IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 August 2023, accepted 15 September 2023, date of publication 19 September 2023,
date of current version 22 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317083

SURVEY

A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges

JUNYUN CUI[1], XIAOYU SHEN[2], AND SHAOCHUN WEN[3]
[1]School of Information, Xián University of Finance and Economics, Xi'an, Shaanxi 710100, China
[2]Amazon Alexa AI, 47495 Berlin, Germany
[3]ASEAN International School, Guangxi Transport Vocational and Technical College, Nanning, Guangxi 530023, China

Corresponding author: Junyun Cui (junyun_cui@xaufe.edu.cn)

This work was supported in part by the Science Foundation of the China (Xi'an) Institute for Silk Road Research under Grant 2019YB05 and Grant 2019YA07, in part by the Innovation Training Project of Shaanxi College Students under Grant S202011560030, in part by the Research Foundation of the Xi'an University of Finance and Economics under Grant 18FCJH02, in part by the Education Department of Guangxi Zhuang Autonomous Region, and in part by the Guangxi Transport Vocational and Technical College.

ABSTRACT Legal judgment prediction (LJP) applies Natural Language Processing (NLP) techniques to predict judgment results based on fact descriptions automatically. The present work addresses the growing interest in the application of NLP techniques to the task of LJP. Despite the current performance gap between machines and humans, promising results have been achieved in a variety of benchmark datasets, owing to recent advances in NLP research and the availability of large-scale public datasets. To provide a comprehensive survey of existing LJP tasks, datasets, models, and evaluations, this study presents the following contributions: 1) an analysis of 43 LJP datasets constructed in 9 different languages, together with a classification method of LJP based on three different attributes; 2) a summary of 16 evaluation metrics categorized into 4 different types to evaluate the performance of LJP models for different outputs; 3) a review of 8 legal-domain pretrained models in 4 languages, highlighting four major research directions for LJP; 4) state-of-the-art results for 11 representative datasets from different court cases and an in-depth discussion of the open challenges in this area. This study aims to provide a comprehensive review for NLP researchers and legal professionals to understand the advances in LJP over the past years, and to facilitate further joint efforts towards improving the performance of LJP models.

INDEX TERMS Legal judgment prediction, natural language processing, survey, benchmark datasets, neural network.

I. INTRODUCTION

Legal Judgment Prediction (LJP) is a crucial task that aims to predict the outcome of legal cases based on their fact descriptions, providing significant benefits for both legal practitioners and ordinary citizens [1]. Currently, this task is primarily performed by legal experts, who require extensive specialized training to process legal cases, as it involves several time-consuming and domain-specific steps, such as finding relevant law articles, defining the charge range, and deciding the penalty term. In Louisiana, every attorney handles up to 50 cases per day, leaving only 1-5 minutes for case preparation [2]. There are 332 thousand cases in progress in Brazil per day, considering only the financial domain [3]. 44 million pending cases cannot be handled on time until April 2023 [4]. The overwhelming demand for legal assistance and the limited number of legal experts have caused significant social problems, such as inadequate or no legal help for low-income Americans [5]. Therefore, there is an urgent need to develop automatic LJP systems that can enhance the working efficiency of legal experts and provide real-time legal consultation, improving public access to justice.

LJP is a long-standing task, and early approaches were based on rules or statistical methods [1], [6], [7], [8]. For instance, factor and linear regression analyses have been

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

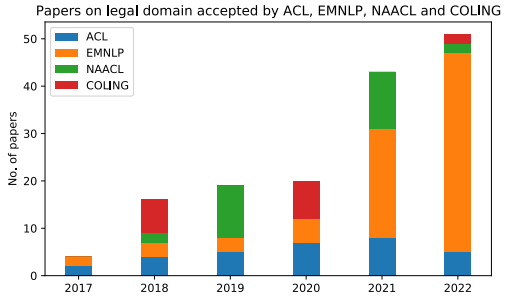

FIGURE 1. The number of legal-domain papers in major NLP conferences.

applied to predict decisions (pro or con) of the US Supreme Court cases depending on the 26 factual elements (patterns) with 14 training and 14 test ''right to counsel'' cases in as early as the 1950s [9], [10]. However, these systems were not robust to noise and could not generalize to other law domains. Later, researchers began to use machine learning techniques trained on a more extensive collection of legal cases [11], [12], [13], [14], [15], [16]. With the recent advancements in neural networks and large-scale pretrained language models based on the Transformer architecture, significant improvements in LJP have been achieved [17], [18], [19], [20], [21], [22]. As illustrated in Fig. 1, LJP has become a hot research topic, with approximately 65% of legal domain papers at major NLP conferences related to LJP. The availability of challenging benchmark datasets plays a crucial role in spurring innovation in LJP [23]. The recent years have witnessed an explosion of public LJP benchmark datasets, including CAIL2018 [24], [25], ECHR-CASES [24], [26], SwissJudgment [27], and JUSTIC [28], [29]. These datasets have inspired numerous LJP models, such as TopJudge [30], MLCP-NLN [31], MPBFN-WCA [32], and LADAN [33]. Although impressive results have been achieved in various benchmark datasets, a clear gap remains between machine and human performance [25]. However, researchers often focus on a few popular LJP datasets, neglecting many other datasets that are less well-known and less studied. In addition, there is a need for systematic categorization/classification of LJP subtasks. While LJP tasks are generally divided into three subtasks (i.e., the decision of applicable law articles, charges, and terms of penalty) [25], [30], [32], [33], this classification method is limited and does not apply to all legal systems and domains [34], [35]. Thus, a comprehensive survey of existing LJP tasks, datasets, evaluation metrics, and models is needed to promote the future development of LJP.

This paper aims to address this gap by providing a comprehensive survey of the LJP task. While a few surveys in the LJP domain have been conducted, they are limited to specific benchmark datasets (such as CAIL2018) or Indian Legal NLP benchmarks [23], [24]. To the best of our knowledge, this survey is the first work that provides a comprehensive survey of the LJP task, introducing 43 publicly LJP datasets in 9 languages and the pros and cons of popular state-of-the-art models.The key contributions of this survey can be summarized as follows:

- **Datasets**. The survey provides a comprehensive analysis of 43 LJP datasets constructed in 9 different languages and 8 pre-training datasets in 4 languages, including their resources, categories, input/output elements, data distribution features, construction methods and statistics.
- **Tasks**. The classification of LJP tasks is proposed including type of tasks, legal systems and law domains, which is based on the differences in the method of action for tasks, the litigation procedure between different legal systems and concepts of different law domains.
- **Metrics**. 16 evaluation metrics are categorized into 4 different types to evaluate the performance of LJP models for different outputs.
- **Models**. The survey overview 4 major research directions (multi-task learning, interpretable learning and few-shot learning) for LJP models, 8 legal-domain pretrained models in 4 languages, and state-of-the-art results for 11 representative datasets from different court cases.
- **Challenges**. This survey highlights the following challenges: (1) Datasets: the need for the monolingual datasets for other 27 official languages and multilingual datasets for all the 36 official languages; (2) Tasks: the current lacunae of LJP tasks that could be filled by more realistic applications; (3) Metrics: the fairness evaluation of LJP results in the future; (4) Models: the need for enhancing interpretability and reasoning capabilities for future LJP models.

The remainder of this survey is structured as follows: Section II compares the litigation procedure differences between common-law and civil-law systems, and gives a taxonomy to classify existing LJP tasks. Section III analyzes the existing LJP datasets. Section IV introduces evaluation metrics for LJP tasks, including their computing methods and categories. In Section V and VI, this survey reviews 4 major research directions for LJP models, 8 legal-domain pre-trained models in 4 languages, and state-of-the-art performance results for 11 representative LJP datasets from different court cases. Section VII discusses possible research directions for LJP studies in the future.

II. TASKS FORMULATIONS

A. BACKGROUND

The legal systems of continental civil-law and common-law are of significant importance in regulating and harmonizing human activity within their respective societies. The former legal system is applicable in France, Germany, Switzerland, Belgium, and the Netherlands in European continental, China, Thailand, and Vietnam in Asiatic countries, and Scandinavian countries and Soviet countries, while the latter one is adopted in the United Kingdom, the United States, Canada, Australia, New Zealand, and India. Additionally, the two legal systems have been functioning together in

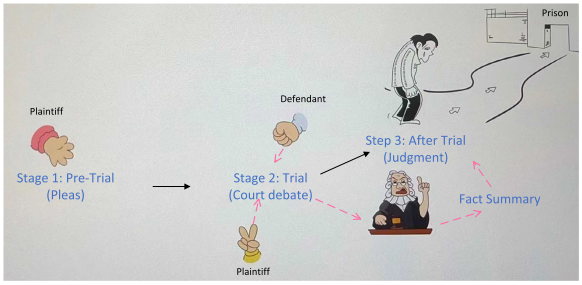

FIGURE 2. The procedure of court judgment. Firstly, the plaintiff submits his pleas. Secondly, fact verdict is made based on the court debate held between the plaintiff and defendant. Finally, the judge make decisions based on the fact verdict.

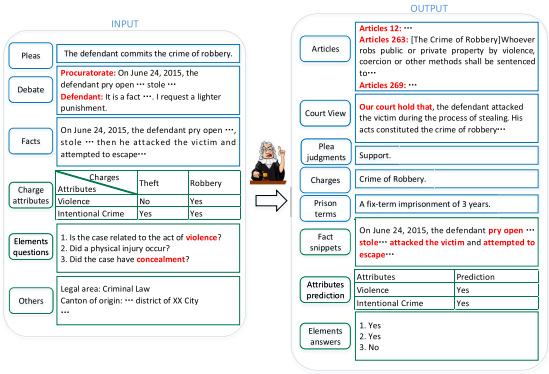

FIGURE 3. An outcomes-based judicial framework example interpretation of notations introduced by a law case life-cycle in a real court setting. Within this framework, the LJP tasks are divided into two categories: main LJP tasks (blue) and auxiliary LJP tasks (green).

countries and regions like Japan, Italy, Louisiana, Quebec, Scotland, and South Africa. The two systems are influenced by their culture, epistemology, civilization, and history. It is acknowledged that there exist notable differences between these legal systems. Roman-law writers have noted that Roman law from the classical period is more similar to the common-law than modern civil-law systems, which are derived from Roman law [36].

In a real court setting, "no claim, no trial" is an important principle in the judicial procedure, which means that "the court will not entertain matters that have not been prosecuted." The claimant's pleadings and court's decisions are crucial components that protect the legitimate interests of the parties involved. A typical litigation procedure comprises three stages. The first one is the pre-trial claim collection stage, where the plaintiff or petitioner submits written materials appealing their case to the court. The trial court debate stage follows, during which the parties, including the plaintiff, defendant, witness, and lawyer, debate before the court, focusing on the factual details of the case. Finally, the after-trial judge sentence stage occurs, where the judge generates the verdict, including judgment. Figure 2 illustrates these stages.

The differences between the civil-law and common-law systems in the litigation procedure can be further examined.

- The first difference lies in the pre-trial claim collection stage. The civil-law system operates on the principle of legality, whereby if sufficient evidence is obtained, the prosecutor is duty-bound to press charges and cannot dismiss a case. In contrast, the common-law system operates on the opportunity principle, which grants the prosecutor discretionary power to decide whether or not to prosecute a case, even with sufficient evidence [37]. Based on the opportunity principle, the portion of federal civil cases resolved by trial fell from 11.5 percent in 1962 to 1.8 percent in 2002, and a similar decline in both the percentage and the absolute number of trials in federal criminal cases [38]. Additionally, in the common-law system, the accused's fate is determined by the trial court, whereas in the civil-law system, it is based on the conclusions drawn from the investigation [39].
- The second difference occurs in the trial court debate stage. Firstly, the common-law system prefers public oral evidence, while the civil-law system favors private written proof before a judge who questions the witness from a neutral standpoint. Secondly, the common-law system employs a voting jury trial, which does not require the jury to justify their fact verdict with reasons, whereas the civil-law system uses a collegial reasoning trial, whereby the judge must give the reasons for the verdict based on established statutory law articles and common sense [36], [40].
- The third difference presents in the after-trial judge sentence stage (precedent in common-law vs. legislation in civil-law) [40]. The civil-law judges search the legislation for the controlling principle and rules governing the subject, which are then applied or interpreted according to the particular facts of the case. Conversely, in the common-law system, judges abide by the provisions of a statute if the text is clear. However, if there is doubt or ambiguity, common-law judges will search for a similar precedent in previous decisions and are guided accordingly. Subsequently, this precedent is applied or interpreted based on the determination of facts by the jury, the summarized evidence and the relevant rules of law.

Given the background of LJP tasks, provided in Table 1 are some notations of LJP tasks for a court case instance, as shown in Fig. 3.

