Mining Supreme Court Cases for Legal Analysis (2019–2024)

Adir Serruya
Department of Software and Information Systems Engineering
Ben Gurion University of the Negev
adirser@post.bgu.ac.il

January 13, 2025

Abstract

The legal domain holds extensive data that can yield valuable insights. This project analyzes publicly available Supreme Court cases from 2019 to 2024 using data mining and natural language processing (NLP) techniques. By uncovering patterns in judicial behavior, case types, and arguments, the study provides a foundation for large-scale legal dataset analysis to aid scholars and practitioners.

1 Introduction

Data mining has revolutionized legal analytics, enabling the discovery of trends and patterns in judicial data. With increasing access to court records, this study analyzes Supreme Court cases from 2019 to 2024 to identify trends, legal arguments, and judge behaviors.

2 Related Work

2.1 Legal Data Mining

Legal data mining provides critical insights into judicial data for tasks like **case law analysis**, **precedent identification**, and **judicial decision prediction** [3]. Techniques such as predictive modeling, clustering, and graph-based methods have been employed to analyze case outcomes, group similar cases, and map relationships between cases and judges [1]. Despite its potential, challenges like data sparsity and domain complexity remain.

2.2 Text Analysis in Law

Natural Language Processing (NLP) plays a vital role in understanding dense legal texts. Techniques include **text classification**, **argument mining**, and **NER**, which extract legal arguments, entities, and sentiments from judicial opinions [4]. Recent advancements like transformer models (e.g., LegalBERT) have further improved legal text analysis, offering structured insights and automating research tasks [2].

2.3 Data Collection

The dataset was scraped from the publicly available Supreme Court portal `https://supreme.court.gov.il/Pages/fullsearch.aspx`, which provides search-based case retrieval with authentication and rate limits. The scraping process required careful planning to ensure comprehensive data collection while adhering to ethical and legal guidelines.

- **Tools and Methods:** The program used a combination of **Requests**, **BeautifulSoup**, and **Selenium** for handling HTTP requests, parsing structured data, and navigating dynamic content. **Regex** was employed to extract details such as case IDs, dates, and names.
- **Pagination and Dynamic Content:** The portal's paginated and JavaScript-dependent content required iterative navigation and session management to ensure complete data retrieval.
- **Ethical Compliance:** Data collection respected rate limits and used only publicly accessible information for academic purposes.

The process took approximately three weeks, resulting in a structured dataset with metadata such as case IDs, judges, parties, dates, and topics, forming the basis for subsequent analyses.

2.4 Data Preprocessing

The raw data from the Supreme Court portal required extensive preprocessing to ensure it was analysis-ready. Key steps included:

- **Data Cleaning and Structuring:**
  - Daily JSON files were processed into structured records.
  - Timestamps were converted to readable dates for **Case Start Date** and **Case Verdict Date**.
  - Free-text fields were cleaned using regex to remove HTML tags, symbols, and diacritics.
- **Metadata Extraction:**
  - Extracted **Prosecutor** and **Defendant(s)** using patterns in Hebrew legal terminology.
  - Identified **Judges** and **Case Topics** through regex and predefined mappings.
  - Parsed verdicts into categories: `Accepted`, `Denied`, `Deleted`, or `Unknown`.
  - Used Named Entity Recognition (NER) to extract geographic and institutional entities.
- **Text Processing:**
  - Normalized Hebrew text by removing diacritics and irrelevant characters while preserving meaning.
  - Tokenized text into individual words for further analysis.
- **Feature Engineering:**
  - Extracted datetime features, such as the duration between **Case Start Date** and **Verdict Date**.
  - Categorized cases by petitioners' political affiliations (`Left Wing`, `Right Wing`, `Neutral`, `Unknown`).
  - Visualized judge involvement with blended color graphs indicating political affiliation ratios.

The preprocessing pipeline was automated and iterative, ensuring consistent and reliable results. This process yielded a clean, structured dataset with enhanced metadata, forming the foundation for robust analyses.

3 Data Analysis

3.1 Dataset Description

The dataset used in this study consists of Supreme Court case data collected over a period of more than four years, spanning from **July 26, 2020** to **November 26, 2024**. It includes a total of **27,005 cases**, with each record containing detailed information about a single case. The dataset comprises the following 15 columns:

Column Name	**Description**
ID	Unique record identifier.
CaseID	Unique case identifier.
CaseNumber	Sequential case number.
Type	Code for legal proceeding type.
IsTechnical	Indicates if the case involves technical matters.
CaseStartDate	Case initiation date.
CaseVerdictDate	Verdict issue date.
Prosecutor	Prosecuting party.
Defendant	Defending party.
FreeText	Case details in text.
Judge	Judges presiding over the case.
Topic	Case category or topic.

