ENHANCING MULTILABEL CLASSIFICATION IN CHARGE PREDICTION USING LABEL CORRELATION AND PROBLEM TRANSFORMATION METHOD

***Nasa Zata Dina*[1], *Sri Devi Ravana*[2*], *Norisma Idris*[3]**

[1, 2] Dept. Information Systems, Faculty of Computer Science and Information Technology, Universiti Malaya,
Kuala Lumpur, Malaysia

[3] Dept. Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti Malaya,
Kuala Lumpur, Malaysia

Emails: nasazata@gmail.com[1], sdevi@um.edu.my[2*], norisma@um.edu.my[3]

ABSTRACT

Legal Judgment Prediction (LJP) has recently gained significant interest from both academic and legal practitioners. The majority of LJP methods focus on single label prediction problem, neglecting the real-world multilabel case. Therefore, this study aimed to classify multilabel legal cases using label correlation and problem transformation methods. Data were collected from a publicly accessible legal document in the European Court of Human Rights (ECHR) and EUR-Lex. Multilabel text classification tasks face challenges such as sample diversity, complexity, and the need for effective utilization of label correlations. In this paper, we propose a model that integrates domain specific text embedding and label correlation. Proposed model leverages label powerset as problem transformation to transform a multilabel problem to a multiclass problem by incorporating domain specific text embedding and label correlation, which enhances classification performance in charge prediction and addresses label omission issues. Extensive experiments on two legal text datasets demonstrate the model's excellent performance. The proposed model substantially outperformed two baseline studies by attaining competitive results of 80.32%-90.09% F1-score and 0.0119-0.0210 Hamming Loss score, respectively. Meanwhile, the baseline models have attained 52%-80% F1-score and 0.0452-0.1479 Hamming Loss score. Proposed model's performance significantly surpasses the baseline models. The significance of this study is the implementation of label correlation in label powerset problem transformation method and the application of domain specific embedding to solve multilabel classification problem in legal domain.

Keywords: Legal documents; Legal judgment prediction; Label correlation; Multilabel classification; Natural language processing, Problem transformation.

1.0 INTRODUCTION

Recent advances in Natural Language Processing (NLP) have led to a surge in interest in text classification [1]. Text classification techniques aim to assign a predetermined label to a given input text [2]. Labeling text data is quite time-consuming but essential for automatic text classification [3]. Especially, manually assigning multiple labels for each document may become impractical when a very large amount of data is needed for training multilabel text classifiers. Although it takes a lot of effort, text data labeling is necessary for automatic text classification [4]. In particular, when training multilabel text classifiers requires a very large amount of data, manually constructing numerous labels for each document requires a lot of effort to create accurate text classification systems [3].

Text classification can be binary, and multilabel. In a binary classification, the target label has only two possible values while multilabel text classification has more than two possible values. Binary classification frequently ignores other crucial information in favor of classifying a document according to its primary subject [5]. The binary classification may not work well for texts that cover multiple categories or labels [6]. In real-life and industrial applications, multilabel classification is an important task in the field of NLP. The difference between multilabel classification and binary classification is the number of labels that can be assigned to one instance. Multilabel classification is more complex than binary classification [7]. The reason is that once the category of an instance is determined, a human judge can identify the label as a single label resulting from a binary classification. However, in the multilabel problem, human judges must determine each instance's possible category. As a result, labelling multilabel data requires a lot more work than labelling single-label data. Research on multi-label classification is limited, despite the importance and usefulness of this problem. The multilabel classification improves the model’s capacity for generalization in addition to offering more thorough and richer information extraction. It is more in line with the requirements of real-world applications, particularly in situations like recommendation systems and decision support systems where it is crucial to identify multiple labels or categories.

