Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation

Xin Zhou*, Yating Zhang*
Alibaba Group
Hangzhou, Zhejiang, China
{eric.zx,ranran.zyt}@alibaba-inc.com

Xiaozhong Liu
Indiana University Bloomington
Bloomington, Indiana, USA
liu237@indiana.edu

Changlong Sun, Luo Si
Alibaba Group
Hangzhou, Zhejiang, China
changlong.scl@taobao.com
luo.si@alibaba-inc.com

ABSTRACT

Various e-commerce platforms produce millions of transactions per day with many transaction disputes. This generates the demand for effective and efficient dispute resolutions for e-commerce transactions. This paper proposes a novel research task of Legal Dispute Judgment (LDJ) prediction for e-commerce transactions, which connects two yet isolated domains, e-commerce data mining and legal intelligence. Different from traditional legal intelligence with the focus on textual evidence of the dispute itself, the new research utilizes multiview information such as past behavior information of seller and buyer as well as textual evidence of the current transaction. The multiview dispute representation is integrated into an innovative multi-task learning framework for predicting the legal result. An extensive set of experiments with a large dispute case dataset collected from a world leading e-commerce platform shows that the proposed model can more accurately characterize a dispute case through buyer, seller, and transaction viewpoints for legal judgment prediction against several alternatives.

CCS CONCEPTS

• **Information systems → Electronic commerce**; • **Computing methodologies → Knowledge representation and reasoning**; **Multi-task learning**; • **Applied computing → Law**;

KEYWORDS

Legal Intelligence, E-Commerce, Multiview Dispute Representation, Multi-task Learning

ACM Reference format:
Xin Zhou*, Yating Zhang*, Xiaozhong Liu, and Changlong Sun, Luo Si. 2019. Legal Intelligence for E-commerce: Multi-task Learning by Leveraging Multiview Dispute Representation. In *Proceedings of Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, July 21–25, 2019 (SIGIR '19),* 10 pages.
https://doi.org/10.1145/3331184.3331212

* These two authors contributed equally.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331212

1 INTRODUCTION

Over the past decades, with the rapid development of information technology, e-commerce has maintained a thriving growth and has become one of the world's most dynamic economic activities. While the major e-commerce platforms, e.g., Amazon, eBay and Alibaba, exemplified such success, by 2018, the market share of e-commerce, as a percentage of all global retail sales, increases to 11.9%. With billions of online transactions, inevitably, new types of disputes arise from the increased business interactions involving the new technology. For instance, according to the annual report of a world leading e-commerce platform, over 20 million disputes were submitted online in 2017, and, oftentimes, the high costs and delays found in dispute resolution and potential litigation. Buyers and sellers need efficient and low-cost mediation and arbitration services in an e-commerce ecosystem. From legal viewpoint, when users are not satisfied with the online dispute resolution (ODR) results, they need to make critical decisions on whether they should file perplexing and expensive lawsuits to protect their own interests. It is clear that most buyers cannot afford the cost of hiring professionals with legal expertise to assist them to win the cases.

In this context, predicting the judgment result of an e-commerce lawsuit case is however not trivial, which can provide the right/fair legal protections for customers and business owners. Behind such efforts, the biggest challenge lies in the difficulty of accurately representing the lawsuit cases which might vary from each other over semantics (e.g., the parties' names, the transaction process, the commodities involved) but share the similar legal logic reasoning in an e-commerce ecosystem. In addition, the relatively sparse historical data of lawsuit cases in e-commerce makes it even harder to automatically learn/represent the trial logic behind it.

Comparing with traditional legal process, ODR provides an important alternative to resolve a claim or dispute via e-commerce platform. ODR can be done entirely on the Internet in a way of low cost, high efficiency, and temporal-geographic flexibility to satisfy the demand of a large amount of cyber-disputes or e-conflicts generated everyday. If buyer and seller are satisfied with the result, they don't have to go to court and start the legal process. However, currently there are no uniform standards or a formal monitoring system for ODR methods. The decision of a dispute is made by empirical experiences. The merits of the ODR are lost if the case ends up being decided merely as a practical matter based on personal jurisdiction determinations.

By observing the shortcomings and challenges of the entire process from claiming an online dispute to potentially file a lawsuit, we are aware of the opportunities of seeking for the legal intelligence for e-commerce by jointly digging into both the large amount of

dispute data and the relatively small set of lawsuit data. For instance, in the e-commerce ecosystem, dispute cases can be judged not only by the fact of the current transaction, but also the historical transaction/dispute records of the target sellers and buyers. The intuition of employing user profiles comes from the fact that the transaction between buyer and seller is usually a "one-shot" deal which makes it difficult for the mediator of e-commerce platform to draw on anything outside of this problematic interaction. Therefore, the prior knowledge of the both parties according to their historical transactions and dispute records could help the mediator to reach the best negotiated agreement as dispute result [20]. Not only in ODR system, but also in justice system, the scholars have suggested that leveraging the more, better, and easier-to-use information from diverse resources, and by removing a litigant's appearance (race, gender, weight, etc.) from a judge's consideration, can render outcomes less subject to implicit biases [4]. Such findings motivate us to investigate the new problem of legal intelligence for e-commerce by leveraging multiview dispute representation learning and multi-task (ODR and judicial predictions) learning. In a joint learning process, various kinds of information, e.g., legal knowledge graph, seller/buyer information and transaction information, are encapsulated to regularize the potential determination of a dispute through the fine-tuning of the dispute representation.

To the best of our knowledge, this work is the first pioneer investigation of e-commerce legal intelligence, which can be important and essential to protect the legal fairness and interests of consumers and sellers in the e-commerce ecosystem. Unlike prior product classification and recommendation efforts, we propose an end-to-end model in a manner of multi-task learning process by jointly training the lawsuit judgment model with the dispute resolution model. Moreover, the proposed model can efficiently cope with the lawsuit data sparseness problem, while ODR task providing critical legal information in the joint learning process.

To sum up, our contributions are as follows:

(1) We propose a novel legal intelligence problem to address e-commerce dispute resolution.
(2) We propose an effective way to quantify a lawsuit case by leveraging multiview dispute representation.
(3) The proposed model is capable of digging the logic reasoning behind a lawsuit and explaining the cause-and-effect relationship between the fact of a lawsuit and its judgment.
(4) We evaluate our proposed model on more than 6,858 lawsuit cases trained with millions of dispute data. The experimental results demonstrate the proposed approach significantly improves the $Micro\_F_1$ and $Macro\_F_1$ scores of the judgment prediction of the lawsuit cases.
(5) To motivate other scholars to investigate this novel and important problem, we share the experiment dataset while removing the sensitive information. To the best of our knowledge, this is the first e-commerce dispute resolution dataset.

