Automatic Classification of Legal Documents

Han Bit (Hailey) Yoon
Department of Computer Science
Stanford University
hbyoon@stanford.edu

Abstract

This paper examines the performance of several machine learning models in the classification of legal documents, based on existing court cases. I find that simple model that take word frequencies as features perform well when detecting the violence level of the crime. However, this method deteriorates when we move to classify its crime type—inchoate crime, statutory crime, or personal/property crime. In this setting, utilizing k-Nearest Neighbors for decision and classification produces the most accurate results.

1 Introduction

There is a great need for lawyers for low-income people who cannot afford legal representation due to the continually rising price of legal service. The United States Attorneys Office (USAO)'s matrix of reasonable hourly rates for attorneys shows that amount has nearly doubled since 2003; the market rate in 2003-2004 is $180 to $380 [1], but the price in 2018-2019 is $307 to $613 [2]. I believe that we can make legal access more affordable by utilizing artificial intelligence (AI). AI technology can assist attorneys to minimize time spent on monotonous tasks and reduce hours billed. To be more specific, it can be achieved through the automatic classification of legal documents.

The primary objective of this project is to automatically classify legal cases into three major crime types. It takes a set of legal documents, which can be a court case, a client inquiry, or a news report. Then, the model produces an output of a list of matching crime types. The result will show "Inchoate Crime" for any attempted and incomplete crime, "Statutory Crime" for a violation of specific state or federal statute (such as selling alcohol to a minor), or "Personal/Property Crime" for offenses against the person or property (such as assault, homicide, and robbery).

2 Literature Review

Recently, there has been a surge of research aimed at benefiting law firms and the legal profession. One of the earliest publications that leveraged artificial intelligence concept was Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning in 1957 [3]. Rissland attempts to use a computer program to perform tasks in legal reasoning and argumentation, such as generating favorable and contrary cases. Another group aims to use intelligent concept extraction to create a judicial search tool [4]. Specifically, the presented model functions as a client-based legal agent and allows the user to find useful legal cases over file systems. Nevertheless, prior work allows the user to only computer a single task per time, which can be improved to take a more extensive set of data and function more efficiently in terms of both accuracy and time. This is the challenge I tackle in this project.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

3 Infrastructure

3.1 Dataset

For the task, I am using the dataset provided by `www.casebriefs.com` , which consists of various legal cases in the United States. To make the problem human-understandable, we only included legal claims with case text with downloadable PDF and clear violations, filtered the publication to be published, and limited to 109 legal cases in total. Here, we need to note that one PDF may contain multiple cases (up to 150 cases). This yields 5,991 pages with 5,391,972 words of case briefs. To reduce the data size, I converted all pdf files into text files, then parse the reshaped set. To get a better understanding, we can look at a sample dataset:

Input – a text file of *United States v. Drew*

... Lori Drew is the Midwestern mother who allegedly participated in a hoax on the social-networking website MySpace that ended in the suicide of junior high student Megan Meier. Meier was formerly friends with Drew's daughter until the two had a falling out. There were also allegations that Meier had acted in negative ways toward Drew's daughter. In retaliation or perhaps as a prank, Drew collaborated with her daughter and Drew's former employee in creating a fake profile of a 16-year-old boy. Using the profile, they friended, befriended, flirted, and started an online relationship with Megan Meier. After some time, the messages became nasty...

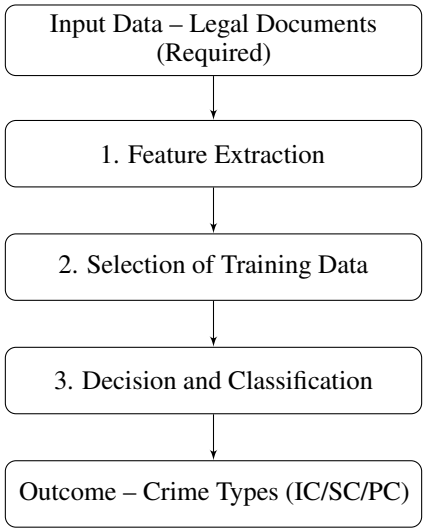
Output – Crime Category

Category: Personal/Property Crime

4 Approach

Automatic Classification of Legal Documents runs three fundamental processes (Figure 1) to achieve its purpose. Figure 1 shows that it requires a user to provide a set of input data, then parses it to detect its crime type. The source code is available at `https://github.com/hay318/Automatic-Classification-of-Legal-Documents`.

Figure 1: Pipeline Overview



4.1 Oracle: Human Judgement

The oracle consists of manual prediction by the author of unprocessed documents. I attained 91% accuracy on 109 randomly-sampled legal cases. I noticed that differentiating inchoate crime against statutory crime and personal/property crime was a major source of human classification error.

4.2 Baseline: Logistic Regression

For the baseline, I used logistic regression with $\text{Loss}_{logistic}(x, y, w) = log(1 + e^{1(w \cdot \phi(x))y})$.

