applied sciences

MDPI

Article

Evaluating Human versus Machine Learning Performance in a LegalTech Problem

Tamás Orosz [1,*], Renátó Vági [1,2], Gergely Márk Csányi [1], Dániel Nagy [1], István Üveges[1,3], János Pál Vadász [1,4] and Andrea Megyeri [5]

1 MONTANA Knowledge Management Ltd., H-1097 Budapest, Hungary; vagi.renato@montana.hu (R.V.); csanyi.gergely@montana.hu (G.M.C.); nagy.daniel@montana.hu (D.N.); uveges.istvan@montana.hu (I.Ü.); vadasz.pal@montana.hu (J.P.V.)
2 Doctoral School of Law, Eötvös Loránd University Egyetem Square 1-3., H-1053 Budapest, Hungary
3 Doctoral School in Linguistics, University of Szeged, Egyetem Street 2., H-6722 Szeged, Hungary
4 Institute of the Information Society, National University of Public Service, H-1083 Budapest, Hungary
5 Wolters Kluwer Hungary Ltd., Budafoki Way 187-189, H-1117 Budapest, Hungary; andrea.megyeri@wolterskluwer.com
* Correspondence: orosz.tamas@montana.hu

Abstract: Many machine learning-based document processing applications have been published in recent years. Applying these methodologies can reduce the cost of labor-intensive tasks and induce changes in the company's structure. The artificial intelligence-based application can replace the application of trainees and free up the time of experts, which can increase innovation inside the company by letting them be involved in tasks with greater added value. However, the development cost of these methodologies can be high, and usually, it is not a straightforward task. This paper presents a survey result, where a machine learning-based legal text labeler competed with multiple people with different legal domain knowledge. The machine learning-based application used binary SVM-based classifiers to resolve the multi-label classification problem. The used methods were encapsulated and deployed as a digital twin into a production environment. The results show that machine learning algorithms can be effectively utilized for monotonous but domain knowledge- and attention-demanding tasks. The results also suggest that embracing the machine learning-based solution can increase discoverability and enrich the value of data. The test confirmed that the accuracy of a machine learning-based system matches up with the long-term accuracy of legal experts, which makes it applicable to automatize the working process.

Keywords: legal tech; data analytics; artificial intelligence; Industry 4.0

Citation: Orosz, T.; Vági, R.; Csányi, G.M.; Nagy, D.; Üveges, I.; Vadász, J.P.; Megyeri, A. Evaluating Human versus Machine Learning Performance in a LegalTech Problem. *Appl. Sci.* **2022**, *12*, 297. https://doi.org/10.3390/app12010297

Academic Editor: Juan Pavón

Received: 13 December 2021
Accepted: 22 December 2021
Published: 29 December 2021

1. Introduction

Finding relevant court decisions is a cornerstone of legal research. It is a time-consuming part of the lawyers' job when preparing for a lawsuit. This mainly involves looking for arguments to convince the court to decide in favor of their clients [1,2]. These manual searches are often inaccurate [3]. Many pieces of research have been published examining the effectiveness of attorney teams. Blair and Maroon showed in their research that, although the attorneys thought that they found 75% of the related documents, they found only about 20% of them [4,5].

One reason for this difficulty is that legal documents contain a detailed description of the case, which uses a wide variety of language and synonyms to describe the same issues. Therefore, the human user has to use many possible combinations and synonyms of the keywords to find the connecting cases.

A good example of this is the case of an employee who committed complicity in smuggling. The employee sued his/her former employer for equal treatment violation and because of the non-payment of wages and cafeteria benefits, and he/she claimed that they terminated employment wrongfully. However, this was not a criminal case. When lawyers

receive such a case, they find themselves in a difficult situation to find similar judgments. If they use words that refer to the illegal smuggling of goods from a foreign country in their search queries, they will obtain mainly criminal and non-labor cases; if they look for termination of employment in general or violation of equal treatment, the result list is also likely to be misleading.

Categorization of the court decisions by their subject matter of the lawsuit can significantly improve the performance of these searches, and many research works have dealt with legal document categorization in the last years [6–11]. However, using human experts for this task is very time consuming and expensive because the documents are relatively long, usually containing thousands of words, and it is a multi-labeling task, meaning that one document can fit into more than one category [12]. Moreover, another research has shown that texts categorized in a binary manner (relevant/irrelevant for specific litigation) by two independent groups of human experts reached only 28% in F_1 score, agreeing on labels in only 70% of the documents [5,13–17]. Hence, human categorization often cannot be handled as a ground truth solution.

Many machine learning-based classification solutions have been published in the literature, but so far, no study has directly compared the performance of ML algorithms to humans in terms of accuracy, and reliability [6,7,18]. Guodong et al. [8] created a method for categorizing Chinese legal documents using Graph LSTM (long short-term memory) network [19–21] combined with domain knowledge extraction [22]. They compared their algorithm with the traditional classification methods of support vector machine (SVM [23,24]) and LSTM. Thammaboosadee et al. [9] made a classifier that uses a two-stage model to identify legal charges and the punishment range, given case facts and attributes, which could exceed 90% precision. However, these researches calculated the absolute accuracy and the absolute performance of the given solutions. Legal firms and companies want to know when the machine learning performance can reach or even surpass human level performance and implement it in their business processes.

Significant research has been conducted in wide variety of other fields that compared the accuracy and performance of human and automated classification ([25–32]). Generally, more and more AI-based solutions are created to replace human activities for industrial applications [29,31,32]. Goh et al. [25] used the support vector machine (SVM) algorithm to classify European Research Council Starting Grant project abstracts and compared the results to human labelers. They found that while the best human classifiers can outperform the algorithm, on average, the algorithm is more accurate and more reliable than human classifiers. The results also showed that using a machine learning algorithm is a cost-effective method to classify different texts. Simundic et al. [27] compared automated detection and visual inspection of preanalytical interference, such as lipemic, icteric, and hemolyzed samples. They found that human inspection is unreliable and automated system should be a standard protocol. Weismayer et al. [28] compared the categorization of TripAdvisor reviews by traditional manual content analysis and fully automated domain-specific aspect-based sentiment analysis tools. They found that the automated tools can analyze the reviews better, and the manual analysis is more time consuming.

