Predicting the Success of Mediation Requests Using Case Properties and Textual Information for Reducing the Burden on the Court

HSUN-PING HSIEH and JIAWEI JIANG, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
TZU-HSIN YANG, Department of Computer Science, University of California, Davis, California
RENFEN HU, Institute of Chinese Information Processing, Beijing Normal University, Beijing, China
CHENG-LIN WU, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

The success of mediation is affected by many factors, such as the context of the quarrel, the personality of both parties, and the negotiation skill of the mediator, which lead to uncertainty for the work of prediction. This article takes a different approach from that of previous legal prediction research. It analyzes and predicts whether two parties in a dispute can reach an agreement peacefully through the conciliation of mediation. With the inference result, we can know whether mediation is a more practical and time-saving method to solve the dispute. Existing works about legal case prediction mostly focus on prosecution or criminal cases. In this work, we propose a long short-term memory (LSTM)–based framework, called LSTMEnsembler, to predict mediation results by assembling multiple classifiers. Among these classifiers, some are powerful for modeling the numerical and categorical features of case information, for example, XGBoost and LightGBM. Some are effective for dealing with textual data, for example, TextCNN and BERT. The proposed LSTMEnsembler aims to not only combine the effectiveness of different classifiers intelligently but also to capture temporal dependencies from previous cases to boost the performance of mediation prediction. Our experimental results show that our proposed LSTMEnsembler can achieve 85.6% for F-measure on real-world mediation data.

CCS Concepts: • **Information systems** → **Information systems applications**; • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Mediation case prediction, long short-term memory, ensemble, textual mining

ACM Reference format:
Hsun-Ping Hsieh, Jiawei Jiang, Tzu-Hsin Yang, Renfen Hu, and Cheng-Lin Wu. 2022. Predicting the Success of Mediation Requests Using Case Properties and Textual Information for Reducing the Burden on the Court. *Digit. Gov.: Res. Pract.* 2, 4, Article 30 (January 2022), 18 pages.
https://doi.org/10.1145/3469233

This work was partially supported by Ministry of Science and Technology (MOST) of Taiwan under grants MOST 108-2221-E-006-142, MOST 108-2636-E-006-013, and MOST 109-2636-E-006-025.
Authors' Addresses: H.-P. Hsieh, J. Jiang, and C.-L. Wu, Department of Electrical Engineering, No. 1 University Road, East Dist, Tainan City, 701401, Taiwan (R.O.C.); emails: hphsieh@mail.ncku.edu.tw, n26100456@gs.ncku.edu.tw, n26090693@mail.ncku.edu.tw; T.-H. Yang, 2063 Kemper Hall, Davis, CA 95616, USA; email: zixyang@ucdavis.edu; R. Hu, No. 19, Xinjiekou Wai Street, HaiDian District, 100875, Beijing, People's Republic of China; email: irishere@mail.bnu.edu.cn.

2639-0175/2022/01-ART30 $15.00
https://doi.org/10.1145/3469233

1 INTRODUCTION

Disputes and conflicts occur often in daily life. Some people choose to settle the matter privately, while others prefer to seek legal means to solve the problem. Private settlement can save a lot of time and labor costs. However, the negotiation process is often hindered due to problems of different perceptions or intense emotions. In addition, whether the private agreement is legally binding is also doubtful. Sometimes, people reach an agreement in private; nevertheless, some might repeal in the near future and the conflict will happen again. As an increasing number of disputes have arisen, it has gradually increased the burden on the court. Thus, the government has organized a mediation committee for local conflicts. The members of the committee are usually good at socializing and experienced in communication and negotiation. Due to the committee's participation in negotiations, the burden on the court might be reduced, and people can reach a satisfactory agreement with little chance of further dispute.

The mediation committee consists of senior members who are widely respected in the local community and usually have experiences in civil service. A mediator will be selected from committees to lead each mediation process and listen to the opinions of both parties from the perspective of a third party. Then, the mediator gives appropriate advice based on the mediator's experiences. The purpose is not only to let the two sides receive a formalized judgment but also to manage the result to an acceptable range for both parties while considering their backgrounds or economic situations. Therefore, the final agreement or compensation usually does not represent a standard legal verdict. Through the coordination of the mediators, both parties are satisfied and the accused will compensate under certain conditions. Conversely, if the negotiation fails, the mediation process will be suspended and the parties may seek a judicial process to solve their problem.

Real-world mediation data is valuable and worthy of investigation. We aim to build an effective framework, called LSTMEnsembler, by analyzing the data and predicting whether a dispute case in the mediation committee will be resolved successfully, which means that two parties reach an agreement peacefully under the conciliation of the mediator. More specifically, a user can submit a request, which includes the description of the event, the backgrounds of the two parties, and the mediator to our framework. The inference model will then respond with the prediction result, which brings practical advice for requesters to consider whether to adopt the more time-saving method, mediation, to solve the conflict. Thus, in this work, we treat such a prediction task as a classification problem, in which the inference result acts as the basis for considering that the mediation will be a success or failure. According to our study and interviews with domain experts, the outcome of mediation is affected by many factors, such as the context of the quarrel, the personality of both parties, and the negotiation skill of the mediator. In addition to proposing an effective prediction framework, another contribution of this work is to extract the features that are highly correlated with the success of mediation from mediation data. Then, the analysis could be used as the foundation for the government to train professional mediators. LSTMEnsembler has already been implemented as a queried system for residents of Tainan City in Taiwan to act as consult for their cases. The system is located in Tainan City Government for realizing digital governance.

Existing works on legal prediction [1, 6, 10, 12, 15, 19, 25, 27] focus only on textual data mining. However, recently, more studies [3, 22, 23, 29, 31] target the multi-modal issue due to heterogeneous data sources. In our work, we take the combination of case information and textual description as input and predict the results of mediation cases. To the best of our knowledge, there is no existing work dealing with mediation data, which is considered sensitive to the negotiation skill and personalities of the parties. Given not only rich and heterogeneous features from textual descriptions and case information but also the features related to mediators, an immediate thought would be to directly train a powerful machine learning model, such as deep neural networks XGBoost or LightGBM, which is commonly used in past text classification works [2, 8, 14, 16, 28], to merge the aforementioned features and make the prediction. However, according to our experiments, such an approach is not effective in view of the accuracy of the prediction, which is not good. The possible reason is twofold. First, the data—including textual, numerical, and categorical features—are too diverse to be directly combined using a single classifier. Second, a single powerful model cannot effectively reflect the evolution and accumulation of

mediators' experiences. In addition, an ensemble approach could be used since it is usually practical and effective for boosting accuracy, that is, combining multiple learning algorithms to obtain better predictive performance than what could be obtained from any of the constituent learning algorithms alone.

