Baltic J. Modern Computing, Vol. 11 (2023), No. 4, pp. 580–590
https://doi.org/10.22364/bjmc.2023.11.4.04

Processing Big Data of Court Decisions ⋆

Vitaliy GOLOMOZIY, Yuliya MISHURA, Iryna IZAROVA, Tetiana IANEVYCH

Taras Shevchenko National University of Kyiv, Ukraine

vitaliy.golomoziy@knu.ua, yuliyamishura@knu.ua, irina.izarova@knu.ua, tetianayanevych@knu.ua

ORCID 0000-0002-3174-9781, ORCID 0000-0002-6877-1800, ORCID 0000-0002-1909-7020, ORCID 0000-0001-8550-8062

Abstract. Large data sets of state registers of judicial decisions are publicly available in almost all European countries, while the analysis of these data, often reaching millions of decisions, is not automated. We are developing a labeled dataset for automated data processing of Ukrainian court decisions, which will allow us to identify patterns and discrepancies in judicial practice in Ukraine using statistical analysis and machine learning methods.

Keywords: court statistics, processing of judicial proceedings, machine learning, effectiveness of justice

1 Introduction

Every person intends to protect her/his rights in case of their violation. Even if there is not a violation at the moment, everyone wants to be sure in the availability of such an opportunity, its reliability and efficiency. This has to be provided by the state system of justice. An effective mechanism of equal access to justice for all is the goal that modern states and open societies around the world seek to achieve. At the same time, the effective functioning of the state justice system is a factor that directly affects on its competitiveness and the successful economic development of the state and society. The transparency of the justice system and the openness of information about the progress of the case and the enforcement of the court decision are the foundation of public trust, which today is extremely necessary to renew and strengthen. The open Ukrainian register of court decisions does not ensure real transparency and openness of information

⋆ The work within the project "Innovative technologies for processing court decisions using machine learning algorithms" K-I-186, financed by an external instrument of assistance of the European Union to fulfill Ukraine's obligations in the European Union Framework Program for Research and Innovation "Horizon 2020".

about the administration of justice, and, therefore, it is not able to strengthen trust of the system in society. The efficiency indicators of the courts, such as relation of the amount of money has to be collected to those that has been collected, are approximately 0.1%, while the courts are overburdened and almost unable to effectively settle and prevent disputes.

Our goal is to develop the monitoring and data collection system based on the indicators, which will allow the fast and flexible detection of changes in the judiciary. In particular, it is proposed to create database using the Unified State Register of Court Decisions (`https://reyestr.court.gov.ua/`) and Open data portal (`https://cutt.ly/zwdojXRT`). This database is supposed to be suitable for the main stakeholders and provide the opportunity for statistical analysis of court decisions and producing the recommendations. With the help of a statistical multifactorial analysis and machine learning algorithms, it is planned to single out the main factors that affect the effectiveness of consideration of private legal cases by the court. At that moment the register of court decisions includes more than 108 million documents. They are actually text files that have to be transformed into the statistical database. And this is a real challenge that cannot be solved without the use of machine learning algorithms. Transformation of the nonstructural text data into the statistical dataset will allow utilizing it for analysis by scientist, journalists, state officers, politicians and anyone wishing to do it and make the judicial system really transparent.

Legal document analysis (we will call it Legal domain) is an important area of modern research. In recent years, many publications were related to various aspects of this problem. The papers (Katz et al., 2023) and (Zhong et al., 2020) summarise methods and approaches used when applying *natural language processing* (NLP) methods specifically to the Legal domain. The following papers (Dragoni et al., 2016), (Chalkidis et al., 2017, 2018), (Kano et al. 2018) and (Sleimi et al., 2018) are related to the extraction problem, such as extracting prohibitions, obligations or other rules from a legal contract or a document.

Generally speaking, in order to perform a statistical analysis of a text document, we have to find some categorical/numerical representation of the text. A typical approach is to use supervised learning methods which usually imply a necessity to hand-label some corpus of legal documents. Such labels could vary depending on the goal of a researcher.

In our case, we are interested in the prediction of a decision of a case trial, whether a case will be tried in an appeal court and what the outcome will be. Thus, our labels are categorical variables describing the aforementioned questions. The algorithms that are used in such situations are supervised classification algorithms, such as logistics regression, decision tree or random forest.

