Hungarian Journal of Legal Studies 64 (2023) 3, 456–471
DOI: 10.1556/2052.2023.00473

AKADÉMIAI KIADÓ

Increasing access to legal information with unsupervised solutions

RENÁTÓ VÁGI[1,2*], ISTVÁN ÜVEGES[1,3], ANDREA MEGYERI[4], ANNA FÜLÖP[4], JÁNOS PÁL VADÁSZ[1,5], DÁNIEL NAGY[1] and GERGELY MÁRK CSÁNYI[1]

[1] MONTANA Knowledge Management Ltd., Budapest, Hungary

[2] Doctoral School of Law, Eötvös Loránd University, Budapest, Hungary

[3] Centre for Social Sciences, Budapest, Hungary

[4] Wolters Kluwer Hungary Ltd., Budapest, Hungary

[5] Institute of the Information Society, National University of Public Service, Budapest, Hungary

ORIGINAL RESEARCH PAPER

Received: April 5, 2023 • Accepted: July 25, 2023
Published online: April 15, 2024

ABSTRACT

Access to justice is a significant area of legal research, especially for Socio-Legal studies. The main research topics of this area are economic or class differences, gender inequalities, or national and ethnic differences in access to justice. However, there is a less discussed aspect of access to justice: the differences between access to legal information and the differences in user groups in terms of comprehending and processing legal information. This is an important topic because there are significant differences among people's abilities to process and understand legal texts, depending on whether we are dealing with a lawyer who is an expert in the given field, a non-expert lawyer, or a citizen with a low or zero (legal) educational level. The paper argues that unsupervised machine learning solutions can help even out these differences. It presents different unsupervised solutions, mainly clustering and topic modelling, which can help to increase access to legal information. Then we present a case study in which we examine these unsupervised tools in the processing of resolutions of the Central Bank in Hungary and anonymized court decisions. The paper argues that these tools can reveal the hidden contextual regularities in unstructured legal texts, facilitating the search for legal texts even for non-legal-experts.

[*] Corresponding author. E-mail: vagi.renato@montana.hu

KEYWORDS

unsupervised machine learning, clustering, topic modelling, resolutions of the Central Bank of Hungary, anonymized court decision

1. INTRODUCTION AND BACKGROUND OF UNSUPERVISED MACHINE LEARNING AND ACCESS TO JUSTICE

Access to justice has been a significant area of legal research for more than half a century. This field began to flourish when it was recognized that the traditional liberal conception of law, i.e. that formal equality before the law itself is sufficient to establish a fair administration of justice, is not sufficient. This is usually described as the so-called gap problem: law in books and law in practice often do not match, and there can be a greater or lesser discrepancy between the two in different areas. Furthermore, legal procedures do not exist in a vacuum; lawyers and legal scholars must recognize their social impact.[1] It may therefore be worth examining how different social groups can access different legal institutions and how to promote equality where such differences exist.

There are two possible directions in the research into access to justice.[2] On one hand, it is possible to investigate whether everyone has equal access to institutions of justice, regardless of economic status, gender, or ethnicity. This could be called as the input side. On the other hand, one can examine whether the final judgments are just, which may be referred to as the output side. The current paper will deal with a relatively under-discussed aspect of the first, the input side.

According to their classic division, Cappelletti and Garth distinguish three waves of reform instruments in countries to promote access to justice. The first wave includes those instruments that seek to even out differences in access to justice in terms of economic, gender, and ethnic differences among citizens.[3] It is not surprising that these are generally the first instruments to be used, since the literature is also dominated by writings that examine economic or class differences,[4] gender inequalities,[5,6] or national and ethnic differences in access to justice.[7]

Following these, the second wave of access to justice reforms, according to Cappelletti and Garth, are the institutional changes within the court system introduced to make the adjudication of different cases fairer and more equitable. Finally, in the third wave, they include procedural changes that pave the way for alternative dispute resolution procedures, which aim to address

[1] Hatıpoğlu-Aydın and Aydın (2016) 73.

[2] Gomes (2019) 360.

[3] Cappelletti and Garth (1978) as cited in Noone and Ojelabi (2020) 110.

[4] Nash (2013).

[5] Carpio (2001).

[6] Hatıpoğlu-Aydın and Aydın (2016).

[7] Gomes (2019).

the flaws in the traditional, institutionalized justice system and increase access to justice.[8] This is also the context in which the legal technology literature has tended to focus on access to justice, with several studies and monographs published in the last decade on how technological innovations facilitate the spread of alternative dispute resolution, such as online dispute resolution for ecommerce services[9] and online courts with extended services[10], for example, for low value civil claims.[11]

However, there is also a fourth, less discussed aspect of access to justice: access to different legal information and the differences in its processing and understanding. Today, it is common in European legal systems that almost all primary sources of law (legislation or various kinds of decisions) are freely available and can be viewed online by anyone. However, a large amount of new legal data is generated every day, making it particularly difficult to process information and track changes. The ability to understand different legal knowledge is also very different, depending on whether we are dealing with a lawyer who is an expert in the given field, a non-expert lawyer, or a citizen without any legal education. That is why we were interested in whether different machine learning tools could help people better understand the large amount of unstructured legal information available.