Table 1: Dataset column descriptions.

3.2 Summary Statistics

- **Number of Cases:** 27,005.
- **Date Range:** The dataset spans from **July 26, 2020** to **November 26, 2024**.
- **Example Record:**

Column	Example Value
ID	8024
CaseID	810059
CaseNumber	272410
Type	1560
IsTechnical	False
StartDate	2024-11-26 11:09:51
VerdictDate	2024-11-26 11:05:53
Prosecutor	אסום - חברה קבלנית בעמ
Defendant	מדינת ישראל
FreeText	בבית המשפט העליון רע"פ...
JudgeStr	כבוד השופטים ל' גליקסמן
Topic	רשות ערעור פלילי

Table 2: Example record from the dataset.

3.3 Topic Distribution

The distribution of case topics provides insights into judicial focus areas. As shown in Figure 1, the majority of cases belong to **Bagatz**, reflecting its importance in addressing critical matters such as politicization and judicial preferences. This analysis focuses primarily on Bagatz cases to uncover patterns in judicial behavior and societal impact.

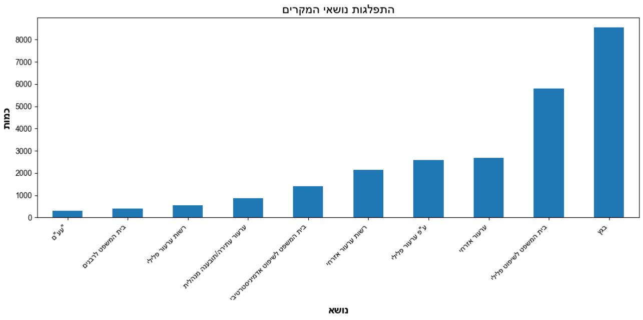

Figure 1: Distribution of Case Topics in Supreme Court Data (2020–2024).

3.4 Bagatz Case Drill Down

Given the prominence of **Bagatz** cases in the dataset, we conducted a detailed analysis of the types of cases handled under this category. Specifically, Bagatz cases can be classified into three types: **Decisions**, **Verdicts**, and **Warrants**.

The distribution of these case types is illustrated in the chart below. As observed, the majority of Bagatz cases are categorized as **Decisions**. However, for the purposes of this analysis, we will primarily focus on **Final Verdicts**, as they provide the most conclusive insights into judicial behavior and decision-making processes.

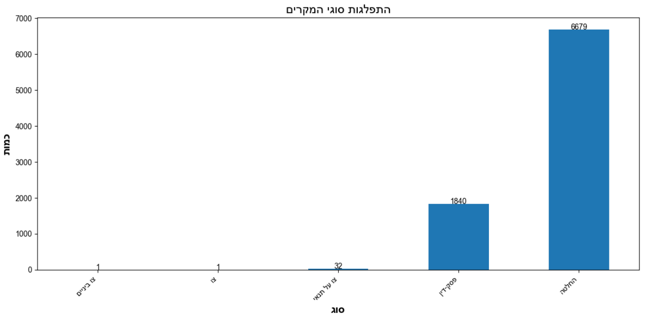

Figure 2: Distribution of Bagatz Case Types in Supreme Court Data (2020–2024).

This focus on Final Verdicts allows us to delve into critical aspects of the judicial process, such as the resolution of high-stakes disputes, the influence of external factors, and the broader implications of judicial outcomes on governance and society.

3.5 Judge Participation in Bagatz Verdicts

To gain deeper insights into the judicial processes within **Bagatz** cases, I analyzed the number of Final Verdicts in which each judge participated. This analysis provides valuable information about the judges' level of involvement and influence in high-stakes cases.

The chart below shows the distribution of the number of Bagatz Final Verdicts by each judge. As seen in the chart, some judges have a significantly higher level of participation, which could indicate their prominence or specialization in handling Bagatz cases.