There are three main issues with multilabel classification problems. First, label confusion may result from the model's inability to predict all relevant labels for certain samples due to the large label feature space's added complexity [8]. Larger label space causes unequal label distribution, with certain labels showing up more frequently than others [9]. Label skew describes the possibility of a large and low number of cases being assigned to more and less frequent label sets, respectively [10][11][12][13][14]. Second, it is a high dimensionality problem where both feature and label spaces have enormous dimensions [15]. The number of labels in a dataset can be more than instances. Therefore, processing time for high-dimensional data takes a lot of time, suggesting potential overfitting in the model [16][17][18][19]. Lastly, errors and noise in labeled data may arise from insufficient or inaccurate manual labeling, which may have an adverse effect on model performance and training [20].

In legal domain, text classification models are crucial for intelligent legal support [21], decision-support assistance [22], and legal documents management [23]. In the end, these models improve judicial impartiality by making legal study and analysis easier. Legal practitioners may obtain more effective, precise, and thorough legal information and services by implementing these ideas and models into practice in the legal domain. Classification in legal domain is the process of assigning legal documents to label based on content, purpose, or relevance. It is essential in document management for courts and law firms, facilitating efficient organization, retrieval, and review. The application of classification in the legal domain is relatively under-explored, although the data about legal cases can be accessed publicly. Classification in legal domain is commonly known as Legal Judgment Prediction (LJP), which offers different multi-tasks [24]. The first task is judgment prediction where a binary classification is implemented to predict the judgment of a legal case by identifying the patterns of past documents [25][26]. The judgment prediction is considered as a binary classification, determining whether the defendants violated the law or not. The second task is charge prediction, where it predicts which law is violated in a legal case. The charge prediction cannot be considered as a binary classification because each legal case has the potential to be applied simultaneously to more than one different charge. Instead of binary classification, charge prediction is part of multilabel classification focusing on assigning multiple violated law labels to related legal cases. In particular, multilabel classification aimed to concurrently associate cases to several labels.

Existing studies [27] [28] [29] [30] [31] [32] [33] mainly focused on developing charge prediction models without considering whether it is a multilabel problem or not. This is demonstrated by the fact that even though charge prediction in the real-world problems is associated with several labels, existing studies solved multilabel problem by implementing binary classification. It causes existing studies to be unable to predict multiple labels at once [34] and the model becomes impractical and ineffective. Other than the inability to predict multilabel at once, the challenges faced in existing studies, including high dimensionality of label space in multilabel problem [27] [28] [29] [34] [35] [36] and the independence problem among labels, can be difficult to resolve since multilabel classifier process multiple labels simultaneously and this affects the performance of classifier and introduces loss [28] [33] [37] [38] [39] [40] [41] [42]. Although some solutions have been proposed, they limit the number of predicted labels to a maximum of three labels [28] [29] ignoring real-world multilabel data in legal domain that has an infinite number of labels to predict and also their charge prediction models are still considered to have poor performance, only 52-79% accuracy. Based on the gaps, there is need for improvement in multilabel classification in the legal domain. Therefore, this study aimed to propose a multilabel classification model ascertaining which charges have been infringed. The main objectives are as follows:

- to develop a ML-based multilabel classification model that can predict multiple charges of a legal case using label correlation and problem transformation,
- to evaluate the model in comparison to the baseline model.

The main contributions of this study are:

- introducing multilabel classification that combines label correlation and problem transformation, and it increases generalizability in multiple contexts, where multilabel classification is applied,
- applying label correlation in problem transformation method to preserve pre-existing label correlations based on its occurrence.

The rest of the study is organized as follows. Section 2 presents the method including the details of the dataset and proposed model. Section 3 contains the evaluation of proposed model, comparing several baseline studies and discussing the results. Moreover, Section 4 presents the conclusion of the study.

2.0 METHOD

This section presents the method of multilabel classification using label correlation and problem transformation to perform multilabel classification task in the legal domain. To validate the effectiveness of the proposed model, we carried out several experiments. We begin by evaluating how well various models perform on two datasets of legal texts. We examine how problem transformation and label correlation affect multilabel classification. Lastly, we employ multilabel legal text datasets to compare proposed model with state-of-the-art models.

