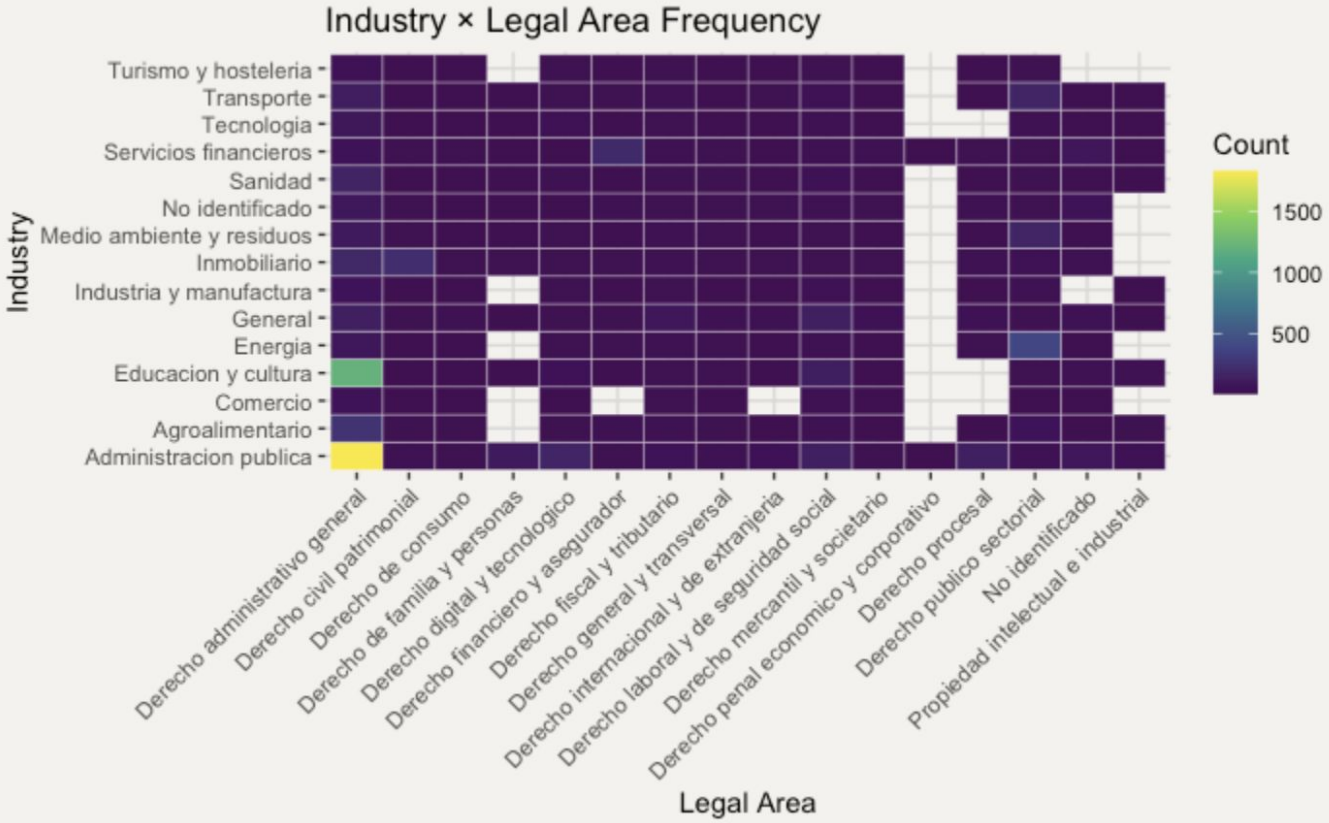


Page Count by Legal Area



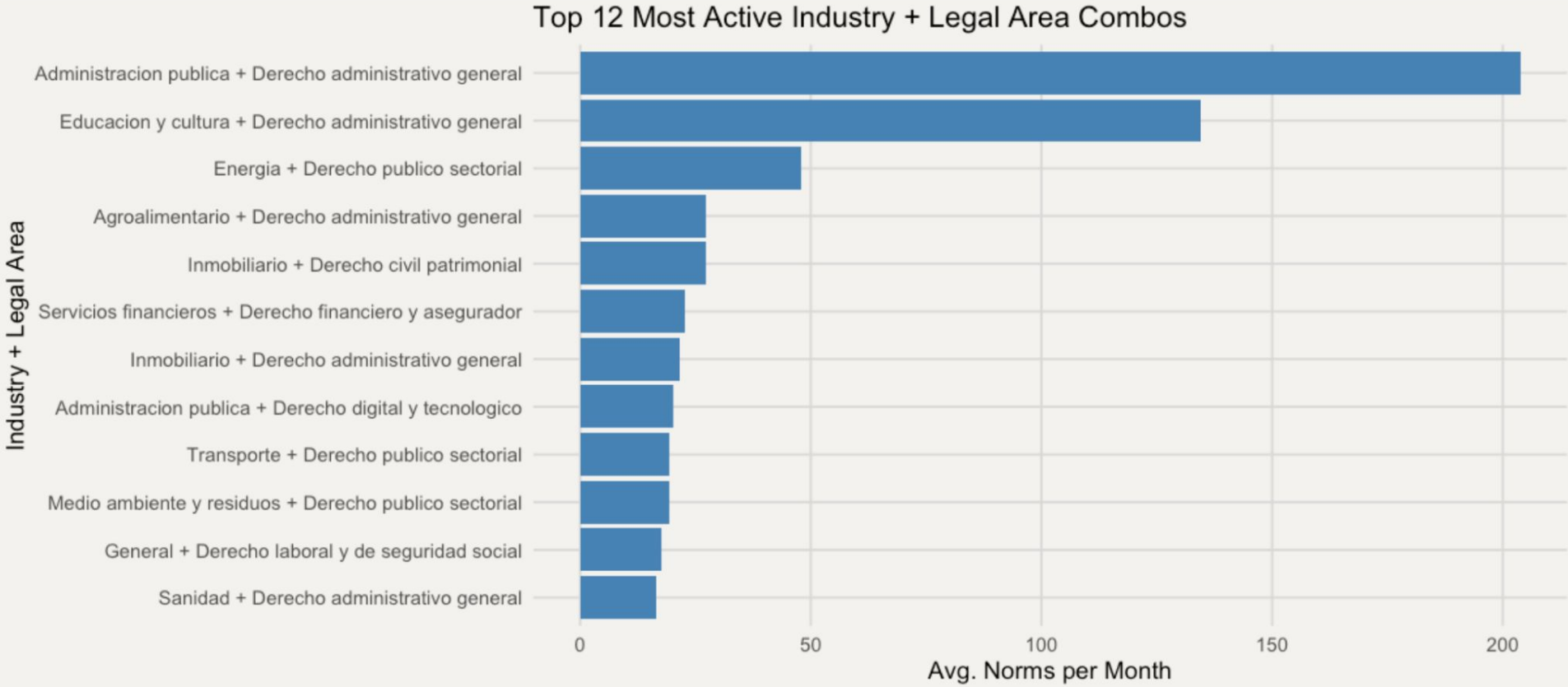
- Longest documents are economic & corporate criminal law by far
- Most have low variation, except for General Law
    - understandable as a broad topic

Legal Area Intersections



1. Public Administration + General Administrative Law
2. Education & Culture + General Administrative Law
3. Energy + Sectorial Public

Intersections Cont'd



Key Insights

- Public Administration, Education, and Energy industries dominate in volume and frequency
- General and Sectoral Public Law are the most frequent legal areas across industries
    - highlights the need for legal tech tools tailored to public regulation.
- Corporate and Economic Criminal Law has the largest median page count by a large margin
- Royal Decrees, Ministerial Orders, and Resolutions are most common types of norm
- In general most documents seem to be published in the first half of a given year

Actionable Analysis

- Prioritize Public Law NLP Models (Administrative & Sectoral Public Law)
- Create more in depth summaries for economic and corporate law
    - Laws are more nuanced and lengthy, increasing risk of misinterpretation or omission in a summary
- After completing analysis of one area, recommend related legal areas or industries with frequent overlap
    - Ie: Company reviewing Energy norms in Sectoral Public Law, recommend reviewing Environmental Law norms
- Agricultural, Energy, and Education & Culture are most active sectors behind the general classes
- Increase reporting & analysis frequency in first half of year to match higher publication volume
- Highlight the overwhelming volume and complexity of regulations companies face
    - Emphasizes necessity and efficiency of AI-powered summaries
    - Ie: Economic & Corporate law avg. 200+ pages per publication

Overall Challenges:

- **Text Noise:** Titles are inconsistent and may lack legal vocabulary and length.
- **Missing Values:** Not always recoverable by rule-based imputation.
- **Sparse Multilabel Structure:** Rows can be linked to multiple classes which can be hard to model without advanced techniques

Improvements:

- Integrate with a dashboard or API.
- Use datasets for laws and regulations outside of Spain.
- Imputation logic should account for polysemy and synonymy in legal documents for greater understanding of context