In real-world tasks, mediation cases are applied one after another and sequentially by the public. Thus, such phenomena inspire us to utilize a long short-term memory (LSTM)–based approach to model the temporal dependency between cases. In this article, we investigate how to build an LSTM framework to assemble multiple classifiers. Among these classifiers, some focus on dealing with the features related to case information and some handle textual data. The experimental results on real-world mediation data have shown that our proposed LSTMEnsembler is valid for combining the features of case information and textual features and more robust than conducting a single powerful classifier.

Our contributions are summarized as follows:

- This work is the first to deal with the mediation classification problem, which aims to increase the efficiency of public use of government resources. We discover that an LSTM-based approach is useful for handling sequential property and temporal dependency of mediation data.
- LSTMEnsembler aims to effectively combine the case information and textual features extracted from mediation applications to predict the results of mediation. The experimental results show that our method outperforms three state-of-the-art machine learning models on real-world law-based data.
- According to our experiments, assembling inference results of different classifiers leads to better performance.

The rest of the article is organized as follows. In Section 2, we describe related works. In Section 3, we introduce the format of our mediation datasets. In Section 4, we introduce the proposed LSTMEnsembler for dealing with case information and textual content. The proposed features and the classifiers in LSTMEnsembler are also introduced in Section 4. We discuss experimental results in Section 5. The development of a mediator recommender system based on LSTMEnsembler is covered in Section 6. Our conclusions are presented in Section 7.

2 RELATED WORK

2.1 Legal Judgment Prediction

There are some works focusing on predicting legal judgment. Kort [6] applied a formula that combined pivotal factors to predict decisions by the Supreme Court. Nagel [25] predicted decisions by assigning correlation coefficients. Keown [12] discussed several mathematical models used in predicting judicial decisions, such as the scheme of nearest-neighbor rule and linear models. These studies focus on the analysis of non-textual information and judges' votes. With the technical development of text mining, increasing numbers of works are focusing on textual information. Sim et al. [27] showed that the information extracted from texts is useful for predicting the votes of the Supreme Court judges. Lin et al. [15] introduced a framework to fetch 21 legal factor labels of robbery and intimidation cases and used the labels for case classification. Aletras et al. [1] used the N-gram model to predict the outcome of cases tried by the European Court of Human Rights. Sulea et al. [26] adopted support vector machines (SVMs) to predict the ruling of the French Supreme Court and the law discipline to which the case belongs. Hu et al. [10] proposed an attribute-attentive charge prediction model that had effective signals for distinguishing confusing charges. This research focuses on analyzing textual information. In addition, some articles discussed whether to use segmented words in Chinese. Meng et al. [19] conducted several natural language processing (NLP) experiments and found that word-based models in Chinese are more prone to overfitting. Song et al. [24] used character-level BERT and deep learning models to classify traditional Chinese medicine cases. In our work, we take the combination of case information and textual description as input and predict the results of mediation cases. To the best of our knowledge, there is no existing work dealing with mediation data, which is considered sensitive to the negotiation skill and personalities of the parties. For case

information, we apply feature engineering methods to data related to mediators and parties. For textual content, we adopt state-of-the-art text mining methods and use a character-based model for extracting textual features. The experiments show that combining these two types of information can boost prediction performance.

2.2 Natural Language Processing–based Detection and Prediction Using Multi-model Approaches

Recently, multi-modal prediction and detection have become popular in NLP and text mining research fields. Zhang et al. [29] model emotion detection with modality and label dependence using multi-modal data, including textual, visual, and acoustic modalities. Cai et al. [3] focus on sarcasm detection in tweets consisting of texts and images in Twitter. The detection makes use of linguistic features, properties of the author, the audience, and the immediate communicative environment. Zhu et al. [31] predict product attributes by fusing textual and image data. Nojavanasghari et al. [22] predict persuasiveness using multi-modal information—including visual, acoustic, and textual features—with a deep fusion architecture. Shu et al. [23] claim that detecting fake news only from content is generally not satisfactory and suggest incorporating user social engagements as auxiliary information to improve fake news detection. Therefore, the study investigates how to jointly consider user profiles and post contents to detect fake news. Among these studies, they usually propose a comprehensive framework, which jointly models the multi-modal features such as content, social network, and user characteristics. However, it is not practical to directly apply a single machine learning model to combine heterogeneous features. For example, an image could be ideally dealt with using convolutional neural network (CNN)–based models [3, 29, 31]. In contrast, a textual data could be analyzed using embedding (e.g., BERT [7, 18, 21] or TextCNN) or recurrent neural network (RNN)–based modules [3, 22, 23, 29, 31]. To the best of our knowledge, few legal prediction or detection studies consider multi-modal prediction and detection; thus, this article is at the forefront of work on the issue.

3 MEDIATION DATASETS

Each mediation record in such data contains not only the backgrounds of both parties and the mediator but also a textual description of the case. In some cases, the attributes of the two parties will affect the success of mediation. For example, according to our preliminary analysis, we observed that there are two kinds of people who prefer to resolve a dispute through mediation rather than going to court. One is the office worker, who thinks that going to court is time-consuming. The other is underage children since their guardians would hope to resolve issues through the mediation committee. Therefore, we must consider the backgrounds of both parties if we want to get an accurate prediction result. The context information in the case also influences the result. For example, according to our investigation, the case that includes "property inheritance" is often more difficult to mediate than the case that includes "car accidents." We hope to determine the critical factors that affect the mediation results through text mining in the event description.

We construct our datasets from 5,776 mediation committee cases collected in Tainan, Taiwan from March 2009 to January 2017. The ground truth of each mediation case belongs to one of two labels: success or failure. Table 1 is an example of our collected mediation data instances. Each mediation is composed of several structured data fields, such as the receipt date, the completed date, the type of case, the detail type of the case, mediator ID, accuser IDs, defendant IDs and the event description. We also have another data source, which contains personal information such as gender, address, and date of birth of all mediators, accusers, and defendants. In addition, we have the occupation of the accusers and defendants. We show an example in Table 2. We further show the data statistics in Tables 3 and 4. The majority of the mediation cases belong to the civil type, most of which concern car accidents. The mediation times refer to the number of meetings convened for a case.

The event description in Table 1 is textual information, which contains two parts. The first part is the appeal, which is recorded using the accuser's testimony. The second part includes the mediation result and the detailed process of the mediation. Since in this work we focus on predicting the results of mediation, we can only use data related to appeal to train the model and make the prediction.

Table 1. An Example from our Mediation Datasets

ID	Receipt data	Completed	Type	Type2	Mediator	Accuser	Defendant	…	Event description
9800151	2009/3/19	2009/5/5	Criminal	Car accident	323XXX	216XXX	356XXX	…	At 3:00 pm, March 9, 2008, 216XXX drove his own car (car number: 999-XU) crashed…The mediation is successful.

