Semantics Extraction and Analytics from SEBI Regulations and Case Files: Industry-academia Collaboration

Natraj Raman[1] **Pulkit Parikh**[2] **Lini Thomas**[2] **Kamalakar Karlapalem**[2]
[1]JPMorgan AI Research [2]IIIT Hyderabad

Abstract

Extracting insights from text documents and developing predictive models for analytics is of critical importance in several domains. However, it is a challenging task owing to the diversity in linguistic characteristics of large scale text corpora, exacerbated by a lack of labeled data. We present here a case-study on extracting semantics from complex legal and regulatory documents and applying them to perform analytical tasks such as violation detection and penalty estimation. Our system was developed in a joint academic-industry collaboration effort and benefited from their complementary research strengths. Specifically, the domain expertise and problem formulation process in the industrial setting were combined with the exploratory research and experimentation rigor of the educational world to develop a system that can help legal actors improve their productivity. We outline our collaboration mechanism, detail the techniques used and functionalities developed, and also discuss the key take-aways that can benefit the research community.

1 Introduction

Text documents are an effective communication medium for humans. A human can read a document and process it to extract its multiple interpretations, and decide on the relevant interpretation. However, with a large number of interrelated documents it gets difficult to extract multiplicity of interpretations, and to identify the relevant interpretations (1). An AI system can efficiently perform semantic processing of the documents to determine and to ascertain these interpretations. Our objective is to develop a framework to facilitate the processing of semantics from legal documents that originate from Securities and Exchange Board of India (SEBI), a regulatory body for the financial markets in India. The regulatory system sees the interplay between regulators, lawyers and companies and requires solving several analytical tasks such as regulation violation detection, penalty estimation, question and answering, and tracking the biography of regulations.

Extracting semantics from legal documents is challenging due to the subjectivity, ambiguity, long form content and complex linguistic patterns prevalent in legal domain. In addition, there are often no readily available labeled data to train a supervised model and the expected outcomes vary depending on the type of user. This requires defining a precise scope, formulating the problem appropriately, collecting annotations in a smart manner, designing sophisticated models and evaluating with appropriate metrics. No one team can possess the wide-ranging capabilities necessary to develop such a complex system and collaborations become pertinent.

In this work, we outline the collaboration effort between an academic research institute and a financial sector company to develop this semantics extraction and analytics system. The two partners provided complementary strengths and benefited from their mutual expertise. For example, the industry partner helped in identifying problems that are relevant for practitioners while the academic partner conducted

Broadening Research Collaborations (NeurIPS Workshop 2022).

a variety of experiments using state-of-the-art language models (2). This fruitful collaboration helped us to accelerate the development of a viable solution, which would have otherwise been complicated.

Furthermore, we detail the Applied Semantics Extraction and Analytics (ASEA) framework that performs the semantic processing of SEBI documents. The functionalities described include an extractive question-answering system to answer queries from a knowledge base of SEBI documents, a mechanism to track amendments to regulations, the creation and automatic assignment of semantic tags to sentences and phrases, and a regulatory domain specific language model.

In summary, this paper highlights the value in a research collaboration that spans outside traditional academic circles and provides insights on what works and avenues for improvement. Section 2 outlines the collaboration structure, section 3 describes the ASEA framework and section 4 discusses the take-aways.

2 Academia-Industry Collaboration Structure

Large firms deal with a lot of text material and AI systems that can simplify and automate the processing of text data is critical to the functioning of businesses. While NLP has made huge advances in recent years (3), there are still unique problems that cannot be solved with general-purpose models and off-the-shelf solutions. Hence formal targeted research is often necessary to address the challenges posed by niche problems. An industrial practitioner must often prioritize for immediate needs and the resources required for conducting exploratory research is scarce. Hence collaborating with the universities is a viable alternative.

In this particular effort, industry researchers partnered with the academic faculties to refine their draft proposal and jointly define a problem scope that is both interesting from a research perspective and valuable from a utilitarian standpoint. This benefited the academic partner by grounding the students' research towards potentially feasible solutions. The collaboration avoided the distraction of problem scoping without actually solving the domain-specific task. For example, extracting even the basic semantics from both SEBI regulations and legal case documents was an onerous exercise. It required legal annotators and domain experts to identify the domain-specific semantics, and developing sophisticated models with limited data was challenging. In the absence of this partnership, we would have over-complicated the problem with generic solutions disregarding the specific requirements.

