



Electoral crime: Using machine-learning algorithms to extract data from unstructured court documents

Luiza Kunz Aires and Hamza Syed

- Automated, machine-learning-based models have come a long way in extracting useful data from unstructured documents.
- Relative to manual extraction by experts, the model has an accuracy of over 70% for two-thirds of the questions.
- Even for complicated legal language, the model seems to understand the context and is able to classify the documents into different legal categories with more than 80% accuracy

Using the data extracted using our Machine-Learning pipeline, we analyse the effects of a guilty verdict in election-related cases. Although the estimates are not precise because of data limitations, the preliminary results, based on manual extraction, can be summarised as follows: a guilty verdict leads to:

- Lower likelihood of being a candidate in future elections
- Lower likelihood of being elected, conditional on being a candidate in future elections

POLICY BRIEF BRA-25112

FEBRUARY 2026

This project was
funded by the IGC

theigc.org

DIRECTED BY



FUNDED BY



Electoral crimes and political accountability in Brazil

Brazil's electoral process is administered by a specialised branch of the judiciary, the Electoral Justice (*Justiça Eleitoral*), which operates at the federal, state, and municipal levels. Electoral courts are responsible not only for organising elections, but also for judging cases related to electoral misconduct.

“Electoral crimes” refer to a broad set of offences defined in the Electoral Code and related legislation, including vote buying, illegal transportation of voters, irregular campaign financing, and electoral fraud. These offences can be punished with fines, suspension of political rights, disqualification from office, or imprisonment. Since 2012, the Clean Slate Law (*Lei da Ficha Limpa*) has imposed an eight-year period of ineligibility on candidates with certain convictions.

While there is substantial research on corruption and electoral accountability, there is limited evidence on two key questions:

1. How effectively electoral rules are enforced.
2. What happens to politicians after they are convicted of electoral crimes

This project addresses these gaps by constructing a novel dataset of electoral crime cases and linking it to administrative electoral records.

Although performance varies across variables (especially for names and more subjective fields), results indicate that automated extraction is sufficiently reliable for large-scale analysis.

Building a new dataset from court documents

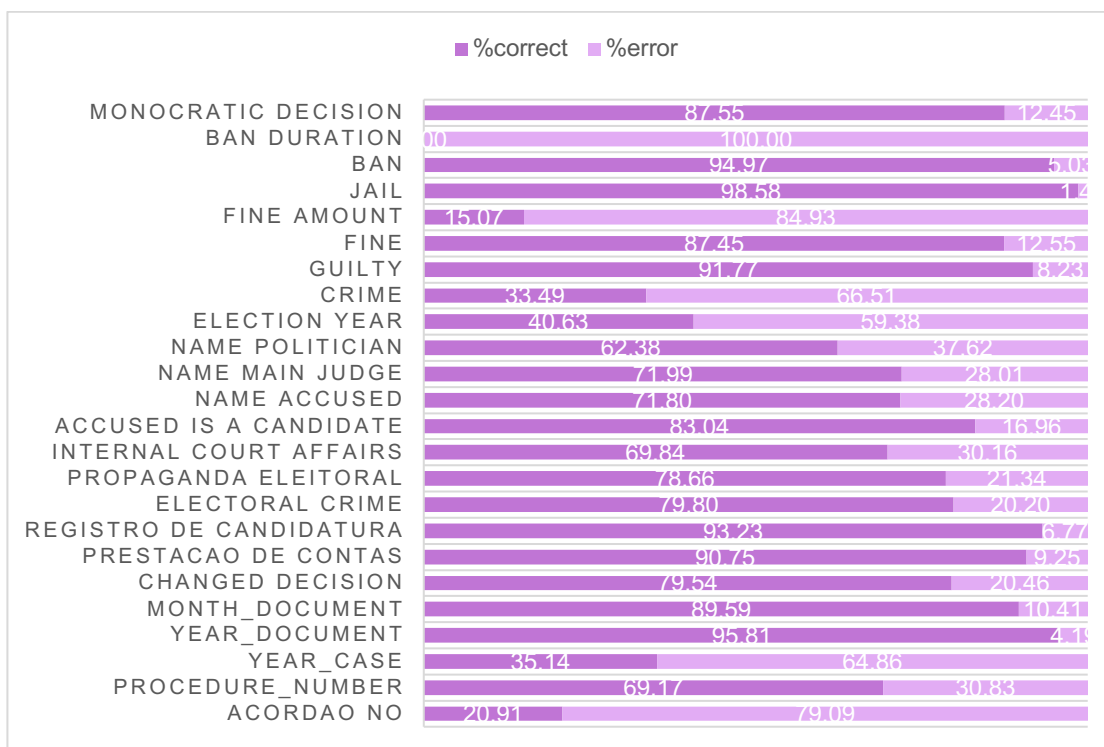
We collect the universe of published electoral court documents from Regional Electoral Courts (*Tribunais Regionais Eleitorais*, *TREs*) between 2008 and 2018. These documents are publicly available but are in unstructured PDF format, varying across states and over time.