There are two significant domains of machine learning classification: supervised classification and unsupervised classification.[12] The essence of supervised classification is to pre-label a certain amount of data, from which the model learns the attributes of each label based on the characteristics of this so-called '(labelled) training data'. Furthermore, during the inference phase, it can automatically decide to which category it belongs to.[13] Two elements are crucial for this type of development. On the one hand, a fixed category system is needed to classify individual documents under one or more categories. On the other hand, an adequate amount of quality training data is needed for each category since the model can learn only from the regularities of the documents in one category. In simple terms, the more data the model can see, the more efficiently it can work.

The availability of these two conditions is not self-evident in many cases since the manual labelling of the appropriate amount of training data often requires significant time and work from experts in the given field. It may also happen that due to the specificities of the domain, there is no uniformly accepted category system, and its creation also requires significant domain expertise. When examining access to justice, we encounter this problem in many cases since it is often a difficulty for people who do not understand the law to find the information that is relevant to them from unstructured court decisions or legislation.

Tools based on unsupervised classification can be a solution to this particular issue. Unsupervised classification differs fundamentally from supervised classification, since in the case of supervised classification, we have training data for each label, and the algorithm learns about the main attributes of each category based on them. However, for unsupervised classification,

[8] Noone and Ojelabi (2020) 110.

[9] Rule (2017).

[10] Susskind (2019).

[11] Civil Justice Council (2015) 20.

[12] Katz (2021) 89.

[13] Katz and Nay (2021) 95.

we do not have any training data or a developed category system, only the set of documents. Unsupervised algorithms can recognize the regularities of the documents and automatically sort them into different groups.[14] The disadvantage of this procedure is that the classification of documents will be less accurate than in the case of a well-trained, supervised model. In addition, we have no control over exactly what categories we will get back, so we need human work to assess what each abstract group means, which can also be labour-intensive. However, their application can be helpful in many areas; on the one hand, during machine learning development, to make the category system as sophisticated as possible, and on the other hand, to support access to information for citizens when they want guidance on their legal case.

Two significant areas of unsupervised machine learning-based classification solutions, which can be used to discover statistical regularities, are *clustering* and *topic modelling*. Clustering means the automatic grouping of documents based on their similarities and differences. Topic modelling helps to better understand the grouping of a given set of documents by collecting the defining words, phrases, and word pairs which caused those documents to be semantically grouped by the algorithm.[15] These solutions are widely used in several areas, including organizing and semantic analysis of emails[16] and newsgroup messages.[17] In addition, their application is becoming increasingly widespread in the legal field. Trappey et al., for example, used various unsupervised tools for grouping trademark cases and analysing their key concepts.[18] Ravi kumar V. and K. Raghuveer performed clustering of Indian judgments to help searching in such a way that certain searches do not have to be run on the entire data set but can be narrowed in advance, which can thereby speed up the search using the algorithm.[19]

We examined unsupervised tools in the case of Hungarian-language legal documents and how efficiently they can be used, specifically in the case of the resolutions of the Central Bank of Hungary and anonymized court decisions. The structure of the study is as follows. In Section 2, we present the characteristics of the data sets and the unsupervised tools used. In Section 3, we present the results of the unsupervised solutions introduced. In Section 4, we analyse the results obtained and then conclude.

2. MATERIALS AND METHODS: LEGAL DATASETS AND UNSUPERVISED SOLUTIONS

2.1. Resolutions of the Central Bank of Hungary

The Central Bank of Hungary (MNB), among its other tasks, supervises the financial intermediary system in Hungary.[20] According to the law defining the organizational structure and

[14] Katz and Nay (2021) 95.

[15] Trappey et al. (2020) 2.

[16] Berger and Merkl (2005).

[17] Natarajan et al. (2007).

[18] Trappey et al. (2020).

[19] Raghuveer (2012).

[20] Fundamental Law of Hungary Article 41 (2).

functioning of the MNB, it acts as an authority while supervising the financial intermediary system. It can conduct authorization procedures, control procedures, proceedings for the protection of consumers' interests, market surveillance procedures, and supervisory control proceedings.[21] In these procedures, the MNB makes decisions, which may take the form of a resolution or ruling.[22] According to the law, the Central Bank of Hungary is obliged to publish the number and subject of the resolution, the name and registered address of the client affected by the proceedings and procedures, the operative part of the resolution, the fact of any review proceedings and the operative part of the final decisions reached in the course of the review.[23] Publishing the resolutions of the Central Bank of Hungary and making them widely available to citizens is also essential because the law requires the MNB to strengthen public confidence in the financial intermediary system during its supervision.[24]

The published decisions are available on a subsection of the MNB website.[25] The site includes a search engine to help search through the resolutions by keyword, date of publication, type of resolution, and type of procedure. In addition, it is the case number of the resolutions that help to orientate among them since, according to the Central Bank of Hungary's publication information, the case numbers of the resolutions are assigned uniformly so that the first letter indicates the type of a resolution, the second letter the area of specialization, then the sector and finally the type of measure. The serial number and the year of the resolution follow them.[26] However, users cannot search according to the case number of the resolutions. Furthermore, as the names indicate, these are mainly used to group documents at a technical and procedural level, not necessarily by content.

In this research, in which we investigated how specific unsupervised machine learning tools can be used to subdivide documents by content meaning beyond legal definitions, we used roughly 10,000 MNB decisions. For each decision, we only had the resolution's text and case number.