This survey compares the performance of humans and machine learning-based algorithms on a multi-labeling task, namely, the classification of jurisprudence documents by their subject matter. The goal of the survey is to highlight when and how a machine learning-based application can be applied in business processes: when these methods can replace humans in data annotation tasks, and how can they improve the quality and the discoverability of a legal database. This experiment differs from the previous ones in the way that the participants had to read relatively long texts, and every document could be categorized into multiple classes. The performance comparison of human versus machine learning methods on the classification of long texts into multiple categories is an open question in research and an interesting question for firms in deciding when and how machine learning-based methodologies can be implemented into their business processes.

2. Materials and Methods

2.1. Research Questions

From the business point of view, the most interesting questions are regarding when the machine learning algorithm-based classifier reaches human-level performance and how these algorithms can be applied in business processes to accelerate the work or increase the discoverability of documents.

A study was designed to answer the following five major questions:

- How much time would the human categorizers need to label the whole dataset (about 170,000 documents)? How could this work be accelerated by the assistance of the computer?
- How much information would a human expert find with or without the aid of the machine learning classifier?
- Are machine learning algorithms more reliable than humans for classifying legal documents?
- Can machine learning algorithms hide the differences between the performance of legal experts and laymen or non-expert lawyers?
- How much is the inter-annotator agreement between legal experts on a specific task?

2.2. Study Design

The legal system in Hungary is a limited precedent-based system, and the judicial practice formally distinguishes six different groups of matters, in other words, law areas: criminal law, military criminal law, administrative law, labor law, civil law, and economic law. The published court decisions counted more than 170,000 documents when the research was done. These documents are relatively long. An average text contains 3330 words. The published case law is entirely in Hungarian, due to the special agglutinative property of the Hungarian language, which makes most natural language processing tasks quite difficult [33].

We selected 220 documents for this survey. These documents were pre-labeled and cross-checked by legal experts. We used this test set as a reference for further evaluations. It was an important point to select a similar amount of documents from the six different groups of matters by the following two different aspects. Firstly, it has only one exact solution and it can be classified easily; secondly, it has many possible categorizations, and it is very hard to find both.

We selected a roughly similar set of documents, where only one label and another set with multiple labels could be added. Moreover, we chose an equal proportion of rare categories, where there was little training data for training the algorithm, and common categories, where there were many documents for training data for the machine learning algorithm. We did this in order to simulate the real working conditions and the effect of the monotonicity of the task, the fatigue, and the different learning patterns of the humans had on performance [5,34]. During the labeling process, the participants had to proceed in the same fixed order. During the sorting of the test sets, we put the hardly categorizable decisions after a similar, simple case.

The participants had three hours to label as many documents as they could. There were 18 participants involved in this study with three different competence levels:

- Laymen: Never received formal legal training in their life, so they were not a student of any law university and had not received any law-related training. They only met law in their everyday life.
- Lawyers: At least fourth-grade law students or people with law degrees.
- Legal editors: Legal database editors employed by Wolters Kluwer Hungary, whose task is to categorize legal documents and manually enrich them with other metadata.

Every group was composed of six people, and they were divided into two subgroups. The first subgroup could use the assistance of the machine learning labeler, while the second subgroup did the labeling independently.

2.3. Evaluation Metrics

The selection of the legal categories followed the Hungarian legal system. We reduced the number of the categories to 167, and every document could receive a maximum of four distinct labels during the labeling process. During the reduction of the category labels, we strived to exclude and merge those categories where the number of the possible elements was under twenty (Figure 1). That was essential to provide enough training data for the machine learning classifier.

Figure 1 shows the estimated sizes of the different label groups in the full dataset. The size of the different areas are not uniform. The number of the elements varies from 30 to 30,000 in the different categories. The information content of a label is in an inverse relation with its element size. This is because when a document with a rare subject matter label is found during a search, it reduces the size of the similar documents set significantly. Hence, the information content of a smaller subject matter label is higher than a very general label to which thousands of documents belong.

We introduced a scoring system to compensate for these differences and measure better the information content. Those labels that have been tagged on more than 200 documents were worth 1 point, between 50 and 200 documents were worth 5 points, which had less than 50 documents were worth 10 points. Every good label counts, and there was no penalty applied for the bad labels during the calculation. Applying this scoring system the area, which represents the value of the information, is in the same range in Figure 1. The total score, which can be calculated in the reference set, was 1020 points.

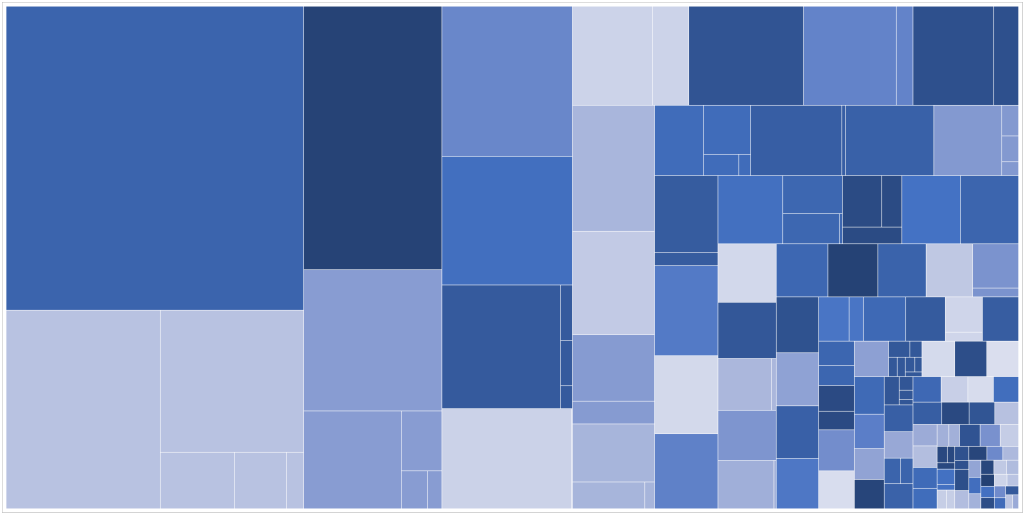

Figure 1. The estimated sizes and the structure of the label set. The area of each rectangle represents the estimated number of documents in the given label set.

