sustainability

MDPI

Systematic Review

A Survey of Research on Data Analytics-Based Legal Tech

So-Hui Park, Dong-Gu Lee, Jin-Sung Park and Jun-Woo Kim *

Department of Industrial and Management Systems Engineering, Dong-A University, Busan 49315, Korea; qkrthgml4632@naver.com (S.-H.P.); dg4210@daum.net (D.-G.L.); pjs0958@donga.ac.kr (J.-S.P.)
* Correspondence: kjunwoo@dau.ac.kr; Tel.: +82-51-200-7687; Fax: +82-51-200-7697

Abstract: Data analytics provides important tools and methods for processing the data generated during legal services. This paper aims to provide a systematic survey of the research papers on the application of quantitative data analytics algorithms in the legal domain. To this end, relevant research papers were collected and used to analyze topics and trends of research on data analytics-based Legal Tech. The key findings of this paper are as follows. Firstly, the number of research papers about Legal Tech has increased dramatically recently. Secondly, the application of supervised learning techniques to legal judgment data is a very popular approach in this research area. Thirdly, preprocessing legal documents is a very important procedure as many legal documents exist in text form. Fourthly, artificial neural networks and their variations are widely used in research on data analytics-based Legal Tech. Fifthly, data analytics-based Legal Tech is a multidisciplinary research topic related to computer science and social science, etc.

Keywords: Legal Tech; legal industry; data analytics; artificial intelligence; Industry 4.0

Citation: 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. https://doi.org/10.3390/su13148085

Academic Editors: Marc A. Rosen, Jae-Ik Shin, Jae-Won Hong and Ji-Hee Jung

Received: 7 May 2021
Accepted: 18 July 2021
Published: 20 July 2021

1. Introduction

Recently, Industry 4.0 has received much attention from both researchers and practitioners. Industry 4.0 technologies, including artificial intelligence (AI), machine learning (ML), robotics, Internet of Things (IoT), wireless communication, big data, and clouds, are greatly affecting entire economies and society [1,2]. Industry 4.0 technologies are recognized as powerful tools for innovating the productivity and competitiveness of a wide range of industries, including manufacturing [3], education [4], and healthcare [5]. The legal industry is no exception, and this paper focuses on the Industry 4.0 technologies applied to legal services. Traditionally, legal services are provided by human experts; however, modern technologies can be used to automate the service procedures of the legal domain. Consequently, Legal Tech has emerged as an important research topic for the legal and IT industries [6,7].

Legal Tech can be defined as modern technologies and IT solutions that can be used to provide some types of legal services [8], and the application of Legal Tech can be classified into eight sub-areas, as shown in Table 1 [9].

Applications in the first sub-area—involving lawyer marketplace, lawyer-to-lawyer outsourcing, and social and referral networks—are used to find appropriate legal service providers conveniently. The second sub-area, document automation and assembly, includes information systems that can be used to create and process electronic documents in the legal domain. The objective of the third sub-area, involving practice management, case management for specific practice areas, and legal billing, is to provide tools and methods for managing the business data of lawyers and judges, which include work schedules, client information, and law articles, etc. The fourth sub-area, legal research, focuses on legal data search services based on technologies that can be used to parse and interpret text data in the legal domain. The fifth sub-area of Legal Tech aims to apply predictive analysis methods to a training set collected from a law court in order to obtain patterns or models that can be used to predict the trial results of future law cases. The sixth sub-area, electronic

discovery, aims to provide technologies that can be used to identify and collect information in the electronic form required for the process of a trial. The seventh sub-area, online dispute resolution, aims to provide alternative procedures for the resolution of disputes, which can be performed outside of a legal court. The objective of the eighth sub-area, data security technologies, is to protect digital data in the legal domain from cyber-attacks and security threats.

Table 1. Sub-areas of Legal Tech.

No.	Sub-Area
1	Lawyer marketplace; lawyer-to-layer outsourcing; social and referral networks
2	Document automation and assembly
3	Practice management; case management for specific practice areas; legal billing
4	Legal research
5	Predictive analytics and litigation data mining
6	Electronic discovery (e-discovery, e-discovery, eDiscovery, or eDiscovery)
7	Online dispute resolution
8	Data security technologies

This paper focuses on data analytics-based Legal Tech applications, primarily identified from sub-areas 4 and 5 in Table 1. Data analytics can be defined as a procedure of creating value by processing, analyzing, and interpreting raw data [10]. Popular approaches and techniques of data analytics are AI, ML, and data mining, which are key Industry 4.0 technologies [11,12]. Moreover, data analytics can be a particularly useful tool for legal services, since many legal service procedures involve various data that need to be examined by human experts [13]. In other words, data analytics is a promising tool for automating and innovating existing legal services. There are several examples of commercial data analytics-based Legal Tech solutions, such as CaseText and Ross [7].

Several survey papers related to Legal Tech are listed in Table 2. Table 1 suggests that Legal Tech is a multidisciplinary research area associated with a wide range of disciplines, from law to IT and engineering. Hence, previous literature reviews on Legal Tech tended to consider research papers with a wide range of objectives. For instance, Chen [14] provided a survey of application areas of various Legal Tech solutions. Hongdao et al. [7] investigated on markets and business models of Legal Tech. Janoski-Haehlen [15] reviewed curriculums and legal education programs that consider Legal Tech. Salmerón-Manzano [16] grouped the research area of Legal Tech into several clusters including computer science, justice, legal profession, legal design, law firms, and legal education. These papers have focused primarily on commercial aspects of Legal Tech or its influences on society and legal industries.

In contrast, Chalkidis and Kampas [17] provided a literature review on a specific area of Legal Tech, deep learning (DL) applications in the legal domain. Moreover, they suggested three application areas for DL-based Legal Tech, including text classification, information extraction, and information retrieval. DL is one of the AI techniques that can be used to analyze the data collected during legal service procedures. In other words, the paper provided a literature review on Legal Tech applications that can be classified as sub-areas 4 and 5 in Table 1. However, there are many other approaches and methodologies that can also be utilized in those sub-areas. To fill this gap, this paper aims to provide a more comprehensive survey of data analytics applications in the legal domain.

The major contributions of this paper are two-fold. Firstly, a wide range of techniques and algorithms including AI, ML, and data mining are considered. Typically, they are applied to perform conventional data analytics tasks such as classification, regression, clustering, and association, etc. This paper provides a systematic survey of Legal Tech from the perspective of such data analytics tasks. Secondly, additional issues related to data

analytics, such as data source and data structure, are also discussed. In tradition, many unstructured data, such as text documents, are generated and utilized during legal service procedures. Unstructured data are not suitable for conventional data analytics methods and algorithms. Thus, the issues related to data, including data types and preprocessing procedures are important for data analytics-based Legal Tech. This paper provides a comprehensive insight into both quantitative methods and data.

Table 2. Survey papers related to Legal Tech.

No.	Research Paper	Scope	Major Topic
1	Chalkidis and Kampas [17]	Deep learning-based Legal Tech	Algorithms and application areas
2	Chen [14]	Overall Legal Tech	Application areas
3	Hongdao et al. [7]	Overall Legal Tech	Markets and business models of Legal Tech
4	Janoski-Haehlen [15]	Overall Legal Tech	Legal Tech and legal education
5	Salmerón-Manzano [16]	Overall Legal Tech	Research areas
6	This paper	Data analytics-based Legal Tech	Data sources, data structure, and algorithms

The remainder of this paper is organized as follows. Section 2 outlines our survey procedure and the research scope of this paper. Section 3 offers some summary statistics of relevant research papers, which suggest recent trends in research on data analytics-based Legal Tech. Section 4 provides discussions on the objectives, algorithms, and data sets of data analytics-based Legal Tech. Finally, Section 5 concludes this paper with some future research directions.

2. Survey Procedure

Figure 1 depicts the overall research procedure of our survey. The first step was to determine the search keywords that would be used to search for research papers relevant to our survey from academic databases. As shown in Table 3, a combination of 2 keywords, keyword 1 and keyword 2, was used for a single search. Keyword 1 was a word chosen from a set of words {'Legal', 'Law'}, while keyword 2 was an element of {'Data Mining', 'Data Analytics', 'Text Mining', 'Classification', 'Machine Learning', 'Deep Learning', 'Prediction', 'Clustering', 'Tech'}. The sets of words for keyword 1 and keyword 2 were obtained based on a pilot study and the opinions of experts in data analytics.

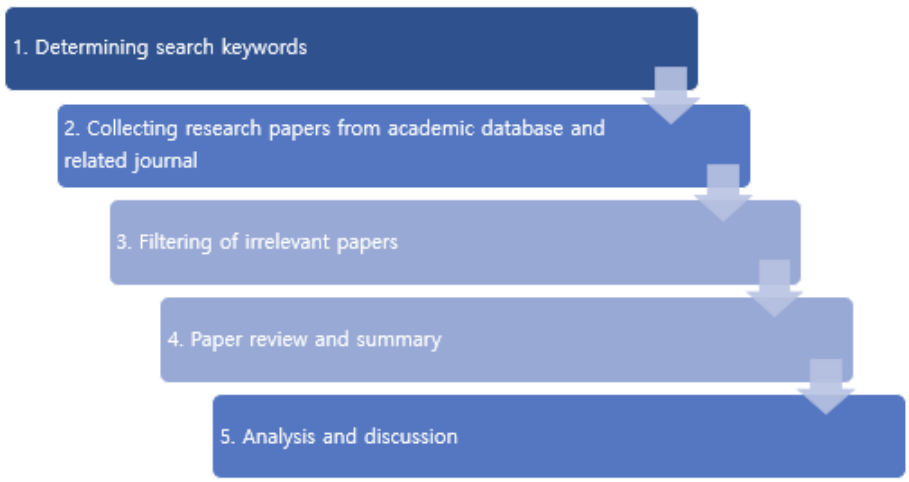

Figure 1. Overall research procedure.

Table 3. Keywords and results of research paper search.