Effective prioritization of the problems is another important area. Tasks such as semantic parsing of text material or search and navigation for Q&A can be used as a building-block for downstream applications and hence were addressed at the beginning. A planned order of execution also helped us focus on problems that have a good impact on the legal community such as penalty estimation and regulation biography, and which are novel and unique in terms of the techniques required to address them. Thus, problem formulation, scoping and prioritization are key areas where collaboration is critical to achieve successful outcomes.

Evaluation metrics and their applicability is an aspect where the industry perspective can differ from the academia. For instance, academic researchers are often interested in obtaining the best available performance metric (say F1 score), while the stability and interpretability of the models are critical for deployed solutions. Hence there is an implicit trade-off between the complexity of the model and the quality of the results. The periodic review of the models by the industry partner helped guide the model developer to focus on research elements beyond traditional metrics and consider the robustness and explainability dimensions.

Constant interaction with industry partners helps the team comprehend the solutions' practicality, rigor, and utility. More importantly, the industry comprehended the state of the art's limitations in addressing specific problems. The collaboration was made smooth by a mutual respect for each partner's priority and ideas. The monthly catch-up meetings helped share thoughts, measure progress, identify issues and perform course corrections appropriately. To over-use the cliché, this collaboration structure ended up being a win-win for both the parties.

3 ASEA Framework

Figure 1 shows the interplay between regulators and companies. While the essence of the interplay depicted is applicable across countries, we use specifics applicable to India to illustrate it. Securities

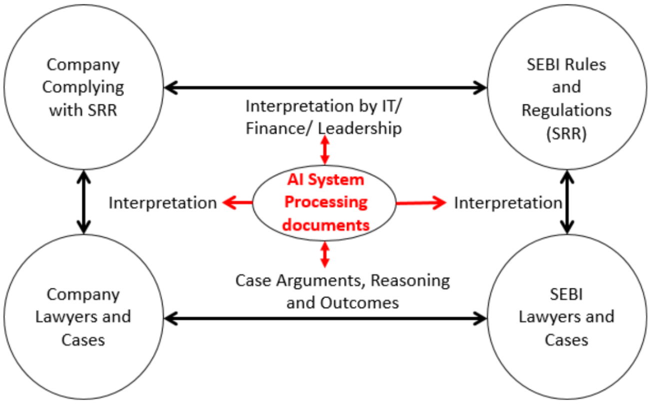
Company Complying with SRR
Interpretation by IT/ Finance/ Leadership
SEBI Rules and Regulations (SRR)
Interpretation
AI System Processing documents
Interpretation
Case Arguments, Reasoning and Outcomes
Company Lawyers and Cases
SEBI Lawyers and Cases

Figure 1: ASEA solutions for Regulators and Lawyers across SEBI documents

and Exchange Board of India, commonly referred to as SEBI, is the regulatory body for securities and commodity market in India. The Indian companies are obligated to adhere to the regulations drafted by SEBI. The regulations need to be interpreted by the information technology departments and/or finance departments of the companies. Moreover, the companies' lawyers interact with their SEBI counterparts about case arguments and outcomes. A well-formulated AI system can aid and improve these interpretations and interactions by processing and deriving insights from SEBI regulations, associated case files and other pertinent documents. Towards this goal, we discuss our framework titled Applied Semantics Extraction and Analytics (ASEA).

3.1 Semantic Segmentation of Case Files through Sentence Classification

Over 1100 Adjudication Orders were scraped from the SEBI website. Out of the scraped documents, 27 adjudication orders have been annotated with the help of a legal expert giving rise to 2052 sentence-label pairs that have been used to train the models. Keeping the tasks of penalty estimation and violation detection in mind, the annotated adjudication orders only pertain to Prohibitions of Insider Trading (PIT) regulations. This dataset is the first of its kind for Indian legal adjudication orders regarding insider trading and can be used by other researchers interested in exploring this line of work. We experimented with a number of machine learning methods to train the sentence classifier. We primarily made use of transformer architectures like uncased L-12 H-768 A-12 BERT, XLNet(4) & Legal BERT, each with a multi layer perceptron that is used on top of the trainable transformer layers for predicting the label.

3.2 Regulation Violation Detection

We pose the problem of regulation violation detection as a multi label classification task. Given the facts of the case and r different regulations, the classification model predicts a one-hot vector of length R, where dimension r being 1 implies that regulation r is violated. As we are interested in detecting regulation violation given a set of facts of the case, we develop a semantic segmentation engine to separate out the different sections of the case-file.