2.1 Datasets

The two datasets used in this study were legal documents, with details as follows:

2.1.1 Dataset 1: EUR-Lex

A collection of documents pertaining to EU law is called EUR-Lex, which is published at http://eur-lex.europa.eu/. The EU Publications Office has annotated each of EU regulations with multilabel. This study used the dataset of Chalkidis et al. [43] which included 57,000 EU legislation documents from EUR-Lex, specifically the English part of the datasets with an average length of 727 words per data.

2.1.2 Dataset 2: ECHR

The second dataset comes from the publicly available data published by the European Court of Human Rights (ECHR). The dataset can be accessed publicly and downloaded at https://echr-opendata.eu/. The number of legal judgment documents used is 2,000, each document with 32,518 words per document. The ECHR was founded in 1959, handling petitions from both individuals and states alleging violations of several rights outlined in the European Convention on Human Rights. In this study, the objective was to predict which law articles were the cases that were violated.

2.2 Study Design

The detailed description of the study design is shown in Fig. 1. The method consists of six stages, namely (a) data preprocessing for noise removal, normalization, and cleaning, (b) label correlation, (c) problem transformation, (d) multilabel classification algorithm, (e) *k*-fold cross-validation, and (f) evaluation metrics.

Based on the analysis, the acquired data was extracted from collective sources. Fig. 2 shows the transformation from a legal document to a data frame in Python. It was claimed further that a particular state had violated the provisions of the European Convention on Human Rights and was heard by the ECHR. Therefore, the main aim was to predict which procedures (multilabel classification) were violated in relation to human rights. The public dataset from the

ECHR was based on claims that human rights laws had been violated. The unmodified dataset form was provided since this study did not focus on algorithmic biases. Additionally, the anonymized version of the dataset had minimal impact on classification outcomes. The selected ECHR dataset included 2,000 public instances that breached several legal articles. Each case contained a text detailing the facts, including additional data elements. Considering this perspective, facts were extracted from every court filing. To obtain more information from the accessible ECHR database, data scraping was adopted. An issue encountered during the process was the inability to delete superfluous content without using regular expressions or regex. The initial deletion of all contents, while retaining THE FACTS through the next phase, was regarded as a relevant legal framework and practice.

2.2.1 Data Preprocessing

Data preprocessing transforms unstructured data into a more easily and effectively processed format in data mining, ML, and other science tasks [44]. It is the process of cleaning and preparing text data, which includes several preprocessing steps. These include (1) Noise removal: the process eliminates URLs, punctuation, special characters, and any blank space, punctuation, and numbers, enabling the proposed model to easily detect the patterns in the text, (2) Lower case conversion is the process of converting all text data to a uniform lower case. This process prompts consistency while removing case-sensitivity from comparisons, and (3) Semantic Word Embedding uses pre-trained word embedding to replace original words in a sentence with their closest neighbors. This enhances the representation of semantic meaning in Natural Language Processing (NLP) tasks, which are converted into numeric vectors. Semantic word embedding applied in this study is Legal-BERT sentence embedding (LBERT). An adaptation of BERT for the legal domain is called LBERT, which includes pre-training on a variety of English legal text fields, such as contracts, court cases, and laws [43][45]. The sub-word vocabulary of LBERT is built from scratch to support legal terminology.