2.2. Anonymized Hungarian Court decisions

According to the Hungarian Fundamental Law, the courts are responsible for the administration of justice.[27] The courts rule on criminal cases and private law disputes and decide cases involving the legality of administrative decisions, among others.[28] The courts are required by law to make public the following decisions:

- The decisions of the Curia adopted on the substance of a matter, its annulment decisions, uniformity decisions, and decisions delivered in uniformity complaint procedures and in appeal proceedings to ensure legality;

[21] Act CXXXIX of 2013 on the Central Bank of Hungary Section 48 (1).

[22] Act CXXXIX of 2013 on the Central Bank of Hungary Section 49/C (1).

[23] Act CXXXIX of 2013 on the Central Bank of Hungary Section 53 (1).

[24] Act CXXXIX of 2013 on the National Bank of Hungary Section 4 (9) d.

[25] Link1.

[26] Link2.

[27] Fundamental Law of Hungary Article 25 (1).

[28] Fundamental Law of Hungary Article 25 (2).

- The decisions of the court of appeal adopted on the substance of a matter;
- The decisions of the general court adopted in administrative actions on the substance of a matter, if the reviewed administrative decision was adopted in a single instance proceeding and no ordinary appeal may be lodged against the court decision.[29]

In addition to these decisions, an anonymized digital copy of all decisions of courts and other public authorities or other bodies that have been overruled or revised by the published court decision must be published as well.[30] This means that not all court decisions are published for the public in Hungary, only those that have been decided by a higher court and all related decisions and rulings.

Approximately 175,000 such documents were already available to the public at the time of the research. Since then, the total number of court decisions available has increased to more than 270,000, which is a significant increase in almost one and a half years. We investigated how we can use the unsupervised tools presented below to automatically group these documents by content, which will help experts and non-legal professionals more easily understand the documents.

2.3. Vectorization solutions

The first step in any text-based machine learning development is the vectorization of documents. For a computer to process and analyse text documents, it is necessary to put the unstructured raw data into a format that the computer can interpret. One way to do this is to use a vectorization process, which involves converting the raw text into a vector of numbers. Machine learning algorithms can then 'read' the text. In this study we tested the TF-IDF, Doc2Vec and Latent Semantic Analysis vectorization methods.

TF-IDF vectorization is a process that has been known for a long time which statistically weights the occurrences of words in a document according to how often the word occurs in the document (term frequency) and how many documents in the whole dataset contain the same word (inverse document frequency).[31] For example, if a term occurs in only a few documents, the IDF value will be higher for that word than for a word with a more general occurrence in the corpus.

Doc2Vec, on the other hand, generalizes the Word2Vec solution to bring each document into numerical form. The idea of Word2Vec[32] is to provide a vector representation that captures the meaning of that word. Similarly, by Doc2Vec vectorization,[33] a vector representation can be calculated for documents that characterize the meaning of the document.

Latent Semantic Analysis[34] (LSA) can complement the TF-IDF solution by considering the positioning of words, since it starts from the assumption that words that are close in meaning are placed in similar pieces of text. It is thus able to reduce the dimensions of the vectorization, which can result in significant computational savings, and can also consider the context of the

[29] Act CLXI of 2011 on the Organization and Administration of the Courts Section 163 (1).

[30] Act CLXI of 2011 on the Organization and Administration of the Courts Section 163 (3).

[31] Bafna et al. (2016) 62.

[32] Mikolov et al. (2013).

[33] Le and Mikolov (2014).

[34] Bellegarda (2005).

documents. This means that, unlike the Doc2Vec solution, it does not just vectorise individual documents but looks at the whole data set.

In the analysis of the resolutions of the Central Bank of Hungary, TF-IDF, Doc2Vec, and LSA vectorization methods were used.

In the case of anonymized court decisions, TF-IDF vectorization was used. However, the disadvantage of TF-IDF vectorization is that it treats documents as a so-called ‘bag-of-words’, ignoring grammar rules and word order. This can be problematic for legal texts, however, as legal concepts can sometimes consist of several words, and words of the same form can carry several meanings depending on their context. Therefore, for this type of document, we have also examined the impact on the model of adding individual words and word pairs to the classical TF-IDF method in vector construction. In this way, the information gained from word order can be retained to a certain extent.

2.4. Clustering

Clustering, as briefly introduced in the introduction, is, in fact, nothing more than the automatic grouping of documents without any pre-trained model, based only on the characteristics of the texts and their similarities and differences. Based on the basic definition of clustering, the aim is that documents in a cluster should be as similar as possible, while documents in different clusters should be as far as possible in the vector space used for representation. It is also crucial, that the measurement of similarity and dissimilarity should be as precise as possible.[35]

As these definitions suggest, clustering involves several tools and solutions depending on the characteristics of the data set. One of the most widely used groups of clustering solutions is Centroid-based Clustering, based on the idea that individual documents are placed in a cluster based on their distance from the centre of the cluster.[36]

Within this framework, one of the first developed and still one of the most widely used solutions is the k-means algorithm. The essence of the k-means clustering[37] is that we can classify the documents into a predefined number of clusters. This method can be complemented with the elbow method to estimate the ideal number of clusters in the given dataset. The elbow, or ‘knee of a curve,’ approach is the most common and straightforward means of determining the appropriate cluster number: it entails repeatedly running the clustering algorithm on the dataset with different values. For every value, we calculate the squared distance between each point (or document, in our case) and its related centroid (the centre of the topics of the given document). The ideal point and the ideal cluster value is the point of inflection of the curve (‘the elbow’), where the graph starts to look like a straight line.[38] After that point, a new topic cannot increase the performance of the clustering.