Keyword 1	Keyword 2	# of Research Papers	
		Initial Search	After Filtering
Legal Law	Data Mining	213	10
	Data Analytics	69	2
	Text Mining	44	8
	Classification	1272	13
	Machine Learning	197	9
	Deep Learning	65	5
	Prediction	670	7
	Clustering	478	2
	Tech	159	1
Total # of research papers		3167	57 (39)

The second step was to collect research papers from academic databases and journals. This paper focuses on research papers published in SCI (Science Citation Index), SCIE (Science Citation Index Expanded), SSCI (Social Science Citation Index), and A&HCI (Arts and Humanities Citation Index)-indexed journals, which can be collected by using the search engine provided by the Web of Science. The Web of Science research engine was chosen because it can be used to search research papers published in high-quality academic journals in a wide range of research fields, such as science, technology, social science, and humanities, etc. In other words, research papers published in conference proceedings or other journals are out of the scope of this paper.

Table 3 shows that 3167 research papers were identified from the Web of Science database. Additionally, research papers published in an individual academic journal related to data analytics-based Legal Tech, Artificial Intelligence, and Law, are collected in the second step in Figure 1. However, many of them were irrelevant for our survey on data-analytics-based Legal Tech. For instance, many research papers in the fields of mineralogy or geology were included in the initial search results due to the use of the term 'Mining' in keyword 2. Thus, the objective of the third step was to filter out irrelevant research papers from the initial search results. To this end, the authors manually examined the titles, abstracts, and author affiliations of the collected papers. Note that this paper focuses primarily on data analytics applications designed to analyze and process data generated during legal and judicial procedures, which include legal judgments, court records, regulations, and law articles, etc. Such data have been analyzed and processed by human experts, such as lawyers and judges; however, this procedure is expected to be automated by using modern Industry 4.0 technologies. On the contrary, research papers focusing on the analysis of data collected from outside of legal and judicial procedures, including behavioral data of citizens, are not considered in this paper. For instance, research topics related to law enforcement, such as crime detection and digital forensic, and patent mining based on technical documents were filtered out at this step. Furthermore, research papers on logical analysis of legal argumentation, architecture of legal information systems, and legal issues related to AI are excluded from our survey study.

Among the research papers collected by using the Web of Science search engine, 57 papers were identified as being relevant for our survey, as shown in Table 3; however, 18 of them were duplicates. Thus, only 39 papers were used in this paper. In addition, 25 relevant research papers were identified from the aforementioned journal, Artificial Intelligence and Law. Consequently, 64 research papers remained for our survey study after the filtering of irrelevant papers and the removal of duplicates.

In the fourth step of our survey procedure, the 64 identified research papers were carefully reviewed by the authors. Specifically, the authors extracted key features of the

research papers, including the publication year, subject area of the journal, country, and continent of the first author's institute, data source, approaches and algorithms for data analytics, etc. Additionally, these features were summarized using tables or charts.

Finally, the results of the review of the relevant research papers were analyzed and discussed in the fifth step. Consequently, remarkable trends and future research directions for data-analytics-based Legal Tech are provided by our survey study.

3. Basic Trends of Research on Data Analytics-Based Legal Tech

3.1. Number of Research Papers Related to Data Analytics-Based Legal Tech

Figure 2 shows recent changes in the number of research papers that deal with data analytics-based Legal Tech. The topic was not so popular in the early 2000s, which might be due to the immaturity of related technologies such as AI, ML, and computer hardware. In contrast, published research papers on data analytics-based Legal Tech have been increasing since about 2010. In particular, research on data analytics-based Legal Tech has shown an exponential increase in the three years from 2018 to 2020. In other words, data analytics-based Legal Tech has recently been emerging as a popular research topic.

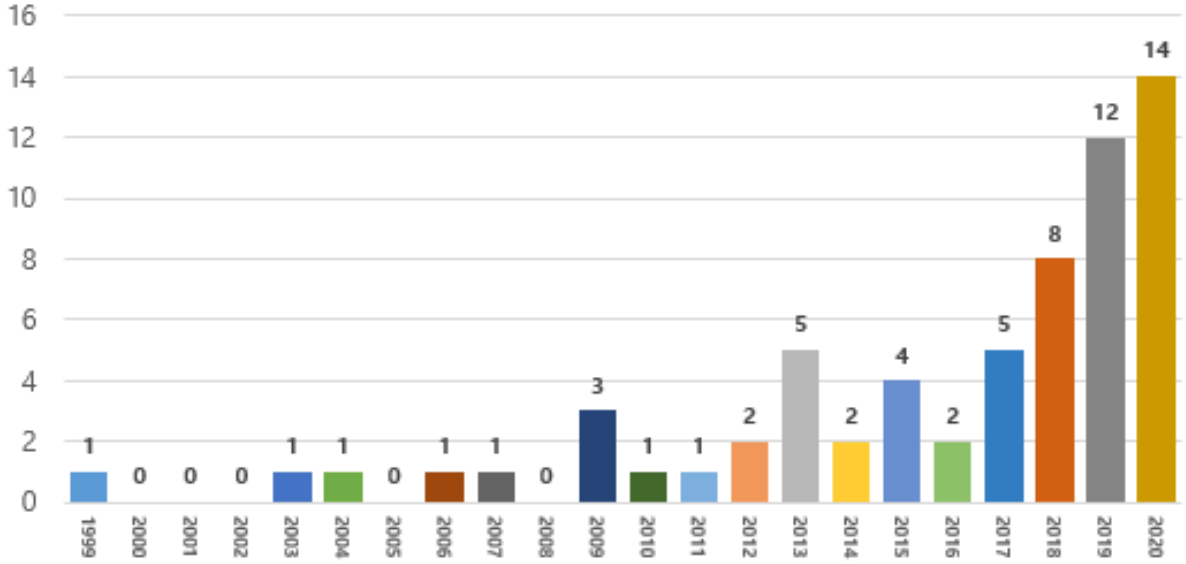

Figure 2. Number of research papers per year.

3.2. Geographical Distribution of Research Papers on Data Analytics-Based Legal Tech

Figure 3 depicts the geographical distribution of the relevant research papers, where half of the papers have been published by a first author from Asia. Studies from Europe (31%) and North America (11%) make up the second and the third largest portions in research on data analytics-based Legal Tech.

Specifically, Figure 4 reveals that the number of research papers from Asia was not significantly higher than that from other continents a couple of years ago. In contrast, there was a dramatic increase in the number of research papers from Asia from 2019. Moreover, Figure 5 shows that most of the research papers from Asia are authored by researchers working for institutes in China. In other words, data analytics-based Legal Tech has recently emerged as a popular research topic, especially in China.

The Chinese government is supporting the development of Industry 4.0 technologies, including Legal Tech, and this seems to contribute to the achievements of Chinese researchers [18]. The Internet court at Hangzhou, China, established in 2017, is the first online court in the world, which provides various services required for judicial proceedings [19]. Other Internet courts in Beijing and Guangzhou were established in 2018, and Chinese media reported that millions of legal cases are handled by this innovative legal

service [20]. In an Internet court, data and documents required for legal proceedings are created, collected, and stored electronically. Such electronic data and documents can be processed more conveniently and efficiently by applying data analytics that can be used to extract meaningful knowledge and useful patterns from large volumes of data. In this context, modern legal services such as Internet courts can provide promising application areas for data analytics-based Legal Tech.

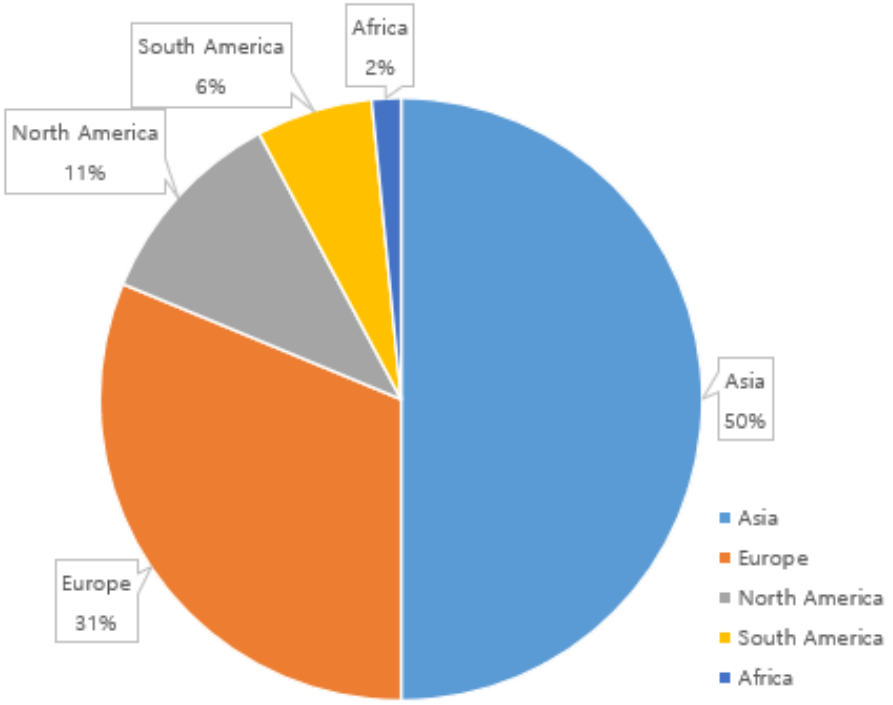

Figure 3. Geographical distribution of research papers on data analytics-based Legal Tech.