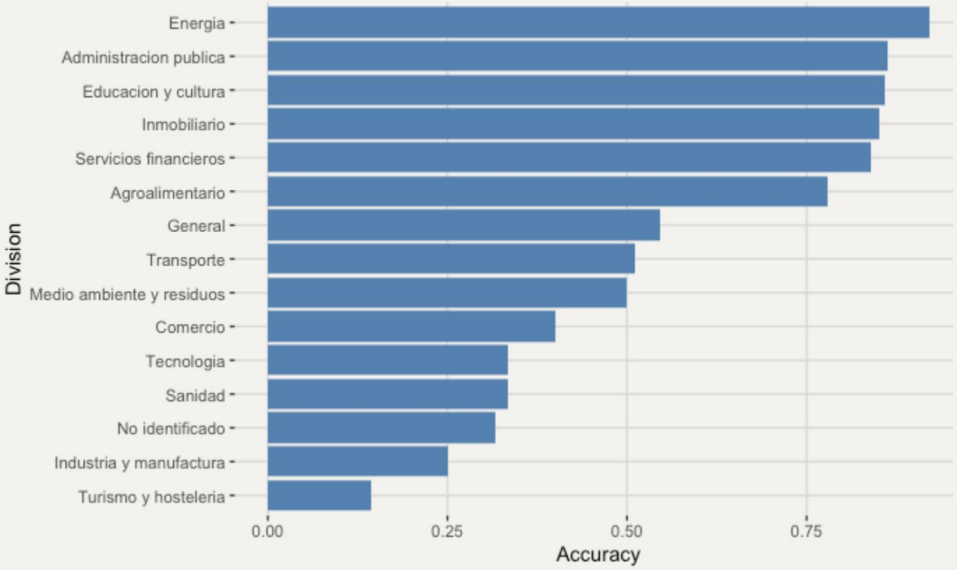
Modeling

Random Forest Results

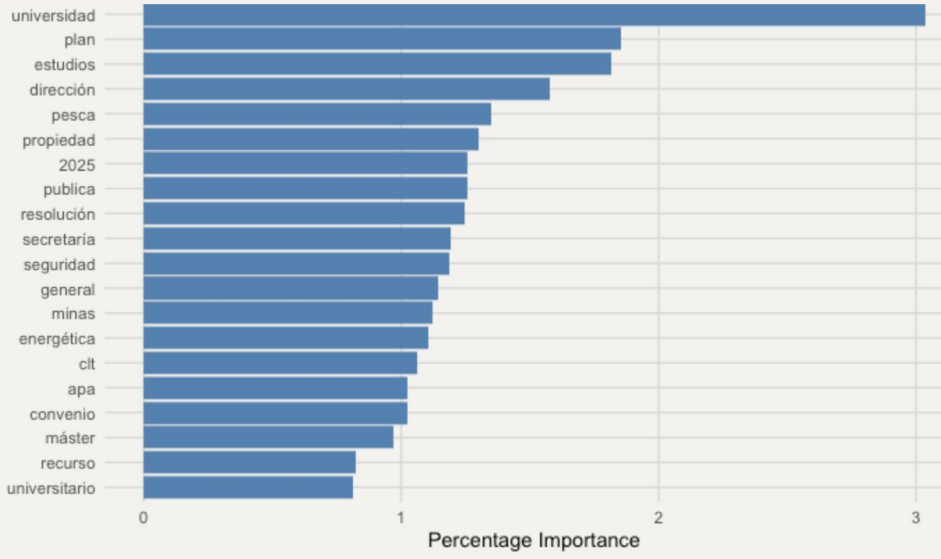
20%/80% Testing/Training Split

Overall Accuracy: 76.38%

Per Class Accuracy:



Most important words for the model fit:



Modeling Takeaways and Challenges:

- **Overall performance:**
  – ~76 % test-set accuracy on titles alone
  – Strong bias toward *Administración Pública* (majority class) inflates global score
- **Per-class insights:**
  – High accuracy (≈85 %) for “Energía” “Administración pública”
  – Poor performance (<30 %) on “Turismo y hostelería” and other underrepresented classes
- **Feature signals:**
  – Top words (“universidad,” “plan,” “estudios,” etc.) drive decisions, but often reflect generic topics
- **Data limitations:**
  – **Titles only**: very short and ambiguous for fine-grained division, constraints learning model

Future Improvements:

- Include abstract or summary column for each publication
- **Address class imbalance:**
  – Oversample minority classes or use class-weighted/loss-sensitive models
  – Stratified CV that leaves out entire publication periods to test temporal generalization
- **Utilize advanced embeddings:**
  – Fine-tune a Spanish BERT or use contextualized transformers

Thank You!