B. TAXONOMY

In this section, we provide an analysis of existing LJP tasks, aiming to offer a better understanding of LJP tasks. We analyzed 43 LJP datasets and identified key dimensions, as shown in Table 2, based on notations and terminologies in Section II-A. Previous researches have mainly divided LJP tasks into three subtasks: the decision of applicable law articles, charges, and penalty terms [25], [30], [32], [33]. However, as shown in Table 2, this classification method has limited coverage of LJP tasks and may overlook important tasks, such as pro or con decisions. Therefore, we propose a

TABLE 1. Notations definition for a court case.

Notation	Description
Pleas	The sentence narratives from the plaintiff for the target dispute.
Plea Judgments	The response sentence from the judge on the pleas with labels like reject, partially support and support.
Debate	The diachronic statement of fact detail from the plaintiff, defendant, witness, lawyer and judge in the real court on the target dispute.
Facts	The admissible facts summarized by the judge by employing evidence
Articles	The applied law articles of a Law case.
Precedents	Prior relevant cases of a Law case.
Arguments	Statements written by a judge or an attorney providing the justification and legal reasoning for a court ruling.
Charges	The specific statement of what crime the party is accused contained in the criminal plea.
Prison terms	The penalty terms for corresponding charges in a criminal case.
Court View	The explanation written by judges to interpret the judgment decision for certain law case.
Facts snippets	The extracted decisive phrases or sentences from facts.
Charge attributes	The predefined attribute knowledge for discriminating confusing charges.
Attributes prediction	A binary classification task used to predict the charge with the attribute or not according to the input facts.
Elements questions	The predefined questions for extracting judgment elements.
Elements answers	The answers for element questions.
Others	The additional metadata that judges can obtain from a case except the items described above, such as the date, the legal area, the canton of origin per case.

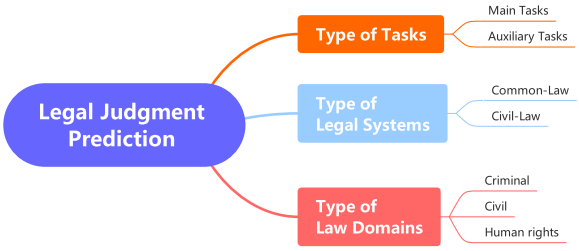

FIGURE 4. Proposed classification method of legal judgment prediction tasks.

new classification method for LJP tasks, as shown in Fig. 4, which includes three attributes: the type of tasks, the type of legal systems, and the type of law domains. Each attribute can be divided into several categories as followed.

- The type of tasks can be classified into two cases: main tasks and auxiliary tasks. Main tasks determine the judgment results, such as article recommendation, precedent prediction, charge prediction, prison term prediction, court view generation, and plea judgment task. Auxiliary tasks improve the judgment results, such as fact snippets extraction, attribute prediction of the confusing charges, and elements answers.
- The type of legal systems can be classified into two main legal systems: common-law and civil-law systems. For the common-law system, fact verdicts provide the voting values for the court's charges, while the trial judge provides the court's results, including law article recommendation, precedent, prison terms, and plea judgments. In contrast, in the civil-law system, the judge justifies the judgments based on given facts, established statutory law articles, and common sense. The trial judge then provides the court's results, including charges, prison terms, and plea judgments.
- Finally, the type of law domains is classified into three prominent cases: criminal, civil, and administrative. Criminal cases include articles, charges, prison terms, and plea judgments. Civil cases include articles, obligation, penalty terms, and plea judgments, while administrative cases include articles, penalty terms, and plea judgments.

1) TYPES OF LJP TASKS

In this section, we provide a taxonomy of LJP tasks based on their outcomes, which enables a comprehensive analysis of the existing LJP tasks from various perspectives.

Article Recommendation Task. It is a significant focus within LJP research, as reflected in the findings presented in Table 2 and Fig. 5. These tasks are divided into three primary categories based on the relationship between case facts and relevant articles, including many-vs-one, many-vs-many, and one-vs-many.

- Many-vs-one. It involves determining whether a particular law article has been violated for given case facts, and are further divided into binary violation and multi-label violation. Binary violation tasks are addressed in works such as echr [13], ECHR-CASES [26], and ECtHR [48], while multi-label violation tasks are tackled in ECHR-CASES [26] and CAIL2018 [25].
- Many-vs-many. Tasks are included such as DPAM [54], FLA [17], and CAIL-Long [55], which aim to find the optimal set of legal articles for each case.
- One-vs-many. MLMN [21] is an example of a task that extracts all the relevant law articles for each fact in a legal case.

Precedent Prediction Task. It refers to the task that automatically searches and recommends relevant precedent cases for supporting the decision of an unseen case description.

TABLE 2. The key dimensions for the format of the existing LJP tasks, either being expert-annotated (highlighted in red) or machine-extracted (highlighted in blue), or both (highlighted in orange) Note: QAjudge refers to the union of dataset QAjudge-CJO, QAjudge-PKU and QAjudge-CAIL.

Dataset	Language	Input			Output		Download
					Rationale	Decision	
SwissJudgment [27]	German / Italian / French	Pleas	Facts		-	Plea judgments	Link
Sulea et al. [14]	French	Pleas	Facts	Others	-	Plea judgments	-
CanAppeal [41]	English	Facts			-	Plea judgments	Link
USClassActions [42]	English	Pleas			-	Plea judgments	Link
JUSTICE [28]	English	Pleas	Facts		-	Plea judgments	-
BStricks_LDC [43]	English	Pleas	Facts	Others	-	Plea judgments	Link
ILDC [44]	English	Pleas	Facts	Others	Fact snippets	Plea judgments	Link
ACI [45]	English	Facts			Fact snippets	Charges	Link
COLIEETask1 [46]	English	Facts			Precedents		Link
AILA [47]	English	Facts			Precedents		-
echr [13]	English	Facts			Articles	-	Link
ECHR-CASES [26]	English	Facts			Articles	-	Link
ECtHR [48]	English	Facts			Articles	-	Link
PhilCases [49]	English	Facts			-	Plea judgments	-
SCDB [50], [51]	English	Facts			-	Plea judgments	Link
HLDC [52]	Hindi	Facts			-	Plea judgments	Link
TSCC [53]	Thai	Facts			Articles	-	Link
MLMN [54]	Chinese	Facts			Articles	Prison terms	Link
DPAM [55]	Chinese	Facts			Articles	-	Link
CAIL2018 [25]	Chinese	Facts			Articles	Charges	Link
TOPJUDGE-CJO [30]	Chinese	Facts			Articles	Charges, Prison terms	-
TOPJUDGE-PKU [30]	Chinese	Facts			Articles	Charges, Prison terms	-
TOPJUDGE-CAIL [30]	Chinese	Facts			Articles	Charges, Prison terms	-
CAIL-Long [56]	Chinese	Facts			Articles	Charges, Prison terms	Link
FLA [17]	Chinese	Facts			Articles	Charges	-
RACP [57]	Chinese	Facts			Fact snippets	Charges	-
MAMD [58]	Chinese	Facts			-	Charges	-
Court-View-Gen [34]	Chinese	Facts			Court View		Link
AC-NLG [35]	Chinese	Facts	Pleas		Court View		Link
CPTP [59]	Chinese	Facts	Charges		-	Prison terms	Link
Criminal-S [60]	Chinese	Facts	Charge attributes		Attributes prediction	Charges	Link
Criminal-M [60]	Chinese	Facts	Charge attributes		Attributes prediction	Charges	Link
Criminal-L [60]	Chinese	Facts	Charge attributes		Attributes prediction	Charges	Link
QAjudge [61]	Chinese	Facts	Elements questions		Articles, Elements answers	Charges	-
Auto-Judge [62]	Chinese	Pleas	Facts	Articles	-	Plea judgments	-
LJP-MSJudge [63]	Chinese	Pleas	Debate data		-	Plea judgments	Link
law-turk [64]	Turkish	Facts			-	Plea judgments	Link
BrCases [65]	Portuguese	Facts			-	Plea judgments	
Homicides [66]	Portuguese	Facts			-	Plea judgments	
Corruption [66]	Portuguese	Facts			-	Plea judgments	
BrCAD-5 [67]	Portuguese	Facts			-	Plea judgments	Link

Furthermore, these precedent prediction tasks are classified into two categories as followed.

- With-decision vs without-decision. It involves a set of prior cases which have been supported for given factual scenarios, and are further into scenarios with decisions and scenarios without decisions. Scenarios with decisions tasks are addressed in works such as AutoLaw [67] and CFLT [68], while scenarios without decisions tasks are tackled in AILA [47], COLIEETask1 [46] and LeCaRD [69].
- Common-law vs civil-law. There are two kinds of prior cases documents: cited-based [46], [47], [67], [68] in the common-law system and expert-based [69] in the civil-law system.

Note that this paper will focus the precedent prediction task on the scenarios without decisions in the common-law system.

Charge Prediction Task. Table 2 reveals the presence of multiple charge prediction tasks and associated datasets in the juridical domain [30], [59], [60]. This study summarizes three cases to address these tasks.

- Case-charge correspondence. It is divided into two categories: (1) One-vs-one, where TOPJUDGE-CJO [30], TOPJUDGE-PKU [30], TOPJUDGE-CAIL [30], Criminal-S [59], Criminal-M [59], and Criminal-L [59] datasets eliminate cases with multiple defendants and charges for judgment prediction. (2) One-vs-many, where MAMD [57] considers cases with multiple defendants charged differently, and FLA [17] only includes cases with one defendant and more than one charge.
- Common-vs-Few shot. It aims to assess the efficacy of the models on few-shot charges, where Criminal-S [59], Criminal-M [59], and Criminal-L [59] datasets include 149 charges with frequencies greater than 10, while FLA [17] selects 50 charges with frequencies greater than 80 and TOPJUDGE-CJO [30], TOPJUDGE-PKU [30], and TOPJUDGE-CAIL [30] consider charges with frequencies greater than 100.

TABLE 3. Abbreviation-Full name correspondence for courts in this paper.

Abbreviation	Full name
BAC	Brazilian Appellate Court
CITUS	Courts in the US
ECHR	European Court of Human Rights
FCC	Federal Court of Canada
FDCOUS	Federal District Court of US
FSC	French Supreme Court
FSCCs	Federal Small Claims Courts under the 5th Regional Federal Court jurisdiction
FSCS	Federal Supreme Court of Switzerland
HCT	Higher Courts of Turkey
PSC	Philippine Supreme Court
SCI	Supreme Court of Indian
SCOTUS	Supreme Court of the US
SKCA	Courts of Appeal for Saskatchewan
SPCC	Supreme People's Court of China
TJSP	Tribunal de Justiça de São Paulo
TSC	Thai Supreme Court
UKC	UK's highest court

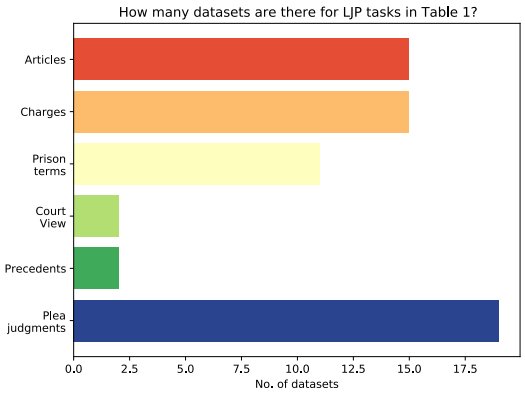

FIGURE 5. The dataset density distribution for specific LJP task.

- Logical knowledge. It focuses on three aspects: (1) Dependencies among LJP tasks, where the TOPJUDGE-CJO [30], TOPJUDGE-PKU [30], and TOPJUDGE-CAIL [30] datasets verify the model's efficiency based on the topological dependencies among law articles, charges, and prison term prediction. (2) Domain knowledge, where the discriminative charges' predefined attributes [59] are utilized for charge prediction, and visualized answers to predefined element questions are employed for interpretable charge prediction [60]. (3) Fact-side representation, where the document-level charge prediction [45] is modeled based on the relationship between sentence-level facts and their corresponding charges.

Prison Term Prediction Task. It suffers from inadequate datasets as shown in Fig. 5. To address this issue, several datasets have been developed, such as CPTP [58], TOPJUDGE-CJO [30], TOPJUDGE-PKU [30], and TOPJUDGE-CAIL [30], which aim to predict the prison terms based on the corresponding charges.

Court View generation Task. This task is challenging due to the complex logical reasoning needed to interpret the judgment of a case. This task is rarely explored as shown in Fig. 5, and only a few studies have investigated it due to the difficulty arising from complex logical reasoning about facts and relevant law article. To this end, two approaches have been proposed, including utilizing the dependencies of charge-label and Court View in Court-View-Gen [34] and using a pleas-aware side facts representation in AC-NLG [35].