The limitation of this solution is that it is perfectly applicable only to single-level category sets. In the case of a hierarchical label structure, it can cause difficulties since it can no longer distinguish between subcategories within each category. This specificity caused us problems in the case of court decisions, as will be shown later.

[35] Xu and Tian (2015) 166.

[36] Google Developers (2022).

[37] Jain (2010).

[38] Onumanyi et al. (2022) 2.

For both the MNB decisions and the court decisions, we applied a k-means clustering procedure.

2.5. Topic modelling

Topic modelling is an unsupervised machine learning tool that can determine the common dominant topic or topics of documents in a group. It makes it possible to determine regularities in unstructured documents by humans without needing pre-trained models.

One of the most frequently used forms of topic modelling is Latent Dirichlet Allocation[39] (LDA), which can find the different topics in each set of documents by the terms defining that group. The assumption on which this solution is based is that documents within a document set belong to different topics and that these topics can be described by the set of words in the document. Thus, the algorithm cannot name the topics, but it can help the expert determine which topics categorize the given set of documents.

We used an LDA solution for the unsupervised machine learning examination of anonymized court decisions.

3. RESULTS

3.1. Automatic clustering of the resolutions of the Central Bank in Hungary

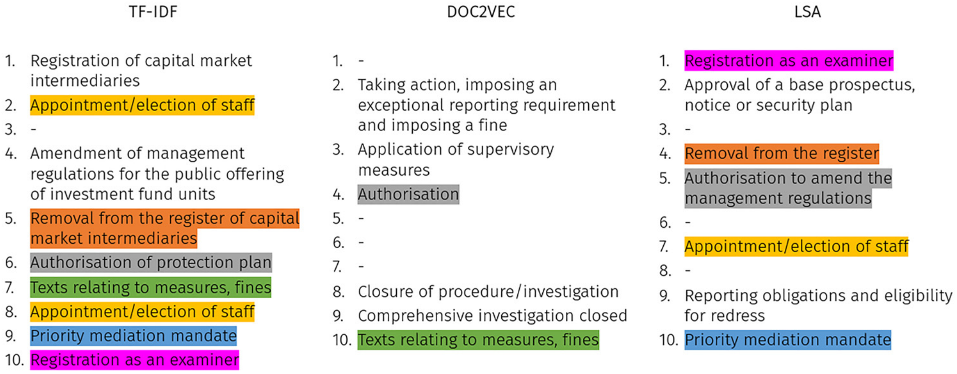
As mentioned above, the search engine for MNB resolutions on the MNB subpage currently provides users with limited options for filtering decisions. These options are also mainly technical, such as the year or type of resolution. In addition to the search, the case number of the resolutions can orient the users. However, the information in the case number indicates the field of expertise and the type of procedure, which do not help non-lawyer or non-expert legal users.

Therefore, we examined how unsupervised tools can help introduce a new way of categorization according to the resolutions' semantic content. We used the already described K-means clustering to divide the dataset into ten groups using different forms of vectorization. Ten was an arbitrary number because, first, we wanted to see how the methods work with default settings. Legal experts then read many of the resolutions in each group. If they found a semantic regularity between the documents in a given cluster that could be used to identify why they belonged to a group, then the group was named accordingly. If they could not discover such a regularity, then that group was not named, and the grouping was assumed to be meaningless and incorrect. Thus, the following partitioning of the document set was produced in each of the vectorization methods used:

Figure 1 shows that a considerable proportion of the groups associated with each vectorization method could be successfully categorized, so this approach can be considered adequate. The clusters without any name in Fig. 1 indicate where the experts found no correlations between documents, and the grouping is therefore considered incorrect, and the colours indicate the correlations between each group.

We also examined which vectorization method is more efficient for the unsupervised clustering of resolutions of the Central Bank of Hungary. We can use two criteria for characterizing

[39] Blei et al. (2003)



	TF-IDF	DOC2VEC	LSA
1.	Registration of capital market intermediaries	-	Registration as an examiner
2.	Appointment/election of staff	Taking action, imposing an exceptional reporting requirement and imposing a fine	Approval of a base prospectus, notice or security plan
3.	-	Application of supervisory measures	-
4.	Amendment of management regulations for the public offering of investment fund units	Authorisation	Removal from the register
5.	Removal from the register of capital market intermediaries	-	Authorisation to amend the management regulations
6.	Authorisation of protection plan	-	-
7.	Texts relating to measures, fines	-	Appointment/election of staff
8.	Appointment/election of staff	Closure of procedure/investigation	-
9.	Priority mediation mandate	Comprehensive investigation closed	Reporting obligations and eligibility for redress
10.	Registration as an examiner	Texts relating to measures, fines	Priority mediation mandate

Fig. 1. K-means clustering of the resolutions of the Central Bank of Hungary according to different vectorization methods

efficiency: on the one hand, we can say that the more efficient a vectorization method is, the fewer groups there will be that experts cannot name by the correlations between the documents in that group. On the other hand, the more groups that appear in other vectorization methods, the better we can consider the solution because these groups can be considered more relevant. In this respect, it can be said that the TF-IDF method performed the best because the legal experts considered the clustering to be substantively correct in 9 out of 10 cases, and 7 of these were clusters that also appeared in other vectorization methods. LSA follows this with 7/10 content-recognizable clustering, with five appearing elsewhere. Doc2Vec performed the worst, with only half of the clusters named by experts, and only two appearing elsewhere.