2 PRELIMINARY: TASKS

The goal of this study is to enable auto-legal lawsuit judgment prediction for e-commerce dispute cases. Comparing with ODR, this task can be more critical (e.g., high-cost) and challenging (e.g., training data sparseness). In this paper, we propose an innovative solution by leveraging multi-task learning, which is able to jointly learn four classification tasks in terms of ODR and legal intelligence. Practically, when a dispute case comes, it goes through the reasoning task at the dispute level that categorizes the reasons of a dispute and then forwards to the task of predicting the result of a dispute. While a case can potentially move to the lawsuit level, it will be first classified by the identified fact and then the model predicts the lawsuit judgment result based on the fact characterized in the prior steps. We handle the above four tasks sequentially by considering the logical hierarchies behind them (see Fig. 1), and hypothesizing that the reason of a dispute can be important to predict its dispute result. It also indicates the fact of the lawsuit case, while the identified fact determines the judgment result.

Main Task: Lawsuit Judgment Prediction. From legal viewpoint, the judgment is considered as a response to the plaintiff's claim. For instance, given the e-commence claims that the buyer requests (1) refund and (2) triple compensation, if the final court judgments are (1) the defendant should return the plaintiff the payment and (2) reject the second claim, then we assign this case with the judgment label `refund`. We define this process as a multi-label classification task which is to optimize a function that can accurately predict the actual value(s) of the labels $\mathbf{y}$ for a given case representation $\mathbf{x}$, where M is the total number of classes[1]. While the training data can be quite sparse for this task, by leveraging multi-task and multiview representation learning, critical information from subtasks will optimize the prediction outcomes. The details will be addressed in Sec. 5.

$$\mathbf{x} \rightarrow \mathbf{y} = (y_1, y_2, ..., y_M) \in \{0, 1\}^M \tag{1}$$

Subtask: Dispute Reason Prediction. As part of the ODR, this is a single-label but multi-class classification task where each dispute case is assigned with one reason out of a list of choices by the customer. Unlike the main task, ODR database provides enough training data for this task. It tries to learn a function that maps a dispute represented as vector $\mathbf{x}$ to a class label $k \in R^K$ where K is the total number of classes[2].

$$\mathbf{x} \rightarrow k, k \in R^K \tag{2}$$

Subtask: Dispute Result Prediction. Given the case itself and the dispute reason the customer selected, the e-commerce platform will notify the consumer with the result, i.e., only refund, return of goods & refund, or reject. Similar to the above task, it is also an ODR single-label and multi-class classification task[3].

Subtask: Lawsuit Fact Prediction. In the real-world e-commerce case, if the customer decides to file a lawsuit, the judges should recognize the legal fact after reviewing all the evidences[4] submitted by the parties including the transaction data, the negotiation record etc., as well as the defence during the trial among the parties. Due to the difficulty of data accessibility of all the evidences, in this work, we only utilize the dispute data which can be obtained from e-commerce platform (e.g., transaction data, dispute record etc.)

[1] M equals 7 in this task based on our data set.
[2] K equals 46 in this task based on our data set.
[3] The total number of classes equals 3.
[4] Evidence is the proof of fact(s) presented at a judicial hearing, which is the key element in convincing the judge or jury that the facts are the proper ones on which to base a final decision [9].

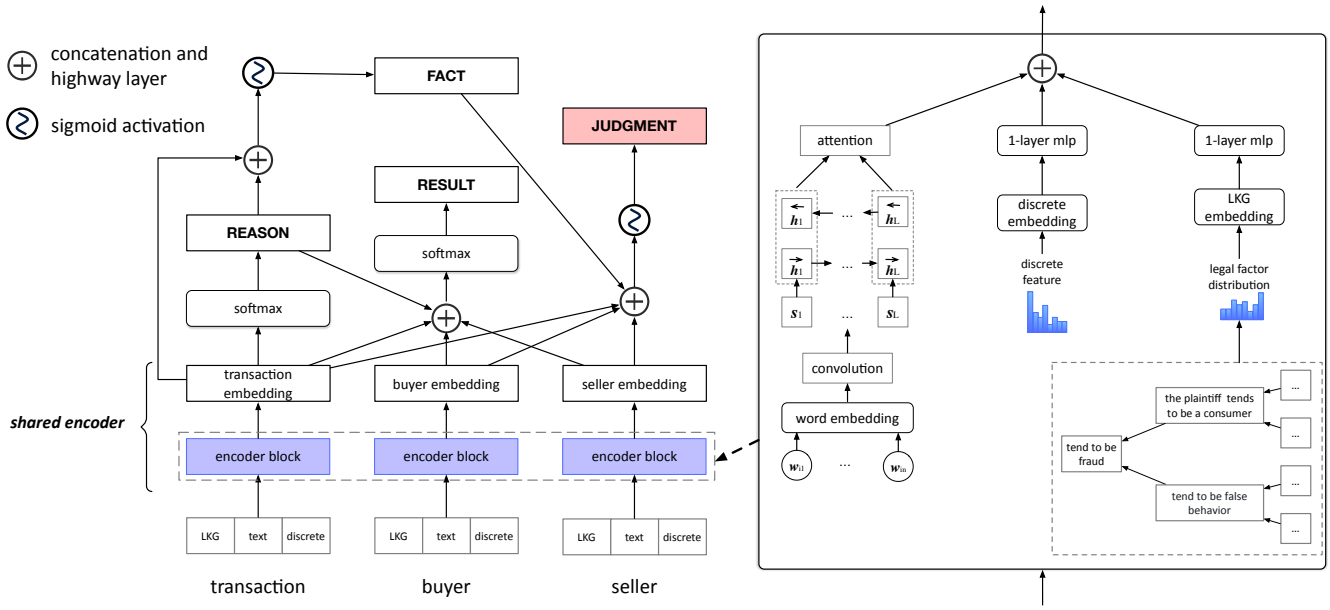

Figure 1: Network architecture of the proposed LDJ model.

to characterize such fact identification process. We formulate this process as a multi-label classification task[5], where each case can be assigned with more than one fact labels[6].

3 MULTIVIEW DISPUTE REPRESENTATION

3.1 Viewpoints

In an e-commerce ecosystem, a dispute case may associate with three types of information, namely, *transaction*, *buyer* and *seller*. From legal perspective, a case can be interpreted via the nodes and edges on the legal knowledge graph (*LKG*). As for the internal aspects, two types of features are involved: meta data and text data. Table 1 shows the features that we employ for the dispute representation from the above three e-commerce viewpoints. Additionally, as the legal representation, we also involve the legal expert knowledge which is represented as a knowledge graph. Figure 2 depicts one portion of the graph itself.