While a simple model with limited subfields performs well with logistic regression, it easily results in overfitting to the training data. In fact, baseline attained 55% accuracy on 109 randomly-sampled legal cases.

4.3 Feature Extraction

I defined and extracted five groups of features from legal documents, shown in Table 1. It extracts critical information within legal documents. The five feature groups play a significant role when deciding the crime type since low correlation will highly likely be IC (attempted crime), the medium level will be SC, and high level will be PC.

Table 1: Feature Groups

Feature Group	Features
Weapons	firearm, gun, sword, knife, ... == 1
Criminal Intent	basic, specific, direct, oblique, ... == 1
Multiple Victims	students, pedestrians, family, children, ... == 1
Victim Weight	w $\leq$ 100
Violence Weight	w $\leq$ 100

4.4 Decision and Classification

Here, I describe the various models and methods used in approach to achieve desired outcomes; specifically, strategies include Naïve Bayes, k-Nearest Neighbors, and Support Vector Machines.

4.4.1 Naïve Bayes

The first approach that I tried was Naïve Bayes since it tends to perform well when there are specific keywords that strongly correlate to certain classes. Also, since the project takes a large volume of inputs, my first attempt focused on using eager learning classifier with fast speed. I was able to classify the documents by the group that gives the highest probability by utilizing Baye's rule:

$$PosteriorProbability = \frac{(Likelihood \cdot ClassPriorProbability)}{PredictorPriorProbabilty}.$$

However, unlike other approaches, it fails to capture the relationship between words since it assumes all words are independent; this attained accuracy lower than the baseline.

4.4.2 Support Vector Machines

I attempted to improve the decision and classification by training data to find the optimal decision boundaries. Support Vector Machines may seem computationally intensive, but the trained data will be beneficial when it detects new unlabelled keywords. As I expected, the Support Vector Machines approach (66%) outperformed the Naïve Bayes approach (42%) .

4.4.3 k-Nearest Neighbors

To create the model that works more efficiently when computing a large volume of input data, I decided to implement k-Nearest Neighbors approach that is effective in case of large number of training examples. This approach requires the longest running time. However, it generates the highest accuracy (74%).

5 Analysis

Table 2: Experiment Results

Model	Accuracy(%)
Baseline	55
Naïve Bayes	42
Support Vector Machines	66
k-Nearest Neighbors	74
Oracle	91

Table 3: k-Nearest Neighbors Result Matrix

Crime Type	IC	SC	PC	Accuracy(%)
IC	9	5	1	51
SC	3	36	4	72
PC	2	2	47	84

Table 2 and Table 3 are the results of this project. In Table 2, kNN shows the highest accuracy (74.33%) out of all three "decision and classification" strategies. Table 3 illustrates that the model successfully categorized most of the statutory crimes (72%) and personal or property crimes (84%). However, inchoate crime shows significantly low accuracy (51%).

After analyzing the results, I noticed that legal cases are easily misclassified to another category when it's only attempted and not committed. To further increase the accuracy of inchoate crime, future training sets should include more non-violent personal/property crimes such as fraud, tax crimes, and gambling. Furthermore, the results suggest that it is not possible to do much better than 74% on the 3-way subfield prediction task because the subfields may overlap. In fact, the accuracy drastically increases to 85% when there is only one field.

It is also apparent that legal cases are easily misclassified to another category when it's only attempted and not committed. Perhaps adding non-violent legal cases to training sets would increase the accuracy. Similarly, unique crime scenarios is a major factor that decreased the accuracy. There are legal cases with multiple criminal charges, but the current model classifies a document into one specific category. To solve this issue, each model generates the most substantial crime category.

6 Discussion and Conclusion

This project attempts to take advantage of various AI concepts to categorize legal documents automatically. The model takes a set of legal documents, then classify its crime-type accordingly. The project would benefit both lawyers and clients in terms of time and cost. The results are intriguing. It indicates that it is possible to obtain reasonably good accuracy when automatically classifying legal documents using keywords within the text. This can be useful when one would like to confirm the crime type of on-going cases.

In the future, I would like to continue improving the predictive power of this model by implementing an enhanced recurrent neural network (RNN). Utilizing an RNN with multiple layers may improve performance. Also, this project can be extended in many ways. It would be interesting to include more subcategories such as federal law versus state law, and generating example court cases would be useful.

References

[1] Justice.gov. (n.d.). *LAFFEY MATRIX — 2003-2012*. [online] Available at: https://www.justice.gov/sites/default/files/usao-dc/legacy/2011/07/06/civil_Laffey_Matrix_2003-2012.pdf [Accessed 20 May 2019].

[2] Justice.gov. (n.d.). *USAO ATTORNEY'S FEES MATRIX — 2015-2019*. [online] Available at: https://www.justice.gov/usao-dc/file/796471/download [Accessed 20 May 2019].

[3] Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 Yale L.J. (1990).

[4] Osborn, J. and Sterling, L. (1999). *The seventh international conference on artificial intelligence and law*. New York: ACM, pp.173-181.