Table 2. An Example of Personal Information

ID	Name	Role	Gender	Birthday	Address	Occupation
9800151	216XXXX	accusers	male	1987/2/28	Tainan City, Central and Western District….	Teacher

Table 3. Simple Data Statistics

Result		Failure	Success
		2,376	3,400
Type I	Civil	1,677	3,018
	Criminal	699	382
Type II	Car accident	1,307	2,687
	Other	1,069	713
Mediation times	One	1,537	2,327
	More than two	839	1,073

Table 4. Simple Data Statistics for Textual Information

	The number of words
Total of different words	13,093
Average number of words in one case	32.97

4 METHODOLOGY

4.1 Overview of LSTMEnsembler

The overview of our system is shown in Figure 1. First, we divide the original data into two parts. One is case information and the other is textual data. Then, these two kinds of data are fed into our feature engineering component. For the data from case information, we first do some preprocessing and feature extraction. After these manipulations, we use the processed data to train some powerful machine learning models, such as XGBoost and

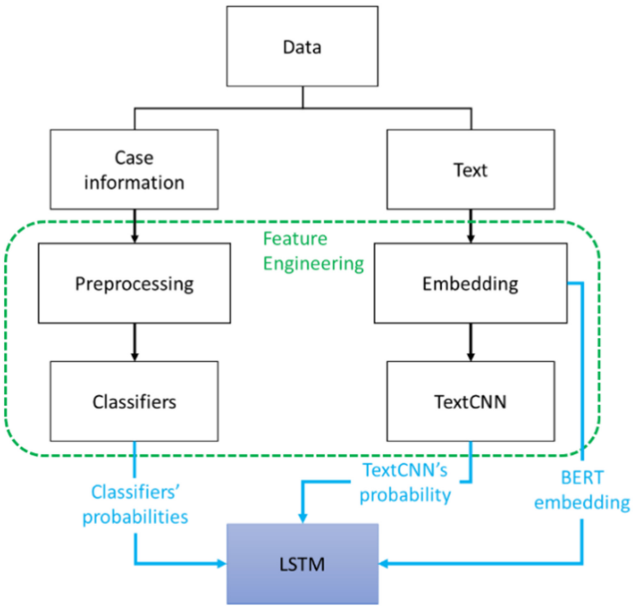

Fig. 1. The overview of LSTMEnsembler.

LightGBM, which will predict the mediation result as a probability distribution. This probability distribution will then be considered as features fed into the LSTM model. For the textual data, we perform a state-of-the-art embedding method, BERT, to extract the embedding vectors of contents. The output embedding vectors will then be utilized in two ways. One is directly feeding them into the LSTM model as feature vectors and the other acts as the features of TextCNN. The reason we perform this operation is to improve the quality of embedding by assembling two different text mining methods. The predictive judgment from TextCNN will also be another feature of LSTM. Lastly, we combine textual vectors, different classifiers' results from CI features, and TextCNN's outcome as LSTM's inputs and obtain the mediation consequence as the output. We will explain each part in detail in the following sections.

4.2 Feature Engineering Framework

We build a feature engineering framework that contains two major modules. The first is feature extraction of case information and the second is a text mining model for processing text information. We will also introduce the classifiers implemented in each part.

4.2.1 Case Information (CI) Features. The case information features are extracted mainly from Tables 1 and 2 except for the event description column. The CI includes the numerical and categorical features related to the mediator, the accusers, and the defendants, and some spatial-temporal attributes of the cases. We introduce them in detail in the following.

General cases and sub-category cases. In our collected data, a type I case has only two kinds: civil and criminal. However, the number of that in type II is 114. According to our observation, there are too many similar kinds just written differently or rarely occurring. In order to reduce the negative effects of low-frequency cases, we divide them into five categories: property, car accident, car accident compensation, injury, and referral from the court based on statistics and opinions of domain experts. The domain experts also suggested that car accidents would be an important indicator for mediation results. Thus, we believe that the sub-category will be an essential categorical feature.

Submission time. Submission time refers to the time period between the occurrence of an event and its application date. We captured the date of the event from the column of text description with regular expressions. A lengthy submission time might indicate that the two parties are quite busy or they have tried to negotiate privately in advance.

Location. For each mediation, the event description contains its address, which refers to the location of the dispute. We believe that the addition of several spatial factors might be beneficial to the prediction model. For example, people living in the country prefer to resolve disputes privately because of the long distance to mediation places. Therefore, in this task, we use villages and districts as location features.

Mediation times. This feature means the number of times convened for this case. Sometimes, the two parties are not satisfied with the preliminary result and will apply for mediation again. Generally, the more time the mediation consumes, the less likely it is to succeed.

The number of participants. We calculate the number of participants in each case since sometimes the number of accusers or defendants is not only one. For example, in the case of a car accident involving multiple cars that collided, multiple drivers would apply for mediation together. We observe that the number of people will negatively affect the mediation result. More people lead to more opinions; therefore, it might be difficult to reach an agreement.

The ratio of male to female for accusers and defendants. We consider that the gender of the two parties might influence the mediation result. Hence, we extract the genders for both accusers and defendants to calculate the ratio of male to female.

The ratio of elders for accusers and defendants. We added a feature to compute the ratio of elders to other age groups in a case. We find that more elderly people make the case difficult to succeed. Experts point out the reason is that elders have more leisure time than other age groups; thus, it is difficult to reach an agreement and more mediation time is required.

The ratio of office workers for parties. Compared with elders, office workers tend not to go to court. We added a feature to compute the ratio of a case containing office workers according to their occupations.

Mediator (ID). This feature refers to the person who presides over the mediation. In our collected data of Tainan, there are 19 mediators in total, most of which are reputable, highly esteemed locals with a passion for helping citizens resolve disputes without remuneration.

Mediator Experience. Since the mediator's experience might affect the result of a mediation, we add a feature that calculates how many similar cases (sub-category) the mediator has handled in the past.

4.2.2 CI Classifiers. In this work, we expect that each classifier is responsible for dealing with its suitable feature set; therefore, several robust classifiers are used to handle different types of features. The CI features are non-text features. We propose to adopt XGBoost and LightGBM to make the predictions because these two classifiers have excellent performance for non-text features according to our experiments.

XGBoost. XGBoost [4] is a scalable tree-boosting system designed for speed and performance. It integrates many previous works on gradient lifting algorithms and has done a lot of optimization in implementation. We choose XGBoost as one of our classifiers because it is a widely used and very popular classifier for Kaggle competitors and data scientists in the industry, as it has been battle-tested for production with many research problems.