3.1.1 E-commerce Viewpoint. Intuitively, *transaction* information can tell the status of a purchase action in terms of the number and price of the goods; the information of the commodity including its negative reviews to reflect the quality of the commodity; the logistic status inferring the current ownership of the goods; online dispute record recording the dialogues of the current transaction among the buyer, seller and the platform administrator to reproduce the target conflicts.

From the perspective of *buyer*, we can indicate if the buyer is a normal consumer or a professional extortioner according to his/her historical purchase and dispute record. Similarly, the credit, reputation, and the dispute(s) a *seller* received can be regarded as the indicator of his/her credibility and the quality of the goods he/she has sold.

[5]The total number of classes equals 15.
[6]The process of fact mining will be described later in Sec. 5.

Table 1: The feature utilized for dispute representation from three e-commerce viewpoints.

Viewpoint	Feature	Type
Transaction	quantity	meta
	price	meta
	commodity's category	meta
	commodity's title	text
	dispute record	text
Buyer	buyer's credit	meta
	buyer's star	meta
	buyer's dispute numbers	meta
	buyer's dispute record	text
Seller	seller's credit	meta
	seller's star	meta
	seller's dispute numbers	meta
	seller's dispute record	text

Theoretically, the *transaction* data could reflect the reason why the customer raised a dispute, and, during the lawsuit, it is also regarded as an evidence to support the legal fact the judges identifies before giving the final judgment. Meanwhile, the user portrait through *buyer* and *seller* data is also an essential step for the personnel of customer service to reach a conclusion of the given dispute. For instance, if the target seller has experienced a number of similar dispute cases initialized by other buyers historically, e-commerce platform may question the integrity of the seller and has the higher chance to draw the conclusion favoring the buyer for the current case. Correspondingly, if this case further goes to the lawsuit procedure, *buyer* and *seller* data are usually treated as the important reference for the judges to build the party portrait together with the recognized legal fact to achieve the final judgment.

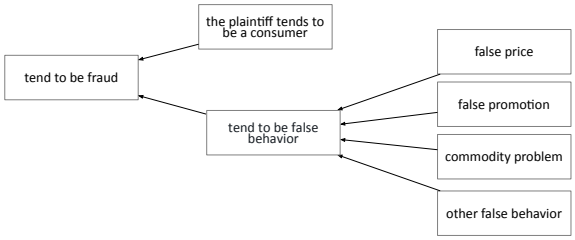

Figure 2: A portion of the Legal Knowledge Graph.

3.1.2 Legal Viewpoint. The *Legal Knowledge Graph*[7] (*LKG*) used in this work is generated by seven judges from three civil courts who have experience of dealing with the online transaction disputes (OTD). Basically, *LKG* is a directed acyclic graph (DAG) where we define the nodes without in-links as *source nodes* (c) and the nodes without out-links as *target nodes* (r). The *LKG* is more like an ontology[8], depicting the trial requirements of the common OTD cases which are summarized according to different scenes. For instance, as Figure 2 shows, an e-commerce fraudulent case can occur in various scenarios: false price, false promotion, commodity problem etc. Each scenario may also contain several sub-scenarios. Furthermore, to sentence a transaction case as a fraudulent, the necessary and insufficient condition is that the target plaintiff (or the buyer) should be confirmed as a real consumer rather than a professional extortioner. In this work, we project each case onto the LKG and the auxiliary case representation will be employed as another important input to the proposed prediction tasks.

3.2 Representation Mechanisms

For each case, we have three types of representations: discrete, word, and legal knowledge graph. We represent them separately according to their characteristics and later concatenate them as the aggregated features.

3.2.1 Discrete Representation. Discrete embeddings are specifically for representing the meta data as shown in Table 1. We first partition the continuous features (e.g., quantity, price, credit) to discretized intervals. Together with the other categorical variables (e.g., category, star), we map them into Euclidean spaces where the mapping is learned by a neural network during the standard supervised training process similar as the way introduced in [15]. It helps to reveal the intrinsic properties of the categorical variables by mapping similar values close to each other in the embedding space.

3.2.2 Word Representation. We use Skip-gram model [27] to train word embeddings. Similar to word embeddings, character embeddings are trained by the same Skip-gram objective. The sequence of character embedding vectors of the word is then fed to a bidirectional LSTM[18]. The final character representation is obtained by concatenating the forward and backward final states. Using the character embeddings can efficiently provide morphological features. The final word representation for each word is the concatenation of its word and character embeddings.

3.2.3 LKG Representations. To explore the legal knowledge via *LKG* for dispute characterization, the proposed method can project the target case to a number of elements on the *LKG*, and the case can be, then, represented by a probability distribution over all the elements on the knowledge graph. For each sentence in the input dispute description text, we calculate the semantic similarity between the text of the sentence (s) and the text on the source nodes (c) and aggregate the score over all the input sentences by taking the maximum value for each source node (c) as the initial projection of the dispute case over the *LKG*.

$$I(c) = \max_{s \in S} E(c, s), c \in C \tag{3}$$

We regard each piece of text as a cluster of words and then compute the average similarity between the two clusters[9].

$$E(c, s) = \frac{1}{n_c n_s} \sum_{i=1}^{n_c} \sum_{j=1}^{n_s} cos(\mathbf{w_{ci}}, \mathbf{w_{sj}}) \tag{4}$$

where the similarity of two words is calculated by the cosine similarity of their word embeddings. n_c and n_s are the numbers of words in the node c and in the sentence s, respectively.

By leveraging the activation function, as denoted in Eq. 5, the legal knowledge closer to the source nodes can be spread/aggregated to the higher level nodes. We utilized the bottom-up design, because the source nodes associate with more detailed legal knowledge for classification.

$$\begin{gathered} a_{i,in} = \sum_{v_j \in N(v_i)} a_{j,out} \cdot w_{ij} \cdot d \\ \text{where } w_{ij} = \frac{1}{N(v_i)} \end{gathered} \tag{5}$$

where $a_{i,in}$ denotes the total activation energy for node v_i, and $N(v_i)$ is the set of v_i's neighbor nodes with incoming links to v_i. $a_{j,out}$ denote the output activation of a node v_j connected to node v_i, and w_{ij} denote the weight of connection between node v_i and v_j. d ($0 < d < 1$) is a global decay parameter to penalize the activation transfer over the longer paths. In the experiment, we set d to 0.85.

The propagation of the similarity score is a depth-first iterative process from the source nodes to the target nodes, via the weighted edges over a directed graph. The final score of each nodes on the graph is normalized[10] before regarded as the *LKG* representation of the current case.