2.4. Machine Learning-Based Classification

Due to the fact that each document could have more than one label, the original problem was decomposed into different multiple binary classification problems [35,36]. Since subject matters belonged to more than one law area, 229 different binary classification models were trained. As a machine learning algorithm to perform the labeling task, support vector machines were chosen, partly because the SVMs tend to perform well in the case of high dimensional vector space [37] and previous studies have also shown the superiority of this algorithm in similar categorization tasks [25,38]. In the case of small categories, text augmentation techniques (EDA, Word Vectors [39]) were used to generate synthetic samples to improve the performance of the training. The machine learning model was developed and deployed via the openly accessible `digital-twin-distiller` computation platform (https://github.com/montana-knowledge-management/digital-twin-distiller, accessed on 12 December 2021), where we used its plugins and the most important natural language processing libraries to accelerate the development [40,41].

The machine learning solution was elaborated by harnessing the following characteristics of legal documents: the legal expressions in texts may refer to the subject matter, and legal references can be helpful to determine the subject matter of a document (e.g., certain acts or paragraphs of acts). Hence, to tackle the problem, as a vectorization process, TF-IDF (term frequency inverse document frequency) vectorization was chosen [42,43]. From the texts, law references were extracted and normalized by using a regular expression-based solution. The law reference extractor returns a list of the law references found in the legal document in the most specific form possible.

The detailed description of the proposed machine-learning-based solution is a subject of another paper [44].

3. Results

3.1. Throughput

The first question of the research was to estimate how long it takes to make the labeling process by hand for all 170,000 documents. The conducted survey measured how many documents the different groups of participants could categorize after 3 h according to their level of competence and whether they received pre-labeled decisions or not. The average number of labeled judgments for the different groups of participants are shown in Figure 2.

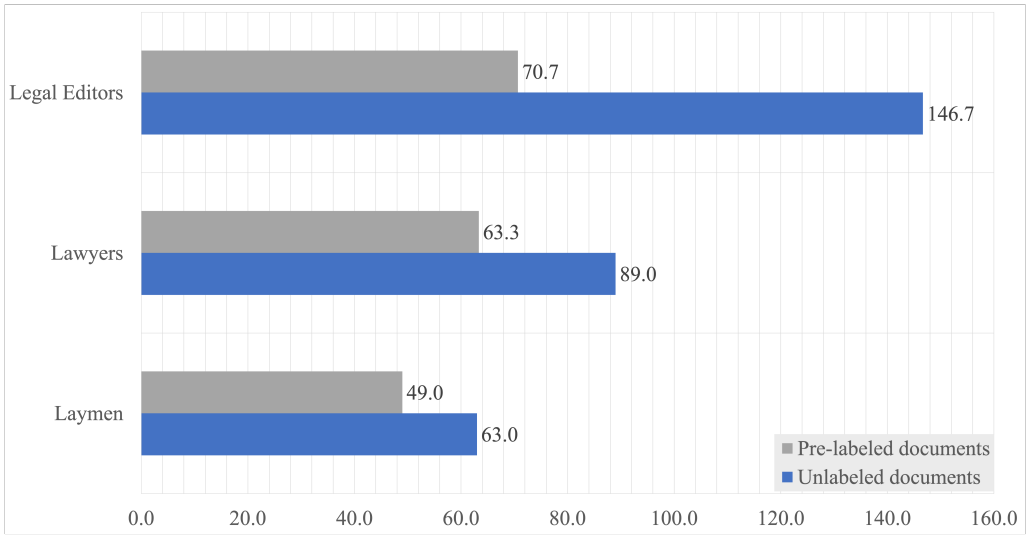

Figure 2. The average number of processed documents in the three examined groups after three hours of work.

The results gave back the expectations that the experienced legal editors processed the most documents, 108.7 on average, almost double that of an inexperienced person, since the laymen processed only 56 documents on average. The result also illustrates that even the most competent participants could categorize approximately 300 judgments per day without computer assistance. It means that if a database provider wanted to label all the available Hungarian judicial decisions, which is approximately 170,000 documents, with human work, it would take more than two years if the company employs a professional editor for this low added value task. If the employer uses laymen, a cheap workforce, this task will take about double the time, about 4.5 years. During these calculations, the accuracy of the work and the discoverability of the data were not considered. If the data provider wants a reference set quality result, they have to employ three professional editors for this task. In this case, about seven years of work is needed to process these documents. However, in this case, the discoverability of the data will be two times better. On the contrary, the applied machine learning algorithm labels a batch of 300 judicial decisions in minutes and only several hours are enough to label all of the datasets.

There was a surprising result, shown in Figure 2. Those participants who received pre-labeled documents labeled slower (61.0 documents on average) than those who re-

ceived unlabeled decisions (99.5 documents). However, these participants working with pre-labeled document sets extracted more information (50%) from the same amount of document. It seems the pre-labeled documents forced the labelers to read the decisions more thoroughly.

3.2. Accuracy

We applied different metrics to compare the results. Firstly, we calculated the accuracy of the labeled documents (Figure 3). This accuracy means the proportion of the documents that were completely or partially labeled correctly by each group. A document was considered partially labeled when at least one correct label was found for a given judgment. The results were based only on the documents that the participants managed to label, not the whole dataset.