LightGBM. LightGBM [11] is a gradient-boosting decision tree algorithm proposed by Microsoft in 2017. Compared with XGBoost, LightGBM can process data with lower memory and higher speed. Therefore, we choose it to combine with XGBoost and enhance our system.

In our initial experiment, the two classifiers can achieve 80% for F-score and 73% for accuracy. The result is shown in Table 5. The performance is not low; however, we found that when the text information (here, we utilize the embedding vectors from BERT) was included, the result of XGBoost and LightGBM dropped to 79%. It seems that XGBoost and LightGBM are not suitable for dealing with textual information. Therefore, we consider

Table 5. The Performance of XGBoost and LightGBM

	F-score	Accuracy
XGBoost (CI)	0.802113	0.732007
XGBoost (CI+Text)	0.799704	0.730277
LightGBM (CI)	0.800711	0.732007
LightGBM (CI+Text)	0.7904	0.723

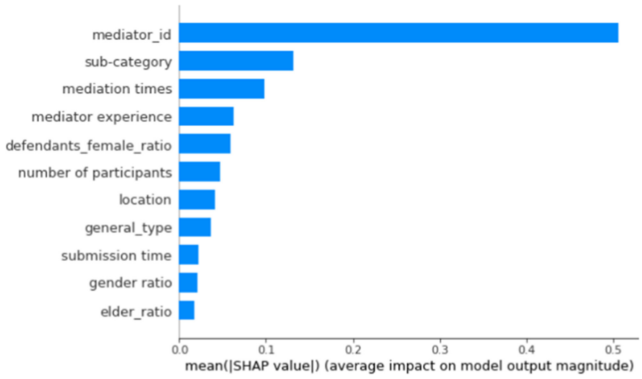

Fig. 2. The feature importance of CI.

adopting other modeling methods to handle this issue. After training and testing, XGBoost and LightGBM will generate the probabilities of success rate for each case, acting as a part of LSTM input features. We also report the feature importance (SHAP [**SH**apley **A**dditive ex**P**lanations] value) of XGBoost for relatively important CI features in Figure 2. SHAP is probably state-of-the-art in machine learning explainability. This algorithm was first published in 2017 by Lundberg and Lee [17]. What SHAP does is quantify the contribution that each feature brings to the prediction made by the model. In Figure 2, we find that the most influential factor is the mediator ID. This phenomenon is matched with the conclusions from interviews with domain experts.

4.2.3 Text Information. Each case has textual information describing the details of the dispute. It is worthy of investigating the content of the dispute because it includes the argument of both parties and the objective fact stated by the third party. In order to increase the diversity of LSTMEnsembler to effectively process text description, we employ two well-known models to generate embedding features from the text description. One is Word2vec and the other is BERT. We describe these models next.

Word2vec: Word2vec [20] is a well-known method for learning word vectors. It can be divided into two models: CBOW and Skip-gram. CBOW predicts word vectors given the context, while Skip-gram predicts context given words. Since all textual descriptions of our data are written in Chinese, we employ Jieba[1] for word segmentation first. Then, we use these two models to get the different vectors of words in each case. After using the Word2vec model to generate the word vector for each case, we average the vectors of all words in each case. Finally, in our experiments, we find that when Skip-gram is used with 150-dimensional vectors, the best result can be obtained. As a result, we adopt the vector generated from Skip-gram as input for TextCNN.

BERT: BERT [5] is a novel model developed by Google. It can be used to deal with a variety of NLP problems and has achieved excellent results in multiple tasks [18, 21]. In this article, we exploit the "BERT-Base, Chinese"

[1]https://github.com/fxsjy/jieba.

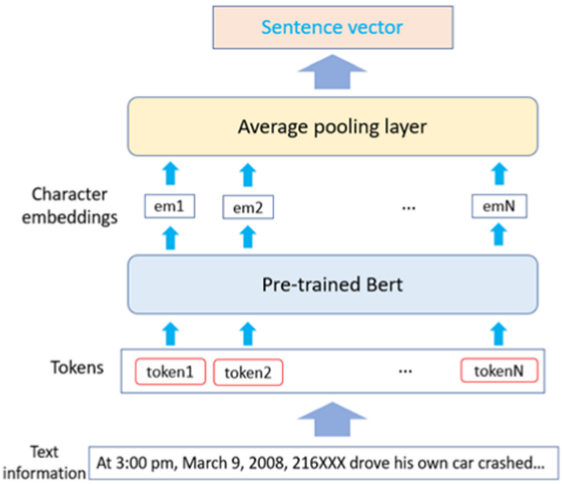

Fig. 3. The architecture of extracting embeddings from BERT.

Table 6. Results of Different Past Cases

Past case	F-score	Accuracy
TextCNN	0.76254	0.72342
BERT	0.64253	0.60244

pre-trained model[2] that Google provides to extract textual features. The architecture is shown in Figure 3. Unlike Word2vec, using a word as a token, we treat each Chinese character as a token in BERT. BERT has worse performance for segmented words according to our experiments. Therefore, there is no need to apply Jieba before applying BERT. After performing BERT embedding, we obtain the embedding vector of each character in the textual information; their dimension is 768, which is set by the pre-trained model. However, we have a large amount of textual information from more than 5,000 cases; it would be too large to use 768 dimensions for each character. Therefore, we add an average pooling layer to average the vectors of all of the characters in each case. Finally, for the textual information in each case, we get a 768-dimensional BERT vector.

4.2.4 TextCNN. TextCNN [13, 30] is a model using a CNN for text classification. TextCNN utilizes three kernels of different sizes to extract critical information from the text. Then, max-pooling is performed to extract the largest value from the feature vector obtained after convolution. Finally, it connects a fully connected softmax layer to output the probability of each category. We use the output embedding vector of Word2Vec as the input of TextCNN and finally take the output of TextCNN, a probability distribution, as a feature of LSTM. The three kernel sizes of TextCNN are set as 2, 3, and 4 in our structure, which is the same as the setting shown in [30].

People know that BERT can also be used for classification. To preliminarily evaluate the classification effectiveness of BERT and TextCNN, we train and compare their performances using only textual contents and show the result in Table 6. However, in Table 6, BERT's performance is not as good as TextCNN's. We suppose that this is because common equipment cannot afford the cost of retraining BERT on our data. Thus, we directly use the BERT model trained on Wiki data by Google. In the end, we choose only TextCNN for text classification. However, we still consider the BERT embedding in our final model.

[2]https://github.com/google-research/bert.