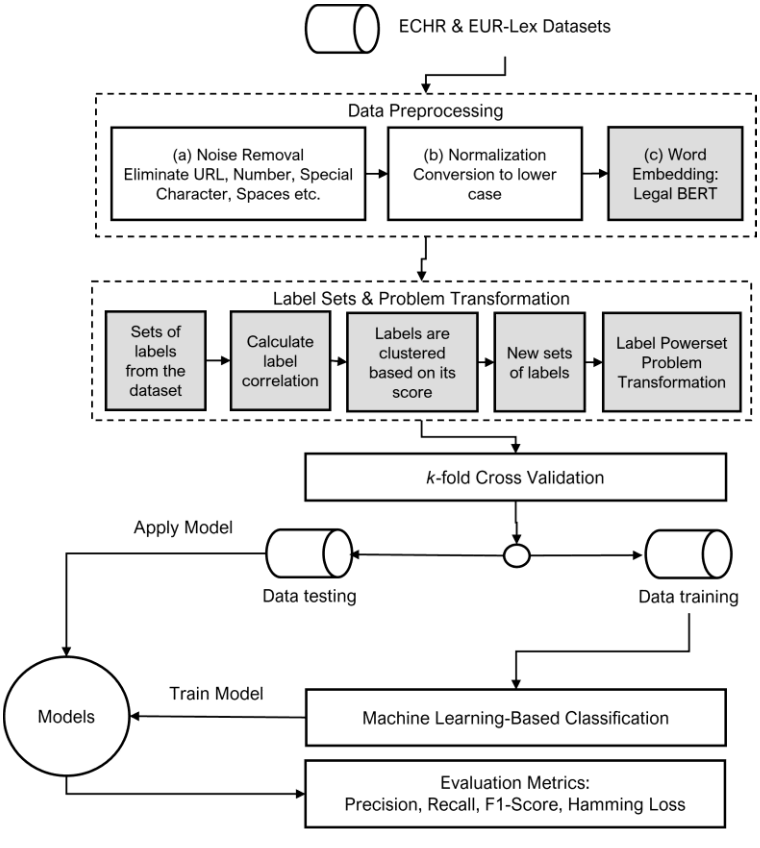

ECHR & EUR-Lex Datasets
Data Preprocessing
(a) Noise Removal Eliminate URL, Number, Special Character, Spaces etc.
(b) Normalization Conversion to lower case
(c) Word Embedding: Legal BERT
Label Sets & Problem Transformation
Sets of labels from the dataset
Calculate label correlation
Labels are clustered based on its score
New sets of labels
Label Powerset Problem Transformation
k-fold Cross Validation
Apply Model
Data testing
Data training
Models
Train Model
Machine Learning-Based Classification
Evaluation Metrics: Precision, Recall, F1-Score, Hamming Loss

Fig. 1: Study Design of multilabel classification in LJP

2.2.2 Label correlation and problem transformation

The summary of label correlation and problem transformation is shown in Fig. 2. Initially, correlation matrix including all pairwise correlations between the dataset's label is created using the Jaccard coefficient. This matrix is subjected to an agglomerative hierarchical clustering, which produces a dendrogram showing the relationship between correlation. To identify model label correlation, the proposed model initially calculates the pairwise similarity degrees of each label using Jaccard, as shown in Table 1 and label space is composed of three labels in this example.