Plea Judgment Prediction Task. The task typically involves predicting the judgments on plaintiffs' pleas based on the fact description. In this regard, three cases have been studied: (1) Judge-summarized facts narrative, which involves predicting the final pro or con decision of judges based on an absolute neutral text summary of the case facts. For instance, Auto-Judge [61] and ILDC [44] have been developed for this purpose. (2) Facts from court debate, which aims to predict the verdicts of plaintiffs' pleas based on the facts from real court debate. LJP-MSJudge [62] is an example of such an approach. (3) Case material except for the plea judgments, where interpretable plea judgments are predicted based on the case material that is masked with the plea judgments [14], [43], [44].

III. DATASETS

LJP poses a significant challenge in terms of dataset availability and potentiality. To address this challenge, two lines of work have been identified, as illustrated in Table 2. The first line of work employs a machine to extract metadata from judgment documents, such as using Regular Expression to identify and mask judgment results [14], [27], [43], [44], or to extract input-output data samples from judgment documents of given court [13], [17], [25], [26], [30], [34], [48], [54], [55], [57], [58], [61]. In contrast, the second line of work involves the recruitment of legal experts to annotate rationale sentences [44], [56] or legal domain knowledge [21], [59], [60], which is not included in judgment documents. Subsequently, data obtained above are employed to predict judgment results [45].

LJP datasets are categorized into two groups: single-task datasets and multi-task datasets, with multi-task datasets

TABLE 4. The statistics of the existing datasets for LJP tasks.

Dataset	Source	Category		#Cases	#Articles	#Charges	#Prison terms
		Law Domain	Legal System				
USClassActions [42]	CITUS	Class action	Common-Law	10,759	-	-	-
BStricks_LDC [43]	UKC	Generic		4,959			
CanAppeal [41]	SKCA			3,670			
JUSTICE [28]	SCOTUS			2,384			
SCDB [51]				28,009			
COLIEETask1 [46]	FCC			5,978			
AILA [47]	SCI			2,964			
ILDC [44]				34,816			
ACI [45]	SCI	Criminal		4,338		20	
HLDC [52]	DCOUP			912,568		-	
PhilCases [49]	PSC		Civil-Law	6,483			
TSCC [53]	TSC			1,207	122		
DPAM [55]	SPCC			17,160	70		
MLMN [54]				1,189	86		5
CAIL2018 [25]				2,676,075	183	202	202
TOPJUDGE-CJO [30]				1,007,744	98	99	11
TOPJUDGE-PKU [30]				175,744	68	64	11
TOPJUDGE-CAIL [30]				113,536	105	122	11
FLA [17]				60,000	321	50	-
RACP [57]				100,000	-	51	
MAMD [58]				164,997		-	
CPTP [59]				238,749		157	226
Court-View-Gen [34]				171,981		51	-
Criminal-S [60]				77,046		149	
Criminal-M [60]				191,960		149	
Criminal-L [60]				383,697		149	
QAjudge-CJO [61]				15,120	19	20	
QAjudge-PKU [61]				14,000	19	20	
QAjudge-CAIL [61]				13,423	19	20	
CAIL-Long [56]				115,849	244	201	240
		Civil		113,656	330	257	-
DPAM [55]				4,033	30	-	
LJP-MSJudge [63]				70,482	-		
AC-NLG [35]				66,904			
Auto-Judge [62]				100,000	62		
echr [13]	ECHR	Human rights		584	3		
ECHR-CASES [26]				11,478	66		
ECtHR [48]				11,532	14		
SwissJudgment [27]	FSCS	Generic		85,268	-		
Sulea et al. [14]	FSC			126,865			
law-turk [64]	HCT			39,147			
BrCases [65]	BAC			4,043			
Homicides [66]	TJSP			591			
Corruption [66]				191			
BrCAD-5 [67]	FSCCs			765,602			

consisting of multiple LJP subtasks and single-task datasets comprising a single subtask. And for brevity, we will use the abbreviation of all the courts in this paper as shown in Table 3.

A. SINGLE-TASK DATASETS

This section discusses various single-task datasets released publicly, which are categorized based on their task-specific outcomes.

1) ARTICLE PREDICTION DATASETS

The HUDOC ECHR[1] and China Judgments Online (CJO)[2] are two publicly available case databases, each containing various legal documents such as judgments, verdicts, conciliation statements, decision letters, notices, etc. These

[1] https://hudoc.echr.coe.int/
[2] https://wenshu.court.gov.cn/

resources provide valuable data for researchers interested in uncovering patterns underlying judicial decisions.

One dataset of note is the echr [13], which represents the first public English legal judgment prediction dataset. It comprises 584 cases from the ECHR and articles 3, 6, and 8 of the European Convention of Human Rights. The dataset is designed to predict whether a given case has violated one of the articles of the Convention. Another similar dataset, the Thai Supreme Court Cases (TSCC) [53], contains 1,207 criminal judgment records and 122 law records from the Supreme Court of Thailand. It is constructed to predict which specific law records have been violated, using sequence models. There is also a dataset, MLMN [21], based on 1,189 crime judgment documents in CJO and 86 criminal law articles for fact-article correspondence annotation. This dataset aims to verify the improvement on articles recommendation accuracy through establishing fine-grained fact-article correspondences.

A substantially more extensive dataset is ECHR-CASES [26], which includes 11,478 cases tried by the ECHR and 66 articles from the European Convention of Human Rights. This dataset not only judges whether a violation of an article of the Convention has occurred but also determines the name of the violated articles. While all these datasets are related to such article prediction tasks, they often do not pertain to civil and administrative cases, as evidenced in Table 4.

2) PRECEDENT PREDICTION DATASETS

Thomson Reuters Westlaw India[3] is a publicly available case database published by SCI.

To the best of our knowledge, from the view of scenarios without decisions, there are only two English precedent prediction datasets in common-law system as follows:

The first dataset dataset is AILA [47]. At first it randomly selects 50 case documents out of a filtered collection of case documents from SCI based on cited articles feature. Subsequently the facts of these 50 case documents are extracted manually as the input of the precedent prediction task. Finally, a total of 2,914 prior cases are obtained based on case titles with their ID in the 50 case documents above. The second one is COLIEETask1 [46]. The latest version of this dataset contains 1,200 query cases among a total of 5,978 cases drawn from an existing collection from FCC case laws provided by Compass Law. In addition, access to the COLIEETask1 is granted upon request.

3) CHARGE PREDICTION DATASETS

Two publicly available case databases, namely Case Information Disclosure (CID)[4] and CJO, have been published by the Supreme People's prosecution and SPCC, respectively. While several datasets have been constructed for charge prediction tasks based on publicly available resources, these datasets tend to have a focus on Chinese law and less emphasis on English (as seen in Table 4).

FLA [17] is a dataset constructed from CJO that comprises 60,000 cases, 50 charges, an average of 383 words per fact description, and 3.81 articles per case. 3.56% of the cases have more than one charge, and 321 distinct articles are included in the dataset. The cases with one defendant are retained, and charges that appear more than 80 times are treated as positive data while vice versa as negative ones. This dataset aims to improve charge prediction following the prediction of the relevant law articles. CAIL2018 [25], on the other hand, is the first large-scale Chinese legal dataset to predict relevant law articles, charges, and prison terms, respectively. It comprises 2,676,075 criminal cases published by the SPCC, 183 criminal law articles, 202 charges, and prison terms. Only cases with a single defendant are retained, and charges and law articles whose frequency is larger than 30 are treated as positive data. MAMD [57], a dataset designed to predict multi-defendant charges, has been constructed from the published legal documents in CID. The dataset contains a total of 164,997 cases, involving fact description, defendants' names, and charges. Cases involving multiple defendants account for about 30%, and in cases involving multiple defendants, the ratio of those charged with the same offense is about 90%. RACP [56], a dataset constructed from CJO, contains 100,000 documents in which rationale sentences were annotated based on the extracted fact description and charge labels. Similarly, ACI [45] collected 4,338 judgment documents with document-level charge-labeled information from the SCI cases. Facts were extracted by legal experts and automated methods for 70 and 4,268 documents respectively, and legal experts annotated sentence-level charges for 120 documents. Furthermore, three datasets [59] have been published with selected case's fact part and extracted charges of judgment documents from CJO, denoted as Criminal-S (small), Criminal-M (medium), and Criminal-L (large). The cases that have more than one charge in a verdict were removed from the datasets.

4) PRISON TERM PREDICTION DATASETS

CJO, a public case database published by the SPCC, contains various types of legal documents, such as verdicts, judgments, conciliation statements, decision letters, notices, etc. Despite the availability of several datasets for prison term prediction, they have primarily focused on Chinese legal cases, with limited consideration given to English cases (see Table 4).

To evaluate the impact of charge-based prison term prediction (CPTP) on the accuracy of predicting the full prison term for each defendant, a dataset called CPTP [58] has been developed. This dataset comprises 238,749 criminal cases, [1,240] months of prison terms, and 157 types of charges. In addition, the fact-article correspondence dataset, MLMN, developed in [21], can be used to enhance the downstream task of legal decision prediction, where the results of judgments are classified into five categories: exempt from criminal punishment, criminal detention,

[3]https://www.westlawasia.com/

[4]http://www.ajxxgk.jcy.gov.cn/html/index.html

TABLE 5. The statistics of the pre-training datasets.

Dataset	Language	Model	Source	#Documents	#Size(MB)	Corpus
Chalkidis et al. [72]	English	LEGAL-BERT [72]	EU legislation	61,826	1,900	-
			European Court of Justice (ECJ) cases	19,867	600	
			UK legislation	19,867	1,400	
			ECHR cases	12,554	500	
			US court cases	164,141	3,200	
			US contracts	76,366	3,900	
LexGLUE [29]		BERT [19]	ECHR [26]	11,000	116	Link
		RoBERTa [73]	ECHR [78]	11,000	116	
		DeBERTa [74]	US Law [79]	7,800	328	
		Longformer [75]	EU Law [80]	65,000	492	
		BigBird [76]	Contracts [81]	80,000	62	
		LEGAL-BERT [72]	Contracts [82]	9,414	3	
		CaseLaw-BERT [77]	Harvard Law case [77]	52,800	86	
CaseHOLD [77]		CaseLaw-BERT [77]	Harvard Law case [77]	3,446,187	37,000	Link
CourtListener [83]		LegalDB [83]	US Board Of Tax Appeal	11,059	8,000	Link
			US Court Of Federal Claims	13,410		
			Court Of Customs And Patents Appeal	2,388		
			Supreme Court Of The United States	31,470		
			Court Of Appeals (First–Eleventh Circuit)	269,622		
Xiao et al. [56]	Chinese	Lawformer [56]	SPCC Criminal cases	5,428,717	17,000	-
			SPCC Civil cases	17,387,874	67,000	
Douka et al. [84]	French	JuriBERT [84]	French Court cases	123,361	6,300	
			Legifrance			
Garneau et al. [85]		CriminelBART [85]	Criminal and Penal Chamber of Quebec cases	9,000	-	
AL-Qurishi et al. [86]	Arabic	AraLegal-BERT [86]	Books	6,000	-	
			Cases	336,000		
			Terms and laws	3,000		
			others	5,000		

fixed-term imprisonment of not more than one year, 1 - 3 years, and not less than three years.

5) PLEA JUDGMENT PREDICTION DATASETS

There are assorted publicly available case databases that are used in various jurisdictions worldwide, including CJO published by the SPCC, entscheidsuche.ch by the FSCS, IndianKanoon by SCI, Court de Cassation by the FSC, Oyez by the SCOTUS, the Saskatchewan Court of Appeal Sentencing Digest Database[5] by the SKCA, and the Canadian Legal Information Institute (CanLII)[6] website by all Canadian case law and judgments.