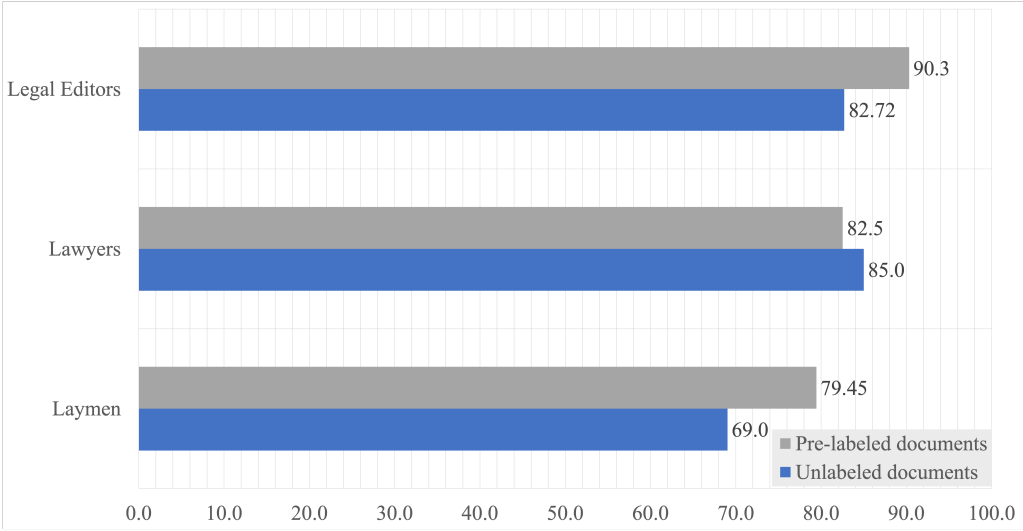

Figure 3. The proportion of at least one correct label found per document partly match plus complete match percentage.

It can be seen that even laymen were capable of finding at least one correct label for a document in 69% of the tagged documents. The laymen who worked with the pre-labeled documents reached the accuracy of those professionals who did not use the unlabeled documents. From this point of view, there is no significant difference between the lawyers and the professional legal editors.

However, we got a different picture when we calculated the accuracy of those documents which can fit at least three categories. Here, we accepted a solution from the participant if they found at least three labels correctly for a given document (Figure 4). It can be seen that those participants who could use the support of the machine learning algorithm reached significantly higher accuracy. It can be seen that there is no significant difference between the laymen and the legal editors in this type of contest. The application of the computer can increase the accuracy of the participants by more than 50%.

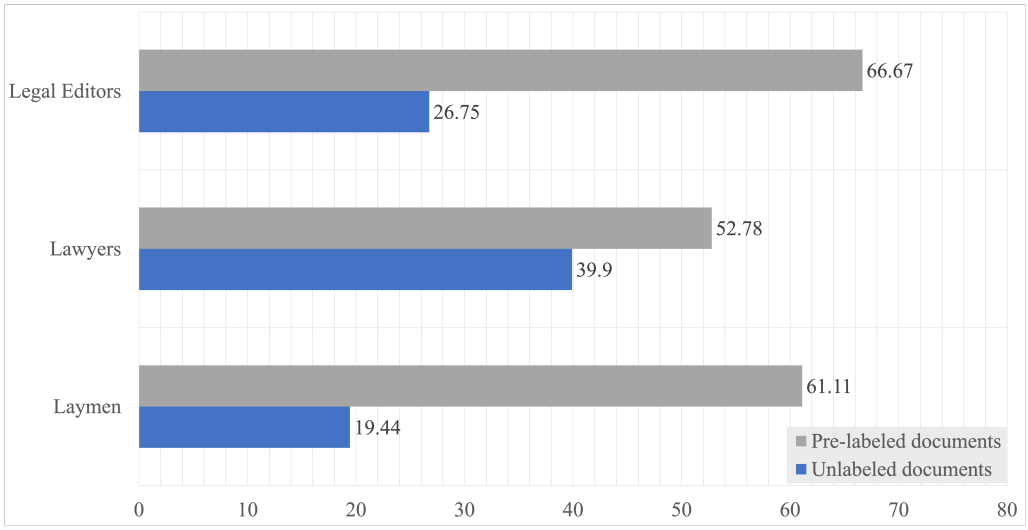

Figure 4. The proportion of the found labels on those documents, which contain at least three good labels.

3.3. Performance

The performance of the different participants was calculated with the aid of the previously introduced scoring sheet (Section 2.3). The score of the legal editors and that of the machine learning algorithm are compared in Figure 5. The machine learning code achieved 488 points from the possible 1020, which seems to be a relatively low performance. If we compare it with the performance of the legal editors, it found 50% more information on the same reference set than the human editors in three hours. There is a surprising result that those editors who could not use the assistance of the computer found more than 40% more information in the dataset than those who used the prelabeled labels. Checking the normalized values on Figure 6, we obtain the previous findings that those editors who used the computer assistance discovered 50% more information than the others and the computer. Figure 6 shows that the machine learning algorithm performance reaches the performance of the human level.

