Applying NLLP and ML to Predict Damages as a Remedy for Contract Breach

Frank Giaoui
Columbia Law School
New York, NY, USA
frank@optimalexsolutions.com

Luv Aggarwal
Columbia University
New York, NY, USA
luv@optimalexsolutions.com

Diego Lobo
New York Law School
New York, NY, USA
diego@optimalexsolutions.com

Joan Gondolo
La Sorbonne Law School
Paris, France
joan@optimalexsolutions.com

Philippe Lachkeur
Institut International de Management
Paris, France
philippe@optimalexsolutions.com

Satvik Jain
Columbia University
New York, NY, USA
sj2995@columbia.edu

ABSTRACT

Motivated by the subjective decision making and lack of strict protocols in damages as a remedy for contract breach, this project uses natural legal language processing (NLLP) and artificial intelligence (AI) techniques to analyze patterns in contract law cases and reduce uncertainty in their outcome.

A 'hybrid' model combining heuristics, NLLP & the results of an LSTM based model into an XGBoost regressor along with contextual information had the best performance for the classification of entity types from unstructured proceedings text. Linear regressors were developed to approximate the Recovery Rate and the Win Rate using a set of 6 engineered features likely to affect the outcome.

ACM Reference Format:
Frank Giaoui, Luv Aggarwal, Diego Lobo, Joan Gondolo, Philippe Lachkeur, and Satvik Jain. 2023. Applying NLLP and ML to Predict Damages as a Remedy for Contract Breach. In *Nineteenth International Conference on Artificial Intelligence and Law 2023 (ICAIL 2023), June 19–23, 2023, Universidade do Minho, Braga, Portugal.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3594536.3595119

1 INTRODUCTION

One of the most consistent problems with lawsuits has been the subjective decision making of jury and no strict adherence to protocols which lead to uncertainty in the outcome of cases commonly seen in the sphere of criminal law where even similar cases and arguments have been given different judgements. Another domain which is plagued by this issue is Contract Law [4]. It has been observed that two claimants seeking compensation for damages under similar circumstances are granted different amounts.

Traditional approaches used heuristic approaches to derive conclusions, however with the advent of artificial intelligence, natural language processing techniques are being employed for information retrieval and for identifying patterns in the text data.

This research on reducing the uncertainty in the outcome of contractual damage lawsuits is being carried out as an independent project by the lead researcher, Frank Giaoui with the technical support of students interns from the Columbia University [2]. To find patterns in such lawsuits, the researchers have to disintegrate each case to its roots and analyze each fact from basics [3].

The current phase 1) automates the creation of a vastly larger legal database suitable for predictive modeling; and 2) extrapolates the correlation between the features and the outcome to validate prior findings and identify new patterns.

2 BACKGROUND

The lead researcher of the team, Frank Giaoui has completed his PhD and JSD dissertation on the 'Valuation of Damages for Contract Breach' [1]. Assisted by a team of researchers, he analyzed 300 cases from US, French and International Law, trying to find patterns in contractual lawsuits. The empirical analysis from the dissertation seeks to formulate strong relationships between certain inferred features and the outcome of the lawsuits. To validate the correlation findings, the team aimed to extrapolate the correlation between the features and the outcome such as 'Probability of Grant' and the 'Grant/Claim Ratio' to a larger corpus of 8000 files of raw texts.

Two of the key features are the 'Claim Value' and 'Sophistication Index'. 'Claim Value' is the amount which the claimant asks as compensation for damages caused by the breach of contract. 'Sophistication Index' of a case is reflective of how elaborately and concretely was the claimant able to justify the amount being claimed. The following topics are covered in the next sections:

(1) Identification of grant & claim sentences, extraction of the quantum and study of the correlation with the outcome
(2) Identification of sophistication sentences, computation of the sophistication index and study of the correlation with the outcome
(3) Development of a mathematical equation to study the impact of 6 features from section 3.4 on the outcome of the case

3 METHODOLOGY AND RESULTS

3.1 Building an annotated corpus

From the sample of 300 case proceedings published by the courts and collected by Westlaw, 96 cases from US law were annotated manually to build a corpus of 8500 sentences across 48 classes for context analysis. As a sentence can belong to multiple classes, they are regrouped based on the type of context analysis to be performed.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06.
https://doi.org/10.1145/3594536.3595119

3.2 Quantum Analysis

3.2.1 Contextual Analysis. A classification model is developed to differentiate between three categories of sentences 'Claim', 'Grant', 'FID' (First Instance Decision) and 'Others' (neither of the categories). NLTK library is used to tokenize each collection of annotated text to sentences of fixed word lengths. This expands the data set to 12,000 sentences. Next, the Gensim library is used to pre-process each sentence into a group of words which gets rid of punctuations or numerical values inside a sentence.