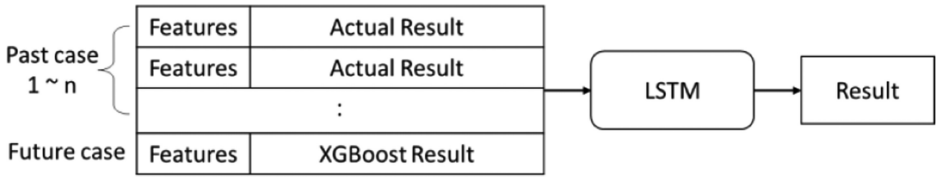

Fig. 4. The structure of LSTM.

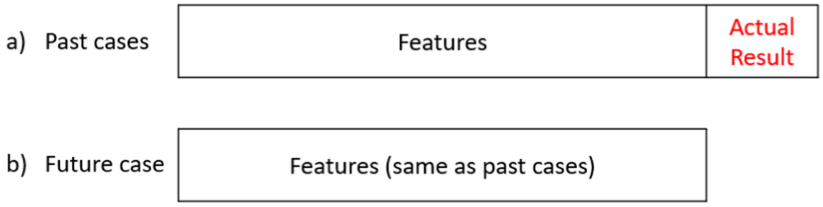

Fig. 5. The structure of past case and future case features.

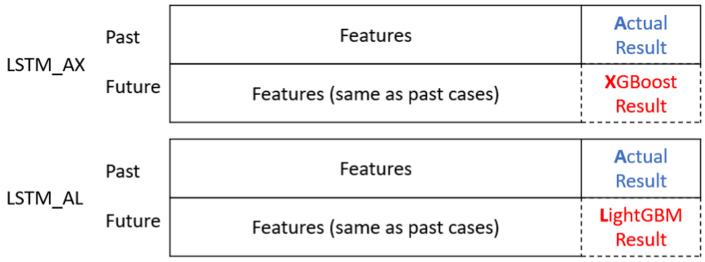

Fig. 6. Balancing the input length of LSTM using XGBoost and LightGBM.

4.3 Long Short-Term Memory Modeling

We find that real-world mediation cases have high temporal dependencies. Mediators can learn from experience and enhance their negotiating ability when involved in increasing numbers of cases as time goes on. We propose adopting LSTM to combine the results from all classifiers and utilize past cases' information to predict the outcome of future cases. LSTM [9] is a powerful recurrent neural network. It can learn temporal dependency between mediation cases. In this work, we mainly use the past n cases' information to predict the outcome of the next future case, where $n \in N$. According to our experiments, prediction accuracy is generally good from $n = 1$ to $n = 10$. The LSTM inference structure of this work is shown in Figure 4.

We should note that directly applying LSTM to make the prediction for the future case will encounter a limitation, in which the size of features must be the same for both past cases and future cases. However, we can see in Figure 5 that the feature size of the past case will be longer than that of the future case because the target case to predict will not have the genuine result. On the other hand, directly ignoring the ground truths of previous cases is also not applicable since these previous "answers" are highly correlated with the ground truth of the future case.

We claim that "Actual Results" must be considered in our LSTM-based approach, but we need to conquer the limitation of LSTM mentioned earlier. We propose utilizing XGBoost's (or LightGBM's) predictive outcome to fill into LSTM's future case to maintain the length. We name the updated LSTM as LSTM_AX (using **A**ctual results for past cases and **X**GBoost's result for future cases) and LSTM_AL (using **A**ctual results for past cases and **L**ightGBM's result for future cases). These structures are shown in Figure 6 and we evaluate whether LSTM_AX

Table 7. Results of Different Past Cases

Past case	F-score	Accuracy
1	0.831168	0.752166
5	**0.841212**	**0.771379**
10	0.840989	0.765625
15	0.81697	0.737847

and LSTM_AL can perform better than single classifiers in Section 5.2. We also verify that including "Actual Results" can help boost the effectiveness of prediction. Finally, we choose LSTM_AX as our main ensemble model of LSTMEnsembler because LSTM_AX has a slightly better performance than LSTM_AL in the experiments. LSTM_AX takes the combination of BERT's textual vectors, the predictive results of XGBoost and LightGBM classifiers, and TextCNN's outcome as input and makes the final prediction, indicating whether the mediation will succeed or not.

5 EXPERIMENTS

For the LSTM structure we use in this work, we adopt a single LSTM layer with 30 neurons and a fully connected sigmoid layer. Our LSTM uses 50-dimensional hidden representations and memory cells. We use a forget bias of 2.5 to model long-range dependencies, the Adadelta method to optimize the parameters, and a learning rate of 0.01. We conduct several experiments with different combinations of predictive results (probability distributions) and BERT's embedding vectors to train an LSTM model and predict the future testing data. We evaluate the effectiveness using F-score and accuracy metrics. To evaluate the effectiveness, we sort the data chronologically and select the first 80% of cases for training, the next 10% of cases for validation, and the remaining 10% of cases for testing.

5.1 LSTM Settings

In the beginning, we aim to find out how many past days considered will benefit our prediction performance of LSTMEnsembler. We conduct the experiments by changing the number of cases in the past for each LSTM's prediction instance. The results are shown in Table 7. It shows that the best result is obtained when the number of past cases is five. Then, the effectiveness of increasing the number of past cases gets worse but not obvious. Therefore, in the following experiments, we take the past five cases into consideration and make a prediction for the next case.

5.2 Overall Evaluation

In this section, we evaluate our proposed LSTMEnsembler compared with other baseline methods and other robust classifiers.

Baseline. We create a baseline to prove that machine learning works for the mediation classification problem. Since mediation cases can be sorted by application time and mediators can accumulate their experience by being involved in an increasing number of cases, we propose a reliable but intuitive baseline. For each new case and its sub-category (i.e., property, car accident, car accident compensation, injury, and referral), we consider the mediator's success rate of past cases in each sub-category. If the mediator's previous success rate exceeds 0.5 for one of the sub-categories, the baseline model guesses that the mediator will succeed for the next case corresponding to that sub-category. Therefore, we sort cases by application time and can generate the prediction for each case based on the mediator's dynamic success rate mentioned earlier. In the beginning, the success rate for each sub-category is initialized as zero. Then, the success rate increases when a successful case is encountered.

Table 8. The Baseline Example of a Certain Mediator

ID	Sub-category	Success rate of past cases	Predict by success rate	Ground truth
0	Car accident	0%	0	1
1	Car accident	100%	1	0
2	Car accident	50%	0	1
3	Property	0%	0	0
4	Car accident	66.67%	1	1

Table 9. Result of Each Classifier

	F-score	Accuracy
Baseline	0.727905	0.648199
LSTMEnsembler	0.855791	0.788561
Personal-LSTMEnsembler	0.812783	0.741362
XGBoost	0.802732	0.735467
LightGBM	0.800711	0.732007

An example of a mediator is shown in Table 8. In the mediator's first case, the baseline method will predict failure because the individual does not have experience serving as a mediator. However, after the individual successfully deals with the first case, we tend to predict that the mediator will succeed in the second case. For the mediator's fourth case of the car accident, the mediator's previous success rate is 66.67% since the individual has experienced two success cases and one failure case. On the basis of the result, we predict that the mediator will succeed again.