Each document may contain information on:

- The reporting judge (relator),
- The accused,
- The alleged crime,
- The verdict,
- The sentence (e.g., fine, ban),
- Whether the decision was unanimous

Combining legal expertise and machine learning

FIGURE 1: Accuracy rates: Gemini Pro vs Manual Extraction



Note: Comparing the accuracy rates of Gemini Pro with manual extraction.

To extract structured data from these unstructured texts, we use a two-stage approach:

1. Manual extraction: Brazilian law graduates manually extract key variables from a sample of documents, creating a validated dataset.
2. Automated extraction: A Large Language Model (Gemini Pro 1.5) processes full court documents using a fixed prompt to extract the same variables

Under the automated extraction, we first pre-process the documents by passing them through an Optical Character Recognition (OCR) and using AI-based expressions to extract the relevant information. This has been completed for the state of Tocantins in Brazil. Accuracy is benchmarked against manual extraction (Figure 1). For approximately 4,000 documents:

- Document classification and key verdict variables achieve high accuracy.
- For many variables, accuracy exceeds 70%.
- Classification of document type exceeds 80% accuracy

Although performance varies across variables (especially for names and more subjective fields), results indicate that automated extraction is sufficiently reliable for large-scale analysis.

We match electoral court cases to administrative candidate data from the *Tribunal Superior Eleitoral (TSE)*, covering all Brazilian elections. The electoral data includes the office sought, municipality, party affiliation, election results and candidate characteristics.

This linkage allows us to observe whether politicians accused or convicted of electoral crimes:

- Run in subsequent elections,
- Are elected.

Preliminary findings

Current results are based on manually extracted data from one state and should be interpreted as preliminary.

Descriptive statistics show that politicians who are found guilty and those who are acquitted are broadly similar in observable pre-case characteristics such as age, education, gender, and prior candidacy experience

Using the severity of the randomly assigned just as an instrumental variable, estimates from the manually extracted data suggest that a guilty verdict:

- Reduces the probability of being a candidate in future elections.
- Reduces the probability of being elected in future elections.
- Lowers the number and share of subsequent candidacies and electoral victories

However, standard errors are large, and estimates are imprecise at this stage.

Expanding the automated extraction to all states is a key next step.

We are working on improving the accuracy of the automated extraction for some of the variables where it is performing poorly. And plan to extend the automated extraction pipeline to the rest of Brazil for a greater sample size and more precise estimation.

Policy implications

1. Enforcement mechanisms can matter for political careers

Preliminary evidence suggests that conviction for electoral crimes reduces future electoral participation and success. This indicates that judicial enforcement may complement electoral accountability mechanisms.

2. Digital judicial records are an underused public resource

Electoral court documents contain rich information but are difficult to analyse at scale. Machine-learning-based extraction can transform unstructured public records into usable datasets for policy analysis.

3. Technology can enhance transparency and research capacity

Even in contexts with complex legal language and heterogeneous documentation, automated models can achieve meaningful accuracy when validated against expert extraction.

4. Scaling up is critical for precision

Full implementation across all states will allow more precise estimation and enable analysis of heterogeneous effects, including differences by crime type and sentence severity.