4 LEGAL DISPUTE JUDGMENT MODEL

This section describes the inference procedure of the proposed Legal Dispute Judgment (LDJ) model. The proposed multi-task framework is depicted in Fig. 1. It has three main components - embedding module, shared encoder, and task-specific decoders. As described in Sec. 3.2, the embedding module consists of three types of embeddings—discrete, word, and LKG embeddings, which are intended to capture the semantic meaning of discrete features,

[7] In this study, we use the Chinese e-commerce law as a case. Other countries may have different e-commerce law systems.
[8] The most commonly accepted definition of an ontology is "an explicit specification of a conceptualization"[14].
[9] The idea is similar to the average linkage of two clusters
[10] The sum of their scores equals to 1

text features and the semantic associations between legal terms, respectively. Additionally, for each e-commerce viewpoint, we build an encoder block to encode different types of data as shown in the sub-graph in Fig. 1. All the tasks share the three encoders in different extent so that the shared representations can be learned. In the decoder part, we leverage the dependencies across the tasks to enable the communication between the task pair.

4.1 Input

The text input $\mathbf{x_t}$ operates in a document-wise way where $\mathbf{x_t} = \{s_1, s_2, ..., s_L\}$ is a sequence of L sentences and each sentence $s_i = \{w_{i1}, w_{i2}, ..., w_{iT}\}$ contains T_i words while each word is represented by its word embeddings (see Sec. 3.2.2). The discrete input $\mathbf{x_d} = \{v_1, v_2, ..., v_K\}$ is a variable set and in a fixed length of K. The LKG input $\mathbf{x_g} = \{a_1, a_2, ..., a_N\}$ is a probability distribution as pre-computed in 3.2.3 and the dimension N equals the number of nodes on the LKG.

4.2 Encoder

To encode the text input, we propose to use a hierarchical network aiming at capturing the document structure behind the text by first building the representations of sentences and then aggregating those into a document representation.

Sentence Representation. We use convolutional neural network (CNN) to compute continuous representations of sentences with semantic composition. CNN for sentence classification [21] has been widely adopted recently due to its simplicity and even a shallow word-level CNN can outperform much deeper and more complex CNN architectures on a wide range of text classification tasks [24]. Another advantage of CNN is its robustness when handling with long text where RNN-based encoder (e.g., LSTM, GRU)[11] usually encounters the gradient vanishing problem. Consider our text input is the dispute record where the customer tended to use long sentences to describe the context of the case itself, CNN is more suitable in such circumstances. To be specific, we employ CNN with multiple convolutional filters of different widths to produce sentence representation. The reason is that they are capable of capturing local semantics of n-grams of various granularities, which are proven powerful for sentence-level classification tasks [38, 46]. For instance, the convolutional filter with a width of 3 essentially captures the semantics of trigrams in a sentence. In our model, we use five convolutional filters whose widths are from 1 to 5. Besides the semantics of unigrams, bigrams and trigrams, due to the complexity of vocabularies in legal and e-commerce scenarios, such as the named entities, accusation names and commodity names, larger widths can encode longer n-grams. Each filter consists of a list of linear layers with shared parameters, then their outputs are fed to a max pooling layer to extract the the most important features out of the tedious complaints in the dispute. We further add Rectifier Linear Unit (ReLU) activation [30] and concatenate the outputs of multiple filters to get the hidden representation as the sentence vector denote as s_i.

Document Representation. At the document-level, we utilize a bidirectional GRU [1, 6] to capture the dependencies between sentences by summarizing the information from adjacent sentences in a more efficient way compared with the bidirectional long short-term memory (LSTM) [7]:

$$\begin{aligned} \overrightarrow{h_i} &= \overrightarrow{GRU}(s_i), i \in [1, L] \\ \overleftarrow{h_i} &= \overleftarrow{GRU}(s_i), i \in [L, 1] \\ h_i &= [\overrightarrow{h_i}, \overleftarrow{h_i}] \end{aligned} \tag{6}$$

h_i is the concatenation of the forward hidden state $\overrightarrow{h_i}$ and the backward one $\overleftarrow{h_i}$. In this way, the annotation h_i summarizes the information of adjacent sentences on both sides while due to the tendency of RNNs to better represent recent inputs, h_i still focuses on the current sentence s_i.

In the legal scenario, one sentence difference in the evidence may change the final judgment decision, so the judges also need to focus on the key elements of the submitted evidences during the process of trial. In such context, we employ an attention mechanism which is often used in the sentence/document classification tasks [44] to highlight such pieces of information that can be important to the algorithm outcome of the whole article and aggregate the representation of those informative sentences to form a document vector.

$$\begin{aligned} u_i &= tanh(W_\omega \cdot h_i + b_\omega) \\ \alpha_i &= \frac{exp(u_i^T u_\omega)}{\sum_t exp(u_i^T u_\omega)} \\ h^t &= \sum_{i \in [1,L]} \alpha_i \cdot h_i \end{aligned} \tag{7}$$

where u_i is a hidden representation of h_i through a one-layer MLP, u_ω is a randomly initialized normalization factor and jointly learned during the training process. α_i is a softmax function. And we compute the document vector h^t as a weighted sum of the sentence annotations.

LKG & Discrete Represenation. To encode the LKG and discrete input, we connect each input to a one layer fully connected neural network with a logistic sigmoid activation function where h^g and h^d are the hidden representation of LKG and discrete input respectively:

$$\begin{aligned} h^g &= sigmoid(W_g \cdot v^g + b_g) \\ h^d &= sigmoid(W_d \cdot v^d + b_d) \end{aligned} \tag{8}$$

We finally concatenate the hidden representation of the three types of features as the output of the encoder.

$$V = [h^t, h^g, h^d] \tag{9}$$

Note that as introduced in Sec. 3.1, our data coming from three e-commerce viewpoints: transaction, buyer and seller, therefore we build three encoders in parallel to deal with three data streams, respectively. For every stream, it incorporates three types of features: text, discrete and LKG to build the dispute representation as described above. We let the three encoders to share one embedding module. The output of the three encoders (transaction layer (V_T), buyer layer (V_B) and seller layer (V_S)) will be further input to the following tasks.

[11] We have also tested LSTM and GRU as the sentence-level text encoder, the comparison between their performances is shown in Sec. 6.3.

4.3 Task-specific Decoders

As mentioned in Sec. 2, our goal in this work is to leverage the large amount of dispute data to enable lawsuit judgment prediction. We thus describe the subtasks one by one and finally reach our main task.