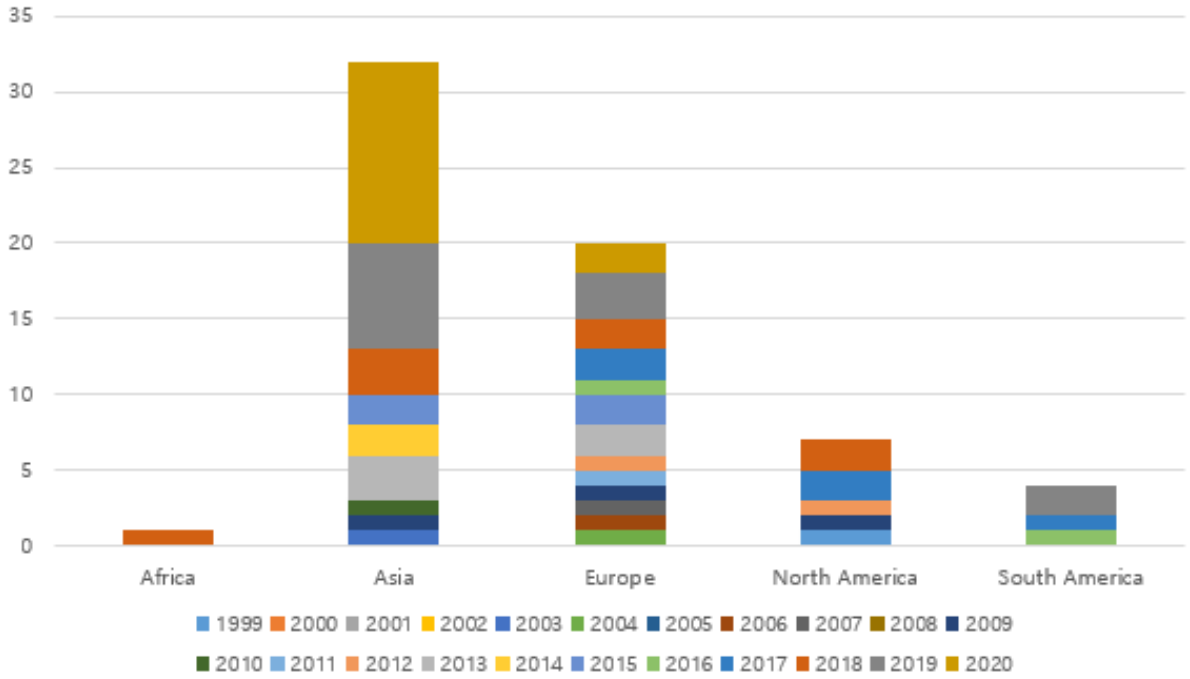

Figure 4. Number of research papers by continent and year.

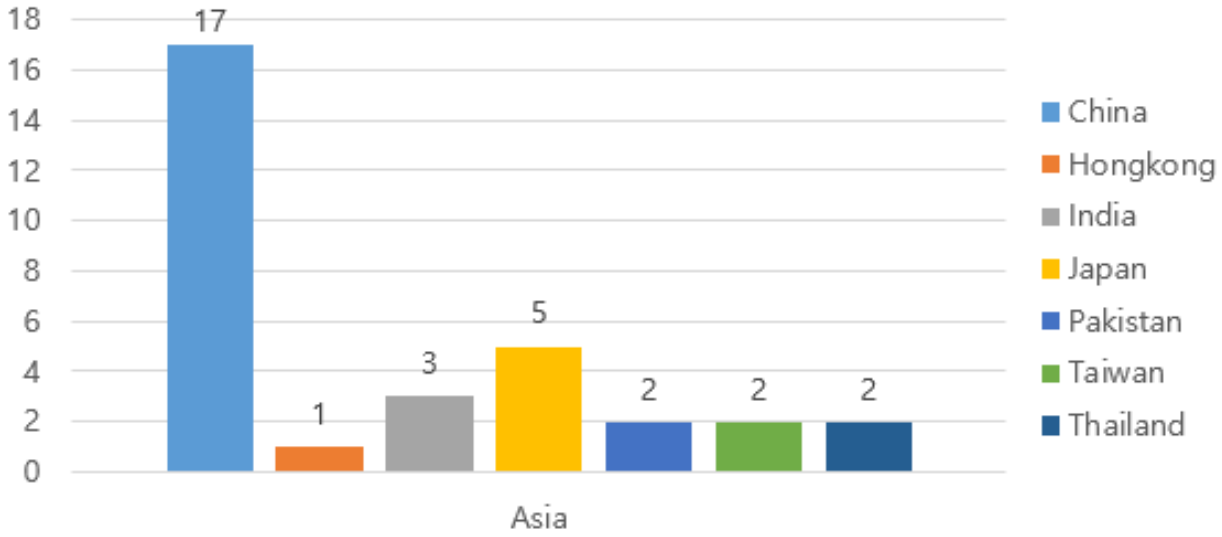

Figure 5. Number of research papers by country in Asia.

3.3. Subject Area of Journals

Figure 6 summarizes the subject areas of the journals that have published relevant research papers. The subject area of a journal can be identified on the Web of Science website, where a single journal can be associated with two or more subject areas. From Figure 6, the following observations can be made.

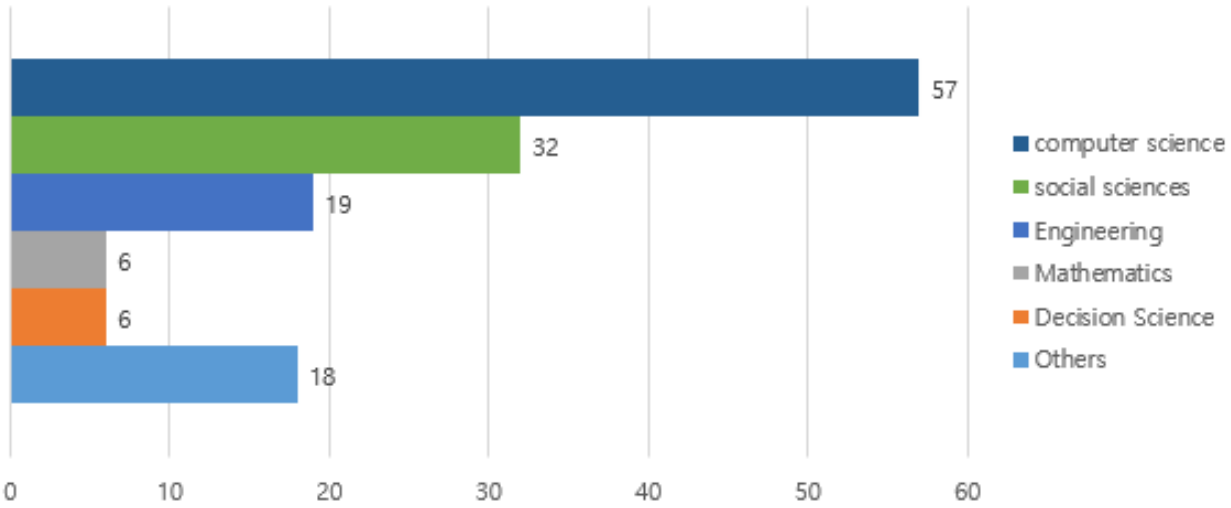

Figure 6. Number of research papers by subject area.

Firstly, the most popular subject area for research papers on data analytics-based Legal Tech is computer science, where 57 out of 64 research papers are published in journals associated with computer science. This implies that the recent remarkable achievements of computer science, such as AI, ML, big data, and cloud, are important technological enablers for Legal Tech. Additionally, Legal Tech is emerging as a promising application domain for modern Industry 4.0 technologies.

Secondly, the second most popular subject area is social sciences, which includes "law" as one of its sub-areas. In other words, Legal Tech is being paid much attention by researchers from the field of law since Legal Tech is expected to have significant impacts on legal industries and services. Moreover, the first and the second most popular subject

areas in Figure 6 indicate that data analytics-based Legal Tech is a kind of multidisciplinary research topic.

Thirdly, engineering, mathematics, and decision sciences are the third, the fourth, and the fifth most popular subject areas, respectively. These subject areas are related to quantitative analysis procedures for extracting meaningful knowledge and useful patterns from large volumes of data, which is an integral part of data analytics-based Legal Tech. In other words, such algorithms and methodologies are useful tools for implementing Legal Tech applications.

Lastly, other subject areas are classified as "Others" in Figure 6. These subject areas include materials science, physics and astronomy, medicine, and environmental science, which are not directly related to data analytics or Legal Tech. This reveals that the research papers on data analytics-based Legal Tech are published in a wide range of academic journals.

3.4. Data Sources for Data Analytics-Based Legal Tech Research

In this paper, the types of data sources for data analytics-based Legal Tech research are grouped into five categories, as shown in Table 4 and Figure 7. Note that two or more data sources can be utilized together in a single research paper.

Table 4. Data sources of data analytics-based Legal Tech research.

Data Source	Description
Court record	Historical records of trials
Civil petition	Documents about requests or demands from the public
Law	Clauses and articles of laws
Legal judgment	Historical records of court decisions and descriptions of cases
Legislative document	Historical records bills or votes on them in congress

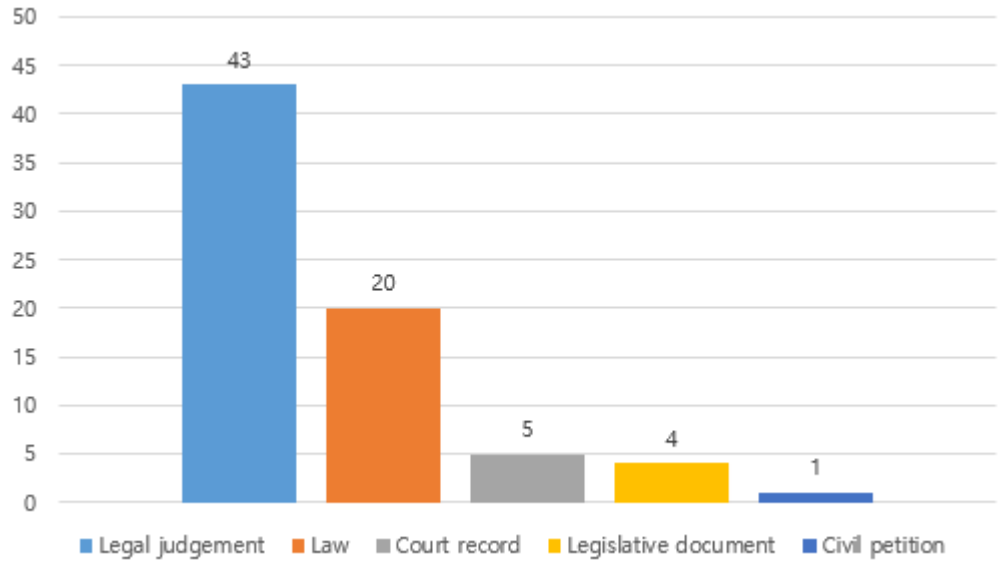

Figure 7. Data sources of relevant research papers.

'Legal judgment' is a historical record about previous cases, which includes court decisions and case descriptions, such as defendant profiles and information about the associated law articles, etc. The majority (67.2%) of the papers applied data analytics techniques to legal judgment data. Trial for lawsuits is the most fundamental service of the legal industry, and it is clear that much data is created, collected, and processed during the process of trials. In this context, legal judgment data are the most popular data source for research on data analytics-based Legal Tech.

'Law' includes clauses and articles of laws that can be found in legal codes, and 31.3% of the relevant research papers utilized this type of data. Law data also play an important role in the process of trials in that they provide a basis for judgments and court decisions.