It is also possible to consider another type of labels. For example, one may be interested in finding pieces of a document describing some particular aspects of a contract, like a specific obligation. In this case we deal with a so-called named *entity recognition problem* or NER. This is a typical problem for a natural language processing. Usually, it is addressed with recurrent neural networks. Such algorithms are processing each entry token (typically a word) sequentially and generate an output symbol for each token. For example, assume we would like to extract every mentioning of a jury in a trial. Having

the sentence: “Your Honor, members of the jury, my name is John and I am representing the defendant in the case”, an NER algorithm will produce a sequence of output symbols like N, N, B, N, N, E, N, N ... Here “N” means that the token is neither beginning nor end of an entity under the question, “B” indicates a beginning and “E” indicates an end of an entity. Thus, in our example, the third word (“members”) is marked as “B” which means it is the beginning of an entity, and the sixth word (“jury”) is marked as “E”. This means that all words in between (“members of the jury”) compose an entity. In such a fashion, we can extract any fragments of text from a document.

Other typical tasks addressed by NLP are text summarization and question answering. Both are valuable and important in the Legal domain. Text summarisation allows generating a short summary of a long document or answering a question related to the text. These tasks are much more complex compared to NER or classification.

As we mentioned before, the typical algorithms for NER were recurrent neural networks. However, after a seminal work (Vaswani et al., 2017) published by a group of researchers from Google, recurrent neural networks have been superseded by transformers that is another class of algorithms that outperformed recurrent neural networks on NER and classification tasks and gave rise to a series of new algorithms, called “generative algorithms”, of which ChatGPT is the most notable example.

In this paper we follow a simpler approach. First, we develop a labeled dataset. The dataset contains both categorical and text data. Then we applied multiple classical machine learning algorithms to analyse categorical data. An analysis involving text data (and thus use of transformers) will be provided in the future works.

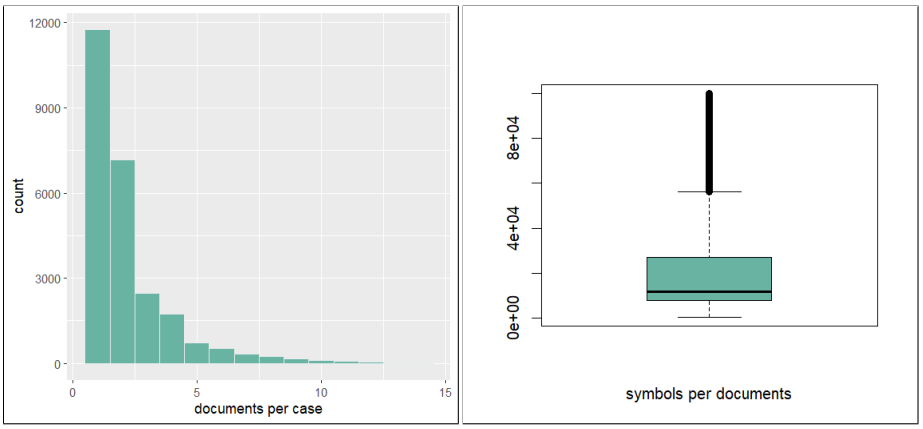
This paper organized as follows. Section 2 describes the structure of the dataset. Section 3 is devoted to the labeling methodology. Section 4 contains results of the statistical analysis.

2 Dataset structure

Developing a proper data set is an important step that determines the outcome of any analysis using machine learning algorithms. In the Legal domain of the natural language processing there are multiple benchmark datasets, such as (Duan et al., 2019), (Katz et al., 2023) and (Merchant and Pande, 2018). Using a benchmark dataset is good practice for evaluating a general-purpose algorithm. However, it is typical to develop a specific dataset for a particular, narrow purpose. See, for example, a dataset aimed to analyse contracts related to construction engineering at (ul Hassan et al., 2021). For our purposes, it is essential to have an algorithm capable for processing texts in Ukrainian. That is why we developed our own dataset.

The corpus of documents related to court trials consists of more than 100 million documents. The corpus is not available as a holistic dataset. Instead we can only download individual documents. We downloaded a sample of about 360 000 documents related to the year 2022 (of about 8 million for that year), which correspond to about 150 000 particular cases. The distribution of the number of documents per case is depicted on the histogram (Figure 1a).