Several datasets for pleas-related tasks are scraped from publicly available resources, but these large-scale datasets mainly focus on DCOUP, FSCCs, SPCC and FSC cases (see Table 4). For instance, Auto-Judge [61] is a dataset containing 100,000 divorce cases, 185,723 pleas and verdicts, and 62 law articles. It is designed to predict the final judgment results based on semantic interactions among facts, pleas, and laws. In order to incorporate actual case inputs from courtrooms rather than judge-summarized case narratives for judgment prediction, LJP-MSJudge [62] has been released, containing 70,482 Private Lending cases collected from CJO, 133,209 claims and verdicts, 4.1 million debates, and 10 fact labels. Additionally, Sulea et al. [14] proposed a dataset comprising 126,865 unique court rulings, which was first used to predict court rulings in French Supreme Court cases. The Indian Legal Documents Corpus (ILDC [44]) includes 34,816 cases and is introduced for court judgment prediction and explanation. Moreover, a labeled dataset of 4,959 UK court cases [40] has also been created for legal judgment prediction. A total of 2,384 SCOTUS cases from Oyez based on a series of manually balanced procedures are also evaluated for predicting judicial judgments. Finally, SwissJudgment [27] is a diachronic multilingual dataset of 85,268 cases from the FSCS, which includes 49,882 cases in German, 31,094 in French, and 4,292 in Italian. USClassActions [42] and its variants are curated from 10,759 class action cases by extracting the plaintiffs' facts and allegations via a rule-based regex extraction system. CanAppeal [41] is generated by linking the SKCA Sentencing Digest Database and CanLII website and using two-step extracting the judgment labels from 3,670 SKCA cases based on regular expressions. To ensure that case outcome prediction is general, consistent, and out-of-sample applicable, [51] relies on 15 features from SCDB, which includes historical decisions over the past 200 years. [49] constructs a dataset of 6,483 criminal appeal cases from the Chan Robles Virtual Law Library[7] and the Lawphil project[8] using regular expression to predict court case outcomes. The decisions of 5 appeal courts are extracted in [63] using regular expression to predict the different Turkish courts verdicts based on fact description. Through retaining yes/no/partial decision labels of cases from the Tribunal de Justiça de Alagoas, 4,043 Brazilian cases are obtained to predict court decisions. BrCAD-5 dataset containing 765,602 appealed lawsuits from 3,128,292

[5]https://www.lawsociety.sk.ca/legal-resources-library/research-tools/
[6]https://www.canlii.org/en/
[7]https://chanrobles.com/
[8]https://lawphil.net/

lawsuits in FSCCs with a label of voting, is released to predict the appeal panel decisions on the lower court decisions.

6) COURT VIEW GENERATION DATASETS

The SPCC has made available CJO, a publicly accessible case database containing various legal documents such as verdicts, judgments, notices, decisions, and conciliation statements.

Despite the availability of several publicly scraped datasets for the task of generating court views, these resources are largely focused on Chinese-language documents, with limited attention paid to English-language materials (see Table 2).

Court-View-Gen [34] is a novel dataset of 171,981 Chinese legal cases, each with one defendant and a single charge, encompassing a total of 51 charge labels. The dataset is specifically designed to facilitate the generation of court views based on charge labels. The data was gathered from published legal documents in the CJO repository.

B. MULTI-TASK DATASETS

Multi-task datasets have been developed to improve legal judgment prediction by providing detailed subtasks (see Table 2).

For example, the QAjudge datasets [60], namely QAjudge-CJO, QAjudge-PKU, and QAjudge-CAIL, have been created using data from China Judgments Online, Peking University Law Online, and Chinese AI and Law Challenge, respectively. These datasets include fact descriptions, applicable law articles, charges, and penalty terms for each case, with multiple defendants and charges, infrequent charges, and articles less than 100 times being filtered out.

Another publicly high-quality dataset, CAIL2018 [25], is of large scale with 2,676,075 criminal cases, 183 criminal law articles, 202 charges, and prison terms. This dataset only retains cases with a single defendant, and charges and law articles with frequency larger than 30.

Recently, a larger scale dataset, CAIL-Long [55], containing 1,129,053 criminal cases and 1,099,605 civil cases, was constructed to predict judgment results. Each criminal case is annotated with charges, relevant laws, and penalty terms, whereas each civil case is annotated with causes of actions and relevant laws.

In addition, generating court views by jointly producing the judgment and rationales can enhance interpretability. For example, AC-NLG [35] has built 66,904 civil legal judgment cases from CJO, each categorized into plaintiff's claim, fact description, and court's view. These cases were further annotated with judgments and rationales to provide a more comprehensive understanding of the court's decision-making process.

C. DATASETS FOR PRE-TRAINED LANGUAGE MODEL

As widely acknowledged, numerous open access repositories have been established to construct unlabeled pretraining corpora, as presented in Table 5. For instance, the SPCC and ECHR have built publicly available case databases, namely CJO and HUDOC ECHR, respectively. Furthermore, US court cases, raw French legal text, and decisions from the judicial and administrative tribunals of Québec can be found on CourtListener,[9] Légifrance, and Société Québécoise d'Information Juridique (SOQUIJ)[10] websites, respectively.

Following the success of pre-trained language models (PLMs) in the general domain, several datasets have been scraped from publicly available legal resources to investigate their adaptation to legal tasks. For example, 12GB pre-trained unlabeled corpora of diverse English legal text [70] from legislation, court cases, and contracts has been scraped from publicly available resources to pre-train different variations of BERT in the legal area. Subsequently, the Legal General Language Understanding Evaluation (LexGLUE [29]) has been selected as a generic benchmark dataset for multiple legal NLP tasks in English. 8GB of long legal documents from the US have also been used as legal domain pre-training corpora to validate the efficiency of the LegalDB model [81]. Case Holdings On Legal Decisions (CaseHOLD [75]), a large-scale pre-training dataset with 3,446,187 legal decisions from Harvard Law case corpus, has been constructed to explore the influence of difficulty and domain specificity on domain pre-training gains.

Despite the usefulness of PLMs when adapted to the legal domain, the main effort has mainly focused on the English language. Therefore, Longformer-based PLMs based on 84GB pre-training corpora of Chinese legal long documents, and the large-scale judgment prediction dataset, CAIL-Long [55], have been constructed to address legal judgment prediction tasks by capturing the long-distance dependency on Chinese legal case documents up to 512 tokens. In addition, a collection of 6.3GB raw French legal text has been gathered to explore the performance of the adaptation of domain-specific BERT models in the French language. Additionally, a collection of 9000 judgments from the Criminal and Penal Chamber of Québec is used to further pre-train BARThez, a specialized model for criminal law.

However, most of these datasets for legal PLMs have focused on a single language corpus and have not adequately considered a multi-language corpus, as shown in Table 5.

IV. EVALUATION METRICS

Fig. 6 provides a framework for evaluating the outputs of LJP tasks, comprising four categories as followed. The first is text classification metrics for evaluating law articles, charges, plea judgments, and element answers. The second is text classification with text regression metrics for evaluating prison terms according to their value distribution. The third is text classification with text generation metrics for evaluating fact snippets based on their generation methods. The fourth is text generation metrics are used for evaluating the generated court view.

[9]https://www.courtlistener.com/api/bulk-info/
[10]https://soquij.qc.ca/a/fr

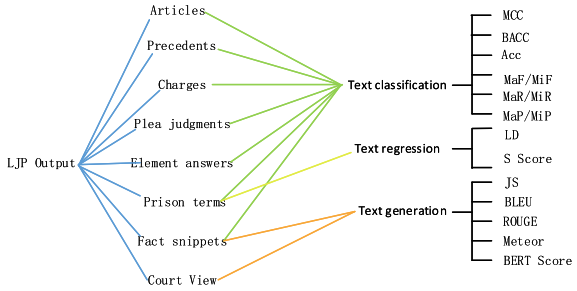

FIGURE 6. Evaluation metrics for different outputs from LJP tasks.

A. EVALUATION OF TEXT CLASSIFICATION

To be specific, for the prediction of articles, charges, prison terms, plea judgments, fact snippets and element answers, we can take them as text classification problems, such as in Fig. 6. For a specific text classification task, suppose there are M categories and N law cases. This text classification task aims to predict the category label for the text description of each law case. Let $y_{ij} \in \{0,1\}(i \in \{1,2,\cdots,M\}, j \in \{1,2,\cdots,N\})$ denote as the ground truth label of the category result. Let $\hat{y}_{ij} \in \{0,1\}(i \in \{1,2,\cdots,M\}, j \in \{1,2,\cdots,N\})$ denote as the predict label of the category result. Then, we can obtain the true positive, false positive, false negative, true negative, precision, recall metrics for the i-th category as follows:

$$
\begin{aligned}
TP_i &= \sum_{j=1}^{N} [y_{ij} = 1, \hat{y}_{ij} = 1], \\
FP_i &= \sum_{j=1}^{N} [y_{ij} = 0, \hat{y}_{ij} = 1], \\
FN_i &= \sum_{j=1}^{N} [y_{ij} = 1, \hat{y}_{ij} = 0], \\
TN_i &= \sum_{j=1}^{N} [y_{ij} = 0, \hat{y}_{ij} = 0], \\
P_i &= \frac{TP_i}{TP_i + FP_i}, \\
R_i &= \frac{TP_i}{TP_i + FN_i}
\end{aligned}
$$

Then let true positive, false positive, false negative and true negative metrics be represented as $TP = \sum_{i=1}^{M} TP_i$, $FP = \sum_{i=1}^{M} FP_i$, $FN = \sum_{i=1}^{M} FN_i$ and $TN = \sum_{i=1}^{M} FP_i$, respectively.

After that, as in Table 6, we can obtain the following evaluation metrics to evaluate the performance of LJP text classification.

- **Macro precision/ Macro recall/ Macro F value**. To evaluate the performance in the macro-level through averaging over each category, macro precision MaP, macro recall MaR and macro F value MaF are follows:

$$
\begin{aligned}
MaP &= \frac{1}{M} \sum_{i=1}^{M} P_i \\
MaR &= \frac{1}{M} \sum_{i=1}^{M} R_i \\
MaF &= \frac{1}{M} \sum_{i=1}^{M} \frac{2P_i \times R_i}{P_i + R_i}
\end{aligned} \tag{1}
$$

- **Micro precision/ Micro recall/ Micro F value/ Acc**. To evaluate the performance in the micro-level through averaging over each law case, micro precision MiP, micro recall MiR, micro F value MiF, accuracy Acc, balanced accuracy (BACC) and Matthews Correlation Coefficient (MCC) are follows:

$$
\begin{aligned}
MiP &= \frac{TP}{TP + FP} \\
MiR &= \frac{TP}{TP + FN} \\
MiF &= \frac{2MiP \times MiR}{MiP + MiR} \\
Acc &= \frac{TP + TN}{TP + FP + TN + FN} \\
BACC &= \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \\
MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}
\end{aligned}
$$

Note that MCC metric above is for binary classification problem, but [85] generalizes it in multiclass problems.

B. EVALUATION OF TEXT GENERATION

For the generation of court view, we can take them as text generation problems, such as BLEU-1, BLEU-2, BLEU-N, ROUGE-1, ROUGE-2, ROUGE-L and BERT SCORE adopted in [35]. The following evaluation metrics to evaluate the performance of LJP text generation are summarized, as shown in Fig. 6.

- **Jaccard similarity**. To evaluate the similarity between two sets, Jaccard similarity is defined as the size of the intersection divided by the size of the union of the two sets, which is follows:

$$
JS = \frac{|\{Candidates\} \cap \{References\}|}{|\{Candidates\} \cup \{References\}|} \tag{2}
$$

where, $\{Candidates\}$ denotes the text set predicted, and $\{Candidates\}$ denotes the text set of the reference.

- **BLEU**. To evaluate the exact form closeness between the candidate sentence and its reference sentences, BLEU is averaged geometrically computed for mutiple modified n-gram up to length N precision (e.g. $n = 1,2,3,4, N = 4$), which is follows:

$$
BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}
$$

where BP, p_n, w_n are an exponential brevity penalty factor, modified n-gram precision and a weighted coefficients of modified n-gram precision for the up to N-gram combining system, respectively.

$$
BP = \begin{cases} 1, & if \quad c > r, \\ e^{(1-r/c)}, & if \quad c \le r. \end{cases} \tag{4}
$$

where c denotes the length of the candidate sentence and r denotes the length of the reference sentences.

TABLE 6. Evaluation metrics used in papers which construct datasets for evaluating the performance of the existing LJP models. Here, EM: exact match rate, Acc@p: error-tolerant accuracy, where p is the maximum acceptable error rate.

