Predicting Judgment decisions using Natural Language Processing

Group 41

April 20, 2018

Introduction

Legal institutions in most countries suffer from significant delay due to large number of cases. This is not only an issue in law & ethics but a research problem which comes under the purview of legal engineering. With modern tools and computational abilities we can look into using Natural Language Processing(NLP) as a guiding mechanism for legal systems and influence the productivity of our over-burdened courts.

Case for India

The courts of india are relatively overburdened. The New Delhi's High Court (2009) observed that the existing backlog of cases would take another **466 years** to come to a verdict! The Hindu estimates the number of cases pending with Indian courts to be around **30 million**. The Law Commission, 1987, estimated that there are 10 judges for every million population of India. With increasing population the cases with the courts has increased drastically while the judge to population ratio did **not** increase to the recommended level. All this together stanches the flow of justice.

With state of the art Machine learning techniques in NLP our aim, is *'to accomplish human-like language processing'* (Elizabeth D. Liddy,2001.) As the applications of NLP continues it has accommodated many human-intensive work. Machine Translation extends the data availability to everyone by breaking the language barrier. As the financial markets are moving towards the algorithmic trading, the analysis of plain text of news becomes important. Again, NLP provides the tools to Extract Information from various sources which can then be used to make trading choices. One most human-intensive work where the NLP is penetrating is Question Answering. The focus remains on recognizing the question, extracting the meaning and then provide the appropriate answer. With AI shifting the problems of human-intensive functions towards machine-intensive computations, its tools can be used to reduce the burden of the legal courts by introduction of NLP based mechanisms in the system.

The paper has the following structure Section 2 specifies the Objective, Section 3 covers Literature Review, Section 4 has data and materials and their sources, Section 5 Methodology. Results are mentioned in Section 6. Section 7 has Conclusion and Scope of further research .

Objective

Predicting the outcomes of cases which are under the jurisdiction of **European Court of Human Rights**, regarding violations of Articles 3,5,6,8. Output will be a binary vector of Violation and No violation.

Literature Review

This kind of analysis comes under the purview of **Legal Engineering**. Legal Engineering is the new field that studies the methodology and applies information science, software engineering and artificial intelligence to laws in order to support legislation and to implement laws using computers. Lawlor predicted that computers may someday be able to analyze and predict outcomes of judicial decisions (Lawlor, 1963). Previous work has focused on prediction of judges' votes given non textual information, (the nature and gravity of the crime or the preferred policy position of each judge) (Lauderdale & Clark, 2012). Aletras N. et al. (2016) predict Article violations for cases in European court of Human rights using linguistic representations: **Facts**, **Circumstances** and **Relevant laws**.

Material and Methods

European Court of Human Rights

The ECtHR was established in 1959 as an international court, by the European Convention Human Rights. The court's jurisdiction is on the applications of individuals or sovereign states alleging violations of the civil and political rights set out in the Convention. Since then it has expanded significantly and now covers forty-seven states in total, (population $\tilde{8}00$ million). By 1998, the Court ha set up as a full-time court and individuals can apply to it directly, if they argue that they have voiced their grievances concerning human rights by exhausting all effective remedies available to them in their domestic legal systems before national courts.

Case processing by the court

Most of applications are lodged are made by individuals. First, the Applications are assessed on some admissibility criteria (chiefly the exhaustion of effective domestic remedies).ng Passi this first stage, the case can be allocated to a single judge, a Committee or a Chamber. Most applications fail the admissibility stage. A text-based predictive analysis is not possible because cases held inadmissible are not reported. The cases analysed are the ones that have passed the first admissibility stage, so the Court gave a decisive hearing on these cases.

Main premise

This kind of analysis requires certain assumptions. Legal institutions limit the access to sensitive documents like individual submissions and lodged complaints. Thus published judgments as used as proxies for the inaccessible material. The following hypothesis is assumed: if there is sufficient similarity in the chunks of text of published judgments and lodged applications and briefs,then this approach can be used to predict outcomes with such kind of cases.

Case structure

The Court Judgments have are suited for text-based analysis. A judgment, generally, contains an account of the procedure followed, case facts, a summary of main legal arguments, and reasons stated by the Court. Judgments are labelled into different sections which aid in standardization of text and text-based analysis. Following sections have been analyzed in detail:

Procedure

This section contains the details of the procedure, starting with the individual complaint and ending with the court verdict, observed before the Court.

PROCEDURE

1. The case originated in an application (no. 47741/16) against the Russian Federation lodged with the Court under Article 34 of the Convention for the Protection of Human Rights and Fundamental Freedoms ("the Convention") by a stateless person, Mr Noe Georgiyevich Mskhiladze ("the applicant"), on 5 August 2016.