References

- Avis, E., Ferraz, C., and Finan, F. (2018). Do government audits reduce corruption? Estimating the impacts of exposing corrupt politicians. *Journal of Political Economy*, 126(5):1912–1964.
- Avis, E., Ferraz, C., Finan, F., and Varjão, C. (2022). Money and politics: The effects of campaign spending limits on political entry and competition. *American Economic Journal: Applied Economics*, 14(4):167–99.
- Bachmann, M. (2020). RapidFuzz: Rapid fuzzy string matching in Python. Available at: <https://github.com/rapidfuzz/RapidFuzz>. Accessed on: October 14, 2025.
- Baltrunaite, A. (2020). Political Contributions and Public Procurement: Evidence from Lithuania. *Journal of the European Economic Association*, 18(2):541–582.
- Besley, T., Folke, O., Persson, T., and Rickne, J. (2017). Gender quotas and the crisis of the mediocre man: Theory and evidence from sweden. *American Economic Review*, 107(8):2204–42.
- Bobonis, G. J., Cámara Fuertes, L. R., and Schwabe, R. (2016). Monitoring corruptible politicians. *American Economic Review*, 106(8):2371–2405.
- Cagé, J., Pennec, C. L., and Mougin, E. (forthcoming). Firm donations and political rhetoric: Evidence from a national ban. *American Economic Journal: Economic Policy*.
- Callen, M. and Long, J. D. (2015). Institutional corruption and election fraud: Evidence from a field experiment in Afghanistan. *American Economic Review* 105(1):354–81.
- Chemin, M. (2012). Welfare effects of criminal politicians: A discontinuity-based approach. *The Journal of Law and Economics*, 55(3):667–690.
- CNJ (2019). Artificial intelligence in the Brazilian judiciary. National Justice Council. Available at: <https://bibliotecadigital.cnj.jus.br/xmlui/handle/123456789/98>. Accessed on: October 14, 2025.
- CNJ (2024). The use of generative artificial intelligence in the Brazilian judiciary: Research report. Technical report, National Justice Council, Brasília, Brazil. Available at: <https://www.cnj.jus.br/wp-content/uploads/2024/09/cnj-relatorio-de-pesquisa-iag-pj.pdf>. Accessed on: October 14, 2025.
- Dal Bó, E. and Finan, F. (2018). Progress and perspectives in the study of political selection. *Annual Review of Economics*, 10(1):541–575.
- Ferraz, C. and Finan, F. (2011). Electoral accountability and corruption: Evidence from the audits of local governments. *American Economic Review*, 101(4):1274–1311.
- Frandsen, B., Lefgren, L., and Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, 113(1):253–77.
- Fujiwara, T. (2015). Voting technology, political responsiveness, and infant health: Evidence from Brazil. *Econometrica*, 83(2):423–464.
- Garbiras-Díaz, N. and Montenegro, M. (2022). All eyes on them: A field experiment on citizen oversight and electoral integrity. *American Economic Review*, 112(8):2631–68.
- Gemini_Team (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL]. Available at: <https://arxiv.org/abs/2403.05530>. Accessed on: October 15, 2025.
- Gulzar, S. (2021). Who enters politics and why? *Annual Review of Political Science*, 24(1):253–275.
- Gulzar, S., Rueda, M. R., and Ruiz, N. A. (2022). Do campaign contribution limits curb the influence of money in politics? *American Journal of Political Science*, 66(4):932–946.

- Lambais, G. and Sigstad, H. (2023). Judicial subversion: The effects of political power on court outcomes. *Journal of Public Economics*, 217, 104788.
- Litschig, S., & Zamboni, Y. (2019). Judicial presence and rent extraction (GRIPS Discussion Paper No. 19-20). National Graduate Institute for Policy Studies.
- Michaelowa, K., Panda, S., and Martin, N. (2019). The teflon effect of political power: When criminal charges do not stick. CIS working paper.
- Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, 115(2):200–249.
- Poblete-Cazenave, R. Do politicians in power receive special treatment in courts? evidence from India. *American Journal of Political Science*.
- Prakash, N., Rockmore, M., and Uppal, Y. (2019). Do criminally accused politicians affect economic outcomes? Evidence from India. *Journal of Development Economics*, 141, 102370
- Wang, Y., Guo, H., Liu, H., & Abraham, A. (2012). A fuzzy matching approach for design pattern mining. *Journal of Intelligent & Fuzzy Systems*, 23(2,3), 53–60.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., et al. (2025). A survey of large language models. arXiv:2303.18223 [cs.CL]. Available at: <https://arxiv.org/abs/2303.18223>. Accessed on: October 15, 2025.