Comparative Methods. We compare LSTMEnsembler with two robust classifiers, XGBoost and LightGBM. The features used in XGBoost and LightGBM include embedding vectors from BERT, TextCNN's inferred probability, and all features of case information. XGBoost will also take LightGBM's inferred probability as its feature and vice versa. We propose that these two classifiers be compared because we would like to verify that our LSTMEnsembler is better than other single classifiers for handling temporal dependency and heterogeneous features. On the other hand, we also include another competitor, called Personal-LSTMEnsembler, which builds an LSTMEnsembler model for each mediator (not for all mediators) and considers only the past experiences of each one to make predictions.

Results. We show LSTMEnsembler and other competitors in Table 9. First, the baseline achieves 72% in F-score and 64% in accuracy. This baseline is utilized for modeling mediators' experiences but does not consider the detailed properties of cases. The result of Personal-LSTMEnsembler shows that it is worse than the proposed LSTMEnsembler method. We believe that it is because there are not enough training samples for some mediators to train the models adequately. Among all of the results, our system achieves the best results and proves that when each classifier does what it is good at, LSTMEnsembler can achieve the best performance.

5.3 Experiments of LSTM

In this section, we would like to evaluate whether it is correct to involve "Actual Results" of Figure 5 in LSTMEnsembler with different feature combinations. To make the comparison more convincing, we include two more balanced input methods: LSTM_PX and LSTM_PL. LSTM_PX means that the "Actual Results" in the past cases are changed to the predictive probability distributions of XGBoost. The replacement is similar to LSTM_PL.

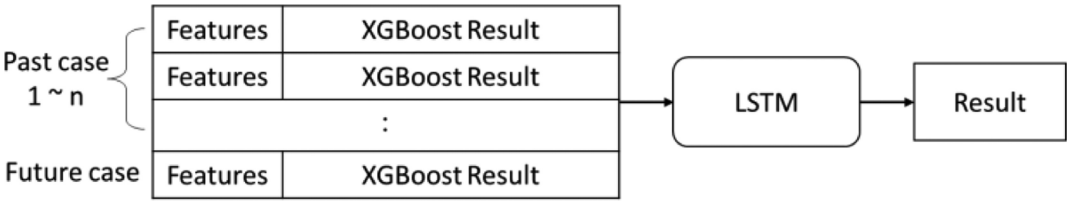

Fig. 7. The LSTM_PX method for balancing length of LSTM using XGBoost.

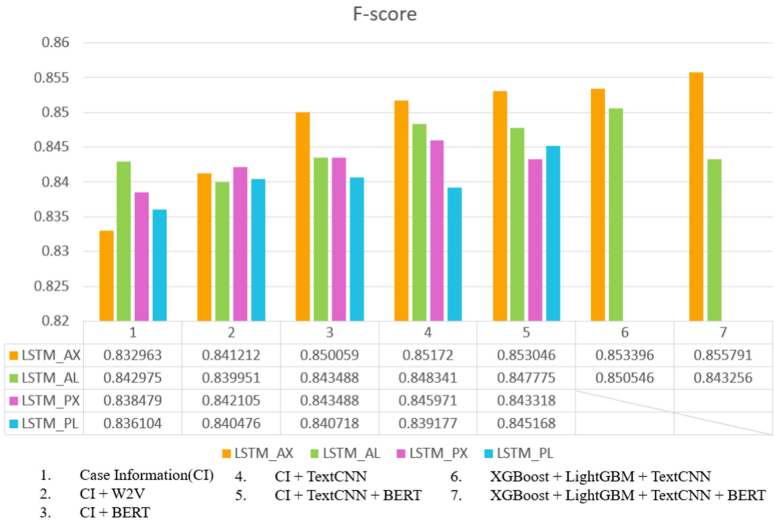

Fig. 8. The result of F-score.

We take XGBoost as an example in Figure 7, which is different from Figure 4; that is, all actual results in the past cases are replaced with XGBoost's inferred results.

We take seven different combinations of input features into the LSTM block in the experiments. The first three combinations simply use CI and text vectors (BERT or Word2Vec) to make predictions. The fourth and fifth consider TextCNN's predictive probability distribution to be a feature of LSTM to enhance the performance. The last two combinations include all classifiers' results but do not consider CI due to the overfitting issue. The seventh includes the BERT vectors. Figures 8 and 9 show the results of these combinations evaluated by F-score and accuracy. For the first three results, we find that adding text information improves the result.

Furthermore, the BERT vectors can enhance more than Word2vec. Thus, we conclude that using characters as a token is more suitable for our work. In addition, when we combine more classifiers' predictive probability distributions, the results improve more. We then add the BERT vectors and find that it improves the F-score, which turns out the best result (the seventh combination) in our framework. Among the four different ways to deal with the ground truth of the case, the LSTM_AX's performance is the best. It proves that using "Actual Results" is useful and better than using classifiers' probability distributions. The experimental result also shows that BERT vectors allow our system to achieve 85.5% in F-score and 78.8% in accuracy. The experiments in this section show that our proposed LSTMEnsembler performs effectively for assembling the inferred results from other classifiers and the BERT's embedding vectors.

Table 10. The Performance of Different Classifiers

		CI	CI+W2V	CI + BERT	CI+ TextCNN	CI+ TextCNN+ BERT
LSTM_AX	F-score	0.8329	0.8412	0.8501	0.8517	**0.853**
	Accuracy	0.7642	0.7714	0.7799	0.7834	**0.7868**
XGBoost	F-score	0.8021	0.7997	0.7995	0.8004	**0.8027**
	Accuracy	0.732	0.7303	0.7268	0.7303	**0.7355**
LightGBM	F-score	**0.8007**	0.7904	0.7888	0.7921	0.7966
	Accuracy	**0.732**	0.723	0.713	0.7216	0.7268

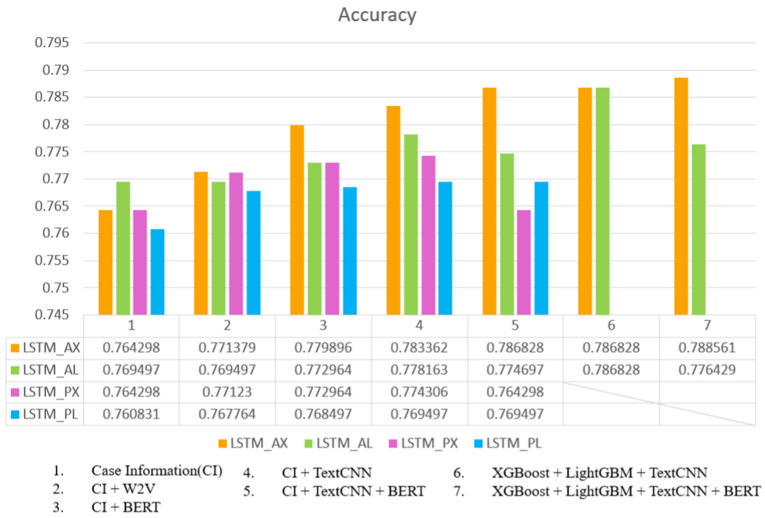

Fig. 9. The result of accuracy.