Subtask: Dispute Reason Prediction. For predicting the reason of dispute, we use the output of transaction layer as input, denoted as $x^{(1)} = V_T = [h_T^t; h_T^g; h_T^d]$ and employ a highway network [37] after it (eq. 10) to solve the training difficulties with the model parameters growing:

$$x^* = C(W_C \cdot x + b_C) \cdot U(W_U \cdot x + b_U) + x \cdot (1 - U(W_U \cdot x + b_U)) \quad (10)$$

where x equals $x^{(1)}$, $U(\cdot)$ is a sigmoid activation function and $C(\cdot)$ is a Rectified Linear Unit (ReLU) activation function. Then $x^{(1)*}$ is fed into a standard softmax classifier with a single ReLU layer which outputs the probability vector $y^{(1)}$ for each of the dispute reason labels.

Subtask: Dispute Result Prediction. Dispute result prediction is performed on top of the REASON layer. As motivated by the fact that leveraging the historical transactions and dispute records of both parties can benefit for reaching the best negotiated agreement as dispute result [20], we stack the output of transaction layer, buyer layer and seller layer together with the output of REASON layer as the concatenated input of RESULT layer, denoted as $x^{(2)} = [V_T^*; V_B^*; V_S^*; y^{(REA)}]$, where V_T^*, V_B^* and V_S^* are the outputs of V_T, V_B and V_S by passing a highway network (eq. 10), respectively. We define the weighted REASON label embedding $y^{(REA)}$ as follows:

$$y^{(REA)} = \sum_{j=1}^{K} p(y^{(1)} = j | h^{(1)}) \cdot \ell(j) \quad (11)$$

where K is the number of REASON labels, $p(y^{(1)} = j|h^{(1)})$ is the probability value that the j-th REASON label is assigned to the current case, and $\ell(j)$ is the corresponding label embedding. For predicting the RESULT labels, we employ the same strategy as REASON classification by passing a highway network (Eq. 10) and we also use a single ReLU hidden layer before the softmax classifier.

Subtask: Lawsuit Fact Prediction. Lawsuit fact prediction identifies the legal fact admitted by the judges, which implies the relationship to the process of dispute reason prediction because some dispute reasons have correspondence to the certain legal facts. For instance, the disputes arising from `fake problem` and `counterfeit brand`[12] are more likely to be recognized as the legal fact of `Brand infringement&fake goods`[13] if it goes to the lawsuit process. Thus, to predict the labels of lawsuit fact, we let the input as the concatenation of the output from transaction layer and from REASON layer, denoted as $x^{(3)} = [V_T^*; y^{(REA)}]$. We also let it pass a highway network (Eq. 10) and then connect to a fully connected layer with sigmoid output. Different to the dispute-level tasks, we apply binary cross-entropy loss over sigmoid activation as the objective loss function which is proved to be more suitable for the aim of multi-label classification compared to using softmax activation [25]. The binary cross-entropy objective can be formulated as:

$$\mathcal{L} = \min_{\Theta} -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} \{y_{ij} log(\sigma(z_{ij})) + (1 - y_{ij}) log(1 - \sigma(z_{ij}))\} \quad (12)$$

where C is the number of labels, n is the number of cases. y_{ij} denotes the ground truth whether the instance i is assigned with the label j. $\sigma(z_{ij})$ is the sigmoid output where $\sigma(x) = \frac{1}{1+e^{(-x)}}$.

Main Task: Lawsuit Judgment Prediction. Here comes to the main task. To predict the lawsuit judgment, we incorporate the information of the transaction itself (V_T^*), the historical experience of the litigants (V_B^* and V_S^*), as well as the FACTs recognised in the prior step (y^{FAC}), because the judgment is the interpretation that the judgment must be supported by the findings of fact [33]. In such context, we define the input of JUDGMENT layer as $x^{(4)} = [V_T^*; V_B^*; V_S^*; y^{(FAC)}]$. To use the output from the FACT layer directly, we use the label embeddings for the FACT layer. More concretely, we compute the class label embeddings for the FACT classification task similar to Eq. 11. The final feature vectors are concatenated through a highway network (Eq. 10) and fed into the JUDGMENT classifier which is a fully connected layer with sigmoid output similar to the FACT classifier. We also use the binary cross-entropy loss to enable the multi-label classification.

The training process has two stages. First, we utilize all the dispute data to train the REASON and the RESULT layers while switching off the FACT and JUDGMENT layers in order to optimize the dispute representation by leveraging the large amount of dispute data. Then, we switch on the four tasks and use the lawsuit data[14] to fine tune all the parameters. We empirically set the weight of losses of the main tasks and three subtasks to 0.6, 0.2, 0, 1, 0, 1 respectively and the final loss equals to the sum of their weighted loss.

5 EXPERIMENTAL SETTINGS

5.1 Datasets

We conduct extensive experiments on two datasets: one is the historical e-commerce dispute data generated from Taobao[15], the other is the lawsuit data provided by Supreme Court[16]. Note that the experiment is based on the Chinese e-commerce and lawsuit data, but the method can be generalized to any other nations and e-commerce platforms as long as the target LKG and training data are available. The details of the two datasets are described as follows.

Dispute (ODR) dataset.

We collect 400K historical dispute records varying from 46 dispute reasons and 3 dispute results. The most frequent dispute reasons are `quality problem`,`wrongly ordered`,`return and exchange` and `wrong product information`. There are three dispute results, i.e., `refund and return` (the seller should refund the payment and the buyer should return the goods), `reject` (the request of dispute is denied), and `only refund` (when buyer hasn't received the product due to delivery problems), so it only requires the seller to refund the payment. To the best of our knowledge, few open e-commerce dataset is available for dispute or legal related research,

[12] Two of the dispute reason labels in our data set.
[13] One of the lawsuit fact labels in our data set.
[14] We use 80% data as training data, 10% for validation, and the rest for test.
[15] It is the world leading e-commerce website.
[16] In this work, the data is provided by Supreme People's Court which is the highest level of court in the mainland area of the people's Republic of China [43].

Table 2: Main Results of All Tested Methods. Improvements of LDJ_Multi over best baseline TextCNN_GRU_Atten on all metrics are significant at $p = 0.01$. LDJ_Multi over LDJ_Multi(Fact+Judge) and LDJ_Multi(Fact+Judge) over TextCNN_GRU_Atten on all metrics are significant at $p = 0.05$.