Average number of symbols per document is 20000 with standard deviation equals to 18600. We can see the symbols distribution on the boxplot (Figure 1b).



(a) Number of documents per case (b) Number of symbols per document

Fig. 1: The features of cases

There is some additional data available. For each case we know an unique case identifier, judgement type, some very high-level categories. For each document we know its date, adjudication date, information about a judge. Everything else has to be extracted from the documents.

Document files are available in `rtf` format, so we had to convert them into `html`. We used `MySQL` database to save documents and cases.

3 Labeling mechanism

In order to use any supervised learning technique, whether classical or deep learning algorithms, we need the labels to train an algorithm. Usually, labelling is the most labour-demanding part of an NLP project (see, Hendrycks et al., 2021, Holzenberger et al., 2020).

To label a dataset, we created the special software which allows manual editing. We have made the labeling of more than 1000 cases and selected some categorical/numeric variables to analyze with the classic machine learning algorithms and some text variables for further analysis using deep learning NLP techniques.

Categorical/numberic variables are:

variable	type	range
decision	categorical	”Rejected”
		”Satisfied completely”
		”Satisfied partly”
		”Returned”
		”Left without consideration”
		”Closed”
appeal	categorical	”Cancel decision of the 1st instance”
		”Leave decision of the first instance in force”
		”Not appealed”
		”Partially annulled”
cassation	categorical	”Cancel decision of the 1st instance”
		”Leave decision of the 1st instance in force”
		”Not appealed”
proceedings	categorical	”General”
		”Simplified”
		”Imperative”
		”Combined”
immediate execution	boolean	No,Yes
court control	boolean	No,Yes
abuse rights	boolean	No,Yes
trial in absentia	boolean	No,Yes
trial termination	boolean	No,Yes
IDP participation	boolean	No,Yes
claim amount	numeric	0-∞
case starting date	date	date
case end date	date	date
expenses	numeric	0-∞

Text variables are:

variable	average length
plaintiff	12300
defendant	12300
claim	12300
claim justification	12300
response	12300
decision	12300
decision motivation	12300

Due to the poor categorization of the original documents, we added our own categories. We identified more than 50 meaningful categories, such as property rights or pension disputes. To preserve the homogeneity of the data, we restricted the cases under consideration.

4 Analysis

Categorial data at our disposal can be analysed using standard statistical tools like contingency tables, histograms, boxplots etc. We can make different statistical testing. But at this point we have only the pilot sample consisting of 1221 observation suffering of errors and absence of responses. So, the analysis presented below can be considered as demonstration of our abilities in the future.

So, we'll have the possibility to study the structure of the data according to different categories we are interesting in. The contingency tables like Table 1, can help to find out some features.

Table 1 is the crosstable for two variables which characterize the type of decision had been made in the first instance and the decision in the appeal court. Its cell content includes:

- Count
- Expected Values
- Chi-square contribution
- Row Percent
- Column Percent
- Total Percent

Statistics for all Table 1 factors within Pearson's Chi-squared test:
$Chi^2 = 358.0522$, d.f. = 24, p = 2.872414e-61.

This means that there is not sufficient evidence for independence of this two variables.

It is very interesting to analyze the consistency of the similar cases with regard to plaintiff's activity, that is, wether plaintiff submit or not the claim for consideration to the appeal court. This task is rather complex and requires the additional information from the case text. But even having only information from the Table 1 we can find out some interesting features. In particular, we can see that:

- out from 1221 cases 661 (54.136%) were not appealed and we don't have information about 331 cases (27.109%), so the real rate of cases appealed can increase up to 81.245%;
- out of 976 cases in which the decisions of courts of the 1st instance were satisfied and partially satisfied and we know it for sure, 18% were revised in the appeal court (177 cases);
- at the same time, among these 18% (177 cases) that were contested in the appeal court, in 17% of cases the decisions were canceled or partially canceled (17+15=32 cases);
- accordingly, 82% of the decisions of the first instance courts were upheld (145 cases).

In many situations it is worth to visualize such information using the plots. For example, the Figure 2 shows the distributions of cases with different decisions of the 1st instance according to the situation in the appeal court.