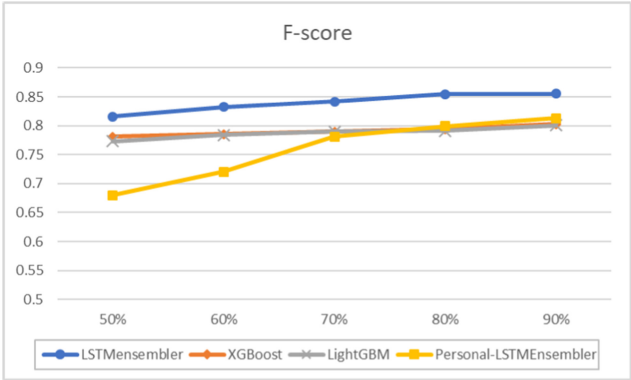

Fig. 10. The F-score results by varying training data size.

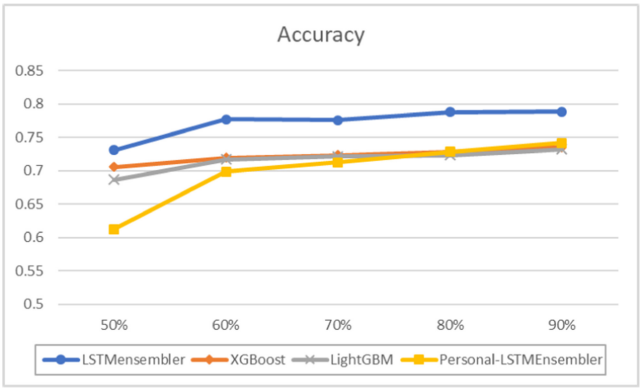

Fig. 11. The accuracy results by varying training data size.

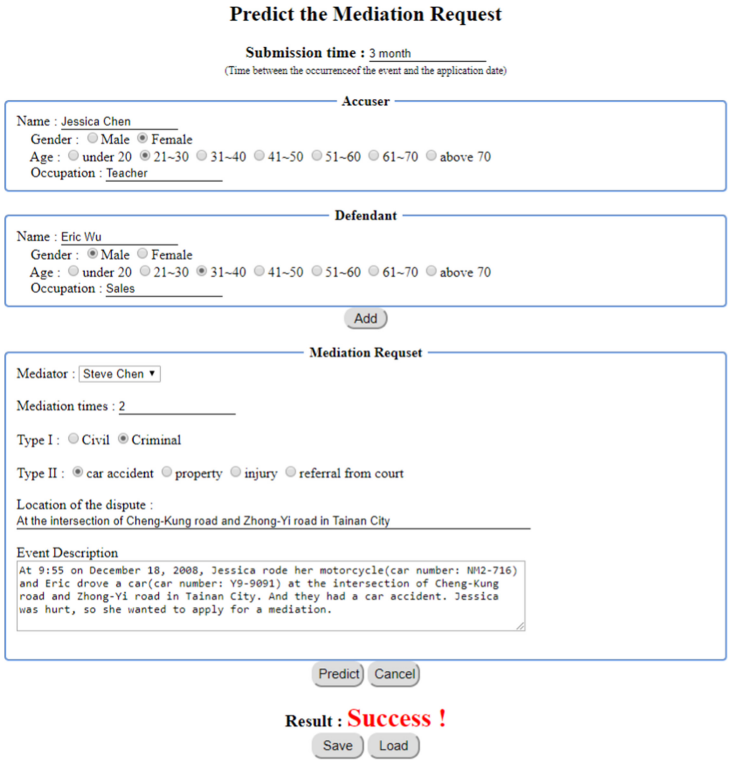

Fig. 12. The system of predicting mediation request.

5.4 Comparison for Different Feature Sets

In this section, we further compare the performances of LSTM_AX, XGBoost, and LightGBM by varying different feature sets but do not consider the predictive probability distributions of each. We would like to verify that modeling temporal dependency between cases can improve performance.

The input feature is the same as the first five feature combinations in Section 5.3. Table 10 shows each performance in F-score and accuracy. XGBoost and LightGBM perform better when considering only CI for input.

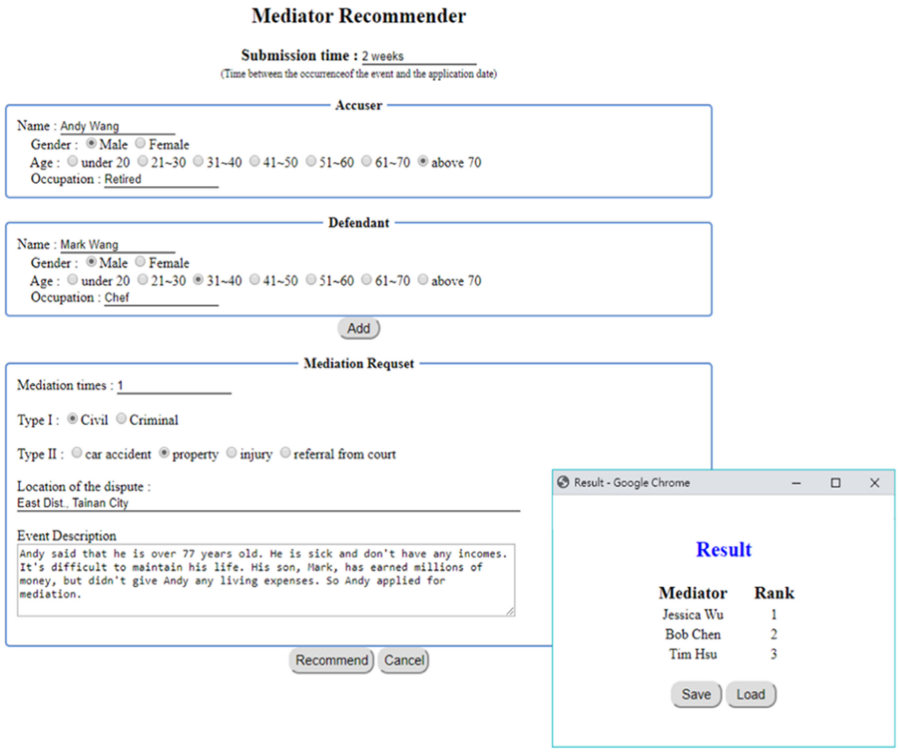

Fig. 13. The system of recommending mediators.