Dataset	Output	Evaluation metrics			
ECtHR [48]	Articles	Acc			
echr [13]					
USClassActions [42]	Plea judgments				
PhilCases [49]					
MAMD [58]	Charges				
ACI [45]		MaF		MaP	MaR
ECHR-CASES [26]	Articles				
MLMN [54]					
RACP	Charges				
	Fact snippets	Acc	MaF	MaP	MaR
SCDB [51]	Plea judgments				
CanAppeal [41]					
BrCases [65]					
Homicides [66]					
Corruption [66]					
JUSTICE [28]					
BStricks_LDC [43]					
Sulea et al. [14]					
Auto-Judge [62]					
CAIL2018	Articles				
	Charges				
	Prison terms				
TOPJUDGE-CJO,TOPJUDGE-PKU,TOPJUDGE-CAI [30]	Articles				
	Charges				
	Prison terms				
QAjudge-CJO,QAjudge-PKU,QAjudge-CAIL [61]	Articles				
Criminal-S,Criminal-M,Criminal-L [60]	Charges				
ILDC [44]	Plea judgments				
	Fact snippets	JS	ROUGE	BLEU	Meteor
DPAM [55]	Articles		MaF	MaP	MaR
SwissJudgment [27]	Plea judgments	MaF/MiF			
TSCC [53]	Articles				
CAIL-Long [56]	Charges				
	Prison terms	LD			
QAjudge-CJO,QAjudge-PKU,QAjudge-CAIL [61]	Elements answers	Acc		MaF	
HLDC [52]	Plea judgments			MiF	
law-turk [64]				BACC	MaF
COLIEETask1	Precedents	MiF		MiP	MiR
FLA [17]	Charges	MaF/MiF		MaP/MiP	MaR/MiR
LJP-MSJudge [63]	Plea judgments			MaP	MaR
Court-View-Gen [34]	Court View	ROUGE		BLEU	
AC-NLG [87]		BERT SCORE	ROUGE		
CPTP [59]	Prison terms	S	EM	Acc@p	
BrCAD-5 [67]	Plea judgments	MCC			

- **ROUGE**. To evaluate n-gram co-occurrence statistics between the computer-generated text and the referenced text created by humans, a family of ROUGE metrics is defined as a recall-based measure, which is follows:

$$ROUGE-N = \frac{\sum_{S\in\{References\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{References\}}\sum_{gram_n\in S} Count(gram_n)}$$

where the numerator of $ROUGE-N$ denotes the number of n-grams co-occurring between a candidate text and a set of reference texts, and the denominator of $ROUGE-N$ denotes the number of n-grams in the set of reference texts.

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ \beta &= \frac{P_{lcs}}{R_{lcs}} \\ ROUGE-L &= \frac{(1+\beta^2)P_{lcs}R_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \end{aligned} \tag{5}$$

where $LCS(X,Y)$ is the length of a longest common subsequence of sequence X of length m and sequence Y of length n.

- **Meteor**. To give an explicit word-to-word matching between the candidate text and its reference texts, Meteor is allowing backing-off from the exact unigram matching to porter stem matching and synonyms, which is follows:

$$Meteor = F3 \times (1 - penalty) \tag{6}$$

where $F3$ denotes a harmonic metric by combining the unigram precision (P, the ratio of number of unigrams mapped to the total unigrams in the candidate text) and the unigram recall (R, the ratio of number of unigrams mapped to the total unigrams in the reference text). And *penalty* denotes a penal function increases as the number of mapped subsequence chunks while decreasing as the number of unigrams mapped.

$$F3 = \frac{10PR}{R + 9P}$$
$$penalty = 0.5 \times \left(\frac{\#chunks}{\#unigrams\_matched}\right)^3$$

where, $\{Candidates\}$ denotes the text set predicted, and $\{Candidates\}$ denotes the text set of the reference.

- **BERT SCORE**.To evaluate the semantically similarity between the candidate sentence and its reference sentence, BERT SCORE is defined based on the contextual embedding for vector representations for the word depending on its surrounding words, which is follows:

$$BERT - P = \frac{1}{|y|}\sum_{y_j \in y} \max_{x_i \in x} \mathbf{x}_i^T \mathbf{y}_j$$
$$BERT - R = \frac{1}{|x|}\sum_{x_i \in x} \max_{y_j \in y} \boldsymbol{x}_i^T \boldsymbol{y}_j$$
$$BERT - F = \frac{2BERT - R \times BERT - P}{BERT - P + BERT - R} \tag{7}$$

where $x = \langle x_1, \cdots, x_m\rangle, \langle \mathbf{x}_1, \cdots, \mathbf{x}_m\rangle$, $y = \langle y_1, \cdots, y_n\rangle, \langle \mathbf{y}_1, \cdots, \mathbf{y}_n\rangle$ denote tokenized reference sentence, sequence of embedding vectors for the tokenized reference sentence, tokenized candidate sentence, sequence of embedding vectors for the tokenized candidate sentence, respectively.

C. EVALUATION OF TEXT REGRESSION

For the prediction of prison term at a continuous interval, we can take them as text regression problems, such as in [58]. Given the difference between the predicted prison term $\hat{y}_i$ and the ground truth value y_i of the the i-th case, the evaluation metrics employed to evaluate the performance of prison term prediction problems are summarized as follows:

- **Log distance**. To evaluate the tiny difference between the predicted prison term and its ground truth value based on distance, Log distance LD is defined as follows:

$$LD = \sum_{i=1}^{N} \frac{|log(y_i + 1) - log(\hat{y}_i + 1)|}{N} \tag{8}$$

- **S score**. As the metrics used in the CAIL2018 Competition [86], S score metric can be obtained to evaluate the similarity between two continuous stochastic variables:

$$S = \sum_{i=1}^{N} \frac{f(|log(y_i + 1) - log(\hat{y}_i + 1)|)}{N} \tag{9}$$

where, function $f(\cdot)$ satisfies:

$$f(v) = \begin{cases} 1.0, & if \quad v \leq 0.2, \\ 0.8, & if \quad 0.2 < v \leq 0.4, \\ 0.6, & if \quad 0.4 < v \leq 0.6, \\ 0.4, & if \quad 0.6 < v \leq 0.8, \\ 0.2, & if \quad 0.8 < v \leq 1.0, \\ 0.0, & if \quad 1.0 < v. \end{cases} \tag{10}$$

V. METHODS

In this section, we present various methods for legal judgment prediction using publicly available datasets, as described in Section III. We recommend the use of multi-task learning (MTL) method for LJP datasets containing multiple subtasks to exploit the dependencies among prediction results (as discussed in Section V-A). For task-specific datasets, we propose the use of pre-trained language models (PTMs, as explained in Sectio V-B) for fine-tuning downstream tasks. In cases where the distribution of judgment results in a few-shot scenario is imbalanced, we suggest using the few-shot learning (FSL) method (as presented in Section V-D).

A. MULTI-TASK LEARNING

Multi-task learning (MTL) has numerous successful usages in NLP tasks, which transfers useful information across relevant tasks by solving them simultaneously so that it has been applied to a wide range of areas, including NLP especially the legal domain.

MTL has become a widely utilized approach in NLP tasks, allowing for the transfer of useful information across related tasks by solving them simultaneously. This has resulted in successful applications of MTL in various domains, including the legal domain.

In this section, we describe a family of MTL methods that focus on utilizing logical dependencies among related tasks.

For instance, Fact-Law Attention (FLA) [17] has been proposed for improving the accuracy of charge prediction in the civil law systems. It employs a two-stack attention-based neural network to jointly model the charge prediction task and the relevant article extraction task in a unified framework. The first stack utilizes a sentence-level and document-level Bi-directional Gated Recurrent Units (Bi-GRU) for fact embedding, while the second stack generates article embedding dynamically for each case based on the fact-side clues.

Subsequently, a novel topological MTL framework, namely TOPJUDGE[11] [30], has been proposed, which is

[11] https://github.com/thunlp/TopJudge

based on the facts of a case and the topological dependencies among the articles, charges, and prison terms. In particular, for the civil law system, the judge first determines the applicable law articles by analyzing the fact description of a given case, then determines the charges based on the instructions of the relevant law articles, and finally confirms the penalty terms based on the aforementioned outcomes.

Furthermore, a Multi-Perspective Bi-Feedback Network (MPBFN) with the Word Collocation Attention (WCA) mechanism was proposed in [32]. This method utilizes the semantic vector of fact with word collocation and number semantic attention mechanism, as well as the judgment results of pre-dependent tasks, to perform forward prediction for follow-up tasks. Meanwhile, the judgment results of follow-up tasks are employed to perform backward verification and evaluate the rationality of pre-order tasks.

An additional category of multi-task learning methods is prevalent because they are capable of separating complex follow-up tasks by leveraging the discriminative information of pre-order tasks. To illustrate, a discriminating and perplexing charges model has been developed by incorporating relevant essential attributes, such as the Few-Shot charge prediction model[12] [59]). Similarly, another approach for distinguishing confusing law articles in the civil law system is the Law Article Distillation based Attention Network, LADAN[13] [33].

In summary, existing MTL works in LJP domain, in which multiple LJP tasks can be trained in a single neural network framework, demonstrate better performance compared to single-task learning as they utilize relevant information sharing. However, these existing works have limitations in terms of legal knowledge diversity such as incorporating judger's values. Therefore, for integrating much more legal knowledge into the existing MTL works, future efforts should be directed towards two areas: (1) **MTL algorithms**. As appropriate MTL algorithms can mitigate negative transfers, researchers can focus more on MTL architecture and optimization. (2) **Task diversity**. As diversified tasks might benefit the MTL system through implicit data augmentation, attention focusing and feature eavesdropping, researchers can focus more on leveraging data and knowledge from multiple tasks with an appropriate task aggregation size.

B. PRE-TRAINED LANGUAGE MODEL

Transformer-based [18] PLMs, such as BERT [19] and its variants ([20], [71], [72], [73], [74], [87], [88], [89], [90], [91], [92]), have achieved state-of-the-art results in several downstream NLP tasks on generic benchmark datasets. Table 5 displays several such approaches for applying BERT-based models in legal domain pretraining to explore state-of-the-art performance in downstream legal tasks.

For example, LEGAL-BERT[14] [70] is a novel family of BERT models that leverage 12 GB of English legal training corpora. Two versions are included in BERT models, namely LEGAL-BERT-FP (adapting standard BERT by additional pretraining on legal domain corpora) and LEGAL-BERT-SC (pretraining BERT from scratch on legal domain corpora). The LEGAL-BERT-SC model is also used in CaseLaw-BERT [75], which employs a case law corpus and custom domain-specific vocabulary. LegalDB, on the other hand, is a DistillBERT-based model that is pre-trained by English legal-specific training corpora. Lawformer [55] is a Longformer-based model that is pre-trained on large-scale Chinese legal long case documents, while JuriBERT [82] is a set of BERT models that uses LEGAL-BERT-SC as a pre-training model on French legal text datasets and adapts CamemBERT by additional pretraining on French legal text datasets.

In addition to developing domain-specific pre-trained language models, researchers have also sought to enhance the performance of PLMs on legal tasks that involve lengthy documents exceeding 512 tokens. One approach to address this challenge is to employ a hierarchical version of BERT called HIER-BERT [70], which combines BERT-BASE with a hierarchical attention network that enables bypassing BERT's length restriction. Another model is Lawformer [55], a Longformer-based pre-trained language model. It utilizes a combination of local sliding window attention and global task-motivated full attention to capture long-range dependencies in processing Chinese legal documents which contain thousands of tokens.

In summary, all these legal-domain PLMs, such as language-specific PLMs and long-document pre-training, outperform generic ones in the various document understanding tasks. However, the construction of large-scale legal corpora with multilanguage and legal knowledge is still challenging due to their confidential nature. Hence, for more effective application of PLMs in LJP, future efforts should be focused on: (1) **Language space**: As a limited number of languages covered by PLMs corpora, researchers can focus more on exploring multilingual, cross-lingual or monolingual PLMs. (2) **Task diversity**: As a lack of natural language generation tasks, generative legal PLMs may be the hot issues in the future. (3) **Parameter optimization**: As exiting PLMs in legal domain require large-scale parameters to handle the large-scale unlabeled corpus, researchers can focus more on task-specific subspaces of PLMs or pruning PLMs. (4) **Calibration**: As the high-stake feature of legal domain, researchers can focus more on the study of the calibration of PLMs.

C. INTERPRETABLE-PERSPECTIVE LEARNING FRAMEWORK

The concept of interpretability in LJP, which refers to the ability of LJP system to explain their prediction, has

[12]https://github.com/thunlp/attribute_charge

[13]https://github.com/prometheusXN/LADAN

[14]https://huggingface.co/nlpaueb

gained significant attention in academia and the legal industry. The lack of interpretability hinders the acceptance of machine-generated judgment results. The interpretability concept [93] has been divided into two categories: introspection explanation and justification explanation. Introspection explanation focuses on explaining how a model arrives at its final output, while justification explanation provides sentences that describe how the evidence is consistent with the system output.

Reinforcement learning method is frequently employed for introspection explanations in LJP. It is used to extract rationales from input fact descriptions that serve as the introspection explanation for charge prediction [56]. For instance, QAjudge[15] [60] is used to interpret judgments based on reinforcement learning. This method involves selecting questions from a given set using a question net, answering questions according to the fact description using an answer net, and generating judgment results based on the answers using a predicted net. Reward functions are designed to minimize the number of questions asked.