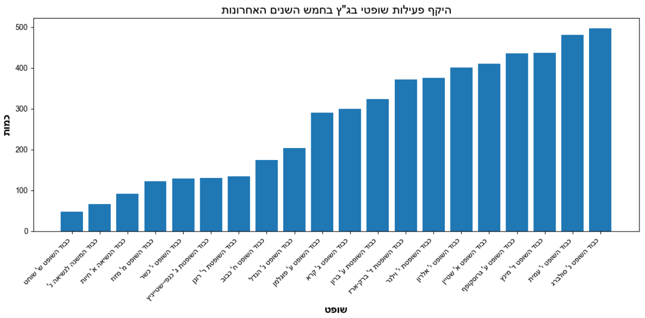

Figure 3: Number of Bagatz Verdicts Participated in by Each Judge (2020–2024).

This distribution helps identify key judges who frequently preside over Bagatz cases, allowing us to further investigate their decision-making patterns, potential biases, and collaboration networks in subsequent analyses.

3.6 Top 5 Petitioners to Bagatz Cases

Another critical aspect of analyzing Bagatz cases is identifying the key petitioners who frequently bring cases before the court. The chart below displays the top 5 petitioners based on the number of Bagatz cases they have filed during the analyzed period (2020–2024).

To provide additional context, the petitioners are color-coded according to their perceived political affiliations:

- **Blue:** Represents petitioners aligned with left-leaning political views.
- **Red:** Represents petitioners aligned with right-leaning political views.
- **Green:** Represents petitioners considered politically neutral.

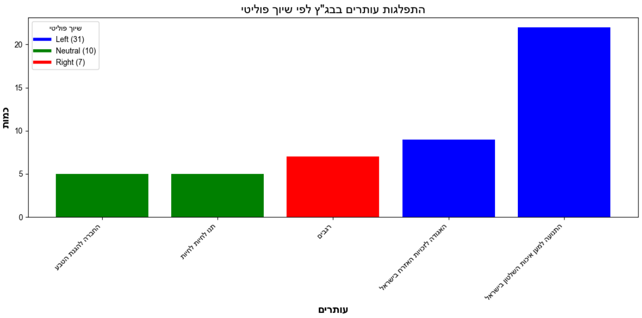

Figure 4: Top 5 Petitioners to Bagatz Cases (2020–2024), Color-Coded by Political Affiliation.

From the chart, it is evident that the majority of Bagatz petitions are filed by politically active entities, Specifically Crushing majority of left winged political entities. This highlights the role of Bagatz as a forum for addressing issues of public and political importance.

3.7 Judge Petitions: Accepted, Denied, or Deleted

The chart below illustrates the distribution of petitions (Accepted, Denied, or Deleted) handled by each judge in Bagatz cases. This visualization helps identify trends in judicial decision-making and highlights judges with a higher frequency of handling specific petition outcomes.

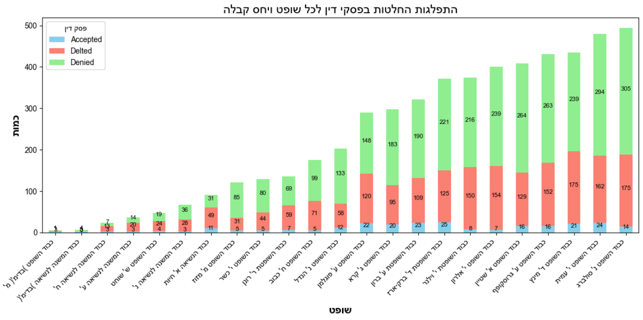

Figure 5: Distribution of Petitions Accepted, Denied, or Deleted per Judge in Bagatz Cases.

3.8 Locations in Bagatz Verdicts

The word cloud highlights frequently mentioned locations in Bagatz verdicts, with larger words indicating higher frequency. A geographical map complements this, providing insights into the distribution of legal significance across regions.

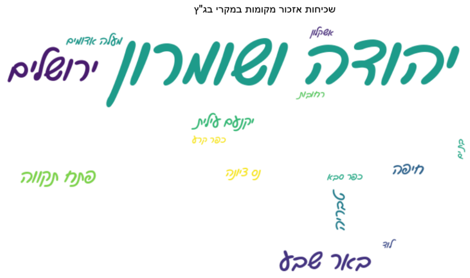

Figure 6: Word Cloud of Bagatz Locations.