Therefore, it can be concluded with a high probability that the TF-IDF method can detect regularities in the text of the resolutions of the Central Bank of Hungary more efficiently than other vectorization methods. It is, therefore, a tool worthy of further use.

3.2. Revealing the characteristics of anonymized Court decisions with clustering and topic modelling

For the court decisions, the task was to develop a category system for the decisions to be able to label them in support of a future supervised machine learning project. This was important because there is no such widely and uniformly used category system for Hungarian court decisions. Furthermore, the category systems used are based on legal dogmatic rules and do not consider the data set's statistical, grammatical, and other regularities. However, these aspects of the category system are essential for supervised machine learning development in order to make the model as efficient as possible. The more efficient the model, the better the results the user gets when searching by these categories, thus improving access to legal information.

Therefore, we used unsupervised tools to answer the following questions to improve the category system we are creating so that not only the aspects highlighted by the legal experts were considered but also statistical regularities:

- How many categories can the dataset be split into?
- What are the main words/terms that characterize these categories?

As a first step, the k-means algorithm was used to determine the number of categories within each jurisdiction that the algorithm considered ideal based on the characteristics of the document set. We can achieve this by using the so-called elbow method that was presented in subsection 2.4. For example, on the entire labour law dataset, this produces the following graph:

Here the elbow-point of the function is the ideal category number according to the algorithm. Figure 2 shows that it is difficult to determine the ideal category number within the complete set of documents in the field of labour law, both because there is no clearly identifiable elbow point and because it is not so exact, given that the ideal number of categories is between 35 and 60.

However, this solution within a predefined main category gives much better results. For example, for the ‘Termination of employment’ category, this is what the function looks like:

Figure 3 shows that the k-means clustering algorithm puts the ideal category number between 3 and 5 (which means 4 in practice), within the main category of Termination of employment. It is interesting to note that if we want to divide this category on a rule basis, based on the Labour Code regulations,[40] we can still divide it into four dogmatic subcategories: Termination by mutual consent, Termination by notice, Termination by dismissal without notice, and Collective redundancies.

An even more illustrative representation is obtained by plotting the documents in space using the k-means algorithm for these four subcategories, which are ideal according to the presented k-means and elbow method:

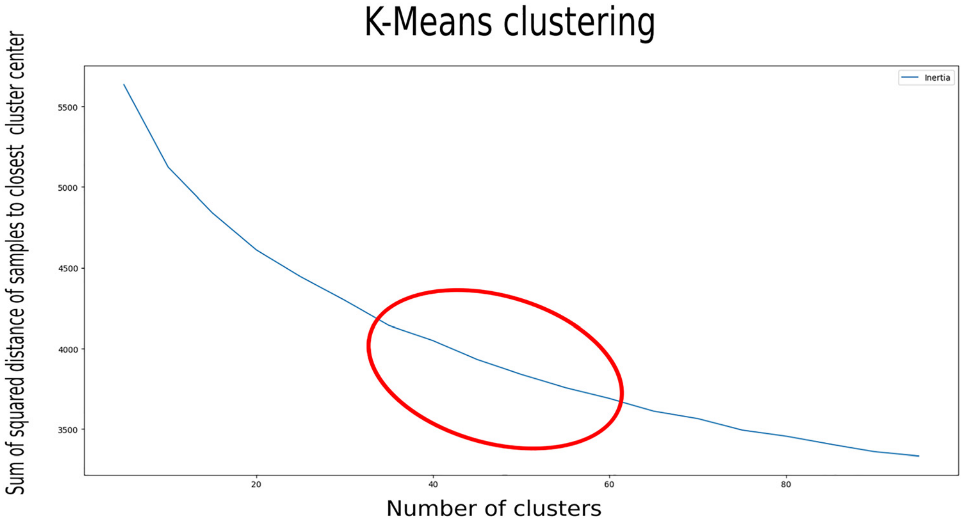


Fig. 2. Estimation of the ideal number of clusters in the whole labour law dataset. The X-axis shows the number of clusters, while the Y-axis shows the summary of the squared distance of the documents to the closest centroid. The squared distance is a calculation between each document and the centre of the topics of the given document. The ideal point and the ideal cluster value is the point of inflection of the curve (‘the elbow’), where the graph starts to look like a straight line

[40] Act I of 2012 on the Labor Code Section 64 (1) and Section 71.

