Research Proposal

Jeongyeop Han

July 2024

1 Project Title

Analyzing the Complexity of Financial Laws and Regulations in Texas Using Machine and Natural Language Processing Techniques

2 Problem Statement

The financial laws and regulations in Texas are complex and multifaceted, making it challenging for businesses and legal practitioners to navigate and comply with them effectively, and this can negatively impact the economy of the state. In fact, there is a correlation between the economy of the country or state and its regulation.

3 Falsifiable Hypothesis

Using advanced NLP techniques and ML models can effectively measure and analyze the complexity of legal texts by evaluating in-depth sentence structures, vocabulary accuracy, and the internal regulatory framework. This approach will provide a more precise assessment of regulatory complexity compared to traditional methods, which then can be used to determine the optimal level of complexity of the regulation.

4 State of the Art

Several methods currently exist to measure the complexity of legal texts, but each method has a limitation. For instance, measuring the volume, a number of norms, to measure the complexities of the legal texts ignores the structural and relational aspects of the texts. Thus, NLP integrated with analysis of network of the legal texts would measure the complexities more accurately. Another problem, though, is that there may be a discrepancy between computed complexities and how people actually perceive them.

5 Solution Direction

This project proposes the use of a combination of advanced NLP techniques and ML models to analyze Texas financial regulations. LLM will be used to validate the complexities measured by NLP techniques. The approach involves:

- **Data Collection and Preparation**:
  - Collect a comprehensive dataset of Texas financial regulations.
  - Preprocess the text data to clean and format it for analysis. This includes tokenization, lemmatization, and removal of stop words.
- **Volume Measurement**:
  - Measure the volume of the legal texts by counting the number of norms and sections.
  - Use these counts as a basic metric of complexity.
- **Linguistic Complexity Analysis**:
  - Employ NLP techniques to analyze the linguistic features of the texts. This includes:
    - **Lexical Diversity**: Calculate metrics such as type-token ratio (TTR) to assess the variety of vocabulary used.
    - **Syntactic Complexity**: Analyze sentence structures using dependency parsing to understand syntactic depth and intricacy.
    - **Readability Scores**: Compute readability scores (e.g., Flesch-Kincaid, Gunning Fog) to determine how easily the texts can be read and understood.
- **Relational Complexity Analysis**:
  - Analyze the network structure of the legal texts by identifying and mapping references and citations within the regulations.
  - Use graph theory metrics (e.g., node degree, centrality measures) to quantify the relational complexity of the texts.
- **Entropy and Information Density**:
  - Compute entropy measures to quantify the unpredictability and information density of the legal texts. This involves:
    - **Shannon Entropy**: Measure the entropy of the distribution of terms within the texts to gauge their informativeness.
    - **Mutual Information**: Assess the amount of information shared between different parts of the text.
- **Validation with LLMs**:

  - Utilize state-of-the-art LLMs (e.g., GPT-4) to validate the computed complexities. This involves:
    - **Legal Reasoning**: Test the LLM's ability to perform legal reasoning on the texts and compare its performance with the computed complexity measures.
    - **Human Perception Alignment**: Compare the LLM's assessments with human evaluations of the texts' complexity to ensure alignment.

- **Evaluation and Refinement**:
  - Evaluate the effectiveness of the proposed methods using quantitative metrics and qualitative feedback from legal experts.
  - Refine the methodologies based on evaluation results to improve accuracy and reliability.

6 Timeline

- **Weeks 1**
  - Project kickoff and initial planning.
  - Begin collecting a comprehensive dataset of Texas financial regulations.
- **Weeks 2-3**
  - Continue collecting the dataset.
  - Begin preprocessing the text data (tokenization, lemmatization, removal of stop words).
- **Weeks 4-5**
  - Complete data collection.
  - Complete preprocessing the text data.
  - Begin measuring the volume of legal texts (counting norms and sections).
- **Weeks 6-9**
  - Continue and complete volume measurement.
  - Start analyzing linguistic features (lexical diversity, syntactic complexity, readability scores).
- **Weeks 10-11**
  - Continue linguistic complexity analysis.

  - Begin analyzing the network structure of legal texts.

- **Weeks 11-12**
  - Complete the analysis of linguistic features.
  - Continue relational complexity analysis.
- **Weeks 13-14**
  - Complete relational complexity analysis.
  - Start computing entropy measures (Shannon Entropy, Mutual Information).
- **Weeks 15-16**
  - Continue computing entropy measures.
  - Begin validating computed complexities using LLMs (e.g., GPT-4).
- **Weeks 17-18**
  - Complete validation using LLMs.
  - Start evaluating the effectiveness of proposed methods using quantitative metrics and feedback from legal experts.
- **Weeks 19-20**
  - Continue evaluation and begin refining methodologies based on feedback.
  - Continue refining methodologies.
- **Weeks 21-22**
  - Complete refinement of methodologies.
  - Start drafting the final report, including all findings and conclusions.
- **Weeks 22-23**
  - Complete the drafting of the final report.
  - Review and revise the final report based on feedback from advisors and peers.
  - Finalize and submit the report.

Literature review of papers related to Financial Laws and Regulation in Machine Learning and Natural Language Processing

Jeongyeop Han
Georgia Tech FinTech Lab / Atlanta, GA, USA
jhan359@gatech.edu

1 Introduction

Measuring the complexities of legal texts is crucial, as well-designed regulations can mitigate market failures, reduce transaction costs, and enhance economic efficiency. Conversely, poorly drafted regulations can increase transaction costs and negatively impact economic performance. The reviewed papers focus on the topics of financial laws and regulations, particularly through the lens of machine learning and natural language processing.