Other data source types include 'Court record', 'Legislative document', and 'Civil petition'. Court record data contain records of testimonies, statements, and discussions collected during trials in court [21,22]. Five research papers analyzed this type of data by applying data analytics techniques. Legislative documents are historical records on the legislative process in congress, which include bill text, legislator profiles, and past vote histories [23]. This type of data is used in four relevant research papers. A civil petition is a request by a civil petitioner to an administrative agency to take a disposition or other specific action. This type of data is used in one relevant research paper.

Traditionally, the types of data in Figure 7 would be processed and interpreted by human experts in order to provide a wide range of legal services. However, data analytics techniques can be used to extract meaningful knowledge and useful patterns from various data, and this allows for providing automated or semi-automated legal services. For instance, AI-based legal decision support systems, including an AI judge that can provide a sentencing recommendation on pending lawsuits, emerged as a promising tool for enhancing judicial decision-making procedures. In general, AI judges are designed to analyze legal judgment data, the most popular data source in Figure 7, and make decisions on pending lawsuits by using decision models extracted from the data. Human judges can make judicial decisions efficiently by using preliminary decisions provided by AI judges [24]. In this way, data analytics can be a powerful tool for innovating legal service procedures by analyzing legal data as shown in Figure 7.

4. Approaches and Algorithms for Data Analytics-Based Legal Tech

4.1. Algorithms and Methods of Data Analytics

The main data analytics techniques include data mining, AI, statistics, etc. In this study, the approaches and algorithms of the relevant research papers are categorized by applying the taxonomies of these techniques.

Data mining can be defined as an automated or semi-automated procedure for extracting knowledge, rules, and patterns from large volumes of data [25]. Typically, data mining techniques are grouped into two categories, namely supervised learning (predictive analytics) and unsupervised learning (descriptive analytics), as shown in Figure 8.

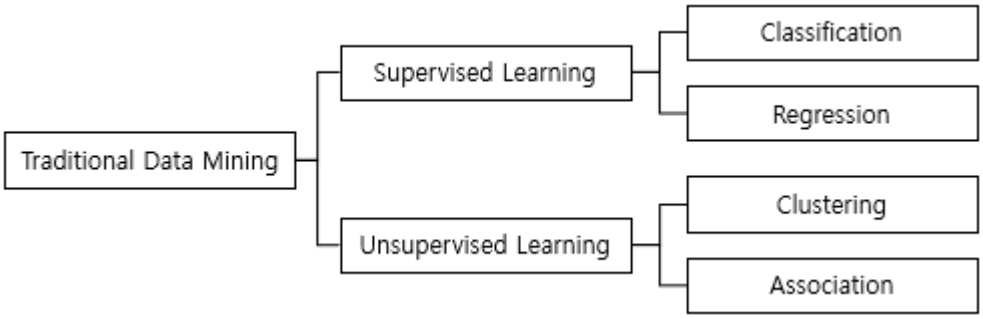

Figure 8. Traditional data mining tasks.

The ultimate goal of supervised learning is to predict the value of a target variable for new input data. To this end, supervised learning techniques are designed to build a model that explains the relationships between a target variable and predictors by analyzing a training set where the values of the target variable are known [26]. In other words, training sets for supervised learning contain both predictors and target variables. A target variable is an output value or dependent variable to be estimated, while predictors are independent variables that can affect the target variable [27]. Supervised learning is subdivided into two types: classification and regression. Classification is used for estimating the value

of a categorical target variable [28]. Examples of classification techniques are decision trees, Bayesian classifiers, and nearest neighbor, etc. [26,29,30]. In contrast, the objective of regression is to estimate a numerical target variable. Examples of regression techniques are linear regression, ridge regression, lasso regression, and artificial neural networks (ANNs), etc. [31].

Unsupervised learning is used to understand and describe the structure of a given data set. Typically, techniques and algorithms of unsupervised learning do not consider a target variable. Unsupervised learning includes clustering analysis and association analysis. The objective of clustering is to find groups of records, such that similar records belong to an identical group while dissimilar records belong to different groups. Examples of clustering algorithms are k-means, DBSCAN, and hierarchical clustering, etc. [26,29,30]. On the contrary, association analysis is used to extract interesting association rules, which represent the cause-effect relationships among the variables, from a transactional data set [32]. Association analysis can be performed by applying the well-known Apriori algorithm and its variations [32,33] or the FP growth algorithm [34].

Supervised learning techniques can be used to analyze the relationship between target variables and predictors related to legal decisions. For instance, those techniques enable the prediction of the trial result if appropriate models and predictor variables are given. In contrast, unsupervised learning techniques are typically used to analyze similarities or correlations between legal documents. For instance, sub-groups of similar cases can be identified by applying clustering analysis to legal judgment data.

Recently, much attention has been paid to AI and ML. AI techniques are used to develop computers or machines that can mimic human intelligence. ML provides algorithms that enable machines to learn from given examples. Thus, the definitions of AI and ML are slightly different from that of data mining. Nevertheless, the aforementioned traditional data mining tasks of classification, regression, clustering, and association can also be performed by applying algorithms and techniques of AI and ML [31,35].

Additionally, two modern techniques, namely text mining and network analysis, are also important approaches for data analytics-based Legal Tech according to our survey. Text mining is the process of extracting useful information from text data. Text data are unstructured data that are hard to process and analyze. Since raw data for Legal Tech often exist in text form, preprocessing for raw data is an important issue of data analytics-based Legal Tech. The ultimate goal of preprocessing is to obtain useful and refined data to be analyzed by data analytics techniques. Well-known preprocessing techniques include feature selection, feature construction, missing value imputation, data integration, and data transformation, etc. These techniques help to obtain high-quality data more suitable for data analytics [29]. Among them, data transformation techniques are very important for Legal Tech, because most traditional data analytics algorithms cannot deal with text data directly.

Text mining algorithms provide non-trivial procedures for transforming text data into structured data [36–38]. Well-known text mining techniques include TF-IDF, a bag of words (BoW), and word embedding, etc. TF-IDF generates structured data in the form of a document-term matrix, where the importance of a term is calculated by using frequency and inverse document frequency. For a single term, TF (term frequency) and IDF (inverse document frequency) denote the number of occurrences and their reciprocal, respectively [39]. Let us consider three short documents: document 1, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified'; document 2, 'Autonomous Weapons Systems and the Law of Armed Conflict'; and document 3, 'On the Right of Citizens to Assemble Peacefully, without Weapons, Freely Conduct Meetings and Demonstrations'. The TF values for the terms within documents 1–3 are summarized in Table 5.

The IDF value for a specific term is calculated by using DF (document frequency), the number of documents containing the term, as follows:

$$\mathrm{IDF} = \log\left(\frac{n}{\mathrm{DF}}\right) \tag{1}$$

where n is the number of given documents. Table 6 summarizes the IDF values for the terms within documents 1–3.

Then, TF-IDF values can be calculated by using (2), and the TF-IDF values for this example are summarized in Table 7.

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} \times \mathrm{IDF} \tag{2}$$

In a TF-IDF matrix, a term is frequently found in given documents if the sum of values in the corresponding column is small. For instance, 'Weapons' and 'and' are frequent terms in Table 7.

Similarly, BoW generates a document-term matrix by using the frequencies of given terms [40]. Assume three documents: document 1, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions, and Legal Issues to be Clarified'; document 2, 'Autonomous Weapons Systems and the Law of Armed Conflict'; and document 3, 'On the Right of Citizens to Assemble Peacefully, without Weapons, Freely Conduct Meetings and Demonstrations'. A BoW-based document-term matrix for the terms within documents 1–3 is shown in Table 8. Note that this document requires the length of each document. In Table 8, a document is represented as a row vector containing frequency values of the given terms. These vectors can be used to calculate similarity or dissimilarity between documents.

Word embedding is used to convert a term into a dense vector containing continuous values, which can be obtained by learning given documents or corpus data [41]. Word embedding enables one to represent a term by a vector with lower dimensionality. Moreover, the dense vector of word embedding can reflect similarity or dissimilarity between terms.

Typically, raw data generated and collected during legal service procedures are in the form of text. However, many data analytics algorithms are designed to handle structured data such as table data. Thus, text mining techniques, including TD-IDF, BoW, and word embedding, are important in that they provide useful preprocessing methods for data analytics-based Legal Tech. The transformed data in Tables 5–8 are in tabular form, where a record is characterized by a number of variables. Since many data analytics algorithms, supervised and unsupervised techniques, assume tabular structured data, this data structure is most commonly used in the fields of AI, ML, and data mining. In contrast to preprocessing, postprocessing techniques are used to interpret and utilize knowledge and patterns obtained by applying supervised and unsupervised learning techniques more effectively. Data summarization and visualization are examples of postprocessing tasks, however, the postprocessing procedure is out of the scope of this paper.

Another important modern data analysis approach is network analysis. Network analysis is the process of understanding the structures of a given network and the relationships between nodes therein [23]. For instance, legal documents and their citation relationships can be represented as nodes and edges of a network. Network analysis techniques can be used to find important nodes or a community of some nodes, which are useful for information visualization and document recommendation [42,43].

Table 5. Example of TF values.

Document	Autonomous	Weapons	and	International	Humanitarian	Law	Advantages	Open	Technical	Questions	Legal	Issues
Document 1	1	1	2	1	1	1	1	1	1	1	1	1
Document 2	1	1	1	0	0	1	0	0	0	0	0	0
Document 3	0	1	1	0	0	0	0	0	0	0	0	0
Document	**to**	**be**	**Clarified**	**Systems**	**the**	**of**	**Armed**	**Conflict**	**On**	**Right**	**Citizens**	**Assemble**
Document 1	1	1	1	0	0	0	0	0	0	0	0	0
Document 2	0	0	0	1	1	1	1	1	0	0	0	0
Document 3	1	0	0	0	1	1	0	0	1	1	1	1
Document	**Peacefully**	**without**	**Freely**	**Conduct**	**Meetings**	**Demonstrations**						
Document 1	0	0	0	0	0	0						
Document 2	0	0	0	0	0	0						
Document 3	1	1	1	1	1	1						

Table 6. Example of IDF values.