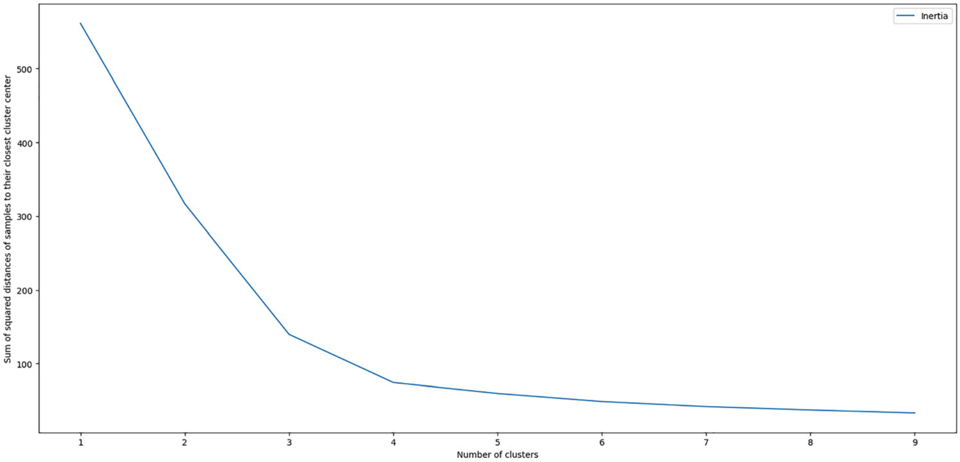


Fig. 3. Estimation of the ideal number of clusters in the Termination of employment main category

Figure 4 shows how the algorithm would divide the main category of Termination of employment into four sub-categories and how these documents are positioned semantically to each other. In this way, it will be possible to distinguish between documents closer to the middle of a group and those which overlap with other groups.

For court decisions, we have complemented this solution with the already presented LDA topic modelling solution, which can visualize the dominant words of each group, thus helping to identify abstract groups. Figure 5 below illustrates this by showing that within the field of labour law, there is a topic group whose defining words are ‘salary grade’, ‘sector supplement’, ‘law

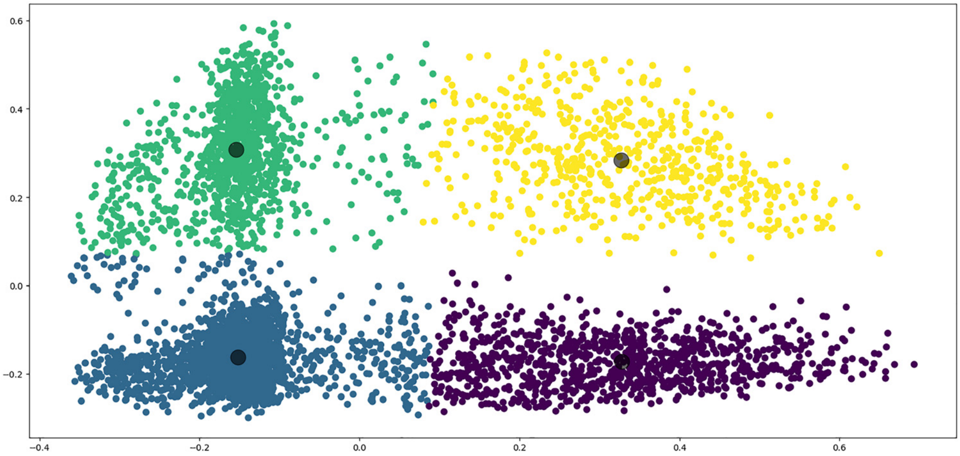


Fig. 4. The k-means algorithm divides the ’Termination of employment’ category into four sub-categories. Each point represents a document according to its location in the vector space

enforcement', and 'law enforcement sector'. This implies that it is worthwhile including a label for those with a law enforcement background, especially concerning their salary, so we created the 'Salary and other allowances for members of the law enforcement' label.

This does not always give valuable results, as the topics are not necessarily specific to one group or may be less meaningful to a human reader. Indeed, as can be seen in Fig. 6 below, this

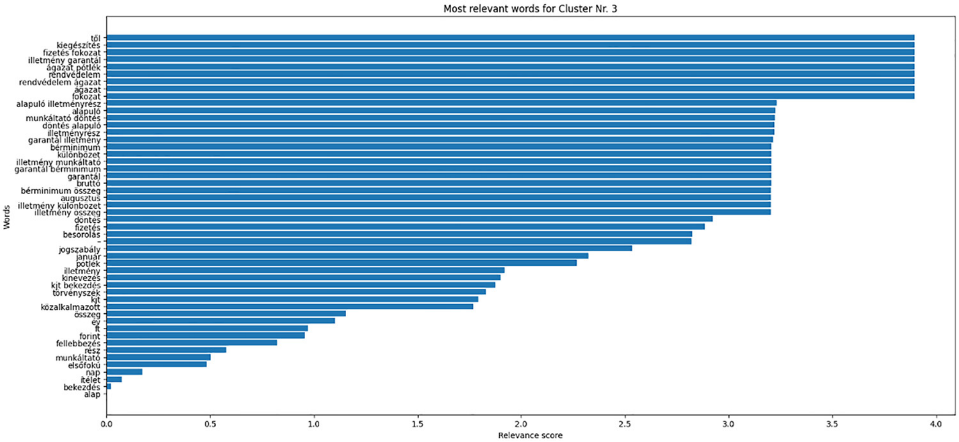


Fig. 5. The most relevant words in the third cluster of the whole labour law dataset. The higher the relevance score of a word, the better it represents the results of topic modelling for the court decisions in this group. The higher the relevance in a cluster, the higher the number of occurrences in the given cluster for that phrase