Specifically, they discuss how natural language processing methods can assist in analyzing the complexities of legal texts. Legal texts can exhibit three forms of complexity: volume (quantity), ambiguity (linguistic), and extensive interdependencies (relational) (Lucio and Mora-Sanguinetti, 2021). Measuring the volume of legal text can be accomplished simply by counting the number of norms. However, assessing other forms of complexities requires careful consideration and can be quite challenging. For instance, linguistic complexity can be measured by analyzing the structure and language, and relational complexity can be measured by analyzing network of the regulation.

On the other hand, GPT-4 was found to be the most outperforming model in a benchmark of legal reasoning and the Multistate Bar Exam (MBE). GPT-4 outperformed the average test taker in the MBE, suggesting that LLMs can be valuable tools for measuring the complexities of legal texts.

2 Body

A primary area of research in legal text complexity involves the use of Natural Language Processing. In fact, The structure and the language of the legal texts contribute to the linguistic complexity of the legal texts.

This involves µ indicator proposed by Muñoz and Muñoz, measuring entropy and number of not frequent words, and measuring the hierarchical depth of elements within the legal text. The greater µ value means better readability (Lucio and Mora-Sanguinetti, 2021).

$\mu_n = \left(\frac{W_{on}}{W_{on-1}}\right)\left(\frac{\overline{L_{en}}}{\sigma^2_{L_{en}}}\right) \times 100$ where W_{on} denotes the number of words, $\overline{L_{en}}$ represents the average number of letters per word, $\sigma^2_{L_{en}}$ indicates the variance in the number of letters per word, and n is number of norms (Lucio and Mora-Sanguinetti, 2021).

“entropy” indicator proposed by Katz and Bommarito (2014) and Shannon (1951) can also be used to determine the linguistic complexities of the legal texts (Lucio and Mora-Sanguinetti, 2021). The "entropy" indicator essentially measures how frequent a word appears in the texts.

Additionally, measuring the proportion of commonly used words within the legal text can provide another useful metric for assessing readability. The structure of the legal text must also be considered when evaluating linguistic complexity. For instance, the deeper depth of the elements within the legal texts means more complex legal texts. Elements are chapters, sub chapters, sections, subsections, and etc (Katz and Bommarito, 2014). However, there are some limitations to these approaches for measuring complexity. For instance, the scale can be somewhat narrow for µ indicator, often clustering around similar values for legal texts, making it challenging to distinguish significantly different levels of complexities.

Analyzing the network of legal texts should also be considered when measuring complexities. One approach is to count the number of links associated with the legal texts. For example, the ratio of links incorporated in the norms adopted during a year to the number of norms adopted by the region in that year can be used to determine relational complexities (Lucio and Mora-Sanguinetti, 2021).

On the other hand, there is an algorithm to measure the complexities developed by Colliard and

Georg. using Halstead measures (Colliard and Georg, 2023). This approach includes measuring volume, potential volume, and level of complexity. Volume refers to the number of operators and operands in the legal texts. Potential volume is the volume of the theoretically shortest program that can solve the problem. Level is the ratio of potential volume to the actual volume.

Although these combined approaches can effectively compute complexities, validation is still necessary. This is because there may be a discrepancy between computed complexities and how people actually perceive them. I believe LLMs can aid in this task. In fact, LLMs are highly capable of legal reasoning, including Issue, Rule, Application, and Conclusion (IRAC) (Guha et al., 2023). Additionally, GPT-4 outperforms average human test-takers on the Multistate Bar Examination (MBE) (Katz et al., 2024). This suggests that GPT-4 could be a suitable candidate for validating these complexities as if it were a human, providing a more comprehensive assessment.

3 Conclusion

In conclusion, the complexities of legal texts can be measured using various approaches, including natural language processing methods, the µ indicator, entropy measures, and Halstead measures. Each approach has its strengths and limitations. While these combined approaches can effectively compute complexities, validation is necessary to ensure their accuracy aligns with human perception.

Future research should focus on validating these methods using LLMs like GPT-4, which have demonstrated exceptional capabilities in legal reasoning and outperform average human test-takers on the Multistate Bar Examination. By leveraging the strengths of LLMs, we can achieve a more comprehensive assessment of legal text complexities.

Further studies should explore how these models can be integrated into real-world applications, ensuring their outputs align with human expertise. This could lead to more efficient and accurate evaluations of legal texts, ultimately contributing to well-designed regulations that can mitigate market failures, reduce transaction costs, and enhance economic efficiency

References

Jean-Edouard Colliard and Co-Pierre Georg. 2023. Measuring regulatory complexity. Research Paper FIN-2020-1358, HEC Paris.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael A. Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 2023 Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*, number 4583531 in Osgoode Legal Studies Research Paper.

Daniel Martin Katz and Michael James Bommarito. 2014. Measuring the complexity of the law: The united states code. *Artificial Intelligence and Law*, 22(4):337–374.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382.

Juan De Lucio and Juan S. Mora-Sanguinetti. 2021. New dimensions of regulatory complexity and their economic cost. an analysis using text mining. Working Paper 2107, Banco de Espana.