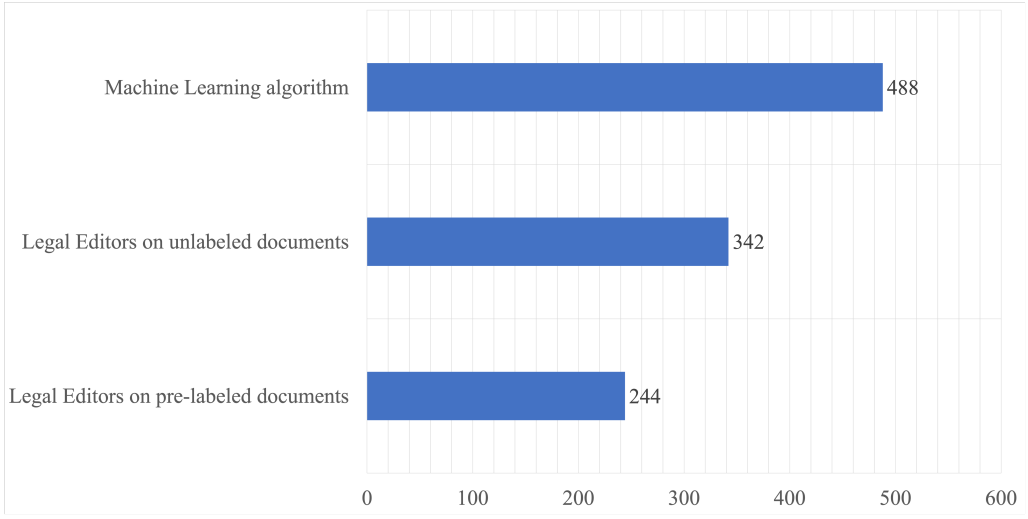

Figure 5. How many points the editors gained based on the extracted information content compared to the machine learning algorithm.

Figure 6. shows how much information the editors could extract from one document compared to the algorithm:

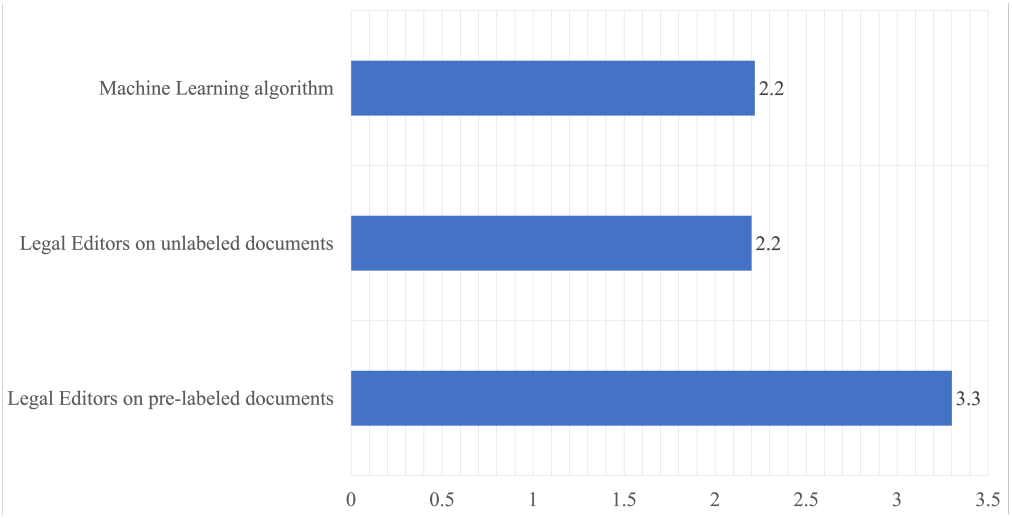

Figure 6. How many points the editors gained from one document in average based on the extracted information content compared to the machine learning algorithm.

Figure 7 shows the best-performing legal editor and the group of legal editors performance without computer assistance gained their points throughout the time of the examination.

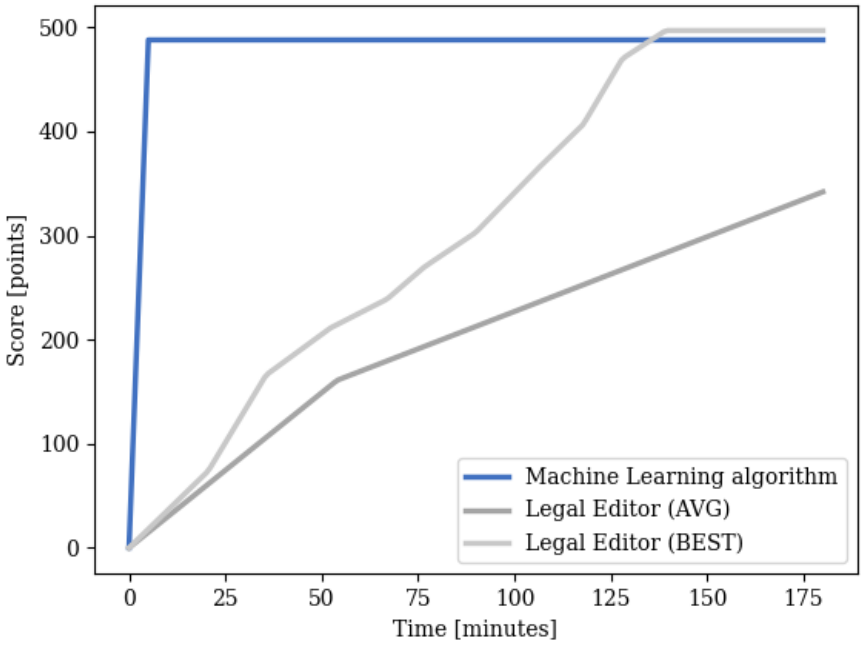

Figure 7. The performance of the best-performing legal editor and the group of legal editors who got unlabeled judgments over time.

Even the best performing legal editor could only retrieve the points after more than two hours, and they scored 3 points more than the machine learning algorithm. We can see the effect of the fatigue on the picture, where the performance of the editor group started to decrease. This means that a machine learning system can be used in ways that are different to a Legaltech business process. It can replace the work of human experts or be used to increase the discoverability of the dataset.

3.4. Inter-Annotator Agreement

The aim of measuring the inter-annotator agreement is to assess an annotation process' reliability (IAA). During this evaluation, we did not use the reference set for the comparison.

