Increasing Public Data Transparency for Immigration Law in Canada

Ismail (Husain) Bhinderwala, Jessica Yu, Ke Gao, Yichun Liu

2025-05-04

Table of contents

1 Executive Summary

This project aims to improve transparency in Canada's immigration law by analyzing publicly available data on inadmissibility decisions and legal outcomes. Using datasets from Immigration, Refugees and Citizenship Canada (IRCC) and federal court decisions curated by the Refugee Law Lab, we will investigate how individuals are found inadmissible to Canada and how such decisions are reviewed in court.

Our team will use data science methods, including exploratory data analysis and basic natural language processing, to uncover patterns and potential biases in decision-making. The final product will be a public-facing dashboard built in Python using Dash, designed to present insights in a clear and interactive way for legal professionals, policymakers, researchers, and the general public. This initiative supports the broader goal of increasing accountability in immigration governance by making complex data more accessible and interpretable.

2 Introduction

2.1 Problem Statement and Importance

Inadmissibility decisions under Canadian immigration law play a crucial role in determining who may enter or remain in Canada. These decisions are typically made when an applicant is found to pose a threat to national security, violate human rights, engage in criminal activity, or commit other serious breaches under the *Immigration and Refugee Protection Act (IRPA)*. One such ground is section A34(1), which relates specifically to security-based inadmissibility.

While these decisions have serious implications for individuals and national policy, the related data remains difficult to access and interpret. Information released by IRCC is often raw, inconsistently structured, and lacking contextual detail. Similarly, legal decisions from federal courts are publicly available but presented in unstructured formats that limit systematic analysis.

This lack of transparency hinders legal practitioners, public interest groups, and even policymakers from identifying trends, systemic issues, or inconsistencies in immigration decision-making. Without clear, interpretable data, it is difficult to advocate for fairer and more accountable processes.

To address these challenges, this project combines legal domain knowledge and data science methods. Drawing inspiration from Professor Sean Rehaag's research (Rehaag (2023)) on judicial decisions in refugee and immigration law, we aim to apply similar techniques to the domain of inadmissibility.

2.2 Tangible Objectives

This project has four concrete goals:

1. **Analyze IRCC inadmissibility and litigation datasets** to identify trends based on time, country of citizenship, type of decision, and applicant status (temporary or permanent).
2. **Apply legal analytics** to federal court decisions involving inadmissibility, using natural language processing to extract case-level information such as outcomes, judges, and legal reasoning.
3. **Develop a public-facing dashboard** using Dash (a Python framework) to allow users, legal professionals, policymakers, and others, to explore key trends and findings interactively.
4. **Promote open and interpretable data use** by transforming difficult-to-access raw data into structured, contextualized insights.

2.3 Final Data Product

The final deliverable will include:

- A **web-based interactive dashboard**, allowing users to filter and visualize patterns in immigration inadmissibility and court decisions.
- **Documentation** describing the data sources, analytical methods, and key limitations.
- **Reproducible Python scripts** for transparency and future use by researchers or advocacy groups.

This product aims to serve multiple audiences, lawyers seeking trends in legal decisions, policymakers monitoring fairness in immigration processes, and data scientists interested in legal data applications.

3 Data Science Techniques

3.1 Data Sources

The project will draw on three key datasets:

1. **IRCC A34(1) Refusals (2019–2024):**
   Structured data that records how often applicants are refused entry under section A34(1) of IRPA. Information includes country of citizenship, year, and residency status (temporary/permanent).

2. **IRCC Litigation Applications (2018–2023):**
   A dataset summarizing legal challenges to immigration decisions. It records case types, outcomes,applicant's nationality, litigation counts,etc.

3. **Canadian Legal Decisions (2001–2024):**
   An unstructured dataset of legal texts from federal court and tribunal decisions related to immigration, compiled by the Refugee Law Lab. Our analysis will filter this data to focus specifically on inadmissibility cases, excluding refugee claims.

Each dataset provides a different perspective, administrative and judicial, allowing a well-rounded examination of inadmissibility in Canada.

3.2 Analytical Approach

3.2.1 Data Preparation and Quality Checks

Before conducting any analysis, we will assess each dataset's completeness and clarity. This involves:

- Identifying missing or inconsistent entries (e.g., inconsistent country names).
- Reviewing metadata to understand how variables were defined.
- Verifying that categories (e.g., outcomes, statuses) are clearly and consistently applied.

This step ensures that both legal and data science conclusions are grounded in well-understood data. We will also report any inconsistencies in data quality to encourage IRCC to improve data transparency.

3.2.2 Structured Data Analysis (IRCC Datasets)

For the two IRCC datasets, we will:

- **Restructure the data** into a consistent format (also known as "tidy" data).
- **Use visual tools** (bar charts, heatmaps, and trend lines) to explore:
  - Year-by-year changes in refusal rates.
  - Differences in outcomes by country or region.
  - Disparities between temporary vs. permanent applicants.
- **Form hypotheses** about whether certain groups are more likely to be found inadmissible.

While this analysis cannot establish causality, it can highlight patterns warranting further legal or policy investigation.

3.2.3 Unstructured Legal Text Analysis

For the court decision texts, we will:

- **Classify cases** based on the type of inadmissibility using regular expression by keyword-matching.
- **Extract key information** such as judge names, outcome (granted or denied), and city of filing using large language models (LLMs).
- **Analyze patterns** across judges, regions, or years to detect potential biases or inconsistencies in rulings.

A stretch goal includes **semantic analysis**, examining the reasoning within judgments to see how legal language evolves over time or varies by case type.

3.2.4 Time Series and Comparative Analysis

Across all datasets, we will conduct basic time-based trend analysis to identify:

- Whether inadmissibility findings have increased or decreased over time.
- If certain court outcomes correspond with changes in policy or global events.
- How legal decisions align (or diverge) from administrative trends.

3.3 Evaluation Metrics and Success Criteria

To assess whether the project has met its goals, we will use the following criteria:

- **Coverage Metrics**: Percentage of cases successfully categorized from legal text.
- **Accessibility**: Can both legal and non-technical users interact with and understand the dashboard?
- **Partner Expectation**: A functional, clear dashboard and evidence-based insight into inadmissibility trends.
- **Reproducibility**: Are our data processing and analysis steps transparent and replicable?

We will also gather feedback from mentors and the capstone partner to ensure that the final product aligns with stakeholder expectations.

4 Timeline

Week	Task Description
Week 1	Set up environment and version control (Dash, Quarto, GitHub); review datasets
Week 2	Exploratory analysis on IRCC datasets; clean and filter legal text dataset
Week 3	Extract metadata from legal decisions; continue EDA on litigation data
Week 4	Integrate court data with dashboard; finalize IRCC analysis
Week 5	Finalize dashboard layout; populate visuals and filters
Week 6	Finalize documentation; incorporate feedback; prepare final submission

Parallel Tasks: Legal and IRCC datasets will be analyzed concurrently by different team members to ensure timely delivery.

References

Rehaag, Sean. 2023. "Luck of the Draw III: Using AI to Extract Data about Decision-Making in Federal Court Stays of Removal." *Queen's LJ* 49: 73.