Autonomous	Weapons	and	International	Humanitarian	Law	Advantages	Open	Technical	Questions	Legal	Issues
0.176	0	0	0.477	0.477	0.176	0.477	0.477	0.477	0.477	0.477	0.477
to	**be**	**Clarified**	**Systems**	**the**	**of**	**Armed**	**Conflict**	**On**	**Right**	**Citizens**	**Assemble**
0.176	0.477	0.477	0.477	0.176	0.176	0.477	0.477	0.477	0.477	0.477	0.477
Peacefully	**without**	**Freely**	**Conduct**	**Meetings**	**Demonstrations**						
0.477	0.477	0.477	0.477	0.477	0.477						

Table 7. Example for TF-IDF values.

Document	Autonomous	Weapons	and	International	Humanitarian	Law	Advantages	Open	Technical	Questions	Legal	Issues
Document 1	0.176	0	0	0.477	0.477	0.176	0.477	0.477	0.477	0.477	0.477	0.477
Document 2	0.176	0	0	0	0	0.176	0	0	0	0	0	0
Document 3	0	0	0	0	0	0	0	0	0	0	0	0
Document	**to**	**be**	**Clarified**	**Systems**	**the**	**of**	**Armed**	**Conflict**	**On**	**Right**	**Citizens**	**Assemble**
Document 1	0.176	0.477	0.477	0	0	0	0	0	0	0	0	0
Document 2	0	0	0	0.477	0.176	0.176	0.477	0.477	0	0	0	0
Document 3	0.176	0	0	0	0.176	0.176	0	0	0.477	0.477	0.477	0.477
Document	**Peacefully**	**without**	**Freely**	**Conduct**	**Meetings**	**Demonstrations**						
Document 1	0	0	0	0	0	0						
Document 2	0	0	0	0	0	0						
Document 3	0.477	0.477	0.477	0.477	0.477	0.477						

Table 8. Example of BoW-based document-term matrix.

Document	Autonomous	Weapons	and	International	Humanitarian	Law	Advantages	Open	Technical	Questions	Legal	Issues
Document 1	1	1	2	1	1	1	1	1	1	1	1	1
Document 2	1	1	1	0	0	1	0	0	0	0	0	0
Document 3	0	1	1	0	0	0	0	0	0	0	0	0
Document	**to**	**be**	**Clarified**	**Systems**	**the**	**of**	**Armed**	**Conflict**	**On**	**Right**	**Citizens**	**Assemble**
Document 1	1	1	1	0	0	0	0	0	0	0	0	0
Document 2	0	0	0	1	1	1	1	1	0	0	0	0
Document 3	1	0	0	0	1	1	0	0	1	1	1	1
Document	**Peacefully**	**without**	**Freely**	**Conduct**	**Meetings**	**Demonstrations**			**Length of document**			
Document 1	0	0	0	0	0	0			16			
Document 2	0	0	0	0	0	0			9			
Document 3	1	1	1	1	1	1			15			

4.2. Input Data and Algorithms

The research papers that applied supervised learning algorithms to legal data are listed in Table 9, while research papers that used unsupervised learning algorithms are shown in Table 10. In addition, research papers that are not contained in Table 9 or Table 10 are listed in Table 11.

The 'Data structure' columns of Tables 9–11 indicate the type of input data for the data analytics algorithms. This paper considers six types of data structure, including bag of words, TF-IDF, word embedding, segmented document, structured document, and text. Data in the form of the first three types—namely, bag of words, TF-IDF, and word embedding—are generated by applying text-mining algorithms. A segmented document can be defined as a set of elements obtained by splitting the contents of a given document. For instance, keywords, sentences, and paragraphs can be used as the elements for creating a segmented document. In structured document-type data, a single document is represented by using a number of features that can be identified from a given document. Examples of features for structured document-type data include information on the victim or defendant, location of judgment, and amount of money involved, etc. [44,45]. Moreover, this type of data is sometimes provided in the form of an XML (Extensible Markup Language) document [44,46,47] or electronic database [42,48,49]. Inherently, structured document-type data are a sort of table data. Thus, this type of data can be processed and analyzed in a convenient way if appropriate features are carefully developed [50,51]. Finally, documents containing plain text are classified as text-type data in Tables 9–11.

Table 9. Research papers dealing with traditional supervised learning tasks.

No.	Task	Author	Data Category	Data Structure	Algorithm
1	Classification	Ji et al. [21]	Court record	Segmented document	ANN&DL
2		Ji et al. [22]	Court record	Segmented document	ANN&DL
3		Li et al. [44]	Legal judgment	Structured document	ANN&DL
4		Sharafat et al. [45]	Legal judgment	Structured document	Hidden Markov model
5		Thammaboosadee and Watanapa [46]	Legal judgment	Structured document	ANN&DL
6		Thammaboosadee et al. [47]	Legal judgment	Structured document	ANN&DL, decision tree
7		Katz et al. [49]	Legal judgment	Structured document	Random forest
8		Mitchell et al. [51]	Legal judgment	Structured document	ANN&DL, Bayesian
9		Ashley and Brüninghaus [52]	Legal judgment	Bag of words	Nearest neighbor
10		Boella et al. [53]	Legal judgment	Bag of words	SVM
11		Boella et al. [54]	Law	TF-IDF	SVM
12		Cheng et al. [55]	Legislative document	Structured document, bag of words	Multiple kernel learning
13		El Jelali et al. [56]	Legal judgment	TF-IDF	SVM, Bayesian, Decision Tree
14		Fang et al. [57]	Legal judgment	Bag of words	Induction network, Relation network
15		Fernandes et al. [58]	Court record	Structured document, word embedding	ANN&DL
16		Fornaciari and Poesio [59]	Court record	Word embedding	SVM
17		Francesconi and Passerini [60]	Legislative document	TF, TF-IDF	SVM, Bayesian
18		Francesconi and Peruginelli [61]	Law	TF, TF-IDF	SVM, Bayesian

Table 9. *Cont.*

No.	Task	Author	Data Category	Data Structure	Algorithm
19		Guo et al. [62]	Legal judgment	Word embedding	ANN&DL
20		Hachey and Grover [63]	Legal judgment	TF-IDF	SVM, Bayesian, Decision Tree
21		Iftikhar et al. [64]	Legal judgment	Structured document	ANN&DL
22		Jin et al. [65]	Legal judgment	Structured document	ANN&DL
23		Lesmo et al. [66]	Law	Structured document	Heuristic
24		Li et al. [67]	Law, legal judgment	Bag of words, TD-IDF	Markov logic network, SVM
25		Li et al. [68]	Law, legal judgment	Bag of words	ANN&DL
26		Li et al. [69]	Legal judgment	TF-IDF	ANN&DL
27		Liu and Chen [70]	Legal judgment	TF-IDF	SVM
28		Liu et al. [71]	Legal judgment	TF-IDF	SVM
29		Ma et al. [72]	Legal judgment	Word embedding	kNN
30		Mahfouz and Kandil [73]	Legal judgment	Structured document	ANN&DL, SVM, Bayesian
31		Medvedeva et al. [74]	Legal judgment, law	TF-IDF	SVM
32		Nanda et al. [75]	Law	TF-IDF	ANN&DL, SVM, Bayesian
33		Nguyen et al. [76]	Law	Word embedding	ANN&DL
34		Pudaruth et al. [77]	Legal judgment	TF-IDF	kNN
35		Qiu et al. [78]	Legal judgment	Word embedding	ANN&DL, SVM, kNN, Bayesian
36		Raghupathi et al. [79]	Legal judgment	TF-IDF	Bayesian
37		Saravanan and Ravindran [80]	Legal judgment	Segmented document	Heuristic
38		Shulayeva et al. [81]	Legal judgment	TF-IDF	Bayesian
39		Tran et al. [82]	Law	Segmented document	Maximum entropy, SVM
40		Waltl et al. [83]	Law	Structured document, TF-IDF	ANN&DL, SVM
41		Yamada et al. [84]	Legal judgment	Segmented document	SVM
42		Yang and Luk [85]	Law	Bag of words	ANN&DL
43		Yao et al. [86]	Legal judgment	Segmented document	ANN&DL
44	Regression	Li et al. [50]	Legal judgment	Structured document	ANN&DL, linear regression
45		Guo et al. [87]	Legal judgment	Text	Lasso regression
46		Guo et al. [88]	Legal judgment	Text	Ridge regression
47		Tran et al. [89]	Legal judgment	Segmented document, word embedding	ANN&DL

Table 10. Research papers dealing with traditional unsupervised learning tasks.

No.	Task	Author	Data Category	Data Structure	Algorithm
1	Clustering	Fang et al. [57]	Legal judgment	Bag of words	Hierarchical clustering
2		Raghupathi et al. [79]	Legal judgment	TF-IDF	K-means
3		Tran et al. [82]	Law	text	Brown clustering
4		Yang and Luk [85]	Law	Bag of words	ANN&DL
5		Moens et al. [90]	Legal judgment	TF-IDF	K-medoid
6		Acharya et al. [91]	Legal judgment	TF-IDF	K-means
7		Sadeghian et al. [92]	Law	Word embedding	K-means
8	Association	Liu et al. [71]	Legal judgment	TF-IDF	Apriori

Table 11. Research papers dealing with other tasks.

No.	Author	Data Category	Data Structure	Algorithm
1	Lettieri et al. [42]	Legal judgment, law, legislative document	Structured document	Summary statistics, data visualization
2	Petrović and Stanković [43]	Legislative document	Segmented document	Network analysis
3	De Luise et al. [48]	Law	Structured document	Heuristic
4	Bartolini et al. [93]	Law	Structured document	Heuristic
5	Boulet et al. [94]	Law	Structured document	Network analysis
6	Boulet et al. [95]	Law	Structured document	Network analysis
7	Chen et al. [96]	Legal judgment	TF-IDF	Heuristic
8	de Araujo et al. [97]	Court record	Structured document	Heuristic
9	Fan and Li [98]	Legal judgment	TF-IDF	Heuristic
10	Hasan et al. [99]	Law	Structured document	Heuristic
11	Herrera et al. [100]	Civil petition	Structured document	Topic modeling
12	Le et al. [101]	Legal judgment	TF-IDF	Heuristic
13	Saravanan et al. [102]	Legal judgment	Structured document	Heuristic

For each research paper, the algorithm(s) applied by the authors can be found in the 'Algorithm' columns of Tables 9–11, where all the ANN-based algorithms, such as multi-layer perceptron (MLP) and DL, are classified as ANN&DL.