For training the models, word embeddings were needed. Prior work done on publicly available embeddings did not give promising results. The authors trained a word2vec model on a legal corpus of 8000 case proceedings. LSTMs and BERTs were trained on labelled sentences, split into stratified train and test set with a 70:30 split.

A drawback of the LSTM model is the loss of information like numerical values and symbols. BERTs are complex to train and require larger train sets whereas efficiency can be achieved using the structure in the legal language. A hybrid model is proposed where 26 features are engineered combining the results of an LSTM model with information such as presence of certain distinguishing words, location of sentences and presence of numbers to train further an XGBoost model. The hybrid model had the best classification performance for the primary classes with F-1 scores > 0.80 per class.

3.2.2 Extracting the grant and claim quantum values. Regular expression rules are used to extract the grant and claim quantum values from the sentences. A negative correlation between the claim and grant quantum values is observed confirming the results of the previous research on the larger sample.

3.3 Sophistication Index Analysis

'Sophistication Index' is used to quantify how elaborate is the argument and justification presented for the claim quantum value demanded by the claimant which can have a direct impact on the grant value awarded against the claim.

3.3.1 Identification of Sophistication Sentence. A model similar to the quantum model was trained to classify between three different categories of legal sentences: 'SOPHISTICATION', 'LAC' (Legal Argument Claimant) and 'Others' (neither of the categories).

It was seen that the hybrid model performs extremely well on the classification task achieving an F-1 score > 0.85 for the classification of 'SOPHISTICATION' sentences.

3.3.2 Sophistication Index Classification. Next step was to extract information from the sophistication sentences identified in the previous step and then use that information to categorize the cases into different indices.

The following was the definition of indices as developed by the legal experts:

(1) Index 4: Multiple unique methodologies present
(2) Index 3: Single unique methodology present
(3) Index 2: No methodology and multiple claim values
(4) Index 1: No methodology and single claim value
(5) Index 0: No methodology and no claim value

The index classification model was tested on the set of 96 cases achieving an F-1 score > 0.84 for each of the indices.

3.3.3 Correlations between Sophistication Index and Outcomes. The Average Grant/Claim ratio and the probability of Grant showed a positive increasing trend from Index 1 to Index 4, which is expected since the amount granted against a given claim and the chance of getting some grant should increase with increase in the sophistication of the claimant's argument.

3.4 Equation Analysis

This section focuses on developing a mathematical model to study the correlation between the outcomes (Probability of Grant and Grant/Claim Ratio) and the following set of 6 features (or a subset of them) as hypothesized by the legal team - claim, sophistication index, business risk, reputation of law firm, length of negotiation, size of law firm.

A linear model is used initially to approximate the 'Grant/Claim Ratio' and 'Probability of Grant' for better interpretability of the importance/coefficients of the features used in the equation.

To be able to assess the quality of the mathematical equation formulated, the R2 score was computed for the different experiments performed. Further, the coefficients of the features were also computed for each feature to determine the importance and contribution of each feature in the final outcome.

Among the various experiments performed, the following were the best R2 scores achieved:

(1) Probability of Grant: The best R2 score achieved using a linear model was 0.998 . This was using the following features: Claim, Sophistication Index & Reputation.
(2) Grant/Claim Ratio: The best R2 score achieved from a linear model was 0.953 using the following features: Claim, Sophistication Index and Length of Negotiation.

4 DISCUSSION

This report is focused on the analysis of two of the key features which have an impact on the outcome of the case, namely the 'Claim Value' and 'Sophistication Index'. The work done by the team has confirmed and further corroborated the hypothesis of the legal team by evaluating the relationship of these features on the outcomes for a set of 8000 cases. Further, the team has also generated promising initial results in the use of linear models to describe how the features/subset of features can determine the Grant/Claim Ratio and Probability of Grant.

Future work for the team would focus on extending the analysis done for Claim and Sophistication Index to other features as well.

REFERENCES

[1] Frank Giaoui. 2019. *Indemnisation du Préjudice Economique.* Harmattan, Paris, France. 724 pages. https://isbnsearch.org/isbn/9782343180670
[2] Frank Giaoui. 2019. Une évaluation innovante des dommages et intérêts pour traduire les faits en règles de droit et réduire l'imprévisibilité judiciaire. In *La Revue des Contrats*. Vol. 1. Lextenso, Paris, France, 164–181. https://www.labase-lextenso.fr/revue-des-contrats/RDC115y3
[3] Frank Giaoui. 2020. Towards Legally Reviewable Damage Awards. *Corporate and Business Law Journal* 173 (2020), 173–229. http://cablj.org/wp-content/uploads/2020/02/W2020-FINAL-F_Giaoui-.pdf
[4] Frank Giaoui. 2022. Breaches of Agreements to Negotiate: A Comparative Analysis of Damages. *American Journal of Trade and Policy* 9, 2 (2022), 77–98. https://doi.org/10.18034/ajtp.v9i2.623