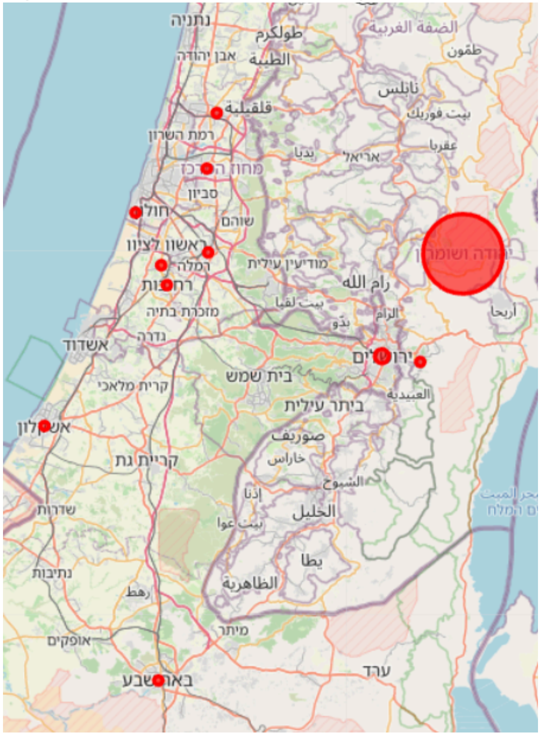

Figure 7: Geographical Distribution of Bagatz Locations.

3.9 Cases in Bagatz Over Time with Rolling Median

Understanding temporal trends in Bagatz cases is crucial for identifying significant shifts in judicial patterns. The chart below displays the number of Bagatz cases over time, accompanied by a rolling median (3-month window) to smooth fluctuations and highlight trends.

As observed, there is a sharp drop in the number of cases at the beginning of 2022. This may indicate a potential **concept drift** or **distribution shift** in the dataset, reflecting a change in judicial focus, case types, or external factors influencing case filings.

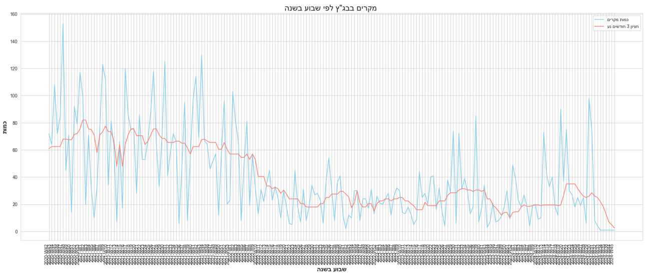

Figure 8: Number of Bagatz Cases Over Time with Rolling Median (3 months). Sharp drop identified in early 2022.

3.10 Case Connections Based on Shared Judges

A graph of Bagatz cases was constructed, with nodes representing cases and edges indicating shared judges. Nodes are color-coded by prosecutor affiliation: **Blue** for left-leaning, **Red** for right-leaning, **Green** for neutral, and **Gray** for unknown. The graph reveals judicial overlaps and connections across cases of varying political affiliations.

3.11 Judge Collaboration Graph

The collaboration graph visualizes interactions between judges in Bagatz cases. Nodes represent judges, connected by edges if they worked on the same case. Node colors indicate the political affiliations of cases handled: **Blue** for predominantly left-leaning, **Red** for right-leaning, **Green** for neutral cases, and **Gray** for mixed or unclear affiliations. This graph reveals patterns such as judges handling cases of specific political leanings and forming collaborative clusters.

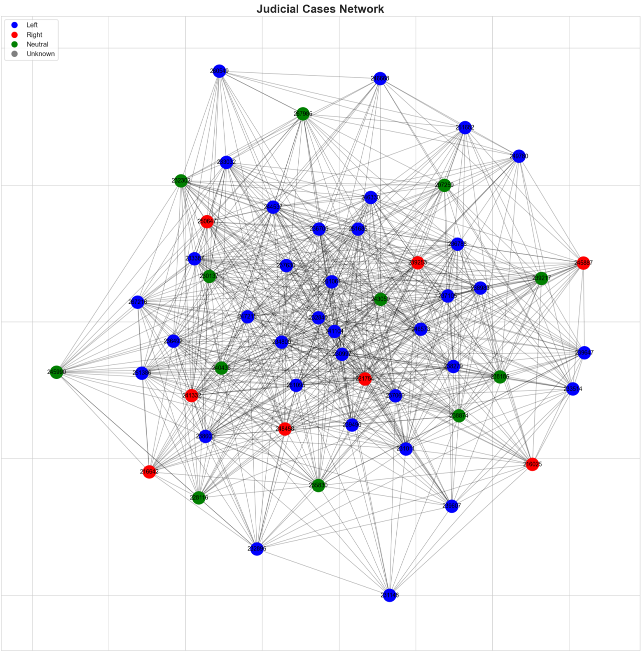

Figure 9: Graph of Case Connections Based on Shared Judges. Nodes are colored by the political affiliation of the prosecutor.