It is also important to analyse the duration of the case consideration in the courts. We have this information for 855 cases totally. It can be visualized with the help of

Table 1: Decisions in the first instance *vs* decisions in the appeal court

Decision of the 1st instance \ Decision in appeal court	NA	Cancel decision of the 1st instance	Leave decision of the 1st instance in force	Not appealed	Partially annulled	Row Total
NA	115 41.206 132.157 75.658% 34.743% 9.419%	0 3.984 3.984 0.000% 0.000% 0.000%	9 22.283 7.918 5.921% 5.028% 0.737%	27 82.287 37.146 17.763% 4.085% 2.211%	1 2.241 0.687 0.658% 5.556% 0.082%	152 12.449%
Rejected	13 20.603 2.806 17.105% 3.927% 1.065%	15 1.992 84.954 19.737% 46.875% 1.229%	22 11.142 10.582 28.947% 12.291% 1.802%	24 41.143 7.143 31.579% 3.631% 1.966%	2 1.120 0.691 2.632% 11.111% 0.164%	76 6.224%
Satisfied completely	126 165.907 9.599 20.588% 38.066% 10.319%	8 16.039 4.030 1.307% 25.000% 0.655%	76 89.720 2.098 12.418% 42.458 % 6.224%	399 331.312 13.829 65.196% 60.363% 32.678%	3 9.022 4.020 0.490% 16.667% 0.246%	612 50.123%
Satisfied partly	75 98.676 5.681 20.604% 22.659% 6.143%	9 9.540 0.031 2.473% 28.125% 0.737%	69 53.363 4.582 18.956% 38.547% 5.651%	199 197.055 0.019 54.670% 30.106% 16.298%	12 5.366 8.201 3.297% 66.667% 0.983%	364 29.812%
Returned	2 2.440 0.079 22.222% 0.604% 0.164%	0 0.236 0.236 0.000% 0.000% 0.000%	0 1.319 1.319 0.000% 0.000% 0.000%	7 4.872 0.929 77.778% 1.059% 0.573%	0 0.133 0.133 0.000% 0.000% 0.000%	9 0.737%
Left without consideration	0 1.084 1.084 0.000% 0.000% 0.000%	0 0.105 0.105 0.000% 0.000% 0.000%	3 0.586 9.934 75.000% 1.676% 0.246%	1 2.165 0.627 25.000% 0.151% 0.082%	0 0.059 0.059 0.000% 0.000% 0.000%	4 0.328%
Closed	0 1.084 1.084 0.000% 0.000% 0.000%	0 0.105 0.105 0.000% 0.000% 0.000%	0 0.586 0.586 0.000% 0.000% 0.000 %	4 2.165 1.554 100.000% 0.605% 0.328%	0 0.059 0.059 0.000% 0.000% 0.000%	4 0.328%
Column Total	331 27.109%	32 2.621%	179 14.660%	661 54.136%	18 1.474%	1221 100 %

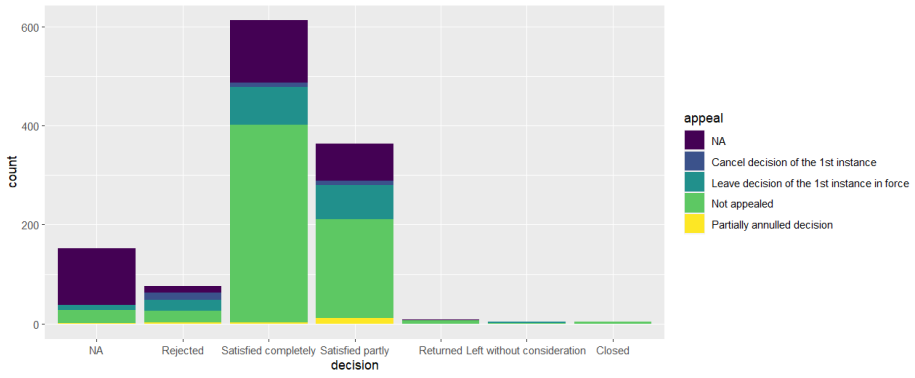


Fig. 2: Distributions of cases according to decision of the 1st instance and in the appeal court.