The three members of each group were considered as an annotator. The reliability of their annotations measured with Krippendorffś alpha (K_α) [45,46], which is a widely used statistical measure, it differs from most of the other IAA methods because it calculates disagreement between the different voters [47]. The K_α statistical measure is selected because it can handle the missing data, various sample sizes, categories. The reliability measure is easy to interpret and does not depend on the number of categories [49**?**], The simplest form of K_α can be calculated by the following formula [45]:

$$K_\alpha = 1 - \frac{D_0}{D_e}, \tag{1}$$

where D_0 stands for the observed disagreement among values, and D_e is for "disagreement one would expect when the coding of units is attributable to chance rather than to the properties of these units" [45]. The result of the K_α calculation is a number between -1 and 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance and negative values indicate inverse agreement. $K_\alpha \geq 0.8$ means usually the acceptance limit. Here the tentative results are also acceptable. The lowest limit for an acceptable agreement is when $K_\alpha \geq 0.667$. This is the minimal requirement to consider an inter-agreement calculation reliable [45,47]. The metric has an important component specific to the actual problem called the difference function, which is used to weight the numerator and denominator. The measuring agreement on set-valued items distance metric seemed to be the most appropriate in our case, due to a large number of possible annotation categories [50].

Figure 8 shows the calculated K_α scores in percentage for each annotator group. There is a surprising result; the group of the laymen who used the computer assistance achieved the highest score in K_α. They achieved 55% reliability, which is significantly higher than the reliability of the professional editors. There are two reasons for this surprising result. Firstly, if we compare the result of the three groups who could not use the assistance of the machine learning methodology, we can see from (Figure 8) that these groups achieved similar reliability, independently from their experience. This reliability score was very low in these cases. The results suggest that the laymen group, who trusted the result of the computer annotation, achieved the highest reliability score in the survey. This indicates that the machine learning-based methodologies produce more consistent solutions for large databases than the humans. However, this highest reliability score is lower than the required minimum ($K_\alpha \geq 66\%$). The second reason is that the categorization group should be revised. The poor agreement between the human experts suggests that the labels are not straightforward and independent from each other, and they did not look for every possible combination due to the time limitations.

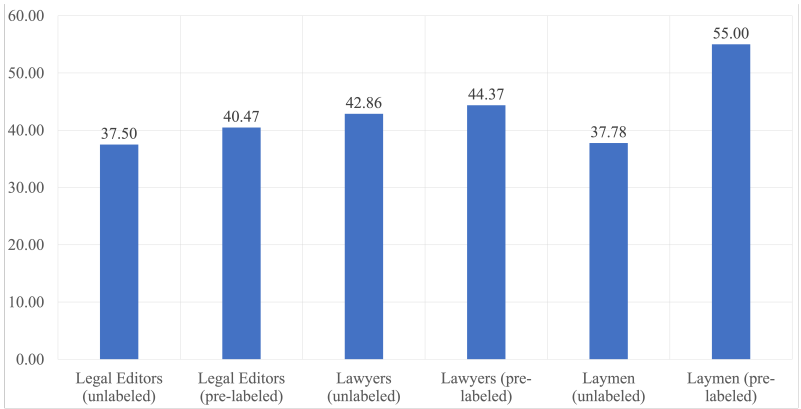

Figure 8. Krippendorff's Alpha scores in different annotator groups with (pre-labeled) and without machine suggested labels (unlabeled).

The percentage agreement between the different groups was also calculated to examine further this reason for the poor reliability between the professional editors (Figure 9). This percentage agreement examines a more permissive measure than the K_α. It shows the ratio of documents, where all three members of the group gave at least one similar label to the given document. The resulting values give back our expectations that the professional legal editors achieved the best results (K_α 50%), and non-specialists were the worst (K_α 23%) in this comparison. It seems to be surprising that those editors who did not use the assistance of the machine learning methodology achieved significantly higher (about 5%) scores than those editors who worked with the computer assistance. These observations seem to justify that the human professional's learning curve is different from the machine learning methodology; they found that the simple labels had much higher accuracy and reliability than the computer. However, the human expert's reliability is worse than the machine learning solution for complex cases. This measurement justifies that the machine learning methodologies can significantly increase the discoverability of large and complex datasets.

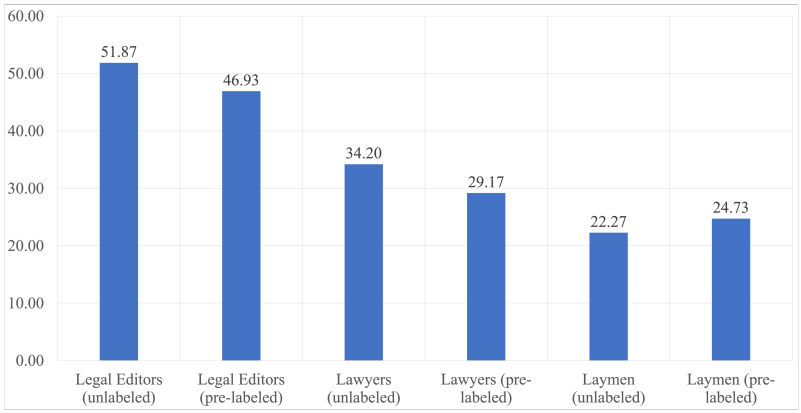

Figure 9. Percentage agreement between the 3 annotators (broken down into annotator groups).