2. The applicant was represented by Ms Olga Pavlovna Tseytlina, a lawyer practising in Saint Petersburg. The Russian Government ("the Government") were represented by Mr G. Matyushkin, Representative of the Russian Federation to the European Court of Human Rights, and then by his successor in that post, Mr M. Galperin.

3. On 6 December 2016 the complaints concerning the conditions of the applicant's detention, the legality of his detention and its review were communicated to the Government and the remainder of the application was declared inadmissible pursuant to Rule 54 § 3 of the Rules of Court.

4. The Government objected to the examination of the application by a Committee. After having considered the Government's objection, the Court rejects it.

Facts

This section includes content which is beyond the purview of written law e.g., legal arguments. Subsequently facts in above are not just referred to as actions and events of the past as formulated by the Court, they may indicate a violation of a Convention article. 'Facts' has following subsections.

THE FACTS

I. THE CIRCUMSTANCES OF THE CASE

5. The applicant was born in 1972.

A. Administrative-offence proceedings against the applicant

6. The applicant arrived in Russia in 1988. He was subsequently convicted of criminal offences on several occasions. He was released on 3 December 2014 after serving his most recent prison sentence.

The circumstances of the case: The factual background of the case and procedure followed before local courts, before the application comes to the Court. It has content relevant to the applicant's story about the experience with state's authorities, and recounting of all actions and events that have allegedly given rise to a violation of the ECHR. The text has been formulated by the Court itself. Thus it should not always be considered as an unbiased factual background of the case. Domestic judgments reflect assumptions on the relevance of events. But it is at least framed in as neutral and impartial way. Thus clear displays of bias, (like failing to mention certain crucial events), are infrequent. Generally disputing parties agree on formulations in majority of cases, hence it is a (crude) proxy for a textual representation of the facts of a case.

Relevant law: It contains all legal provisions that can be relevant to deciding the case. Generally these comprise of provisions of domestic law.

The law

It concerns the merits of a case, through legal argument. Based on the no. of issues an application raises, it is divided into subsections that investigate an alleged violation of a specific Convention article. Court avoids a detailed examination, it frequently decides ones central to the arguments made.

THE LAW

I. ALLEGED VIOLATION OF ARTICLE 3 OF THE CONVENTION

32. The applicant complained under Article 3 of the Convention about the conditions of his detention in Krasnoye Selo. Article 3 reads as follows:

"No one shall be subjected to torture or to inhuman or degrading treatment or punishment."

Alleged violation of article x

It consists of the following two subsections

Parties' submissions: The Parties Submissions generally articulates the primary arguments made by the applicant and the state. Since most cases the facts are taken for granted,being proven by domestic courts, this part is mostly about legal arguments used by the parties.

Merits: It provides legal reasons that aim to justify the specific outcome reached by the Court. Usually, the Court reasons with a wider set of rules, doctrines and principles that have already been established in its past case-law. Thus it is mostly composed of legal arguments and factual information.

Operative provisions

This section consists of the outcome of the case regarding whether a decision on violation of an article took place or not.Can also have division of legal costs.

```
FOR THESE REASONS, THE COURT, UNANIMOUSLY,

1.  Declares the complaints concerning the conditions of detention from
15 December 2015 to 22 June 2017, and the lawfulness, arbitrariness and
review of detention from 26 January 2016 to 22 June 2017 admissible;

2.  Declares the remainder of the application inadmissible;

3.  Holds that there has been no violation of Article 3 of the
Convention;

4.  Holds that there has been a violation of Article 5 § 1 of the
Convention;
```

Data

Data-set has been created using the cases related to Articles 3, 5, 6 & 8. The selection of these article is based on a few criterion. Firstly, we could extract most data provided in the articles. No. of cases available were sufficient to run the chosen models. Both the conditions are satisfied by the chosen data.

Number of Cases verses Article of Convention:

We extract all cases of each article available with HUDOC. We only consider cases in English and parse them into various articles as described in the tables below. Firstly we parse the cases into the articles 3,5,6 or 8. The parsing results into total of 1151, 946, 1319 & 762 cases for articles 3, 5, 6 & 8 respectively. Table 1 shows number of cases and the right protected by each article.

Article No.	Article	No. of Cases
3	Prohibits torturous & degrading treatment	1151
5	Right to liberty & security	946
6	Right to a fair trial	1319
8	Right to privacy	762

Table 1: Table of no. of cases for each article

In the second parsing, we obtain total number of cases of Violation of Article and Not Violation of the articles for each article. This result is tabulated in table 2.

Description of Textual Features

Textual features are derived from each section of a case. The features constitute of tfidf vector and N-grams that is word sequences, word clusters and semantic

	Violations	Not Violations
Article 3	591	560
Article 5	509	437
Article 6	754	565
Article 8	411	351

Table 2: Violation and Not Violation cases for each article

structures.