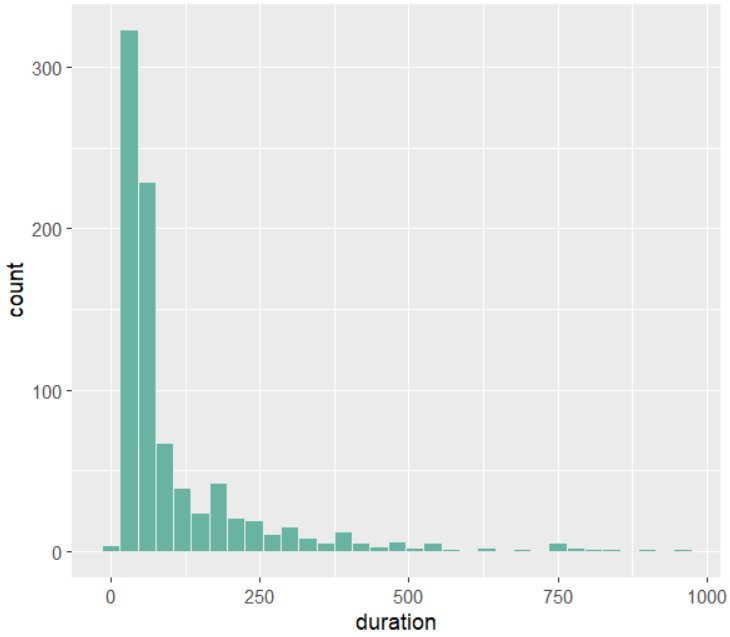
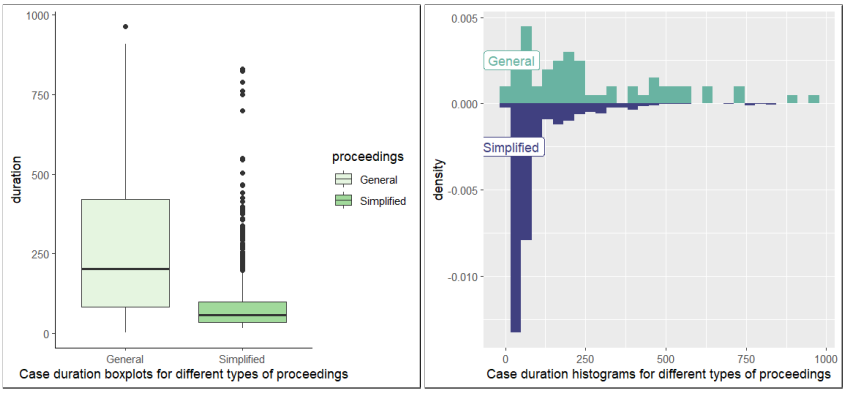


Fig. 3: Duration of the case consideration in the courts

histogram on the Figure 3. It is interesting wether there is a substantial difference in the case duration for different types of proceedings. There is no any case in those available 855 cases with the Imperative or Combined type of proceeding so we can compare only the cases which have the General and Simplified proceedings. Let's look at their boxplots (Figure 4a) and compare their duration histogram (Figure 4b)



(a) Boxplots for General and Simplified proceedings

(b) Histograms for General and Simplified proceedings

Fig. 4: Duration boxplots and histograms for General and Simplified proceedings

Table 2: Duration quantiles for the General and Simplified proceedings

General						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	2.0	81.0	201.0	268.2	420.0	963.0
Simplified						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	15.00	33.00	56.00	94.65	98.00	830.00

As it can be seen from the boxplots and histograms, we may expect that there are substantial differences in case duration for the General and Simplified proceedings. In particular we see that for General proceedings only 25% of cases last less than 81 days. For Simplified proceeding, 75% of cases have duration less than 98 days. We may expect to have in average almost 4 times less case duration for the Simplified proceedings than for the General ones.

Running the Welch Two Sample t-test results in:

t = 5.7993 days, df = 62.231, p-value = 2.401e-07 days

alternative hypothesis: true difference in means between group General and group Simplified is not equal to 0

95 percent confidence interval:[113.7405, 233.3828] days

Sample estimates: time differences in days

268.2131 – mean in group General

94.6515 – mean in group Simplified.

So, these data do not provide the sufficient evidence for equality of mean durations for the General and Simplified proceedings.

Moreover, for the specific , pre-selected tasks, we plan to design the interface generating some standard statistics. This will make it possible, for example, to find out whether there are specific circumstances (judicial errors) in the decisions of the court of the first instance taking into account the cases of the specified category that were reviewed in appeal and/or cassation. Such results will provide an opportunity to check the effectiveness and consistence of the judicial system.