Tables 9–11 provide the following observations. Firstly, preprocessing plays a significant role in most research papers. In other words, input data for data-analytics-based Legal Tech are generally obtained by applying preprocessing techniques to raw legal documents. Several research papers in Tables 9 and 10 proposed methodologies that can be applied to text-type data; however, those methodologies often contain their own preprocessing procedures that transform text-type data into a more structured form [87,88]. Thus, it can be concluded that it is difficult to directly use legal documents in the form of text to develop data analytics-based Legal Tech applications.

Secondly, a structured document is also a popular data structure. It is well known that data quality has a significant impact on the usefulness of analysis results [29]. Moreover, structured documents of high quality can be obtained by creating meaningful features or variables that describe the contents of the raw data well. Such features can be created automatically by using relevant tools such as natural language processing (NLP) [64]. In this context, it is expected that structured document-type data will continue to be widely adopted by Legal Tech applications.

Thirdly, a significant number of research papers applied supervised learning algorithms to legal judgment data. One reason is that legal judgment is quite an important decision-making process in the legal industry. The other reason is the structure of legal judgment data. In order to apply a supervised learning algorithm, input data should contain both a target variable and predictor variables, such that predictors affect the target variable. In legal judgment data, trial results such as the length of imprisonment and amount of penalty are affected by other information such as the associated law articles and defendant profiles [87]. In other words, trial results and other information can be used as target variables and predictors, respectively. Thus, legal judgment data can be regarded as a good data source for supervised learning tasks. This led the supervised learning task to be dealt with more frequently than unsupervised learning and other tasks.

Fourthly, among the traditional unsupervised learning tasks, clustering is dealt with more often than association, as shown in Table 10. For instance, clustering analysis is a useful tool for discovering a group of legal documents to be focused on [79,91].

Furthermore, the research papers listed in Table 11 generally provide methodologies and applications for information extraction. The topics of the research papers include entity recognition [43,48,93,99], similarity-score-based recommendation or information retrieval [96,98], information visualization [42], and opinion mining [100], etc.

Lastly, 21 of 64 (32.8%) research papers utilized ANN&DL algorithms, which revealed the good performance and wide applicability of ANN and its variations.

The tasks and algorithms of previous research papers are summarized in Table 12. Among supervised learning tasks, the classification task is more frequently tackled than regression is. In other words, most research papers that utilize supervised learning techniques consider categorical target variables. For instance, length of imprisonment, which can be used as a target variable related to legal judgment, can be discretized into two intervals, $[0,$ 1 year $]$ and $[$ 1 year, $\infty]$, in order to apply classification algorithms. Since trial results, such as length of imprisonment, are sometimes specified by using intervals in law articles, classification is frequently adopted in research on data analytics-based Legal Tech.

Table 12. Tasks and algorithms of previous research papers.

Task		# of Research Papers	The Most Frequently Used Algorithm
Supervised Learning	Classification	43	ANN&DL
	Regression	4	ANN&DL
Unsupervised Learning	Clustering	10	K-means
	Association	1	Apriori
Other		13	Heuristic

The most widely used algorithm for supervised learning tasks is ANN&DL. An ANN is a network of artificial neurons (nodes) and connections between them, used to generate output values for given input values. The nodes within an ANN form two or more layers. The input layer contains input nodes that indicate input values, while the output layer consists of output nodes that produce output values. Moreover, the layers between the input layer and output layer are called hidden layers. Typically, an ANN with many hidden layers is complex and time-consuming to train, although the hidden layers can contribute to obtaining output values appropriate for the given input values [35,103]. However, modern computer hardware and efficient activation functions allow for utilizing a number of hidden layers for a wide range of practical purposes. An ANN with many hidden layers is called a deep neural network (DNN), and DL is a set of algorithms that use a DNN [104]. Table 12 shows that ANN&DL techniques are also widely used in the legal domain.

The most frequently used clustering and association algorithms are *k*-means and Apriori, respectively. *K*-means is a well-known clustering algorithm that is used to find

centroid-based and non-overlapping k clusters from a given data set. The number of clusters, k, should be prespecified by the analyzer, and a single cluster should contain records similar to each other [26,29,30]. In the legal domain, a clustering algorithm is often used to find clusters of similar documents. Association analysis is rarely applied in the legal domain. The Apriori algorithm is a traditional algorithm used for association analysis that is designed to extract useful, interesting association rules from a given transaction data set [32]. An association rule indicates cause-and-effect relationships or correlations between items within a given transaction data set, and a single association rule is useful if and only if its support and confidence measures simultaneously satisfy minimum threshold values [29]. Typically, association analysis is used to identify a set of items frequently found together in identical transactions. In the legal domain, Liu et al. [71] applied the Apriori algorithm to analyze citation relationships between statutes.

4.3. Target Variable

As discussed in the previous section, supervised learning techniques that consider a target variable is widely used in research on data analytics-based Legal Tech. In a training set for supervised learning, both predictors and target variable values should be known, and supervised learning algorithms are generally designed to build models that reflect the relationship between predictors and the target variable. If a model is obtained, it is used to estimate the value of the target variable for a new data object, where only the predictors are known. The target variables of the research papers listed in Table 9 are summarized in Table 13. Note that a single research paper can consider two or more types of target variables.

Table 13. Type of target variables.

Type	# of Research Papers
Trial result	25
Document type	7
Element type (semantics of element)	13
Law Article	6
Others	2

In Table 13, the most frequently considered target variable type is the trial result, which includes the length of imprisonment, amount of penalty, guiltiness of defendant, and validity of the patent, etc. [44,51,65,73,79]. This type of target variable is very popular in research on data analytics-based Legal Tech since it is the primary output of the most important legal service, i.e., legal procedures.

Element type as a target variable is used to specify the type of entity identified in legal documents. For instance, some noun phrases can indicate information, such as the name of the person, location, and time, in legal documents [21,45,64]. Ji et al. [22] used ANN&DL techniques to classify the type of paragraphs in court record data. Examples of application areas of classification models for element type target variables include annotation of legal documents and transformation of legal documents into structured documents.

Document type is the third most popular target variable type. Typically, legal experts have to examine a large volume of legal documents in order to provide legal services. Sometimes, a legal expert or a department of an organization will specialize in specific types of documents. Similarly, different types of legal documents are often processed in different ways. Thus, the research papers that focused on document type as the target variable aimed to improve the efficiency of legal service procedures by classifying the involved legal documents into appropriate groups. Examples of document type target variables are case type, complaint type, accusation type, and topic type, etc. [57,62,72,77,78].

Law article as a target variable denotes law articles associated with a specific case. In other words, models for this type of target variable are able to find law articles relevant for a given case conveniently [71]. Moreover, the information about law articles determined by data analytics techniques can be used to estimate trial-result-type target variables [68,86].

5. Conclusions and Further Remarks

5.1. Key Findings

This paper provides a systematic survey of research on Legal Tech, which provides innovative legal services in the Industry 4.0 era. The key findings of our survey are as follows: Firstly, the number of published research papers on data analytics-based Legal Tech dramatically increased in recent years, especially in Asia. Previous surveys suggested that the Legal Tech industry is rapidly growing across the world [7,16]. In contrast, this paper reveals that a specific sub-area of Legal Tech can be a popular research topic in specific regions.

Secondly, many of the associated research papers applied supervised learning techniques to legal judgment data. The most popular type of such application is trial result prediction, where the trial result, such as length of imprisonment and amount of penalty, is used as a target variable. Legal Tech applications for trial result prediction can help human judges to make legal decisions more efficiently. Furthermore, clients can take advantage of such applications in order to assess their chances of winning a case. Another interesting application of supervised learning techniques to legal judgment data is law article prediction, which aims to identify law articles relevant to a case. Law article prediction can also be used as a preprocessing procedure for trial result prediction in that law articles related to a case typically provide valuable information that affects the trial result. In this manner, data analytics enable the provision of legal services more efficiently and accurately [14].

Thirdly, the structured document is the most popular data structure for research on data analytics-based Legal Tech. It is difficult to process and analyze raw documents in text format. Thus, many researchers extract useful features from legal documents and utilize them to convert the documents into tabular form. Such features have to be identified and collected by applying nontrivial techniques such as text mining and NLP.

Fourthly, ANN&DL is the most frequently used algorithm type in research on data analytics-based Legal Tech. ANN&DL can be applied to both classification and regression analyses. Additionally, unsupervised learning tasks can be dealt with by using ANN&DL techniques. In other words, ANN&DL techniques have wide applicability. Moreover, ANN&DL techniques often produce good performances even if the training set is very complex. In this context, ANN&DL is a promising approach for developing data analytics-based Legal Tech applications.

Fifthly, data analytics-based Legal Tech is a multidisciplinary research topic related to computer science, social science, engineering, and mathematics, etc. Especially, Legal Tech emerged as a promising application domain for quantitative methods and algorithms, which were popular research topics for researchers from computer science and engineering.

5.2. Future Research Topics and Challenges

The authors conclude this paper by listing several future research topics and challenges for data analytics-based Legal Tech. The first topic is legal issues related to the applications of data analytics in the legal industry. For instance, misjudgments of data analytics-based Legal Tech applications can cause significant loss of stakeholders and spark arguments over who is responsible. Furthermore, the decision-making model in data analytics-based Legal Tech applications can contain some biases related to discrimination. Thus, reliability, fairness, and scope of application will be relevant research topics for data analytics-based Legal Tech.

The second topic is business and service models based on data analytics-based Legal Tech. Modern Legal Tech enables innovative business and service models in the legal industry [7]. Data analytics-based Legal Tech can also contribute to providing new legal

services. For instance, trial result prediction can be used to evaluate the difficulty of a lawsuit, which might affect the pricing of related legal services. The intelligent pricing policy can enable new business models in the legal industry. Acceptance intention of such business and service models is also an important future research topic.

Thirdly, the language of legal documents should be carefully considered in research on data analytics-based Legal Tech. Previous research papers generally consider only a single language. In particular, the structure of the specific language can affect the performance of the preprocessing procedure. An application successfully applied to one language can fail to process legal documents written in another language. Thus, diversity of language is an important challenge for data analytics-based Legal Tech.