We experiment with several different machine learning classification set ups. The recent years have shown how effective transformer based architectures are for various NLP tasks, and thus we also develop a transformer based multi label classifier. However, these models perform poorly on domain specific tasks such as those in the scientific, medical or legal domain. A possible solution in these closed domains is to fine-tune the model on domain specific data, and approaches such as Sci-BERT (5), BioBERT(6) have shown that fine-tuning can greatly improve performance in these closed domains. Thus, we fine-tune BERT for the SEBI domain to create “SEBI-BERT". We fine tune on a corpus consisting of SEBI regulations, SEBI case-files as well as a collection of financial and SEBI related news articles. The facts of the case can be large, and thus will not fit in the BERT context span. To solve this, we use the sliding window technique. More details can be found in (7).

3.3 Regulation Biography

SEBI provides us information about all regulatory documents including their amended versions and additional supporting documents related to the domain of investment banking in India. This mine of information can be best understood when analysed, organised and tagged using various NLP methods. Our work offers a comprehensive view of a regulation as well as its metamorphosis over time in the form of (1)visualisation that identifies what changes have been made between two successive versions of a regulation document (2)provides additional information extracted from SEBI documents like Annual reports and Concept papers that help understand the rationale behind the amendments made (3)provides tags in order to categorise the kind of amendments, etc using rule based methods (4)identifies references to the regulations in news Articles and SEBI related case files. It also identifies references of case files in news articles. 22091 news articles relevant to SEBI, 4974 news articles relevant to SEBI cases and 7406 adjudication orders were used to build this module. More details can be found in (8).

4 Challenges, Benefits and Learnings

We discuss the challenges faced, the benefits of the partnership and our key takeaways below.

- **Lack of Labeled Data**: The unavailability of labeled data for evaluation, let alone training was a recurring issue. The complexity inherent in legal language and the diversity in tasks required thoughtful consultation and careful annotation treatment, available only through wider expertise.
- **Avalanche Effect**: A small change in the input often resulted in massive deviations of the output results. For instance, two case-files may have large overlap between the facts of the case, but a single difference in the fact can result in one case-file violating a regulation while the other not violating. Hence models had to be well calibrated and thoroughly reviewed.
- **Domain Specificity**: The legal domain uses precise terminologies and even a small misunderstanding can have enormous consequences. Standard NLP techniques often do not perform as well, and domain specific solutions had to be built even for primitive tasks such as entity recognition. These require sustained efforts from a sizable group of researchers.
- **Overlapping Tasks**: Large systems such as ASEA often have overlapping sub-problems and it is important to design modules that are re-usable. This can amortize the efforts spent on a particular module and improve the overall robustness. However, this may result in arriving at a trade-off between performance and complexity.
- **Competing Priorities**: With new advances in AI appearing frequently, researchers can get distracted by the urge to consider techniques that are narrow and complicated. On the other hand, industry practitioners may want solutions that are widely applicable. Collaborations must carefully manage the divergence in priorities and interests.
- **Targeted Solution**: Our approach of introducing solutions to specific problems such as penalty estimation, violation detection and regulation biography tracking, all in a framework based on extracting semantics for analytics, can serve as a template for breaking down problems in related areas and indulging in research that has broad interest.
- **Flexible Planning**: When developing complex systems, despite thoughtful approaches, unexpected outcomes are a reality. For instance, the number of required annotations may prove to be insufficient or the model doesn't sufficiently generalize. Hence any collaboration effort must account for such *unknown unknowns*, and be flexible and ready to improvise.
- **Complementary Strengths**: The cross-fertilization of problem comprehension in industry with the ability to produce state-of-the-art research in academia can be mutually productive and beneficial. Hence collaborations benefit from partners with symbiotic pairings.

Acknowledgments

This work has been supported by J.P. Morgan AI Faculty Research Award. Any opinions, findings, and conclusions in this paper are those of the authors only and do not necessarily reflect the views of the sponsors.

References

[1] I. Spasić, F. Sarafraz, J. A. Keane, and G. Nenadić, “Medication information extraction with linguistic pattern matching and semantic rules,” *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 532–535, 2010.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[5] I. Beltagy, K. Lo, and A. Cohan, “Scibert: Pretrained language model for scientific text,” in *EMNLP*, 2019.

[6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, Sep 2019. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btz682

[7] U. Narayan, P. Parikh, K. Karlapalem, and N. Raman, “Detecting regulation violations for an indian regulatory body through multi label classification,” in *Companion Proceedings of the Web Conference 2022*, ser. WWW ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 610–614. [Online]. Available: https://doi.org/10.1145/3487553.3524640

[8] S. S. Buggana, D. Saravanan, S. Kanchi, U. Narayan, S. Mangale, L. T. Thomas, K. Karlapalem, and N. Raman, “SEBI regulation biography,” in *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini, Eds. ACM, 2022, pp. 598–603. [Online]. Available: https://doi.org/10.1145/3487553.3524638