5 Conclusions

The use of automated processing of big data sets of state registers of court decisions and court statistics data will provide the opportunity to use the special software for identification of persistent patterns and reasons for the inefficient functioning of the judicial system; predicting changes in the number of cases, the composition of participants, the amount of court costs and other circumstances that directly affect the proper functioning of the judicial system; ensure equal access to justice for all. It is not possible to obtain such data with the help of exclusively human resources, therefore, the possibilities of using machine learning algorithms should be more actively explored and applied; at the same time, the results of our research demonstrate the need for a more detailed analysis of the circumstances of the case and the determination of quantitative and qualitative indicators of the effectiveness of the case, the creation of a dataset that includes several thousand court cases (respectively, tens of thousands of decisions) to ensure reliable forecasting.

The labelled dataset we developed is aimed at boosting research in the area of legal text analysis and improving the overall functioning of judicial systems in different countries, where open access to national case law is provided.

References

Dragoni, M., Villata, S., Rizzi, W., Governatori, G. (2016). *Combining NLP Approaches for Rule Extraction from Legal Documents*, 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016), Sophia Antipolis, France, available at `https://hal.science/hal-01572443`.

Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., Wang, H., Liu, Z.(2019) *Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension*, available at `https://arxiv.org/abs/1912.09156`.

Chalkidis, I., Androutsopoulos, I., Michos, A. (2017) *Extracting contract elements*, Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law (June 2017), 19–28.

Chalkidis, I., Androutsopoulos, I., Michos, A.(2018) *Obligation and prohibition extraction using hierarchical rnns*, available at https://arxiv.org/abs/1805.03871.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Androutsopoulos, I. (2019) *Large-scale multi-label text classification on eu legislation*, available at https://arxiv.org/abs/1906.02192.

ul Hassan, F., Le, T., Lv, X. (2021) Addressing Legal and Contractual Matters in Construction Using Natural Language Processing: A Critical Review, *Journal of Construction Engineering and Management*, ASCE, **147**, No. 9.

Hendrycks,D., Burns, C., Chen, An., Ball, S. (2021) *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*, NeurIPS 2021 Datasets and Benchmarks Track (Round 1), available at https://arxiv.org/abs/2103.06268.

Holzenberger, N., Blair-Stanek, A., Van Durme, B. (2020) *A dataset for statutory reasoning in tax law entailment and question answering*, Materials of NLLP@KDD2020, August24th, SanDiego, US, available at https://arxiv.org/abs/2005.05257.

Kano, Y., Kim, M., Yoshioka, M., Lu, Y., Rabelo, J., Kiyota, N., Goebel, R., Satoh, K.(eds) (2018) *Coliee-2018: Evaluation of the competition on legal information extraction and entailment*, In: Kojima, K., Sakamoto, M., Mineshima, K., Satoh, K. New Frontiers in Artificial Intelligence. JSAI-isAI 2018, Lecture Notes in Computer Science, vol 11717, Springer, Cham. https://doi.org/10.1007/978-3-030-31605-1_14.

Katz, D. M., Hartung, D., Gerlach, L., Jana, A., Bommarito II, M. J. (2023) *Natural Language Processing in the Legal Domain*, available at SSRN: https://ssrn.com/abstract=4336224 or http://dx.doi.org/10.2139/ssrn.4336224.

Leivaditi, S., Rossi, J., Kanoulas, E. (2020) *A benchmark for lease contract review*, preprint available at https://arxiv.org/abs/2010.10386.

Merchant, K., Pande, Y. (2018) *NLP Based Latent Semantic Analysis for Legal Text Summarization*, 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, pp. 1803-1807, doi: 10.1109/ICACCI.2018.8554831.

Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Dann, J. (2018) *Automated Extraction of Semantic Legal Metadata using Natural Language Processing*, 2018 IEEE 26th International Requirements Engineering Conference (RE), Banff, AB, Canada, pp. 124-135, doi:10.1109/RE.2018.00022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017) *Attention is all you need*, Part of Advances in Neural Information Processing Systems, **30**, available at https://proceedings.neurips.cc/.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun M. (2020) *How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, available at https://aclanthology.org/2020.acl-main.466.

Received August 7, 2023 , revised September 29, 2023, accepted October 9, 2023