In addition, multi-task learning methods mentioned in Section V-A and correspondence-based methods among relative tasks can also be used for introspection explanations in LJP. For example, a fine-grained fact-article correspondence method is proposed for recommending relevant law articles to a given legal case [21], since the existing recommended articles do not provide specific information about the facts to which they are relevant. Charge-based prison term prediction (CPTP) [58] has been proposed to make the total prison term prediction more interpretable based on fine-grained charge-prison term correspondence feature selection and aggregation.

From the justification explanation aspect, Court Views [34] has been considered the explanation for the prediction of charges.

In summary, existing works mainly rely on the input elements or outputs of related tasks as explanations for the final results, and have made a series of progress. However, they are not capable of providing thorough understandable explanations due to the black-box nature of state-of-the-art technologies like PLMs. Therefore, to advance the intelligibility of LJP works for humans, there are two directions which need future effort: (1) **Interpretable features**. As model input data and step-by-step outputs are key factors leading to models performance degradation over time, researchers can focus more on data drift detection and explanations of intermediate steps. (2) **Interpretable model**. As a correspondence between model components and data features might enhance the interpretability of black-box models, researchers can focus more on probing the link between models and features.

[15]https://github.com/thunlp/QAjudge

D. FEW-SHOT LEARNING FRAMEWORK

Few-shot learning has garnered significant attention in recent times, as current works predominantly focus on high-frequency judgment results rather than few-shot judgment results to ensure substantial training data.

Several few-shot learning methods have been proposed to predict few-shot judgment results.

For example, discriminative attributes of charges have been leveraged in one such approach [59] to provide additional information for few-shot charges. Another method, known as the Sequence Enhanced Capsule (SECaps) model [94], is based on the focal loss to predict few-shot charges. Moreover, an attentional and counterfactual-based natural language generation (AC-NLG) method [35] has been proposed, wherein the counterfactual decoder is employed to address the imbalance problem in judgments.

In summary, due to limited training data for low-frequency label cases, LJP works often neglect few-shot cases and focus on common label cases. However, with the impressive performance and rich information of few-shot learning methods based on PLMs in open-domain settings, there is potential to improve few-shot performance in LJP tasks. Future efforts should consider: (1) **Data diversity**. As few-shot performance can be sensitive to the data diversity, researchers can focus more on the selection of few-shot instances. (2) **Task diversity**. As diversified tasks might benefit the few-shot learning performance on new tasks through implicit data augmentation, attention focusing and feature eavesdropping, researchers can focus more on leveraging data and knowledge from multiple tasks with an appropriate task aggregation size.

VI. RESULTS & OBSERVATIONS

This section presents a comparative analysis of empirical results obtained using various NLP models on LJP datasets. Specifically, we focus on datasets obtained from 10 different sources, including the ECHR, SPCC, DCOUP, FSCCs, FSC, UKC, SCOTUS, SCI, TSC and FSCS. While we introduced 43 LJP datasets and 8 unlabeled pre-trained corpora, we limit our analysis to 11 datasets from different source Court cases to examine experimental results more closely. The remaining datasets have either limited size or a smaller number of experiments compared to the 11 datasets of interest.

In order to have a better insight of these experiment results, as shown in Table 7, Table 8, Table 9, Table 12, Table 14 and Table 15, we have classified the existing LJP models into four categories. The first is domain-independent supervised model, which is trained with labelled dataset and applicable in a domain with generic terminology. The second is domain-independent unsupervised model, which is trained with unlabeled dataset and applicable in a domain with generic terminology. The third is domain-dependent supervised model, trained with labelled dataset and applicable in a domain with highly specific terminology. The fourth

TABLE 7. The current results for ECHR-CASES.

Method	Domain	Label	Articles	
			Binary	Multi-label
			MaF	MaF
BOW-SVM	Open	Supervised	70.9±0.0	50.4±0.0
BIGRU-ATT			78.9±1.9	56.2±1.3
HAN			80.2±2.7	59.9±0.5
HIER-BERT		Unsupervised	**80.1±1.1**	**60.0±1.3**
LEGAL-BERT-FP 100k ALL LEGAL	Law		**88.3**	63.9
LEGAL-BERT-FP 500k ALL LEGAL			88.0	60.3
LEGAL-BERT-FP 100k SUB-DOMAIN			87.9	60.5
LEGAL-BERT-FP 500k SUB-DOMAIN			88.0	**65.2**

TABLE 8. The current results for CAIL-Long dataset.

Method	Domain	Label	Criminal					Civil	
			Charges		Articles		Prison terms	Articles	
			MiF	MaF	MiF	MaF	LD	MiF	MaF
BERT	Open	Unsupervised	94.800	68.200	81.500	52.900	1.286	61.700	31.600
RoBERTa			94.700	69.300	81.100	53.500	1.291	60.200	29.900
L-RoBERTa	Law		94.900	70.800	81.100	53.400	1.280	61.200	31.300
Lawformer			**95.400**	**72.100**	**82.000**	**54.300**	**1.264**	**63.000**	**33.00**

TABLE 9. The current results for CAIL-2018 dataset.

Method	Domain	Label	ACC			MaP			MaR		
			Charges	Articles	Prison terms	Charges	Articles	Prison terms	Charges	Articles	Prison terms
FLA+MTL	Open	Supervised	92.76	93.23	57.63	76.35	72.78	48.93	68.48	64.30	45.00
CNN+MTL			95.74	95.84	55.43	86.49	83.20	45.13	79.00	75.31	38.85
HARNN+MTL			95.58	95.63	57.38	85.59	81.48	43.50	79.55	74.57	40.79
Few-Shot+MTL			96.04	96.12	57.84	88.30	85.43	47.27	80.46	80.07	42.55
TOPJUDGE			95.78	95.85	57.34	86.46	84.84	47.32	78.51	74.53	42.77
MPBFN-WCA			95.98	96.06	58.14	**89.16**	85.25	45.86	79.73	74.82	39.07
LADAN+MTL			**96.45**	96.57	59.66	88.51	86.22	**51.78**	**83.73**	**80.78**	45.34
LADAN+TOPJUDGE			96.39	**96.62**	59.70	88.49	**86.53**	51.06	82.28	79.08	45.46
LADAN+MPBFN			96.42	96.60	**59.85**	88.45	86.42	51.75	83.08	80.37	**45.59**

is domain-dependent unsupervised model, which is trained with unlabeled dataset and applicable in a domain with highly specific terminology.

A. ECHR CASES

Table 7 presents the evaluation outcomes of recent studies on ECHR-CASES. BOW-SVM is a frequently used classification baseline model using Support Vector Machine (SVM) featured by the bag of words. BIGRU-ATT is the standard sequence model Bi-directional Gated Recurrent Units (Bi-GRU, [95]) with attention, and Hierarchical Attention Network (HAN, [96]) is also a sequential model. HIER-BERT is a Hierarchical Transformer model that can bypass BERT's length limitation, and LEGAL-BERT-FP 100k/500k ALL LEGAL/SUB-DOMAIN are transformer models with running additional pre-training steps (e.g., up to 100k or 500k) of BERT-base on legal-domain corpora (such as all legal corpora or just ECHR-CASES). As the tendency, LEGAL-BERT-FP 100k ALL LEGAL widely outperforms the other methods in binary violation classification, and LEGAL-BERT-FP 500k SUB-DOMAIN and LEGAL-BERT- FP 100k ALL LEGAL show excellent performance. These results suggest that BERT can be adapted to a new domain by further pre-training.

B. SPCC CASES

1) CAIL-LONG DATASET

Table 8 presents the evaluation outcomes of recent studies on CAIL-Long. The PLMs BERT [19] and RoBERTa [97], [98] are widely employed for text classification tasks in the generic domain with length limitations though. To address this issue, L-RoBERTa and Lawformer were both pre-trained by the same legal corpus. Lawformer is a Longformer-based PLM designed for processing long legal documents. Notably, the Lawformer model, which is capable of capturing long-distance dependencies, outperforms other models in criminal and civil cases. This suggests that Lawformer is a suitable pre-trained model for finetuning legal documents in Chinese that exceed 512 tokens.

2) CAIL-2018 DATASET

Table 9 presents the evaluation results of recent studies conducted on CAIL2018. The adopted methods have been explained in detail in Section V. The LADAN approach, which automatically extracts discriminative features from fact descriptions of confusing articles, demonstrates superior accuracy (Acc), macro-precision (MaP), and macro-recall (MaR). The MPBFN-WCA method, which uses bi-directional dependencies among LJP subtasks other

TABLE 10. The current results for DCOUP cases.

Method	Domain	Label	District-wise		All Districts	
			Acc	MiF	Acc	MiF
Doc2Vec+SVM	Open	Supervised	0.72	0.69	0.79	0.77
Doc2Vec+XGBoost			0.68	0.59	0.67	0.57
IndicBert-First512		Unsupervised	0.65	0.62	0.73	0.71
IndicBert-Last512			0.62	0.60	0.78	0.76
TFIDF+IndicBert			0.76	0.74	**0.82**	**0.81**
TextRank+IndicBert			0.76	0.74	**0.82**	**0.81**
Salience pred.+IndicBert		Supervised	0.76	0.74	0.80	0.78
Multi-Task		Unsupervised	**0.78**	**0.77**	0.80	0.78

TABLE 11. The current results for FSCCs cases.

Method	Domain	Label	MCC
Human experts	Law	-	0.1253
ULMFiT forward	Open	Unsupervised	0.3238
ULMFiT backward			0.3544
ULMFiT bidirectional			**0.3688**
BERT+LSTM			0.3127
Big Bird			0.2649

than forwarding dependencies, outperforms the TOPJUDGE model. However, both MPBFN-WCA and TOPJUDGE methods performs poorly in predicting few-shot charges, as reflected in the MP, MR, and F1 metrics. Few-Shot approach shows comparable performance with LADAN for charge prediction, but was less accurate in predicting prison term due to the limitation of predefined attributes.

C. DCOUP CASES

Table 10 presents the evaluation results of recent studies on DCOUP cases. To test the generalization of methods in these studies, HLDC dataset is divided in two settings: district-wise and all districts. For the first setting, case documents to training, validation or testing dataset are from different districts each other. The all districts setting is trained and tested on documents from all districts. Furthermore, Doc2Vec and IndicBERT are classical embedding-based model and transformer-based contextualized embedding model, respectively. Besides, all the TFIDF, TextRank and Salience Pred. are summarization methods to extract top 50% sentences based on sentences scores. Note that Salience Pred. is a supervised approach based on the cosine similarity with judge's summary and the sentence of facts but has a relatively poor performance in these studies. As can be observed, the performance of methods in district-wise dataset setting is generally lower than all-districts one. The model IndicBert is better at summarization than truncation for long case documents.

D. FSCCs CASES

Table 11 presents the evaluation results of recent studies on FSCCs cases. Note that experts' analysis data is different from other models in the study but having highly same proportion of appeals affirming vs reversing. In the study, ULMFiT model is trained by reading the Portuguese Wikipedia text from left to right (dubbed ULMFiT forward), from the right to left (dubbed ULMFiT backward) or both ways on the same line (dubbed ULMFiT bidirectional). After that, ULMFiT models is fine-tuned based on the 3,128,292 first instance court decisions. Finally, a binary classifier is trained by ULMFiT fine-tuning models. BERT+LSTM in these studies is a pipeline architecture of Portuguese BERT, unidirectional LSTM and a linear layer. And Big Bird in the study is Big Bird transformer model trained based on both the Portuguese Wikipedia and the BrWaC dataset frome scratch, then further-trained with Portuguese text for 142,800 steps. After that, Big Bird is fine-tuned based on 3,128,292 first instance court decisions for one epoch and passed it to the classification layer. As can be observed, ULMFiT bidirection method achieves best MCC score, which indicates there is still room for using RNNs in LJP, and all models outperforms expert on MCC score.

E. FSC, UKC, SCOTUS, AND SCI CASES

Table 12 presents the evaluation results of recent studies on FSC, UKC, SCOTUS, and SCI cases. There are various classical feature-based machine learning models utilizing word/sentence embedding and TFIDF, including Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (kNN), Naive Bayes, Perceptron, multi-layer perceptron (MLP), calibrated classifier (Calib. Classifier), and Random Forest. Moreover, transformer models RoBERTa and XLNet are also included. The experimental results show that transformer-based models, such as XLNet with BiGRU, outperform classical and sequential models on the ILDC dataset. On UKC cases, the best performing classical model is TFIDF with LR, whereas kNN outperforms the other classical models on SCOTUS cases. Furthermore, classical models perform excellently on FSC cases. These results suggest that transformer-based sequence models have the potential to enhance the performance of classical models. In addition, Table 13 illustrates the exact matching results of machine-extracted fact snippets and of annotated references by experts on the ILDC dataset, highlighting the importance of developing an efficient and explainable model for the LJP task to guide future research.