4. Conclusions

The paper has shown the result of an experimental study that compares the human performance with a machine learning-based solution on a fixed length, low added value monotonous task, a legal text classification, where the solutions are not exactly defined. Enriching the unstructured legal documents with specific labels is necessary to help the lawyers, judges and prosecutors to find similar cases. However, the classification of these texts is very time consuming, and it requires an unacceptable amount of time from the legal editors. The development time and cost of a machine learning-based solution mainly depends on the complexity of the problem. The motivation behind the experiment was to estimate the performance of the legal editors on this task because if a machine learning solution can reach human performance, it can be worth replacing human work. This assumption can significantly reduce the requirements and the cost of the implementation. During the experiment, the performance of three competence groups were examined: legal editors, lawyers and laymen. Every group was divided into two parts. The first group could use the results of the machine learning algorithm as assistance. The second group completed the labeling without assistance. The results showed that the proposed machine learning solution, which found 48% of the information in the reference dataset, significantly outperformed the average of the legal editors in the whole test. Surprisingly, those participants who used the computer assistance were slower, but their precision increased by more than 50%. Moreover, the computer assistance increased the score of the laymen participants significantly. They achieved comparable performance to expert participants. The results show that the application of a machine learning algorithm in solving a legal tech problem can have positive impacts. It can improve the workflow by replacing the human in the loop and reducing the cost of the production, it can improve the quality of the data or decrease the learning curve of new colleagues working on data

enrichment. The study results show that the applied machine learning algorithm can reach the average performance of human experts. Moreover, machine learning methodologies can be advantageous for those monotonous tasks where finding the correct solution needs deep focus and unique expertise, or it is hard to define the exact solution, as in the case of law. Another insight gained by this study is that the label set should be reviewed from a legal perspective, and other domain knowledge should be taken into account to increase the agreement between the legal experts and create a new ontology for the labeling system. This new ontology and the newly trained models can further increase the legal database's discoverability, usability, and value.

Author Contributions: Conceptualization, T.O., C.G., D.N. and V.R.; methodology, D.N., V.R.; software, G.M.C., D.N., T.O.; validation, A.M., G.M.C. and J.P.V.; formal analysis, G.M.C.; investigation, G.M.C.; resources, J.P.V.; data curation, A.M.; writing—original draft preparation, T.O., R.V.; writing—review and editing, T.O., G.M.C.; visualization, I.Ü., G.M.C., D.N.; supervision, T.O.; project administration, D.N.; funding acquisition, D.N., A.M. All authors have read and agreed to the published version of the manuscript.

Funding: Project No. 2020-1.1.2-PIACI-KFI-2020-00049 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the 2020-1.1.2-PIACI KFI funding scheme.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Parikh, V.; Mathur, V.; Metha, P.; Mittal, N.; Majumder, P. LawSum: A weakly supervised approach for Indian Legal Document Summarization. *arXiv* **2021**, arXiv:2110.01188.
2. Heller, J.; Arredondo, P., AI in Legal Research: How AI Is Provideing Everyone Acces to Information and Leveling the Playing Field for Firms of All Sizes. In *AI for Lawyers: How Artificial Intelligence Is Adding Value, Amplifying Expertise, and Transforming Carriers*; Waisberg, N., Hudek, A., Eds.; Wiley: Hoboken, NJ, USA, 2021.
3. Walters, E.; Asjes, J., Fastcase, and the Visual Understanding of Judicial Precedents. In *Legal Informatics*; Katz, D.M., Dolin, R., Bommarito, M.J., Eds.; Cambridge University Press: Cambridge, UK, 2021; pp. 357–406.
4. Blair, D.C.; Maron, M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM* **1985**, *28*, 289–299. [CrossRef]
5. Roitblat, H.L.; Kershaw, A.; Oot, P. Document categorization in legal electronic discovery: Computer classification vs. manual review. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 70–80. [CrossRef]
6. Park, S.H.; Lee, D.G.; Park, J.S.; Kim, J.W. A Survey of Research on Data Analytics-Based Legal Tech. *Sustainability* **2021**, *13*, 8085. [CrossRef]
7. Chalkidis, I.; Kampas, D. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artif. Intell. Law* **2018**, *27*, 171–198. [CrossRef]
8. Li, G.; Wang, Z.; Ma, Y. Combining Domain Knowledge Extraction With Graph Long Short-Term Memory for Learning Classification of Chinese Legal Documents. *IEEE Access* **2019**, *7*, 139616–139627. [CrossRef]
9. Thammaboosadee, S.; Watanapa, B.; Chan, J.H.; Silparcha, U. A Two-Stage Classifier That Identifies Charge and Punishment under Criminal Law of Civil Law System. *IEICE Trans. Inf. Syst.* **2014**, *E97.D*, 864–875. [CrossRef]
10. Ashley, K.D.; Brüninghaus, S. Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law* **2009**, *17*, 125–165. [CrossRef]
11. Ma, Y.; Zhang, P.; Ma, J. An Ontology Driven Knowledge Block Summarization Approach for Chinese Judgment Document Classification. *IEEE Access* **2018**, *6*, 71327–71338. [CrossRef]
12. de Maat, E.; Krabben, K.; Winkels, R. Machine Learning Versus Knowledge Based Classification of Legal Texts. In Proceedings of the 2010 Conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference, Amsterdam, The Netherland, 12 August 2010; IOS Press: Amsterdam, The Netherland 2010; pp. 87–96.
13. Barnett, T.; Godjevac, S.; Renders, J.M.; Privault, C.; Schneider, J.; Wickstrom, R. Machine learning classification for document review. In *DESI III: The ICAIL Workshop on Globaal E-Discovery/E-Disclosure*; Citeseer: Princeton, NJ, USA 2009.
14. Borko, H. Measuring the reliability of subject classification by men and machines. *Am. Doc.* **1964**, *15*, 268–273. [CrossRef]
15. van Rijsbergen, C. *Information Retrieval*, 2nd ed.; Butterworths: London, UK, 1979.
16. Tonta, Y. A study of indexing consistency between Library of Congress and British Library catalogers. *Libr. Resour. Tech. Serv.* **1991**, *35*, 177–185..