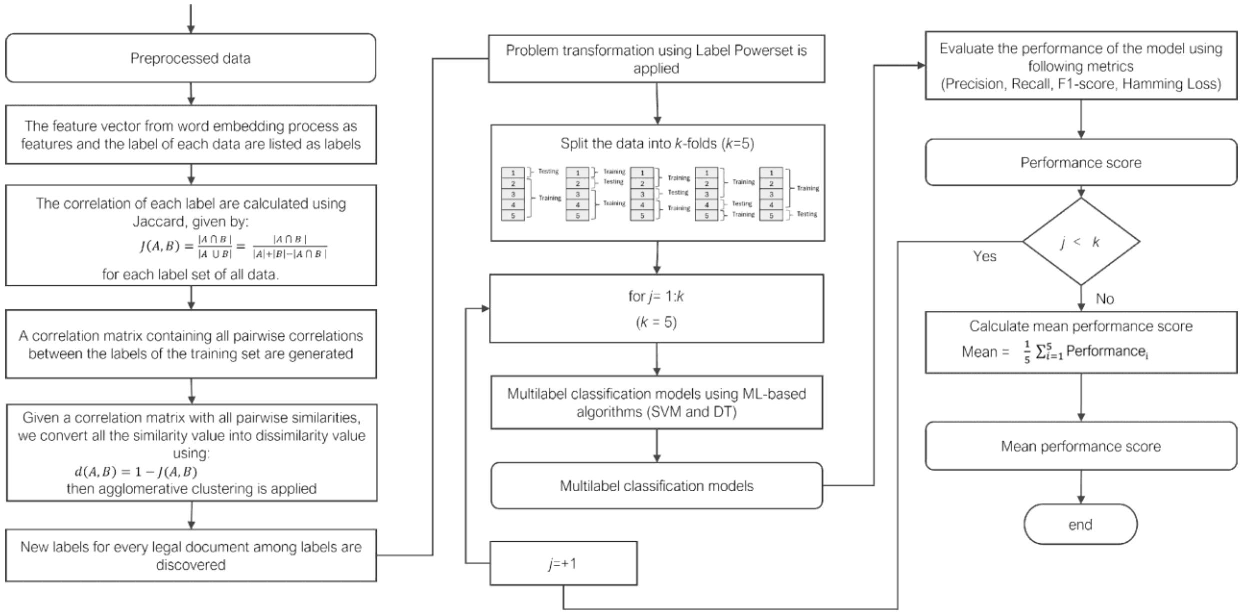

Preprocessed data
The feature vector from word embedding process as features and the label of each data are listed as labels
The correlation of each label are calculated using Jaccard, given by:
$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}$
for each label set of all data.
A correlation matrix containing all pairwise correlations between the labels of the training set are generated
Given a correlation matrix with all pairwise similarities, we convert all the similarity value into dissimilarity value using:
$d(A,B) = 1 - J(A,B)$
then agglomerative clustering is applied
New labels for every legal document among labels are discovered
Problem transformation using Label Powerset is applied
Split the data into k-folds (k=5)
1 2 3 4 5
Testing
Training
for j= 1:k
(k = 5)
Multilabel classification models using ML-based algorithms (SVM and DT)
Multilabel classification models
j=+1
Evaluate the performance of the model using following metrics (Precision, Recall, F1-score, Hamming Loss)
Performance score
$j < k$
Yes
No
Calculate mean performance score
Mean = $\frac{1}{5}\sum_{i=1}^{5}$ Performance$_i$
Mean performance score
end

Fig. 2: Summary of multilabel classification using label correlation and problem transformation method

Table 1: An example of label space with three labels

	Y_i = *Label*$_i$		
X_i=*Instance*$_i$	Y_1=*Label*$_1$	Y_2=*Label*$_2$	Y_3=*Label*$_3$
X_1	1	0	1
X_2	1	1	1
X_3	0	1	0
X_4	1	0	0
X_5	1	0	1

This study finds a hybrid partition by exploring the label correlations. Initially, the Jaccard coefficient is used to construct a correlation matrix containing all pairwise correlations between the labels of the training set as seen in Eq. (1).

$$J(A,B) = \frac{|A \cap B|}{|A \cup \quad B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} \tag{1}$$

Label set Y_i linked to the *i*-th instance X_i of multilabel dataset is represented by each row in Table 1. Therefore, a binary matrix M represents label space, with cell $M_{i;j}$ receiving a value of 1 when instance X_i is given to class Y_j and 0 otherwise. The Jaccard coefficient, as shown in Eq. (1), is used to determine the pairwise similarity of two instances based on labels. Based on the results, a similarity matrix is obtained as shown in Table 2 after calculating Jaccard coefficient to quantify all label pairwise similarities. The Jaccard coefficient has a value between 0 and 1 that measures the similarity between two sets of data. The 0 value means the sets have no overlap while the 1 value means the sets are identical.