F. FSCS CASES

Table 14 presents the evaluation results of recent studies on FSCS cases. The baseline classifiers include Majority,

TABLE 12. The current results for FSC, UKC, SCOTUS and SCI cases.

Dataset	Source	Method	Domain	Label	Plea judgments			
					Acc	MaF	MaP	MaR
Sulea et al. [14]	FSC	**SVM**	Open	Supervised	**96.90**	**97.00**	**97.10**	**96.90**
BStricks_LDC [43]	UKC	**TFIDF+LR**			**69.05**	**69.02**	**69.05**	**69.02**
		LDA+kNN			57.76	57.63	58.01	57.88
		Word2Vec+RF			64.21	64.17	64.17	64.18
		Count+RF			66.13	66.12	66.12	66.13
JUSTICE [28]	SCOTUS	Perceptron			65.00	65.00	65.00	65.00
		SVM			60.00	60.00	60.00	60.00
		LR			61.00	61.00	61.00	61.00
		Naive Bayes			59.00	59.00	59.00	59.00
		MLP			64.00	64.00	64.00	64.00
		kNN			**68.00**	**67.00**	**69.00**	**68.00**
		Clalib. Classifier			62.00	62.00	63.00	62.00
ILDC [44]	SCI	Doc2Vec+LR			60.91	62.00	63.03	61.00
		GloVe+BiGRU+att.			60.75	64.35	68.26	60.87
		RoBERTa		Unsupervised	71.26	71.77	72.25	71.31
		XLNet+BiGRU			**77.78**	**77.79**	**77.80**	**77.78**
		XLNet+BiGRU+att.			77.01±0.52	77.07±0.01	77.32	76.82

TABLE 13. The evaluation results for Fact snippets vs experts in ILDC dataset.

Expert	Fact snippets vs Experts					
	JS	R-1	R-2	R-L	BLEU	Meteor
Expert 1	**0.333**	0.444	**0.303**	0.439	0.160	0.220
Expert 2	0.317	**0.517**	0.295	0.407	**0.280**	**0.300**
Expert 3	0.328	0.401	0.296	0.423	0.099	0.180
Expert 4	0.324	0.391	0.297	**0.444**	0.093	0.177
Expert 5	0.318	0.501	0.294	0.407	0.248	0.279

Stratified, and Linear (BoW), where Majority selects the majority class, Stratified randomly predicts judgment labels, and Linear (BoW) is a linear classifier based on TFIDF features. Long BERT and Hierarchical BERT are BERT-based models that address the limited sequence length issue of the standard BERT model. Long BERT introduces additional positional embedding, while Hierarchical BERT separates and encodes input tokens with a standard BERT encoder and aggregates all segment encodings with an additional Bidirectional Long Short-Term Memory (BiLSTM) to form the classification representation. The results indicate that Majority method performs best among the baseline and BERT-based models in MiF, while both long BERT and hierarchical BERT models (German and French with 20K+ training samples) outperform the Italian ones with 3K training samples in MaF, considering the error margin. These results highlight the impact of imbalanced class distribution, dataset scale, and the type of BERT model on model performance.

G. TSC CASES

Table 15 presents the evaluation results of recent studies on TSC cases. SVM and Naive Bayesare are two commonly-used non-neural classification baseline models. As the tendency, the sequence model BiGRU with attention outperforms non-neural models on the TSCC dataset.

In conclusion, the experiments conducted on ECHR-CASES, CAIL-Long, ILDC, and FSCS cases reveal that further pre-trained BERT-based models trained by legal corpora have the potential to predict judgments better than classical/sequence models or those based on BERT models trained by generic corpora. Moreover, the experiments on CAIL-2018 demonstrate that the combination of two MTL methods outperforms other models in solving multiple tasks simultaneously. However, the evaluation results on the ILDC dataset suggest that interpretable learning for the LJP task is still challenging. In addition, the experiments on FSCS cases reveal that BERT-based models without using few-shot learning on imbalanced label distributions can underperform those trained by balanced label distributions.

VII. DISCUSSIONS & RECOMMENDATIONS

This section provides an analysis of various datasets and solutions for LJP, and presents our findings based on the analysis.

- Diversified benchmark datasets. Although there are existing public LJP datasets available in 9 languages from 12 countries, it is worth noting that, according to Wikipedia, there are 36 official languages spoken across 211 countries worldwide.
- More realistic applications. As the performance of LJP systems continues to improve, it becomes increasingly important to develop datasets that capture more nuanced and complex aspects of real court judgments. While several large-scale judge-summarized fact description-based datasets, such as CAIL2018 and CAIL-Long in Table 2, have been built, they may not adequately represent the case logic or ensure prediction correctness due to their failure to consider the admissibility of

TABLE 14. The current results for FSCS cases.

Method	Domain	Label	Plea judgments					
			de		fr		it	
			MiF	MaF	MiF	MaF	MiF	MaF
Majority	Open	Supervised	**80.3**	44.5	**81.5**	44.9	**81.3**	44.8
Stratified			66.7±0.3	50.0±0.4	66.3±0.2	50.0±0.4	69.9±1.8	48.8±2.4
Linear(BoW)			65.4±0.2	52.6±0.1	71.2±0.1	56.6±0.2	67.4±0.5	53.9±0.6
Native BERT		Unsupervised	74.0±4.0	63.7±1.7	74.7±1.8	58.6±0.9	76.1±3.7	55.2±3.7
Multilingual BERT			68.4±5.1	58.2±4.8	71.3±4.3	55.0±0.8	77.6±2.4	53.0±1.1
long Native BERT			76.5±3.7	67.9±1.8	77.2±3.4	68.0±1.8	77.1±3.9	**59.8±4.6**
long Multilingual BERT			75.9±1.6	66.5±0.8	73.3±1.9	64.3±1.5	76.0±2.6	58.4±3.5
hierarchical Native BERT			77.1±3.7	**68.5±1.6**	80.2±2.0	**70.2±1.1**	75.8±3.5	57.1±6.1
hierarchical Multilingual BERT			76.8±3.2	57.1±0.8	76.3±4.1	67.2±2.9	72.4±16.6	55.5±9.5

TABLE 15. The current results for TSC cases.

Method	Domain	Label	Plea judgments	
			MiF	MaF
Naive Bayes	Open	Supervised	57.32	57.28
SVM			64.23	60.45
BiGRU			58.54	57.99
BiGRU+self-att.			59.76	59.11
BiGRU+att.			**66.68**	**63.35**

evidence and facts in a real courtroom setting. The only two known LJP datasets derived from actual courtrooms are LJP-MSJudge [62] and USClassActions [42]. Furthermore, evaluating the fairness of LJP results and their consistency with the actual legitimate intent from real courtrooms would be a proactive issue for the future, particularly in cases where proof rules are inadequate or the text is ambiguous. Hence, we propose the recommendations for new datasets stating the facts summary given by appeals as input, especially when the rules of proof are inadequate and the text of evidence is doubt or ambiguous.

- Adaptive interpretability. Table 12 demonstrates that on the ILDC dataset, XLNet with BiGRU, the pre-trained language model, achieves superior performance over classical and sequential models [44]. However, it remains a challenge to explain the black-box work in LJP tasks given the uniqueness of the logic used in practical judgments for each legal domain.
- Complex legal logical reasoning. A comprehensive legal judgment predictor should consider the case from various perspectives [62]. As demonstrated in Table 9, enhancing the prediction accuracy of prison terms is an urgent issue, despite the accuracy of predicting articles and charges being over 90%. It is essential to investigate methods for integrating legal knowledge from diverse perspectives and enhancing machine reasoning capabilities in NLP.
- Comparative performance of LJP models. With the exception of pre-trained language models (PLMs), the majority of LJP models are categorized as domain-independent supervised models, including SVM, LR, kNN, Naive Bayes, Perceptron, MLP, Calib. Classifier, RF and BiGRU. As indicated in Table 7, Table 12, and Table 14, PLMs that use large-scale unlabeled data in an open domain and are categorized as domain-dependent unsupervised models, such as HIER-BERT, RoBERTa, XLNet, and Longformer, have outperformed numerous domain-independent supervised models. Additionally, for the legal domain, PLMs that use large-scale unlabeled data in the legal domain and are categorized as domain-dependent unsupervised models, including LEGAL-BERT, L-RoBERTa, and Lawformer, have outperformed domain-independent unsupervised models, as demonstrated in Table 7 and Table 8.

VIII. CONCLUSION

In this survey, we commence by presenting a classification method of LJP tasks based on the difference of two mainstream legal systems. Subsequently, we provide a comprehensive comparison and analysis of 43 LJP datasets in 9 languages and 8 legal-domain PLMs in 4 languages. Furthermore, we categorize the LJP models into 4 different types using 16 evaluation metrics. Besides, we highlight 4 major directions for LJP models and latest experimental results of 11 representative datasets from different court cases. Drawing on our observations, we offer recommendations for future datasets and tasks, as well as the following suggestions for future models: investigating the link between data features and model components, utilizing multiple tasks or data samples appropriately, and expanding the language diversity of datasets or PLMs corpora.

REFERENCES

[1] R. C. Lawlor, "What computers can do: Analysis and prediction of judicial decisions," *Amer. Bar Assoc. J.*, vol. 49, no. 4, pp. 337–344, Apr. 1963.
[2] R. A. Oppel Jr. and J. K. Patel, "One lawyer, 194 felony cases, and no time," New York Times, New York, NY, USA, Tech. Rep., 2019.
[3] L. Garcia-Navarro and P. Moura, "Brazil: The land of many lawyers and very slow justice," Tech. Rep., 2014.
[4] H. C. N. J. Data, "National judicial data grid (district and Taluka courts of India)," 2023.
[5] *The Justice Gap: Measuring the Unmet Civil Legal Needs of Low-Income Americans*, LS Corp., South Korea, 2017.

[6] S. Nagel, "Weighting variables in judicial prediction," *MULL, Mod. Uses Log. Law*, vol. 2, no. 3, pp. 93–97, 1960.

[7] S. S. Ulmer, "Quantitative analysis of judicial processes: Some practical and theoretical applications," *Law Contemp. Problems*, vol. 28, no. 1, pp. 164–184, 1963.

[8] R. Keown, "Mathematical models for legal prediction," *Computer/LJ*, vol. 2, no. 1, p. 829, 1980.

[9] F. Kort, "Predicting supreme court decisions mathematically: A quantitative analysis of the 'Right to Counsel' cases," *Amer. Political Sci. Rev.*, vol. 51, no. 1, pp. 1–12, Mar. 1957.

[10] F. Kort, "The quantitative content analysis of judicial opinions," *Political Res., Org. Design*, vol. 3, no. 7, pp. 11–14, Mar. 1960.

[11] C.-L. Liu and C.-D. Hsieh, "Exploring phrase-based classification of judicial documents for criminal charges in Chinese," in *Proc. Int. Symp. Methodologies Intell. Syst.* Cham, Switzerland: Springer, 2006, pp. 681–690.

[12] D. M. Katz, "Quantitative legal prediction—Or—How I learned to stop worrying and start preparing for the data-driven future of the legal services industry," *Emory Law J.*, vol. 62, p. 909, Dec. 2012.

[13] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective," *PeerJ Comput. Sci.*, vol. 2, p. e93, Oct. 2016.

[14] O.-M. Sulea, M. Zampieri, M. Vela, and J. van Genabith, "Predicting the law area and decisions of French supreme court cases," 2017, *arXiv:1708.01681*.

[15] M. M. S. Fareed, A. Raza, N. Zhao, A. Tariq, F. Younas, G. Ahmed, S. Ullah, S. F. Jillani, I. Abbas, and M. Aslam, "Predicting divorce prospect using ensemble learning: Support vector machine, linear model, and neural network," *Comput. Intell. Neurosci.*, vol. 2022, Jul. 2022, Art. no. 3687598.

[16] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Appl. Sci.*, vol. 12, no. 13, p. 6424, 2022.

[17] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," 2017, *arXiv:1707.09168*.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[20] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[21] J. Ge, Y. Huang, X. Shen, C. Li, and W. Hu, "Learning fine-grained fact-article correspondence in legal cases," 2021, *arXiv:2104.10726*.

[22] Y. Huang, X. Shen, C. Li, J. Ge, and B. Luo, "Dependency learning for legal judgment prediction with a unified text-to-text transformer," 2021, *arXiv:2112.06370*.

[23] P. Kalamkar, J. Venugopalan, and V. Raghavan, "Indian legal NLP benchmarks : A survey," 2021, *arXiv:2107.06056*.

[24] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does NLP benefit legal system: A summary of legal artificial intelligence," 2020, *arXiv:2004.12158*.

[25] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "CAIL2018: A large-scale legal dataset for judgment prediction," 2018, *arXiv:1807.02478*.