	Framework	Training set	Method	MiP	MiR	MiF_1	MaP	MaR	MaF_1
Baseline	Single Task	Lawsuit	BSVM(BOW)	0.494	0.523	0.508	0.488	0.352	0.356
		Lawsuit	BSVM(embed.)	0.557	0.425	0.482	0.494	0.355	0.411
		Lawsuit+Dispute	Semi-supervised BSVM(BOW)	0.499	0.577	0.535	0.467	0.411	0.405
		Lawsuit+Dispute	Semi-supervised BSVM(embed.)	0.537	0.428	0.477	0.611	0.435	0.504
		Lawsuit	TextCNN_GRU	0.658	0.743	0.698	0.622	0.670	0.581
		Lawsuit	Bi-LSTM_GRU_Atten	0.701	0.734	0.717	0.635	*0.723*	*0.661*
		Lawsuit	GRU_GRU_Atten	*0.721*	0.729	0.725	0.640	0.691	0.645
		Lawsuit	TextCNN_GRU_Atten	0.718	*0.744*	*0.731*	*0.654*	0.720	0.658
Ours	Multi-Task	Lawsuit	LDJ_Multi(Fact+Judge)	0.733	0.761	0.747	0.658	0.742	0.688
		Lawsuit+Dispute	LDJ_Multi	**0.749**	**0.820**	**0.783**	**0.676**	**0.781**	**0.726**

we release all the experiment data to motivate other scholars to further investigate this novel but important problem[17]. To break the language barrier and address the privacy issue, we convert all the words and labels into indices and we also provide the word embeddings pre-trained on millions of dispute records[18].

Lawsuit Dataset. The lawsuit dataset is crawled from the website "China Judgments Online"[19] on which the judicial documents are released by the Supreme Court and can be publicly accessible online. We narrow down to the online dispute cases related to Taobao disputes by using keyword search (e.g., 'Taobao', 'civil cases', 'online dispute' etc.). In total, we collect 6,858 cases. For each lawsuit case, the transaction ID appeared in the judicial document can be used to locate the target disputed transaction as well as its corresponding dispute record in the 400K Taobao ODR dataset. This method works well because most of the plaintiff has gone through an ODR process on Taobao platform before filing a lawsuit. As the judgement documents follow the template, we use the regular expression to extract the legal fact and the judgement result of each case for classification tasks. There are 7 kinds of judgments, namely `refund`,`compensation for loss`,`triple compensation`,`ten times compensation`,`single compensation`,`all reject`,`platform joint responsibility`. There are 15 types of lawsuit facts of which `identification label problem`, `illegal additives`,`undocumented production` and `exaggerating false propaganda` are most popular one. The joint dataset of dispute data and lawsuit data can be found on the project site.

5.2 Training Details

Word2Vec [27] is used to pre-train the word embeddings, which are then used to initialize the embeddings in the model. As mentioned in Sec. 5.1, the word embeddings are trained on millions of dispute records[20]. The dimensionality of word, char and discrete embeddings are set to 100, 50 and 10 respectively. We use multiple filters with window sizes of $\{1, 2, 3, 4, 5\}$ and output sizes of $\{32, 32, 64, 128, 256\}$ for the CNN-based sentence encoder. The GRU dimension is set to 100. In this case a combination of forward and backward GRU gives us 200 dimensions for the document annotation. Based on these settings, we optimize the proposed LDJ model using Adam Optimization [22] with a learning rate of 0.001, and perform the mini-batch gradient descent with a batch size of 64 for 50 epochs in total[21].

6 RESULT DISCUSSIONS

6.1 Evaluation Metrics

We use $Micro\_F_1$ and $Macro\_F_1$ (MiF_1 and MaF_1 for short) as the main metrics to evaluate the main task result. A macro-average can measure the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average is able to aggregate the contributions of all classes to compute the average metric. In a multi-label multi-class classification setup, micro-average is preferable if there exists class imbalance (i.e there are many more examples of one class than of other classes).

6.2 Baselines

As we could not find any prior work that performs multi-task learning on legal judgment tasks, we evaluate the proposed model by comparing with several state-of-the-art methods on the single task of judgment prediction. In the tested models, all the proposed viewpoint features are used.

Traditional machine learning based methods. **BSVM(BOW)** and **BSVM(embed.)** are two multi-label classification methods [10]. The former one uses bag of uni-grams for text representation and the latter one applies the same word embeddings used in our proposed methods. Each label prediction is regarded as a binary classification problem, then a ranking mechanism is employed for binary classification with SVM classifier[22]. To leverage the dispute dataset, we also test the above two baselines with semi-supervised setting by automatically labeling dispute cases for self-training[23] [36].

Deep Learning based methods. In this part, we build several combinations of deep learning techniques as baselines to verify the effectiveness of each components used in the proposed model.

[17] https://github.com/zhouxinhit/auto_dispute_judge
[18] the vocabularies are converted to indices as well.
[19] https://wenshu.court.gov.cn/. This is an open dataset.
[20] The minimum of term frequency is set to 5
[21] The iterations will stop if 15 epochs without improvement.
[22] For implementation, we adopt the publicly available library from LIBSVM [5].
[23] Top 20,000 most confident pseudo-labeled dispute data was added into training set for model retraining. We repeated this process until the best performed model was finally found.

TextCNN_GRU uses a convolution network for sentence representation [21], and, in order to compare with our model in the document-level representation, we add a bi-GRU for document encoder. **GRU_GRU_Atten** [44] is a hierarchical attention networks for document classification which uses bi-GRU for sentence and document encoder, and attention mechanism is applied to emphasize the important pieces of information. **Bi-LSTM_GRU_Atten** constructs document modeling with gated recurrent neural network [38] with a bi-LSTM based sentence encoder and a GRU-based document encoder. Attention mechanism is also added for comparison reason. **TextCNN_GRU_Atten** is regarded as our sub-model but in single-task framework. All the above four baselines employ a binary cross-entropy loss for multi-label classification.

6.3 Overall Performance

To comprehensively validate the proposed model for the main task—Lawsuit Judgment Prediction, we report results from four perspectives: (1) comparison against baselines, (2) the benefit of training under multi-task framework, (3) the effectiveness of each multiview points and (4) the influence of the usage of e-commerce dispute data.