Table 2: Pairwise similarity

	Label$_1$	*Label$_2$*	*Label$_3$*
Label$_1$	1	0.77	0.62
Label$_2$	0.57	1	0.11
Label$_3$	0.13	0.47	1

After the Jaccard coefficient is calculated, it is converted into a Jaccard distance by subtracting from 1. This distance will also range from 0 to 1, but with 1 stating the two groups are entirely different and have no members in common. The formula of Jaccard distance can be seen in Eq. (2).

$$d(A,B) = 1 - J(A,B) \tag{2}$$

Clustering methods are built using the distance (dissimilarity) and similarity. Given the nature of quantitative data, distance (dissimilarity) is the best way to determine how the variables relate to one another. On the other hand, similarity is recommended when working with qualitative data. Clustering approaches rely on distance metrics like the Jaccard distance. Selecting the right distance measure for a given dataset can be challenging. Similarity or distance metrics are the main tools used by distance-based clustering algorithms to place similar data points in the same clusters and dissimilar or distant data points in distinct clusters. A machine learning model performs better when the distance metric is effective. Agglomerative hierarchical clustering techniques, in which every instance is regarded as a cluster and the clusters are subsequently combined to form larger clusters, are suggested in this work. This continues until all the clusters are merged into one large cluster that contains all the instances. Agglomerative hierarchical clustering is known as a bottom-up approach that does not need to determine the number of clusters. The bottom-up approaches treat each instance as a single cluster at first, then merge the cluster pairs one after the other until all of the clusters are combined into a single cluster that includes every instance. As a problem transformation technique, the clusters are utilized to identify label combinations for the label powerset once they have been merged.

By treating every label combination as a distinct class, label powerset method combines several labels into a single dataset. Multilabel classification can be achieved by classifying an instance according to a collection of label. Moreover, an instance is trained to be given to any class using a multiclass classifier. Table 3 shows label powerset as one of problem transformation.

Table 3: Illustration of Label Powerset

	Y$_i$ = Label$_i$			
X$_i$=Instance$_i$	*Y$_1$=Label$_1$*	*Y$_2$=Label$_2$*	*Y$_3$=Label$_3$*	*New Label*
X$_1$	1	0	1	1
X$_2$	1	1	1	2
X$_3$	0	1	0	3
X$_4$	1	0	0	4
X$_5$	1	0	1	1

2.2.4 Classification Algorithms

To classify the law articles that were violated on each legal document, this study used Multi-SVM and DT classifiers. Multi-SVM extends binary classifiers to handle multiple class labels, enabling accurate classification into predefined categories. Moreover, DT is a supervised ML algorithm that uses principles that are similar to human decision-making to partition datasets. The study concentrated on DT classification, which is also applied to regression applications.

2.2.5 *k*-fold cross-validation

One resampling method for evaluating machine learning models is cross-validation. The number of groups into which a given data sample is to be divided is shown by a single parameter, *k*. Therefore, the method is frequently referred to as *k*-fold cross-validation. The initial stage is to shuffle and divide the dataset into *k*-folds of similar size. Each iteration conducted at this stage uses a different fold of data for validation and the remaining *k*-1 for training. Particularly, 5-fold cross-validation is used in this study for both training and validation. Later, after shuffling the dataset and dividing the dataset into five sets, then one set is taken for validation and another four set for training. The next stage is to repeat the process for all five sets. The results of the *k*-fold cross validation are given for same session and cross session, respectively. Thus, the process is repetitive until all datasets are evaluated. *k*-fold CV results are normally re-iterated with the mean score of the values of the classification. Fig. 3 provides an example of *k*-fold cross-validation.

To accurately assess the classifier's performance when generalizing to new data, it is critical that we use separate datasets to train and test the classifier. It has been shown that *k*-fold cross-validation appears to be the optimal choice to assess the classifiers' performance in generalization to new data. When the dataset is small, *k*-fold cross-validation is an important technique to assess the robustness of a model. The advantages of *k*-fold cross validation are (1) *k*-fold cross-validation has stable accuracy to solve the random precision issue. In other words, stable accuracy can be achieved because the model is trained on a dataset split into multiple folds. It improves model robustness: It uses all data for training and validation, reducing bias and variance in estimates; (2) *k*-fold cross-validation prevents the overfitting of the training data set by using multiple training and testing cycles, it minimizes the risk of overfitting to a particular data split.