However, when the text vector is included, their results get worse. In contrast, by involving text information in our LSTM block, we can improve performance and obtain a better result.

5.5 The Performance of Varying Training Data Size

We evaluate the performance of our proposed LSTMEnsembler by varying training data size from 50% to 90% compared with other methods. Figure 10 shows the F-score and Figure 11 shows the accuracy. The result shows that the performance of LSTMEnsembler can remain an 81% F-score and 73% accuracy when the training data size is dropped to 50%. However, Personal-LSTMEnsembler drops fast when the training data size is less than 60%. XGBoost and LightGBM also have stable performances but cannot gain the best effectiveness.

6 APPLICATIONS

We have implemented a system of predicting mediation requests for the Tainan City Government in Taiwan. The interface is shown in Figure 12. In addition, we would like to further extend our LSTMEnsembler to develop an intelligent tool that can recommend suitable mediators based on users' mediation requests. This function is quite useful for committees to select the right mediator for different cases. The developed system interface is shown in Figure 13. Furthermore, it saves time for decision-makers to use a ranking list as a reference, which displays the ranking of each candidate mediator based on their predictive probability of success.

For each new mediation request, the system applies the LSTMEnsembler model to predict whether the case will be successfully mediated by different mediators. Finally, we get the recommended ranking list by sorting the success rate of each mediator in the same case.

7 CONCLUSION AND FUTURE WORK

In this work, we focus on predicting the success of mediation cases based on the case information and textual descriptions. For the effectiveness of the experiments, we use LSTM to assemble different classifiers' inferred results. The experiments show that the combination of the features generated by the two kinds of information indeed contributes to the predicting ability of the model. In additioin, the experiments show that the performances improve when the ensemble approach is applied and each classifier does its own job. For the text-mining task, we use Chinese characters as a token to extract text features. The experimental results show that the text vector has a positive contribution to the model skill, especially the BERT model. Furthermore, among all combinations we use, we find that assembling three different classifiers' predictive results with the BERT vectors achieves the best performance. In the future, we would like to increase the explainability of our framework and raise the performance. One direction is to develop an attention mechanism that learns the importance of each aspect of feature representation. How to generate a reliable recommended ranking of mediators for each case is also an interesting and practical topic.

ACKNOWLEDGMENTS

We are grateful to Tainan City Government for providing the mediation data.

REFERENCES

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. 2016. Predicting judicial decisions of the European court of human rights: A natural language processing perspective. *Peer J Computer Science* 2 (2016), e93.

[2] Y. Bi, S. Wang, and Z. Fan. 2020. A hybrid BERT and LightGBM based model for predicting emotion GIF categories on Twitter. *arXiv preprint* arXiv:2008.06176.

[3] Y. Cai, H. Cai, and X. Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2506–2515.

[4] T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*. 785–794.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] F. Kort. 1957. Predicting Supreme Court decisions mathematically: A quantitative analysis of the ''right to counsel'' cases. *American Political Science Review* 51, 1 (1957), 1–12.

[7] Z. Gao, A. Feng, X. Song, and X. Wu. 2019. Target-dependent sentiment classification with BERT. *IEEE Access* 7 (2019), 154290–154299.

[8] Z. He, Z. He, J. Wu, and Z. Yang. 2019. Feature construction for posts and users combined with LightGBM for social media popularity prediction. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2672–2676.

[9] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[10] Z. Hu, X. Li, C. Tu, Z. Liu, and M. G. Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*. 487–498.

[11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30 (2017), 3149–3157.

[12] R. Keown. 1980. Mathematical models for legal prediction. *Computer/LJ* 2:829.

[13] Y. Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[14] Z. Li, Q. Zhang, Y. Wang, and S. Wang. 2020. Social media rumor refuter feature analysis and crowd identification based on XGBoost and NLP. *Applied Sciences* 10, 14 (2020), 4711.

[15] W.-C. Lin, T.-T. Kuo, T.-J. Chang, C.-A. Yen, C.-J. Chen, and S.-D. Lin. 2012. Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction. *ROCLING*. 140–141.

[16] Y. Liu, X. Wang, and W. Long. 2019. Detection of false Weibo repost based on XGBoost. In *IEEE/WIC/ACM International Conference on Web Intelligence—Companion* (2019) 97–105.

[17] S. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4768–4777.

[18] Z. Lv, D. Liu, H. Sun, X. Liang, T. Lei, Z. Shi, F. Zhu, and L. Yang. 2019. AUTOHOME-ORCA at SemEval-2019 Task 8: Application of BERT for fact-checking in community forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 870–876.

[19] Y. Meng, X. Li, X. Sun, Q. Han, A. Yuan, and J. Li. 2019. Is word segmentation necessary for deep learning of Chinese representations? arXiv preprint arXiv:1905.05526.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.
[21] M. Munikar, S. Shakya, and A. Shrestha. 2019. Fine-grained sentiment classification using BERT. In *Artificial Intelligence for Transforming Business and Society (AITB), Vol. 1*. IEEE, 1–5.
[22] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L. P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 284–288.
[23] K. Shu, S. Wang, and H. Liu. 2018. Understanding user profiles on social media for fake news detection. 2018. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'18)*. 430–435.
[24] Z. Song, Y. Xie, W. Huang, and H. Wang. 2019. Classification of traditional Chinese medicine cases based on character-level BERT and deep learning. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. IEEE, 1383–1387.
[25] S. S. Nagel. 1963. Applying correlation analysis to case prediction. *Texas Law Review* 42 (1963), 1006.
[26] O. M. Sulea, M. Zampieri, M. Vela, and J. V. Genabith. 2017. Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306.
[27] Y. Sim, B. R. Routledge, and N. A. Smith. 2015. The utility of text: The case of amicus briefs and the Supreme Court. In *29th AAAI Conference on Artificial Intelligence*.
[28] A. Talun, P. Drozda, L. Bukowski, and R. Scherer. 2020. FastText and XGBoost content-based classification for employment web scraping. In *International Conference on Artificial Intelligence and Soft Computing*. Springer, Cham, 435–444.
[29] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 3584–3593.
[30] Y. Zhang and B. C. Wallace. 2016. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.
[31] T. Zhu, Y. Wang, H. Li, Y. Wu, X. He, and B. Zhou. Multimodal joint attribute prediction and value extraction for e-commerce product. *arXiv preprint* arXiv:2009.07162.

Received October 2020; revised March 2021; accepted June 2021