Tfidf

The Term frequency–inverse document frequency (Tfidf) model (Jonas, 1972) is a popular word representation of test used in Information Retrieval and NLP. A tfidf model is the product of two statistics, term frequency and inverse document frequency. A high weight in tf-idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Tfidf features have been shown to be efficient in various NLP tasks Ramos 2003, T Joachims 1996. For our case we compute the tfidf values for all 4178 cases as well as article wise tfidf matrices. A sample feature matrix would be of 4178*(vocabulary size). We extract tfidf feature matrix for the Procedure (**Procedure**), Circumstances (**Circumstances**), Facts (**Facts**), Relevant Law (**Relevant Law**), Law (**Law**) and the Full case (**Full**) respectively.

N-gram

As with tfidf feature representation, N-gram features have been used for semantic representation of cases. N-grams have been shown to work in various NLP tasks Cavnar, Trenkle 1994; Lin, E Hovy 2003. We have computed all N-grams (where N = 5) for all cases (and article wise) and extracted the same features as for tfidf (Procedure, Circumstances, Facts, Relevant Law, Law and the Full case)

Topics and Important Words

We have two approaches in this section. First we cluster the N-grams (to reduce dimensionality) using the results in Aletras 2016. The paper has clustered N-gram features based on cosine similarity and labelled them in **Topics**. The topics are helpful in predicting individual article violations and non-violations.

In the **first approach** we use the weights provided by the results and proceed with our analysis. In our **second approach** we find the weights for all words using linear kernel SVM and we get a predictive score for an article violation. A positive score indicates article violation and vice-versa.

Classification Models

Neural Network

Neural networks are quite powerful learning models that can learn any non linear function. We use feed-forward neural networks which are non-linear learners. The non-linearity of the network, as well as the ability to easily integrate pre-trained word embeddings, often lead to superior classification accuracy (Goldberg 2016). We train a feed forward NN to learn the binary labels for tfidf and N-gram feature representations. The labels are +1 for violation and -1 for non violations.

Support Vector Machines

The objective of predicting article violations is a binary classification task. Thus we have to look for a violation of specific Article of the Convention. We use two textual features; tfidf and N-grams, to train Support Vector Machine (SVM) classifiers (Vapnik, 1998). SVM are state of the art machine learning algorithm known for high accuracy specially using small data sets (Joachims, 2002; Wang & Manning, 2012). A linear kernel is used to identify important features, indicative of each class using feature weights (Chang Lin, 2008).All the violation cases are labelled as +1, while no violation is 1. Positive weights indicate violation, while negative weights indicate no violation.

Prediction Scores

Using the weights calculated, we formulate a predictive score for each paragraph (via aggregation of the weights of constituent words). As with the word weights, a positive paragraph score suggests article violation while a negative score suggests no violation.
Calculating the prediction scores for all paragraphs in a document we use three approaches to interpret the aggregate score to a single prediction:

- **Min-Max**
  We sort the prediction scores for each paragraphs and use the sum of min and max.
  $$maxS_N + minS_N$$
  where N is the set of paragraphs
- **Aggregate sum**
  We take a linear summation of prediction scores of all paragraphs. A positive sum predicts a violation and vice-versa.
  $$S = \sum_N S_{para}$$
- **Weighted sum**
  We take a weighted mean of the prediction paragraph score using word size as the weight of each paragraph.
  $$S = \frac{\sum_N S_{para} * (wordsize)}{|N|}$$

A positive score indicates violation and a negative score indicates non violation

Prediction Results

Predictive accuracy

We calculate the prediction scores of all three feature representation with their respective classifier models on each article violation. Accuracy is measured as:

$$Accuracy = \frac{TV + TNV}{V + NV}$$

where TV and TNV are the number of cases classified correctly regarding article violation and Not violation. V and NV are the total number of cases with a violation and no violation respectively.

Table 3 summarizes the accuracies with respect to the topmost(most frequent) features corresponding to each of the three articles(3, 6 & 8). It is observed that the subsection "PROCEDURE" performs better as compared to the other two subsections in consideration and the full case representation. However in the case of Article 8 the subsection "FACTS" was the most accurate. Moreover it can be seen that the accuracies in general are low for all the articles and subsections, this can be attributed to the fact that the topmost features are not so predictive of the violation/non-violation of the articles. Adding to this the data set used in [1] was very less and hence the weights obtained might be questionable too.