Fourthly, the application of unsupervised learning techniques will be increased in the future. Thus far, these techniques are less popular than supervised learning techniques in research on data analytics-based Legal Tech. However, they are successfully applied in a wide range of application areas, including e-commerce, customer relationship management (CRM), marketing, manufacturing, education, and healthcare, etc. Similarly, the legal industry can be another promising application area for unsupervised learning and other techniques.

In this paper, previous studies on Legal Tech are reviewed from the perspective of data analytics, one of the most important Industry 4.0 technologies. The authors believe that this paper provides meaningful insights into the concepts, approaches, and research topics of data analytics-based Legal Tech.

Author Contributions: Conceptualization, J.-W.K.; validation, J.-W.K.; investigation, S.-H.P., D.-G.L. and J.-S.P.; data curation, S.-H.P., D.-G.L. and J.-S.P.; writing—original draft preparation, S.-H.P.; writing—review and editing, J.-W.K.; supervision, J.-W.K.; project administration, J.-W.K.; funding acquisition, J.-W.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea, NRF-2019S1A5C2A03080978. The APC was funded by NRF-2019S1A5C2A03080978.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5C2A03080978).

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

References

1. Koh, L.; Orzes, G.; Jia, F.J. The Fourth industrial revolution (industry 4.0): Technologies disruption on operations and supply chain management. *Int. J. Oper. Prod. Manag.* **2019**, *39*, 817–828. [CrossRef]
2. Piccarozzi, M.; Aquilani, B.; Gatti, C. Industry 4.0 in Management Studies: A Systematic Literature Review. *Sustainability* **2018**, *10*, 3821. [CrossRef]
3. Chong, H.R.; Bae, K.H.; Lee, M.K.; Kwon, H.M.; Hong, S.H. Quality strategy for building a smart factory in the fourth industrial revolution. *J. Korean Soc. Qual. Manag.* **2020**, *48*, 87–105.
4. Richert, A.; Shehadeh, M.; Plumanns, L.; Groß, K.; Schuster, K.; Jeschke, S. Educating engineers for industry 4.0: Virtual worlds and human-robot-teams: Empirical studies towards a new educational age. In Proceedings of the 2016 IEEE Global Engineering Education Conference (EDUCON), Abu Dhabi, United Arab Emirates, 10–13 April 2016; pp. 142–149.
5. Ajayi, O.O.; Bagula, A.B.; Ma, K. Fourth industrial revolution for development: The relevance of cloud federation in healthcare support. *IEEE Access* **2019**, *7*, 185322–185337. [CrossRef]
6. Kerikmäe, T.; Hoffmann, T.; Chochia, A. Legal technology for law firms: Determining roadmaps for innovation. *Croat. Int. Relat. Rev.* **2018**, *24*, 91–112. [CrossRef]
7. Hongdao, Q.; Bibi, S.; Khan, A.; Ardito, L.; Khaskheli, M.B. Legal Technologies in Action: The Future of the Legal Market in Light of Disruptive Innovations. *Sustainability* **2019**, *11*, 1015. [CrossRef]
8. Ebrahim, T.Y. Automation & predictive analytics in patent prosecution: USPTO implications & policy. *Ga. St. UL Rev.* **2018**, *35*, 1185.

9. Praduroux, S.; de Paiva, V.; di Caro, L. Legal tech start-ups: State of the art and trends. In Proceedings of the Workshop on 'Mining and Reasoning with Legal Texts' Collocated at the 29th International Conference on Legal Knowledge and Information Systems, Nice, France, 14 December 2016.
10. Wang, D. Building Value in a World of Technological Change: Data Analytics and Industry 4.0. *IEEE Eng. Manag. Rev.* **2018**, *46*, 32–33. [CrossRef]
11. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
12. Kayembe, C.; Nel, D. Challenges and opportunities for education in the fourth industrial revolution. *Afr. J. Pub. Affairs* **2019**, *11*, 79–94.
13. Moses, L.B.; Chan, J. Using big data for legal and law enforcement decisions: Testing the new tools. *UNSW Law J.* **2014**, *37*, 643–678.
14. Chen, D.L. Judicial analytics and the great transformation of American Law. *Artif. Intell. Law* **2019**, *27*, 15–42. [CrossRef]
15. Janoski-Haehlen, E. Robots, blockchain, ESI, oh my!: Why law schools are (or should be) teaching legal technology. *Legal Ref. Serv. Q.* **2019**, *38*, 77–101. [CrossRef]
16. Salmerón-Manzano, E. Legaltech and Lawtech: Global Perspectives, Challenges, and Opportunities. *Laws* **2021**, *10*, 24. [CrossRef]
17. Chalkidis, I.; Kampas, D. Deep learning in law: Early adaptation and legal word embeddings trained on large corpora. *Artif. Intell. Law* **2019**, *27*, 171–198. [CrossRef]
18. Wang, R. Legal technology in contemporary USA and China. *Comput. Law Secur. Rev.* **2020**, *39*, 105459. [CrossRef]
19. Sung, H.C. Can Online Courts Promote Access to Justice? A Case Study of the Internet Courts in China. *Comput. Law Secur. Rev.* **2020**, *39*, 105461. [CrossRef]
20. Guo, M. Internet court's challenges and future in China. *Comput. Law Secur. Rev.* **2021**, *40*, 105522. [CrossRef]
21. Ji, D.; Gao, J.; Fei, H.; Teng, C.; Ren, Y. A deep neural network model for speakers coreference resolution in legal texts. *Inf. Process. Manag.* **2020**, *57*, 102365. [CrossRef]
22. Ji, D.; Tao, P.; Fei, H.; Ren, Y. An End-to-end joint model for evidence information extraction from court record document. *Inf. Process. Manag.* **2020**, *57*, 102305. [CrossRef]
23. Borgatti, S.P.; Mehra, A.; Brass, D.J.; Labianca, G. Network Analysis in the Social Sciences. *Science* **2009**, *323*, 892–895. [CrossRef]
24. Kugler, L. AI judges and juries. *Commun. ACM* **2018**, *61*, 19–21. [CrossRef]
25. Edwards, M.; Rashid, A.; Rayson, P. A Systematic Survey of Online Data Mining Technology Intended for Law Enforcement. *ACM Comput. Surv.* **2015**, *48*, 1–54. [CrossRef]
26. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: Burlington, MA, USA, 2005.
27. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3323–3331.
28. Kesavaraj, G.; Sukumaran, S. A Study on classification techniques in data mining. In Proceedings of the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, India, 4–6 July 2013; pp. 1–7.
29. Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Addison-Wesley: Boston, MA, USA, 2005.
30. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*; Morgan Kaufmann: Burlington, VT, USA, 2011.
31. Ratner, B. *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*; CRC Press: Boca Raton, FL, USA, 2017.
32. Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, 25–28 May 1993; pp. 207–216.
33. Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A.I. Fast discovery of association rules. *Lect. Notes Artif. Int.* **1996**, *12*, 307–328.
34. Pei, J.; Han, J. Constrained frequent pattern mining: A pattern-growth view. *SIGKDD Explor.* **2002**, *4*, 31–39. [CrossRef]
35. Nilsson, N.J. *Principles of Artificial Intelligence*; Morgan Kaufmann: Burlington, VT, USA, 2014.
36. Piatetsky-Shapiro, G.; Fayyad, U.; Smith, P. From data mining to knowledge discovery: An overview. *Lect. Notes Artif. Int.* **1996**, *1*, 35.
37. Simoudis, E. Reality check for data mining. *IEEE Ann. Hist. Comput.* **1996**, *11*, 26–33. [CrossRef]
38. Tan, A.H. Text mining: The state of the art and the challenges. In Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Beijing, China, 26–28 April 1999; Volume 8, pp. 65–70.
39. Ramos, J. Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First Instructional Conference on Machine Learning; 2003; Volume 242, pp. 29–48. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf (accessed on 19 July 2021).
40. Wallach, H.M. Topic modeling: Beyond bag-of-words. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 977–984.
41. Levy, O.; Goldberg, Y. Dependency-based word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA, 22–27 June 2014; pp. 302–308.
42. Lettieri, N.; Altamura, A.; Malandrino, D. The legal macroscope: Experimenting with visual legal analytics. *Inf. Vis.* **2017**, *16*, 332–345. [CrossRef]