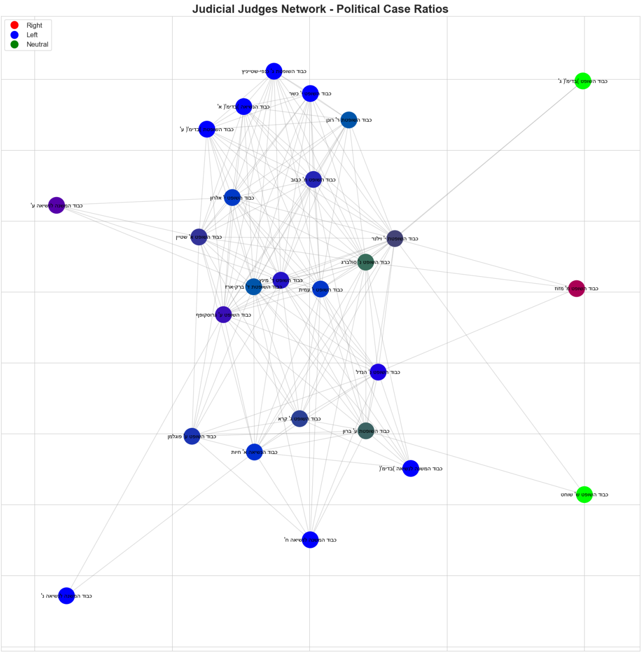

Figure 10: Graph of Judge Collaborations in Bagatz Cases. Node colors represent the mixture of political affiliations among the cases handled.

4 Conclusions

This study demonstrates the utility of data mining and natural language processing (NLP) techniques in analyzing Supreme Court cases. By scraping and processing over 27,000 cases spanning five years (2019–2024), we provided insights into judicial trends, case types, and judge behaviors. The following summarizes my key findings and contributions:

- **Data Collection and Processing:** my methodology for ethically scraping, structuring, and preprocessing Supreme Court data enabled the creation of a clean, feature-rich dataset. This dataset serves as a foundational resource for further analysis in legal research.
- **Judicial Trends and Patterns:** The analysis revealed notable trends in case types, petitioners' political affiliations, and the geographic distribution of cases. For example, a significant concentration of *Bagatz* cases highlighted the prominence of politically sensitive and high-stakes matters.
- **Graph-Based Analysis:** The visualization of case connections through shared judges and the collaboration graph of judges provided new perspectives on judicial networks. These insights are particularly valuable for understanding how specific judges contribute to the legal ecosystem and interact with others.

Key Challenges

Several challenges were encountered during the project:

- **Scraping Dynamics:** Navigating dynamic content and rate limits posed significant technical hurdles, requiring advanced tools such as Selenium and careful session management.
- **Hebrew Language Complexity:** Legal texts in Hebrew, with their dense and technical nature, demanded robust NLP pipelines for tasks like named entity recognition (NER) and argument extraction.
- **Bias and Ambiguity:** The classification of petitioners' political affiliations and sentiment in judicial texts involved subjective judgment and external validation, potentially introducing biases.

Acknowledgment

I would like to express our gratitude to the following resources and tools that significantly contributed to the success of this study:

- **Wikipedia:** For providing accessible and well-documented information that aided in initial research and contextual understanding.
- **ChatGPT:** For assistance in refining, restructuring, and validating the content, methodologies, and presentation of this study.
- **Hugging Face:** For their extensive collection of pre-trained NLP models and libraries, which were instrumental in processing and analyzing legal texts.
- **Nominatim:** For providing reliable and open-source geocoding services to map and analyze location-based data in Bagatz cases.
- **Government Open Resources:** Specifically, the document Net Law Info for the detailed mapping of legal abbreviations and Supreme Court terminologies.

Links

- **GitHub Repository:** The project's codebase, including relevant scripts and models, is available on GitHub. Visit: Github Repo.

Contact

For inquiries regarding the data, scraping code, or methodologies used in this study, please feel free to contact me via email at: adirser@post.bgu.ac.il.

References

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. Predicting the judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. 2020.

[3] Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. *PLoS ONE*, 12(4):e0174698, 2017.

[4] Manfred Stede and Jodi Schneider. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 9(3):1–105, 2016.