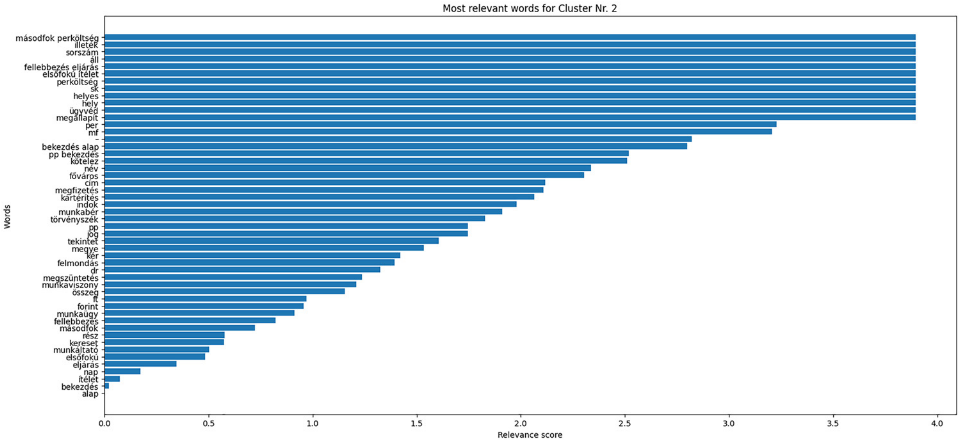


Fig. 6. The most relevant words in the second cluster of the whole labour law dataset. The higher the relevance score of a word, the better it represents the topics of the court decisions in this group

approach also creates a category for mostly procedural terms such as ‘duty on appeal’, ‘fees’, and ‘litigation costs’. This is logical for the LDA algorithm but does not help the user.

In summary, unsupervised tools can help understand the content regularities in unstructured decisions and provide a logical partitioning of the data set that helps to ensure that only part of it needs to be read by someone who is not an expert in the field.

4. DISCUSSION

Unsupervised machine learning tools are suited to reveal the hidden contextual regularities in unstructured legal texts. Different clustering solutions can automatically create groups within text sets, and topic modelling solutions can be used to visualize the subjects of these abstract groups by extracting each group’s most defining words and phrases. These solutions can be applied to a wide range of legal texts. For example, for the resolutions of the Central Bank of Hungary, we only used the algorithm to split the documents into ten groups with different vectorization methods. After that, legal experts reviewed those groups. They can also be used even for court decisions, where we used prior legal dogmatics knowledge and tried to find statistical regularities with these tools.

These pieces of information can then be used in a variety of ways. On the one hand, there is the technical use, in which the categories obtained or refined by unsupervised methods and the most characteristic documents associated with each category can be used to develop a classification model based on supervised learning. In addition, this information can also help legal experts by revealing statistical regularities that they would not have thought of, and thus help them better understand the data set they are working with. Finally, it can also facilitate the search for legal texts by non-experts by giving structure to the document set. It is an essential aspect of increasing access to legal information, as technological tools exist not only to make legally relevant documents, such as legislation and decisions, available to all, but also to help people understand them.

Unsupervised tools, as they are data-driven and domain- and language-independent, can be of great help in uncovering the underlying characteristics of a large dataset, which can significantly help in further developments and tasks. All that is needed is to identify the most effective unsupervised solutions for discovering statistical and content characteristics of a given set of documents.

On the one hand, we were able to use these methods in practice to determine the topics of the resolutions of the Central Bank of Hungary, of which we had no prior knowledge. On the other hand, in the case of anonymized court decisions, we were able to use unsupervised learning methods to support a supervised machine learning categorization task. We used them to create, alongside the regularities of legal terminology, a category system that most accurately and effectively covers the content of decisions. For example, we used k-means and the elbow-method to check whether the number of categories the experts determined is consistent with the ideal category number according to the algorithm. In addition, we used topic modelling to create a label that the experts did not initially consider an essential topic, although the characteristics of the cluster showed that it was worth including as a separate category, as occurred, for example, with the ‘Salary and other allowances for members of the law enforcement’ label.

5. CONCLUSION

Access to legal information is a relatively under-discussed topic of access to justice research. However, one cannot say it is less important than topics such as economic or class differences, gender inequalities, or national and ethnic differences. There are significant differences between people's ability to process and understand legal texts, and we need to even these out if we want to promote equality across the justice system.

As we showed, unsupervised machine learning solutions can help increase the understanding and processing of documents. Clustering can group the different documents automatically based on their similarities and differences. At the same time, topic modelling can help understand the documents and their abstract groups by collecting the words, phrases, and word pairs that define the given documents semantically.

These solutions can also be applied to legal texts and help increase access to legal information. As we examined, in the case of the resolutions of the Central Bank of Hungary, k-means clustering can help introduce a new way of categorization according to the resolutions' semantic content. The legal experts, who were given the documents of the different automatic clusters, found that a considerable proportion of the clusters could be considered adequate. We also compared the efficiency of the various vectorization methods, and found that the tf-idf vectorization was the method best able to detect regularities in the text of the resolutions.

In the case of the anonymized court decisions, which was a significantly larger dataset than the resolutions of the Central Bank of Hungary, we examined deeper statistical regularities than with the latter institution. With these court decisions, k-means clustering and Latent Dirichlet Allocation methods were used successfully. We found that these tools can help understand the content regularities in unstructured decisions and provide a logical partitioning of the data set. In general, one could say that the unsupervised machine learning tools are excellent data-driven and domain- and language-independent solutions that can help find new insights into the dataset, which then can be used for further tasks and developments.

PRIMARY SOURCES

Fundamental Law of Hungary.

Act CLXI of 2011 on the Organization and Administration of the Courts.

Act I of 2012 on the Labor Code.