43. Petrović, Đ.; Stanković, M. Use of linguistic forms mining in the link analysis of legal documents. *Comput. Sci. Inf. Syst.* **2018**, *15*, 369–392. [CrossRef]
44. 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]
45. Sharafat, S.; Nasar, Z.; Jaffry, S.W. Data mining for smart legal systems. *Comput. Electr. Eng.* **2019**, *78*, 328–342. [CrossRef]
46. Thammaboosadee, S.; Watanapa, B. Identification of criminal case diagnostic issues: A modular ANN approach. *Int. J. Inf. Tech. Decis.* **2013**, *12*, 523–546. [CrossRef]
47. 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**, *97*, 864–875. [CrossRef]
48. De Luise, M.D.L.; Pascal, A.; Saad, B.; Álvarez, C.; Pescio, P.; Carrilero, P.; Díaz, J. Intelligent Chatter Bot for Regulation Search. *Open Phys.* **2016**, *14*, 473–477. [CrossRef]
49. Katz, D.M.; Bommarito, M.J.; Blackman, J. A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE* **2017**, *12*, e0174698. [CrossRef]
50. Li, S.; Zhang, H.; Ye, L.; Su, S.; Guo, X.; Yu, H.; Fang, B. Prison Term Prediction on Criminal Case Description with Deep Learning. *Comput. Mater. Contin.* **2020**, *62*, 1217–1231. [CrossRef]
51. Mitchell, J.; Mitchell, S.; Mitchell, C. Machine learning for determining accurate outcomes in criminal trials. *Law Probab. Risk* **2020**, *19*, 43–65. [CrossRef]
52. Ashley, K.D.; Brüninghaus, S. Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law* **2009**, *17*, 125–165. [CrossRef]
53. Boella, G.; Di Caro, L.; Humphreys, L.; Robaldo, L.; Rossi, P.; van der Torre, L. Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law. *Artif. Intell. Law* **2016**, *24*, 245–283. [CrossRef]
54. Boella, G.; Di Caro, L.; Leone, V. Semi-automatic knowledge population in a legal document management system. *Artif. Intell. Law* **2019**, *27*, 227–251. [CrossRef]
55. Cheng, Y.; Agrawal, A.; Liu, H.; Choudhary, A. Legislative prediction with dual uncertainty minimization from heterogeneous information. *Stat. Anal. Data Min.* **2017**, *10*, 107–120. [CrossRef]
56. El Jelali, S.; Fersini, E.; Messina, E. Legal retrieval as support to eMediation: Matching disputant's case and court decisions. *Artif. Intell. Law* **2015**, *23*, 1–22. [CrossRef]
57. 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]
58. Fernandes, W.P.D.; Silva, L.J.S.; Frajhof, I.Z.; Konder, C.N.; Nasser, R.B.; de Carvalho, G.R.; Almeida, G.F.C.F.; Barbosa, S.D.J.; Lopes, H.C.V. Appellate Court Modifications Extraction for Portuguese. *Artif. Intell. Law* **2019**, *28*, 1–34. [CrossRef]
59. Fornaciari, T.; Poesio, M. Automatic deception detection in Italian court cases. *Artif. Intell. Law* **2013**, *21*, 303–340. [CrossRef]
60. Francesconi, E.; Passerini, A. Automatic Classification of Provisions in Legislative Texts. *Artif. Intell. Law* **2007**, *15*, 1–17. [CrossRef]
61. Francesconi, E.; Peruginelli, G. Integrated access to legal literature through automated semantic classification. *Artif. Intell. Law* **2009**, *17*, 31–49. [CrossRef]
62. Guo, X.; Zhang, H.; Ye, L.; Li, S. RnRTD: Intelligent approach based on the relationship-driven neural network and restricted tensor decomposition for multiple accusation judgment in legal cases. *Comput. Intel. Neurosc.* **2019**, *2019*, e6705405. [CrossRef]
63. Hachey, B.; Grover, C. Extractive summarisation of legal texts. *Artif. Intell. Law* **2006**, *14*, 305–345. [CrossRef]
64. Iftikhar, A.; Jaffry, S.W.U.Q.; Malik, M.K. Information Mining From Criminal Judgments of Lahore High Court. *IEEE Access* **2019**, *7*, 59539–59547. [CrossRef]
65. Jin, C.; Zhang, G.; Wu, M.; Zhou, S.; Fu, T. Textual content prediction via fuzzy attention neural network model without predefined knowledge. *China Commun.* **2020**, *17*, 211–222. [CrossRef]
66. Lesmo, L.; Mazzei, A.; Palmirani, M.; Radicioni, D.P. TULSI: An NLP system for extracting legal modificatory provisions. *Artif. Intell. Law* **2013**, *21*, 139–172. [CrossRef]
67. Li, J.; Zhang, G.; Yu, L.; Meng, T. Research and design on cognitive computing framework for predicting judicial decisions. *J. Signal Process. Sys.* **2019**, *91*, 1159–1167. [CrossRef]
68. Li, S.; Zhang, H.; Ye, L.; Guo, X.; Fang, B. MANN: A Multichannel attentive neural network for legal judgment prediction. *IEEE Access* **2019**, *7*, 151144–151155. [CrossRef]
69. Li, X.; Kang, X.; Wang, C.; Dong, L.; Yao, H.; Li, S. A Neural-Network-Based Model of Charge Prediction via the Judicial Interpretation of Crimes. *IEEE Access* **2020**, *8*, 101569–101579. [CrossRef]
70. Liu, Y.H.; Chen, Y.L. A two-phase sentiment analysis approach for judgement prediction. *J. Inf. Sci.* **2018**, *44*, 594–607. [CrossRef]
71. Liu, Y.H.; Chen, Y.L.; Ho, W.L. Predicting associated statutes for legal problems. *Inf. Process. Manag.* **2015**, *51*, 194–211. [CrossRef]
72. 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]
73. Mahfouz, T.; Kandil, A. Litigation Outcome Prediction of Differing Site Condition Disputes through Machine Learning Models. *J. Comput. Civ. Eng.* **2012**, *26*, 298–308. [CrossRef]
74. Medvedeva, M.; Vols, M.; Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. *Artif. Intell. Law* **2020**, *28*, 237–266. [CrossRef]

75. Nanda, R.; Siragusa, G.; Di Caro, L.; Boella, G.; Grossio, L.; Gerbaudo, M.; Costamagna, F. Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives. *Artif. Intell. Law* **2019**, *27*, 199–225. [CrossRef]
76. Nguyen, T.-S.; Nguyen, L.-M.; Tojo, S.; Satoh, K.; Shimazu, A. Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artif. Intell. Law* **2018**, *26*, 169–199. [CrossRef]
77. Pudaruth, S.; Soyjaudah, K.M.S.; Gunputh, R.P. An innovative multi-segment strategy for the classification of legal judgments using the k-nearest neighbour classifier. *Complex Intell. Syst.* **2018**, *4*, 1–10. [CrossRef]
78. Qiu, M.; Zhang, Y.; Ma, T.; Wu, Q.; Jin, F. Convolutional-neural-network-based Multilabel Text Classification for Automatic Discrimination of Legal Documents. *Sens. Mater.* **2020**, *32*, 2659–2672. [CrossRef]
79. Raghupathi, V.; Zhou, Y.; Raghupathi, W. Legal Decision Support: Exploring Big Data Analytics Approach to Modeling Pharma Patent Validity Cases. *IEEE Access* **2018**, *6*, 41518–41528. [CrossRef]
80. Saravanan, M.; Ravindran, B. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artif. Intell. Law* **2010**, *18*, 45–76. [CrossRef]
81. Shulayeva, O.; Siddharthan, A.; Wyner, A. Recognizing cited facts and principles in legal judgements. *Artif. Intell. Law* **2017**, *25*, 107–126. [CrossRef]
82. Tran, O.T.; Ngo, B.X.; Le Nguyen, M.; Shimazu, A. Automated reference resolution in legal texts. *Artif. Intell. Law* **2014**, *22*, 29–60. [CrossRef]
83. Waltl, B.; Bonczek, G.; Scepankova, E.; Matthes, F. Semantic types of legal norms in German laws: Classification and analysis using local linear explanations. *Artif. Intell. Law* **2019**, *27*, 43–71. [CrossRef]
84. Yamada, H.; Teufel, S.; Tokunaga, T. Building a corpus of legal argumentation in Japanese judgement documents: Towards structure-based summarisation. *Artif. Intell. Law* **2019**, *27*, 141–170. [CrossRef]
85. Yang, C.C.; Luk, J. Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 671–682. [CrossRef]
86. Yao, F.; Sun, X.; Yu, H.; Yang, Y.; Zhang, W.; Fu, K. Gated hierarchical multi-task learning network for judicial decision prediction. *Neurocomputing* **2020**, *411*, 313–326. [CrossRef]
87. Guo, X.; Zhang, H.; Ye, L.; Li, S. TenLa: An approach based on controllable tensor decomposition and optimized lasso regression for judgement prediction of legal cases. *Appl. Intell.* **2020**, *51*, 2233–2252. [CrossRef]
88. Guo, X.; Zhang, H.; Ye, L.; Li, S.; Zhang, G. TenRR: An Approach Based on Innovative Tensor Decomposition and Optimized Ridge Regression for Judgment Prediction of Legal Cases. *IEEE Access* **2020**, *8*, 167914–167929. [CrossRef]
89. Tran, V.; Le Nguyen, M.; Tojo, S.; Satoh, K. Encoded summarization: Summarizing documents into continuous vector space for legal case retrieval. *Artif. Intell. Law* **2020**, *28*, 441–467. [CrossRef]
90. Moens, M.-F.; Uyttendaele, C.; Dumortier, J. Abstracting of legal cases: The potential of clustering based on the selection of representative objects. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 151–161. [CrossRef]
91. Acharya, H.R.; Bhat, A.D.; Avinash, K.; Srinath, R. LegoNet-classification and extractive summarization of Indian legal judgments with capsule networks and sentence embeddings. *J. Intell. Fuzzy Syst.* **2020**, *39*, 2037–2046. [CrossRef]
92. Sadeghian, A.; Sundaram, L.; Wang, D.Z.; Hamilton, W.F.; Branting, K.; Pfeifer, C. Automatic semantic edge labeling over legal citation graphs. *Artif. Intell. Law* **2018**, *26*, 127–144. [CrossRef]
93. Bartolini, R.; Lenci, A.; Montemagni, S.; Pirrelli, V.; Soria, C. Automatic classification and analysis of provisions in Italian legal texts: A case study. In Proceedings of the OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Agia Napa, Cyprus, 25–29 October 2004; pp. 593–604.
94. Boulet, R.; Mazzega, P.; Bourcier, D. A network approach to the French system of legal codes—Part I: Analysis of a dense network. *Artif. Intell. Law* **2011**, *19*, 333–355. [CrossRef]
95. Boulet, R.; Mazzega, P.; Bourcier, D. Network approach to the French system of legal codes part II: The role of the weights in a network. *Artif. Intell. Law* **2018**, *26*, 23–47. [CrossRef]
96. Chen, Y.L.; Liu, Y.H.; Ho, W.L. A text mining approach to assist the general public in the retrieval of legal documents. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 280–290. [CrossRef]
97. De Araujo, D.A.; Rigo, S.J.; Barbosa, J.L.V. Ontology-based information extraction for juridical events with case studies in Brazilian legal realm. *Artif. Intell. Law* **2017**, *25*, 379–396. [CrossRef]
98. Fan, H.; Li, H. Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. *Autom. Constr.* **2013**, *34*, 85–91. [CrossRef]
99. Hasan, I.; Parapar, J.; Barreiro, A. Improving the extraction of text in pdfs by simulating the human reading order. *J. Univers. Comput. Sci.* **2012**, *18*, 623–649.
100. Herrera, M.; Candia, C.; Rivera, D.; Aitken, D.; Brieba, D.; Boettiger, C.; Donoso, G.; Godoy-Faúndez, A. Understanding water disputes in Chile with text and data mining tools. *Water Int.* **2019**, *44*, 302–320. [CrossRef]
101. Le, T.T.N.; Shirai, K.; Le Nguyen, M.; Shimazu, A. Extracting indices from Japanese legal documents. *Artif. Intell. Law* **2015**, *23*, 315–344. [CrossRef]
102. Saravanan, M.; Ravindran, B.; Raman, S. Improving legal information retrieval using an ontological framework. *Artif. Intell. Law* **2009**, *17*, 101–124. [CrossRef]

103. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, e00938. [CrossRef]
104. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Rayes, M.P.; Shyu, M.L.; Chen, S.C.; Lyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]