Comparison against baselines. The performance of all tested methods is reported in Table 2. We have the following observations from the results: (1) It is not surprising to see that the traditional machine learning based methods didn't perform well in terms of all the evaluation metrics. It indicates the importance of learning good dispute representations through deep neural models for better judicial judgment prediction. (2) Semi-supervised approach can significantly (p-value< 0.01) enhance the $Macro\_F_1$ score by 18.2% compared with the two corresponding SVM-based methods. This observation shows the possibilities of using e-commerce dispute data to support the task in the judicial field. (3) **TextCNN_GRU_Atten** beats the other three deep learning methods over most of the metrics which indicates the effectiveness of Attention mechanism (i.e., **TextCNN_GRU_Atten v.s. TextCNN_GRU**) and the selection of sentence encoder among bi-LSTM, GRU and TextCNN. Although **Bi-LSTM_GRU_Atten** performs slightly better over $Macro\_F_1$ score, consider the imbalance of the labels in judgments (the metrics $Micro\_F_1$ is more preferable), **TextCNN_GRU_Atten** is regarded as the best baseline for further comparison with the proposed methods. (4) **LDJ_Multi(Fact+Judge)**[24] is designed to jointly learn the tasks of Lawsuit Fact and Judgment together by training with only lawsuit data. It outperforms the best baseline over all the metrics (p-value< 0.05), which demonstrates the benefit of training under multi-task framework. Moreover, when further injecting the dispute data and jointly learning with all the four tasks, **LDJ_Multi** is statistically significant (p-value< 0.01) over the best baseline.

The effectiveness of each multiview points. To access the contribution of different multiview points, we conduct ablation tests for best baseline and best proposed model respectively. Table 3 reports the $Micro\_F_1$ and $Macro\_F_1$ when training on all features and when training on all features except the particular one. Note that since the *transaction* data is basic element of the dispute case, so we mainly test the significance of the other viewpoints. According to the results shown in Table 3, all the feature sets contribute positively to the results. To be specific, the features *buyer* and *seller* have largest impact - their removal causes 45% increase in the error (RIE) in single framework and 7% RIE in multi-task framework (p-value< 0.01). As for the viewpoint *LKG*, it shows more impact under multi-task framework (10% in RIE) than single-task framework (2% in RIE).

[24]The REASON embedding is removed from the input of the FACT layer in this method.

Table 3: Multiview ablation test.

Framework	Method	MiF_1	MaF_1
Single Task	Best_Baseline	**0.731**	**0.658**
	Best_Baseline − LKG	0.723	0.651
	Best_Baseline −Buyer − Seller	0.694	0.624
Multi-Task	LDJ_Multi	**0.783**	**0.726**
	LDJ_Multi − LKG	0.761	0.698
	LDJ_Multi − Buyer −Seller	0.689	0.603

The influence of the usage of e-commerce dispute data. In this part, we present the influence of the increased usage of dispute data on the performance of lawsuit judgment prediction (see Fig. 3). We can observe that our multi-task framework can achieve quite limited performance when training only by 5,458 lawsuit data[25] with its corresponding dispute data. However the big data plugin from e-commerce ODR platform can be quite significant to cope with training data sparseness problem and enables our multi-task model to achieve better performance, especially over $Macro\_F_1$ score (the distance between the two lines gets closer as the increase of dispute data used), which indicates the improvement of the classifier in predicting those sparse categories.

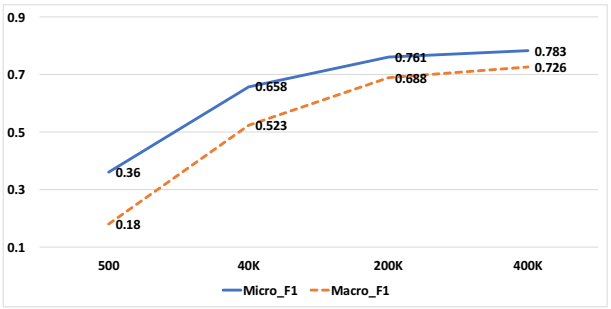

Figure 3: Multi-Task performance with e-commerce data increasing

Error Analysis For the bad cases[26], 70.8% of the errors come from the lawsuit fact prediction. It is clear that lawsuit fact prediction is a challenging but critical task for judgment prediction. From learning viewpoint, the most confusing labels can be: (`undocumented production`, `label problem`, `exaggerating false propaganda` and `other quality problem`). Meanwhile, we also find that the semantic confusion among the judgment labels can be another difficulty that may threaten the algorithm performance, e.g., some class labels can be somehow similar. To improve the model in the future, investigating the distinguishable features or enhance the legal knowledge graph to differentiate the case associated fact/judgment labels can be a promising approach.

[25]80% data of the entire lawsuit data are used for training.
[26]Since the judgment prediction is a multi-label classification task, if the labels of a case are not comprehensively predicted, it is defined as a bad case.

7 RELATED WORK

Legal Intelligence. Judicial decision prediction is not a novel topic which has been raised since 1960s that an article appeared in *The American Behavioral Scientist* entitled "Using Simple Calculations Predict Judicial Decisions [28]". Since then a number of researchers from legal field started to explore the possibilities and methodologies to approach such problem [16, 23, 29, 42]. For instance, some advocates claimed that the computers can help find and analyze the law as well as helping lawyers and lower court judges to predict or anticipate the judicial decision [23]. Correspondingly, some methodologies were adopted to predict the probability of a favorable decision experimented on specific types of cases [16, 29]. In the recent work, Oard et al. studied the information retrieval for e-discovery [31, 32]. Wang et al. [40] proposed a model for crime classification. Despite those opposition voices who were skeptical to the judgments made by machine instead of human [42], the advantages of objectivity and justice brought by the automatic judgment prediction should not be ignored [11]. However, their approaches never have a chance to really implement, especially for e-commerce ecosystem, because of data barrier/sparseness, algorithm limitation, and lack of computational legal knowledge. In this study, we pioneer this problem by using multiview dispute data along with sophisticated multi-task learning and deep dispute representation learning.

Online Dispute Resolution in E-Commerce. 20 years ago, the scholars have predicted the growth of online disputes while e-commerce was becoming an increasingly important place for transactions. There is reason to believe that dispute resolution systems and services are needed to be online [20]. With the development of e-commerce as well as techniques, Online dispute resolution (ODR) system has become mature nowadays which is a form of online settlement that uses alternative methods for dispute resolution. Today almost all the e-commerce platforms operate on their own ODR systems. Though ODR can be done in a way of low cost and high efficiency, there is no recommendation provided for the customers who are not satisfied with the resolution result and about to file a lawsuit. In this work, we leverage the large amount of dispute data provided by the e-commerce platform to enable lawsuit judgment prediction. On the other hand, legal judgment prediction is also a way of making the resolution of e-commerce disputes more legitimate [12].