Act CXXXIX of 2013 on the National Bank of Hungary.

LITERATURE

Bafna, P., Pramod, D. and Vaidya, A., 'Document clustering: TF-IDF approach' (2016) IEEE, 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 61-66.

Bellegarda, J. R., 'Latent semantic mapping [information retrieval]' (2005) 22 IEEE signal processing magazine 70-80.

Berger, H. and Merkl, D. 'A comparison of support vector machines and self-organizing maps for e-mail categorization' in Simoff, S. J., Williams, G. J., Galloway, J. and Kolyshkina, I. (eds), Proceedings

4th Australasian Data Mining Conference - AUSDM05, 5 - 6th December, 2005, Sydney, Australia (University of Technology Sydney 2005) 189-203 <https://static.googleusercontent.com/media/research.google.com/hu//pubs/archive/32722.pdf> accessed 31 March 2023.

Blei, D. M., Ng, A. Y. and Jordan, M. I., 'Latent dirichlet allocation' (2003) 3 Journal of machine Learning research 993-1022.

Cappelletti, M. and Garth, B., 'Access to justice: the world-wide movement to make rights effective' in Capelletti, M. and Garth, B. (eds), *Access to Justice: A World Survey* (Sitjhoff & Noordhoff 1978) page numbers.

Carpio, R. S., 'Women and access to justice. The case of Ecuador' in Puymbroeck, R. V. (ed), *Comprehensive legal and judicial developments: Towards an agenda for a just and equitable society in the 21st Century* (World Bank 2001) 99–107.

Civil Justice Council, 'Online Dispute Resolution for Low Value Civil Claims report' (2015) <https://www.judiciary.uk/wp-content/uploads/2015/02/Online-Dispute-Resolution-Final-Web-Version1.pdf> accessed 31 March 2023.

Gomes, S., 'Access to law and justice perceived by foreign and Roma prisoners' (2019) 9 Race and Justice 359-79.

Google Developers, 'Clustering Algorithms' (2022) <https://developers.google.com/machine-learning/clustering/clustering-algorithms> accessed 29 March 2023.

Hatıpoğlu-Aydın, D. and Aydın, M. B., 'The gender of justice system: Women's access to justice in Turkey' (2016) 47 International Journal of Law, Crime and Justice 71-84.

Jain, A. K., 'Data clustering: 50 years beyond K-means' (2010) 31 Pattern recognition letters 651-66.

Katz, D. M., 'AI + Law: An Overview' in Katz, D. M., Dolin, R. and Bommarito, M. J. (eds), *Legal Informatics* (Cambridge University Press 2021) 87-93.

Katz, D. M. and Nay, J. J., 'Machine Learning and Law' in Katz, D. M., Dolin, R. and Bommarito, M. J. (eds), *Legal Informatics* (Cambridge University Press 2021) 94-98.

Le, Q. and Mikolov, T. 'Distributed representations of sentences and documents' (2014) 32 Proceedings of the 31st International Conference on Machine Learning, PMLR 1188–96.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 'Efficient estimation of word representations in vector space' (2013) arXiv preprint 1-12 https://arxiv.org/abs/1301.3781 accessed 29 March 2023.

Nash, R., 'Financing access to justice: Innovating possibilities to promote access for all' (2013) 5 Hague Journal on the Rule of Law 96–118.

Natarajan, P., Prasad, R., Subramanian, K., Saleem, S., Choi, F. and Schwartz, R., 'Finding structure in noisy text: topic classification and unsupervised clustering' (2007) 10 International Journal of Document Analysis and Recognition (IJDAR) 187-98.

Noone, M. A. and Ojelabi, L. A., 'Alternative dispute resolution and access to justice in Australia' (2020) 16 International Journal of Law in Context 108-27.

Onumanyi, A. J., Molokomme, D. N., Isaac, S. J. and Abu-Mahfouz, A. M., 'AutoElbow: An automatic elbow detection method for estimating the number of clusters in a dataset' (2022) 12 Applied Sciences <https://www.mdpi.com/2076-3417/12/15/7515> accessed 29 March 2023.

Raghuveer, K., 'Legal documents clustering using latent dirichlet allocation' (2012) 2 International Journal of Applied Information Systems 34-37.

Rule, C., 'Designing a Global Online Dispute Resolution System: Lessons Learned from eBay' (2017) 13 U. St. Thomas LJ 354-69.

Susskind, R., *Online courts and the future of justice* (OUP 2019).

Trappey, C. V., Trappey, A. J. and Liu, B. H., 'Identify trademark legal case precedents-Using machine learning to enable semantic analysis of judgments' (2020) 62 World Patent Information <https://www.sciencedirect.com/science/article/abs/pii/S0172219019300638?via%3Dihub> accessed 29 March 2023.

Xu, D. and Tian, Y. 'A comprehensive survey of clustering algorithms' (2015) 2 Annals of Data Science 165-93.

LINKS

Link1: 'Published decisions' <https://alk.mnb.hu/bal_menu/hatarozatok/hatarozatok_keresese> accessed 29 March 2023.

Link2: 'Az MNB határozatok számozása' (The numbering of MNB decisions) <https://www.mnb.hu/felugyelet/engedelyezes-es-intezmenyfelugyeles/hatarozatok-es-vegzesek-keresese/az-mnb-hatarozatok-szamozasa> accessed 29 March 2023.