[26] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," 2019, *arXiv:1906.02059*.

[27] J. Niklaus, I. Chalkidis, and M. Stürmer, "Swiss-Judgment-prediction: A multilingual legal judgment prediction benchmark," 2021, *arXiv:2110.00806*.

[28] M. Alali, S. Syed, M. Alsayed, S. Patel, and H. Bodala, "JUSTICE: A benchmark dataset for Supreme Court's judgment prediction," 2021, *arXiv:2112.03414*.

[29] I. Chalkidis, A. Jana, D. Hartung, M. J. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," Tech. Rep., SSRN 3936759, 2021.

[30] H. Zhong, Z. Guo, C. Tu, C. Xiao, Z. Liu, and M. Sun, "Legal judgment prediction via topological learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, Oct./Nov. 2018, pp. 3540–3549. [Online]. Available: https://aclanthology.org/D18-1390

[31] D. Wei and L. Lin, "An external knowledge enhanced multi-label charge prediction approach with label number learning," 2019, *arXiv:1907.02205*.

[32] W. Yang, W. Jia, X. Zhou, and Y. Luo, "Legal judgment prediction via multi-perspective bi-feedback network," 2019, *arXiv:1905.03969*.

[33] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," 2020, *arXiv:2004.02557*.

[34] H. Ye, X. Jiang, Z. Luo, and W. Chao, "Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions," 2018, *arXiv:1802.08504*.

[35] Y. Wu, K. Kuang, Y. Zhang, X. Liu, C. Sun, J. Xiao, Y. Zhuang, L. Si, and F. Wu, "De-biased court's view generation with causality," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Association for Computational Linguistics, Nov. 2020, pp. 763–780. [Online]. Available: https://aclanthology.org/2020.emnlp-main.56

[36] P. G. Stein, "Roman law, common law, and civil law," *Tul. L. Rev.*, vol. 66, p. 1591, Feb. 1991.

[37] K. Ligeti, "The place of the prosecutor in common law and civil law jurisdictions," *The Oxford Handbook of Criminal Process*. 2019.

[38] M. Galanter, "The vanishing trial: An examination of trials and related matters in federal and state courts," *J. Empirical Legal Stud.*, vol. 1, no. 3, pp. 459–570, Nov. 2004.

[39] B. McKillop, "The position of accused persons under the common law system in Australia (more particularly in New South Wales) and the civil law system in France," *Univ. New South Wales Law J.*, vol. 26, no. 2, pp. 515–539, 2003.

[40] J. Dainow, "The civil law and the common law: Some points of comparison," *Amer. J. Comparative Law*, vol. 15, no. 3, p. 419, 1966.

[41] I. Almuslim and D. Inkpen, "Legal judgment prediction for Canadian appeal cases," in *Proc. 7th Int. Conf. Data Sci. Mach. Learn. Appl. (CDMA)*, Mar. 2022, pp. 163–168.

[42] G. Semo, D. Bernsohn, B. Hagag, G. Hayat, and J. Niklaus, "ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US," 2022, *arXiv:2211.00582*.

[43] B. Strickson and B. De La Iglesia, "Legal judgement prediction for UK courts," in *Proc. 3rd Int. Conf. Inf. Sci. Syst.*, Mar. 2020, pp. 204–209.

[44] V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, S. K. Guha, A. Bhattacharya, and A. Modi, "ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation," 2021, *arXiv:2105.13562*.

[45] S. Paul, P. Goyal, and S. Ghosh, "Automatic charge identification from facts: A few sentence-level charge annotations is all you need," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1011–1022. [Online]. Available: https://aclanthology.org/2020.coling-main.88

[46] M.-Y. Kim, J. Rabelo, R. Goebel, M. Yoshioka, Y. Kano, and K. Satoh, "COLIEE 2022 summary: Methods for legal document retrieval and entailment," in *Proc. 16th Int. Workshop Juris-Inform.*, 2022, pp. 3–16.

[47] P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya, and P. Majumder, "FIRE 2019 AILA track: Artificial intelligence for legal assistance," in *Proc. 11th Forum Inf. Retr. Eval.*, Dec. 2019, pp. 4–6.

[48] M. Medvedeva, M. Vols, and M. Wieling, "Judicial decisions of the European court of human rights: Looking into the crystal ball," in *Proc. of Conf. Empirical Legal Stud.*, 2018, pp. 1–24.

[49] M. B. L. Virtucio, J. A. Aborot, J. K. C. Abonita, R. S. Avinante, R. J. B. Copino, M. P. Neverida, V. O. Osiana, E. C. Peramo, J. G. Syjuco, and G. B. A. Tan, "Predicting decisions of the Philippine Supreme Court using natural language processing and machine learning," in *Proc. IEEE 42nd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, vol. 2, Jul. 2018, pp. 130–135.

[50] R. D. Sharma, S. Mittal, S. Tripathi, and S. Acharya, "Using modern neural networks to predict the decisions of Supreme Court of the United States with state-of-the-art accuracy," in *Proc. 22nd Int. Conf. Neural Inf. Process. (ICONIP)*, 2015, pp. 475–483.

[51] D. M. Katz, M. J. Bommarito II, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the United States," *PLoS One*, vol. 12, no. 4, pp. 1–18, 2017.

[52] A. Kapoor, M. Dhawan, A. Goel, A. H. Arjun, A. Bhatnagar, V. Agrawal, A. Agrawal, A. Bhattacharya, P. Kumaraguru, and A. Modi, "HLDC: Hindi legal documents corpus," in *Proc. Findings Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3521–3536. [Online]. Available: https://aclanthology.org/2022.findings-acl.278

[53] K. Kowsrihawat, P. Vateekul, and P. Boonkwan, "Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism," in *Proc. 5th Asian Conf. Defense Technol. (ACDT)*, Oct. 2018, pp. 50–55.

[54] P. Wang, Z. Yang, S. Niu, Y. Zhang, and S. Z. Niu, "Modeling dynamic pairwise attention for crime classification over legal articles," in*Proc. 41st Int. ACM SIGIR Conf.*, 2018, pp. 485–494.

[55] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *AI Open*, vol. 2, pp. 79–84, Jan. 2021.

[56] X. Jiang, H. Ye, Z. Luo, W. Chao, and W. Ma, "Interpretable rationale augmented charge prediction system," in *Proc. 27th Int. Conf. Comput. Linguistics, Syst. Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, Aug. 2018, pp. 146–151. [Online]. Available: https://aclanthology.org/C18-2032

[57] S. Pan, T. Lu, N. Gu, H. Zhang, and C. Xu, "Charge prediction for multi-defendant cases with multi-scale attention," in *Computer Supported Cooperative Work and Social Computing* (Communications in Computer and Information Science). Singapore: Springer, 2019, ch. 59, pp. 766–777.

[58] H. Chen, D. Cai, W. Dai, Z. Dai, and Y. Ding, "Charge-based prison term prediction with deep gating network," 2019, *arXiv:1908.11521*.

[59] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 487–498. [Online]. Available: https://aclanthology.org/C18-1041

[60] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, and M. Sun, "Iteratively questioning and answering for interpretable legal judgment prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, 2020, pp. 1250–1257.

[61] S. Long, C. Tu, Z. Liu, and M. Sun, "Automatic judgment prediction via legal reading comprehension," in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Cham, Switzerland: Springer, 2019, pp. 558–572.

[62] L. Ma, Y. Zhang, T. Wang, X. Liu, and S. Zhang, "Legal judgment prediction with multi-stage case representation learning in the real court setting," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2021, pp. 993–1002.

[63] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, "Natural language processing in law: Prediction of outcomes in the higher courts of Turkey," *Inf. Process. Manage.*, vol. 58, no. 5, Sep. 2021, Art. no. 102684.

[64] L.-F. Andre, A.-C. Hector, S. Orivaldo, and O.-L. Livia, "Predicting Brazilian court decisions," *PeerJ Comput. Sci.*, vol. 8, p. e904, Mar. 2022.

[65] V. G. F. Bertalan and E. E. S. Ruiz, "Predicting judicial outcomes in the Brazilian legal system using textual features," in *Proc. Workshop Digital Humanities Natural Lang. Process. Int. Conf. Comput. Process. Portuguese (DHandNLP and PROPOR)*, 2022, pp. 22–32.

[66] E. J. D. Menezes-Neto and M. B. M. Clementino, "Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts," *PLoS One*, vol. 17, no. 7, pp. 1–20, 2022.

[67] R. Zev Mahari, "AutoLAW: Augmented legal reasoning through legal precedent prediction," 2021, *arXiv:2106.16034*.

[68] F. Dadgostari, M. Guim, P. A. Beling, M. A. Livermore, and D. N. Rockmore, "Modeling law search as prediction," *Artif. Intell. Law*, vol. 29, pp. 3–34, Mar. 2021.

[69] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma, "LeCaRD: A legal case retrieval dataset for Chinese law system," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2342–2348.

[70] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of law school," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, Association for Computational Linguistics, Nov. 2020, pp. 2898–2904. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.261

[71] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[72] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.

[73] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[74] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big Bird: Transformers for longer sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17283–17297.

[75] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset," 2021, *arXiv:2104.08671*.

[76] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, and P. Malakasiotis, "Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases," 2021, *arXiv:2103.13084*.

[77] H. J. Spaeth, L. Epstein, J. A. Segal, A. D. Martin, and S. C. Ruger, J. Theodore, and Benesh, *Supreme Court Database*, document Version 2020, Release 1, Washington Univ. Law, Washington, DC, USA, 2020. [Online]. Available: http://supremecourtdatabase.org/

[78] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer," 2021, *arXiv:2109.00904*.

[79] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak, "LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts," in *Proc. 12th Lang. Resour. Eval. Conf. (LREC)*, European Language Resources Association, 2020, pp. 1228–1234.

[80] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, and P. Torroni, "CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service," *Artif. Intell. Law*, vol. 27, no. 2, pp. 117–139, Jun. 2019.

[81] P. Bambroo and A. Awasthi, "LegalDB: Long DistilBERT for legal document classification," in *Proc. Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Feb. 2021, pp. 1–4.

[82] S. Douka, H. Abdine, M. Vazirgiannis, R. El Hamdani, and D. R. Amariles, "JuriBERT: A masked-language model adaptation for French legal text," 2021, *arXiv:2110.01485*.

[83] N. Garneau, E. Gaumond, L. Lamontagne, and P.-L. Déziel, "Criminel-BART: A French Canadian legal language model specialized in criminal law," in *Proc. 18th Int. Conf. Artif. Intell. Law*, Jun. 2021, pp. 256–257.

[84] M. Al-qurishi, S. Alqaseemi, and R. Souissi, "AraLegal-BERT: A pretrained language model for Arabic legal text," in *Proc. Natural Legal Lang. Process. Workshop*, 2022, pp. 338–344.

[85] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.

[86] H. Zhong, C. Xiao, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, and J. Xu, "Overview of CAIL2018: Legal judgment prediction competition," 2018, *arXiv:1810.05851*.

[87] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[88] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[89] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[90] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[91] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "MT5: A massively multilingual pre-trained text-to-text transformer," 2020, *arXiv:2010.11934*.

[92] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020.

[93] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 3–19.

[94] C. He, L. Peng, Y. Le, and J. He, "SECaps: A sequence enhanced capsule model for charge prediction," in *Proc. Int. Conf. Artif. Neural Netw.*, 2019, pp. 227–239.
[95] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[96] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.
[97] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "ROBERTa: A robustly optimized BERT pretraining approach,"2019, *arXiv:1907.11692*.
[98] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," 2019, *arXiv:1906.08101*.

JUNYUN CUI received the Ph.D. degree from the State Key Laboratory of ISN, Xidian University, in 2012. From 2015 to 2017, she was a Senior Engineer with the Xi'an Microelectronics Technology Institute. In 2017, she joined the Department of Computer Science, Xi'an University of Finance and Economics. Her research interests include channel encoding, spatiotemporal information optimization, and NLP applications.

XIAOYU SHEN received the B.S. degree in software engineering from Nanjing University, Nanjing, China, and the Ph.D. degree in computer science from the Max Planck Institute for Informatics, Saarbrücken, Germany. He is currently a Scientist with Amazon Alexa AI. His research results have been authored or coauthored international journals and conferences, including ACL, EMNLP, AAAI, INTERSPEECH, and WWW. His research interests include variational inference, text generation, and question answering. He was a recipient of the Best Demo Paper Award from COLING 2020.

SHAOCHUN WEN received the bachelor's degree in English translation and interpreting from Beijing Foreign Studies University and the master's degree in cultural studies from Belarusian State University. She is currently a Lecturer with the Guangxi Transport Vocational and Technical College. Her research interests include applied linguistics and legal translation.