Deep learning and representation learning. The success of machine learning algorithms generally depends on data representation. Although specific domain knowledge can be used to help design representations, learning with generic priors can also be used, and the quest for AI is motivating the design of more powerful representation-learning algorithms implementing such priors[2]. In the field of NLP, the representation learning has penetrated to almost every granularity of text from character to the document representation. For instance, word2vec [27] and Glove [34] are the two widely adopted *word embedding* techniques which can be traced back to the *distributed representations* introduced by Hinton [17], and developed in the context of statistical language modeling [3]. Above the word, the sentence can be also represented as a low-dimensional vector through convolutional network [21] or RNN-based network [6, 18] by considering the sequential information within the sentences. At the document-level, the representation is learned through the hierarchies among the text [13, 38, 39, 44, 44]. In this work, we introduce a multiview dispute representation technique by not only effectively coping with document-level text representation but also leveraging multiple types of domain-specific features (e.g., Legal Knowledge Graph (LKG) and discrete features of e-commerce data).

Multi-task learning for e-commerce. The usage of multi-task learning (MTL) models has become ubiquitous for many machine learning applications in areas ranging from natural language processing, speech recognition, computer vision to drug discovery [35]. The idea of learning multiple tasks simultaneously is to improve the generalization performance by leveraging the information from the related tasks. In most cases, we prioritize the main task, and the goal of choosing related task, as an auxiliary, in MTL is to learn representation beneficial to the main task. For example, Collerbert et al. [8] introduced a single convolutional neural network architecture that is trained jointly on several NLP tasks, like part-of-speech tags, chunks, named entity tags, semantic roles, semantically similar words and so on. Liu et al. learned a multi-task DNN for multiple-domain query classification and ranking for web search[26]. Masaru et al. jointly learned sentence extraction and document classification [19]. Yu et al. used whether a sentence contains a positive or negative sentiment word as auxiliary tasks for sentiment analysis [45]. More recently, Wang et al. [41] and Zhao et al. [47] investigated natural language generation by using jointly user behavior mining and product information extraction. Unlike prior studies, we focus on the legal intelligence problem by leveraging e-commerce tasks. Moreover, in the proposed framework, we learn more tasks jointly by considering the logical hierarchies behind. Two different datasets are employed for model training, and big data can provide critical information for the main task via auxiliary tasks.

8 CONCLUSIONS

As an interdisciplinary study, performing legal dispute prediction can be practically useful while bridging two isolated domains, e-commerce data mining and legal intelligence. In this work, we introduce a delicately designed multiview dispute representation technique and provide an end-to-end solution for lawsuit judgment prediction by jointly learning with three subtasks. The empirical findings validate how learning all tasks jointly improves the performance over state-of-the-art approaches. Additionally, the usage of dispute data from e-commerce platform proves to significantly enhance the performance of judicial prediction. Through results and error analysis, we show the significance of each facet in dispute representation and also anticipate to discover more distinguishable features for future work.

9 ACKNOWLEDGMENTS

This work is supported by National Key R&D Program of China (2018YFC0830200;2018YFC0830206).

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[4] Maximilian A Bulinski and JJ Prescott. 2015. Online Case Resolution Systems: Enhancing Access, Fairness, Accuracy, and Efficiency. *Mich. J. Race & L.* 21 (2015), 205.

[5] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[8] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 160–167.

[9] Lloyd Duhaime. [n. d.]. "evidence", Duhaime's Law Dictionary. available from http://www.duhaime.org/dictionary/dict-e.aspx. Accessed: 2019-01-14.

[10] André Elisseeff and Jason Weston. 2002. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*. 681–687.

[11] Heidi Li Feldman. 1993. Objectivity in Legal Judgment. *Mich. L. Rev.* 92 (1993), 1187.

[12] Fred Galves. 2009. Virtual justice as reality: making the resolution of E-commerce disputes more convenient, legitimate, efficient, and secure. *U. Ill. JL Tech. & Pol'y* (2009), 1.

[13] Shang Gao, Arvind Ramanathan, and Georgia Tourassi. 2018. Hierarchical Convolutional Attention Networks for Text Classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*. 11–23.

[14] Thomas R Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition* 5, 2 (1993), 199–220.

[15] Cheng Guo and Felix Berkhahn. 2016. Entity Embeddings of Categorical Variables. *CoRR* abs/1604.06737 (2016). arXiv:1604.06737 http://arxiv.org/abs/1604.06737

[16] Charles M Haar, John P Sawyer Jr, and Stephen J Cummings. 1977. Computer power and legal reasoning: A case study of judicial decision prediction in zoning amendment cases. *Law & Social Inquiry* 2, 3 (1977), 651–768.

[17] Geoffrey E Hinton et al. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, Vol. 1. Amherst, MA, 12.

[18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[19] Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2101–2110.

[20] Ethan Katsh, Janet Rifkin, and Alan Gaitenby. 1999. E-Commerce, E-Disputes, and E-Dispute Resolution: in the shadow of eBay law. *Ohio St. J. on Disp. Resol.* 15 (1999), 705.

[21] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[22] Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: Amethod for stochastic optimization. In *International Conference on Learning Representations*.

[23] Reed C Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal* (1963), 337–344.

[24] Hoa T Le, Christophe Cerisara, and Alexandre Denis. 2018. Do convolutional networks need to be deep for text classification?. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

[25] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 115–124.

[26] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. (2015).

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[28] Stuart Nagel. 1960. Using simple calculations to predict judicial decisions. *American Behavioral Scientist* 4, 4 (1960), 24–28.

[29] Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.* 42 (1963), 1006.

[30] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[31] Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *Artificial Intelligence and Law* 18, 4 (2010), 347–386.

[32] Douglas W Oard, William Webber, et al. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2–3 (2013), 99–237.

[33] Federal Rules of Civil Procedure. [n. d.]. Rule 52. Findings and Conclusions by the Court; Judgment on Partial Findings. available from https://www.law.cornell.edu/rules/frcp/rule_52. Accessed: 2019-01-14.

[34] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[35] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[36] H Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.

[37] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387* (2015).

[38] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1422–1432.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[40] Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling Dynamic Pairwise Attention for Crime Classification over Legal Articles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 485–494.

[41] Yongzhen Wang, Heng Huang, Yuliang Yan, and Xiaozhong Liu. 2019. User-Centric Quality-Sensitive Training! Social Advertisement Generation by Leveraging User Click Behavior. In *Proceedings of the 2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

[42] Frederick Bernays Wiener. 1962. Decision prediction by computers: Nonsense cubedâĂŤand worse. *American Bar Association Journal* (1962), 1023–1028.

[43] Wikipedia. [n. d.]. Supreme People's Court. available from https://en.wikipedia.org/wiki/Supreme_People's_Court. Accessed: 2019-01-14.

[44] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

[45] Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 236–246.

[46] Ye Zhang, Stephen Roller, and Byron Wallace. 2016. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. *arXiv preprint arXiv:1603.00968* (2016).

[47] Lujun Zhao, Kaisong Song, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2019. Review Response Generation in E- Commerce Platforms with External Product Information. In *Proceedings of the 2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.