17. Voorhees, E.M. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manag.* **2000**, *36*, 697–716. [CrossRef]
18. Fang, Y.; Tian, X.; Wu, H.; Gu, S.; Wang, Z.; Wang, F.; Li, J.; Weng, Y. Few-Shot Learning for Chinese Legal Controversial Issues Classification. *IEEE Access* **2020**, *8*, 75022–75034. [CrossRef]
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
20. Brueckner, R.; Schulter, B. Social signal classification using deep BLSTM recurrent neural networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4823–4827.
21. Rahman, L.; Mohammed, N.; Azad, A. A new LSTM model by introducing biological cell state. In Proceedings of the 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 22–24 September 2016; pp. 1–6. [CrossRef]
22. Agrawal, R.; de Alfaro, L.; Polychronopoulos, V. Learning From Graph Neighborhoods Using LSTMs. *arXiv* **2016**, arXiv:1611.06882.
23. Vapnik, V.N. *Statistical Learning Theory*; Wiley Interscience: Berlin, Germany 1998.
24. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
25. Goh, Y.C.; Cai, X.Q.; Theseira, W.; Ko, G.; Khor, K.A. Evaluating human versus machine learning performance in classifying research abstracts. *Scientometrics* **2020**, *125*, 1197–1212. [CrossRef]
26. Schumacher, J.; Zazworka, N.; Shull, F.; Seaman, C.; Shaw, M. Building Empirical Support for Automated Code Smell Detection. In Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '10, Bozen, Italy, 16–17 September 2010; Association for Computing Machinery: New York, NY, USA, 2010. [CrossRef]
27. Simundic, A.M.; Nikolac, N.; Ivankovic, V.; Ferenec-Ruzic, D.; Magdic, B.; Kvaternik, M.; Topic, E. Comparison of visual vs. automated detection of lipemic, icteric and hemolyzed specimens: Can we rely on a human eye? *Clin. Chem. Lab. Med.* **2009**, *47*, 1361–1365. [CrossRef]
28. Weismayer, C.; Pezenka, I.; Gan, C.H.K. Aspect-Based Sentiment Detection: Comparing Human Versus Automated Classifications of TripAdvisor Reviews. In *Information and Communication Technologies in Tourism 2018*; Stangl, B., Pesonen, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 365–380.
29. Nasiri, S.; Khosravani, M.R. Machine learning in predicting mechanical behavior of additively manufactured parts. *J. Mater. Res. Technol.* **2021**, *14*, 1137–1153. [CrossRef]
30. Chen, H.; Wu, L.; Chen, J.; Lu, W.; Ding, J. A comparative study of automated legal text classification using random forests and deep learning. *Inf. Process. Manag.* **2022**, *59*, 102798. [CrossRef]
31. Khosravani, M.R.; Nasiri, S. Injection molding manufacturing process: Review of case-based reasoning applications. *J. Intell. Manuf.* **2020**, *31*, 847–864. [CrossRef]
32. Niewiadomski, P.; Stachowiak, A.; Pawlak, N. Knowledge on IT tools based on AI maturity–Industry 4.0 perspective. *Procedia Manuf.* **2019**, *39*, 574–582. [CrossRef]
33. Farkas, R.; Szarvas, G.; Kocsor, A. Named entity recognition for Hungarian using various machine learning algorithms. *Acta Cybern.* **2006**, *17*, 633–646.
34. Firestone, C. Performance vs. competence in human–machine comparisons. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26562–26571. [CrossRef] [PubMed]
35. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142.
36. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
37. Ghaddar, B.; Naoum-Sawaya, J. High dimensional data classification and feature selection using support vector machines. *Eur. J. Oper. Res.* **2018**, *265*, 993–1004. [CrossRef]
38. Khor, K.; Ko, G.; Walter, T. Applying machine learning to compare research grant programs. In Proceedings of the STI 2018 Conference Proceedings, Leiden, The Netherlands, 12–14 September 2018; Centre for Science and Technology Studies (CWTS): Leiden, The Netherlands, 2018; pp. 816–824.
39. Csányi, G.; Orosz, T. Comparison of data augmentation methods for legal document classification. *Acta Tech. Jaurinensis* **2021**. [CrossRef]
40. Orosz, T.; CsÁnyi, G.; Nagy, D. Mesterséges Intelligenciát alkalmazó szövegbányászati eszközök készítése a distiller keretrendszer segítségével–Jogi szövegek automatikus feldolgozása: Development of Artificial Intelligence-based Text Mining Tools with the distiller-framework–in case of Legal Documents. *Energetika-Elektrotechnika–Számítástechnika és Oktatás Multi-konferencia* **2021**, *XXI*, 62–69.
41. Orosz, T.; Gadó, K.; Katona, M.; Rassõlkin, A. Automatic Tolerance Analysis of Permanent Magnet Machines with Encapsuled FEM Models Using Digital-Twin-Distiller. *Processes* **2021**, *9*, 2077. [CrossRef]
42. Luhn, H.P. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.* **1957**, *1*, 309–317. [CrossRef]
43. Jones, K.S. A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* **1972**.

44. Orosz, T.; Csányi, G.M.; Vági, R.; Nagy, D.; Üveges, I.; Vadász, J.P.; Megyeri, A. Building a Production-ready Multi-label Classifier for LegalDocuments with Digital-Twin-Distiller. *Appl. Sci.* **2021**, submitted.
45. Krippendorff, K. Computing Krippendorff'rs Alpha-Reliability. Computer Science, Annenberg. 2011. Available online: https://repository.upenn.edu/asc_papers/43/ (accessed on 23 October 2021)
46. Artstein, R., Inter-annotator Agreement. In *Handbook of Linguistic Annotation*; Ide, N., Pustejovsky, J., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 297–313. [CrossRef]
47. Glen., S. Krippendorff's Alpha Reliability Estimate: Simple Definition. Available online: https://www.statisticshowto.com/krippendorffs-alpha/ (accessed on 23 October 2021).
48. Krippendorff, K. Reliability in Content Analysis. *Hum. Commun. Res.* **2004**, *30*, 411–433. [CrossRef]
49. Hayes, A.F.; Krippendorff, K. Answering the Call for a Standard Reliability Measure for Coding Data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
50. Passonneau, R. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, 22–28 May 2006.