Case Structure	Metric	Article 3	Article 6	Article 8
Procedure	Max/min	65.7	58.32	53.05
	Sum Compare	66.78	57.18	50.13
	Size Weighted Sum	66.51	55.74	49.73
Facts	Max/min	60.64	56.92	54.88
	Sum Compare	60.99	56.54	54.19
	Size Weighted Sum	60.18	56.39	53.92
The Law	Max/min	54.59	56.62	51.85
	Sum Compare	53.87	57.22	48.83
	Size Weighted Sum	53.96	56.92	48.84
Full Doc	Max/min	57.16	56.86	53.14
	Sum Compare	54.12	57.24	50.26
	Size Weighted Sum	55.25	57.31	49.88

Table 3: Accuracies for SVM weights extracted from [1]

Table 4 compiles the accuracies of subsections and different metrics using the most predictive features for violation and non-violation of articles and their corresponding weights obtained by training a model using SVM classifier and a linear kernel. The table suggests that the "PROCEDURE" subsection performs better in all cases while the performance of the "FACTS" subsection and "FULL" case representation was comparable. Also the performance of trained SVM weights and corresponding features was far better as compared to those obtained from [1] **Table 5** summarizes the performance of the "tf-idf" representation using Feed forward neural networks and SVM as classifiers.It can be

Case Structure	Metric	Article 3	Article 5	Article 6	Article 8
Procedure	Max/min	81.22	80.86	81.67	71.07
	Sum Compare	81.57	81.08	81.90	71.27
	Size Weighted Sum	81.40	80.65	81.52	72.07
Facts	Max/min	66.87	71.80	68.34	71.11
	Sum Compare	68.59	74.33	71.23	75.37
	Size Weighted Sum	68.86	74.10	71.68	73.31
The Law	Max/min	73.12	71.47	72.07	70.79
	Sum Compare	78.35	76.74	78.38	76.44
	Size Weighted Sum	75.02	73.76	77.54	74.49
Full Doc	Max/min	72.98	75.89	79.22	70.99
	Sum Compare	75.15	73.28	80.00	74.14
	Size Weighted Sum	72.54	79.06	79.57	74.27

Table 4: Accuracies for calculated SVM Weights

noted that the SVM classifier performed better than Neural Network although the accuracies were not that far.Also the performance of "Full" case representation was quite high in general.

Table 6 shows the results for the "N-grams" representation for the article 8

Classifier	Metric	Article 3	Article 5	Article 6	Article 8
NN	Procedure	92.48	92.92	93.41	92.97
	Facts	90.92	90.05	91.69	91.00
	The Law	91.80	91.23	92.77	92.04
	Full Doc	92.16	92.24	93.60	93.12
SVM	Procedure	96.31	96.41	95.74	94.28
	Facts	92.96	92.63	94.14	91.76
	The Law	94.86	94.04	95.05	93.41
	Full Doc	94.17	94.82	95.67	94.48

Table 5: Binary classification for Tfidf representations

with two different values of epochs

	No. of epochs	Accuracy
Article 8	2	63.54
	5	66.43

Table 6: Binary classification for N-Gram representations

Discussion

In contrast to the analysis done by Aletras 2016 which concluded that relevant facts has the highest predictive performance that resonates with he principles of legal realism, we find that with more enriched word representations like word specific prediction weights and tfidf features, **Procedure** outperforms other sections. Various explanation can fit this observation. The most relevant one seems to be that the section Procedure has the most concise description of the facts of the case, hence the most weighted words happen to be in this section, that is the section is **fact dense**. other explanations could be that outcome for a case is biased by the pre judicial treatment of the lodged complaints and the ruling of the domestic courts are good predictors of the outcome.

Conclusions

We present an introductory approach to predicting judicial decisions a legal institution using only the textual information from accessible sections of a case. It is framed as binary classification problem, where the training data consists of various word representations extracted from given cases and the output is the prediction of violation of a specific article of the convention. The model, in general, achieved a strong predictive performance. Along with the prediction, a number of relevant qualitative patterns can be deduced from the above results. The most significant observation of all seems to be that **the factual background of the case** (formulated by the Court) has the strongest predictive capacity in the outcome. Even as crude proxy and all the shortcomings, a robust correlation between the outcomes of cases and **insert** resonates well with other research work and backs assumptions of Legal Engineering

In retrospect, we find that this kind of approach opens doors to future work with enriched word representations and more access to legal documents from various kinds of courts all around the globe. Data access issues lay a significant obstacle in extending this kind of introductory analysis. It is in the interest of judicial institutions to increase access to legal documents under privileged binding contracts to encourage further research in the field.

Further extensions

Circumstances subsection is a crude proxy for non-legal facts and Law subsection is a crude proxy for legal reasons and arguments. Improvised proxies can be formulated which offer more predictive power in predicting outcome of a case. The objective of **sentence prediction** has not been fulfilled in this project. The data set that have been used does not have much information about the sentence/imprisonment/damage recovery etc. The regression models are not able to learn the features given the information provided by the data set.

References

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

[2] Reed C Lawlor. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344, 1963.

[3] Benjamin E Lauderdale and Tom S Clark. The supreme court's many median justices. *American Political Science Review*, 106(4):847–866, 2012.

[1] [2] [3]