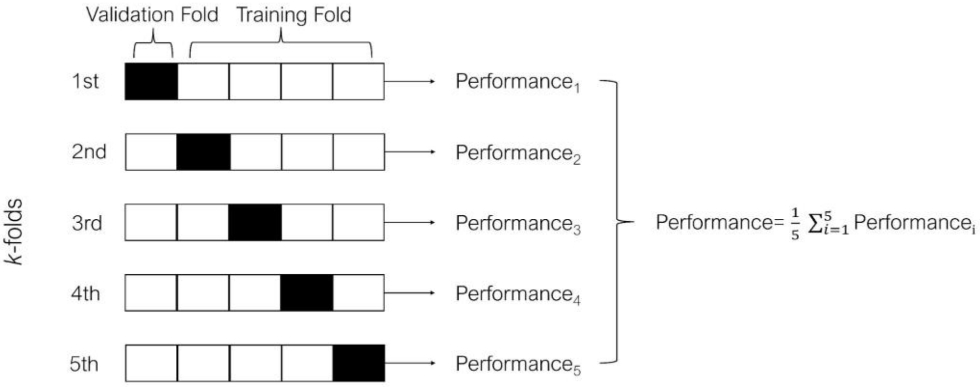


Fig. 3: Illustration of *k*-fold cross-validation

2.2.6 Evaluation Metrics

A total of four performance metrics, namely recall, accuracy, F1-score, and Hamming Loss, are used to evaluate the model's performance. It is essential to specify the type of errors and prediction that a classifier is capable of before establishing the metrics. False Positives (FP) are results that are incorrectly predicted to be positive. False Negatives (FN) are results that are incorrectly predicted to be negative. True Positives (TP) are results that are correctly predicted, and True Negatives (TN) are correctly predicted to be negative. Subsequently, the metrics can be identified based on the information provided.

The ratio of correctly predicted samples to all positive samples is known as precision. During the analysis, Eq. (3) is used to calculate this metric, which shows the percentage of FP rates. The number of TP is divided by positive prediction obtained from the precision score. A 100% precision shows that all of the model's prediction is true without FP.

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

Recall measured the proportion of positive samples accurately classified as TP among the total positive subjects within the same actual class. This measurement is implemented to show the model ability to identify FN, as presented in Eq. (4). The number of TP is divided by prediction that should be given positively yields the recall score. When a model has 100% recall score, no FN is predicted. In light of these results, all negative predictions are considered accurate.

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

The recall and precision metrics combined into a single measurement known as the F1-score. This measurement is calculated by using Equation (5), primarily as an evaluation metric for imbalanced datasets. The precision and recall scores are added together to obtain an F1-score, which is often used for classifier comparison. Therefore, F1-score produces a harmonic mean between the two measuring units.

$$F1 - score = 2 \times \frac{(Recall \times Precision)}{(Recall + Precision)} \quad (5)$$

The percentage of labels that are mispredicted is measured by the Hamming Loss, which is calculated using the average number of incorrectly classified label per instance. The formula of Hamming Loss is shown in Eq. (6).

$$Hamming\ Loss = \frac{1}{N \times L} \sum_{i=1}^{N} \quad \sum_{j=1}^{L} \quad 1(y_{ij} \neq \hat{y}_{ij}) \quad (6)$$

where:

N = the number of instances,
L = the number of labels,
y_{ij} = the true label of the jth label for the ith instance,
$\hat{y}_{ij}$ = the predicted label for the j^{th} jth label of the i^{th} instance,
1 = the indicator function that returns 1 when the argument is true and 0 when false.

A low score of Hamming Loss shows better performance of classification algorithm. Since the optimal value of Hamming Loss is 0, there is no mistake. In multilabel classification, when a single instance of data might belong to multilabel, prediction of label is assessed using Hamming Loss. However, only the single label is penalized by Hamming Loss in multilabel classification, considering both missing and prediction errors. By calculating the percentage of incorrectly predicted labels, Hamming Loss plays a critical role in assessing multilabel classification models. It is also essential in complex scenarios where multiple labels are assigned to instances.

3.0 RESULTS AND DISCUSSION

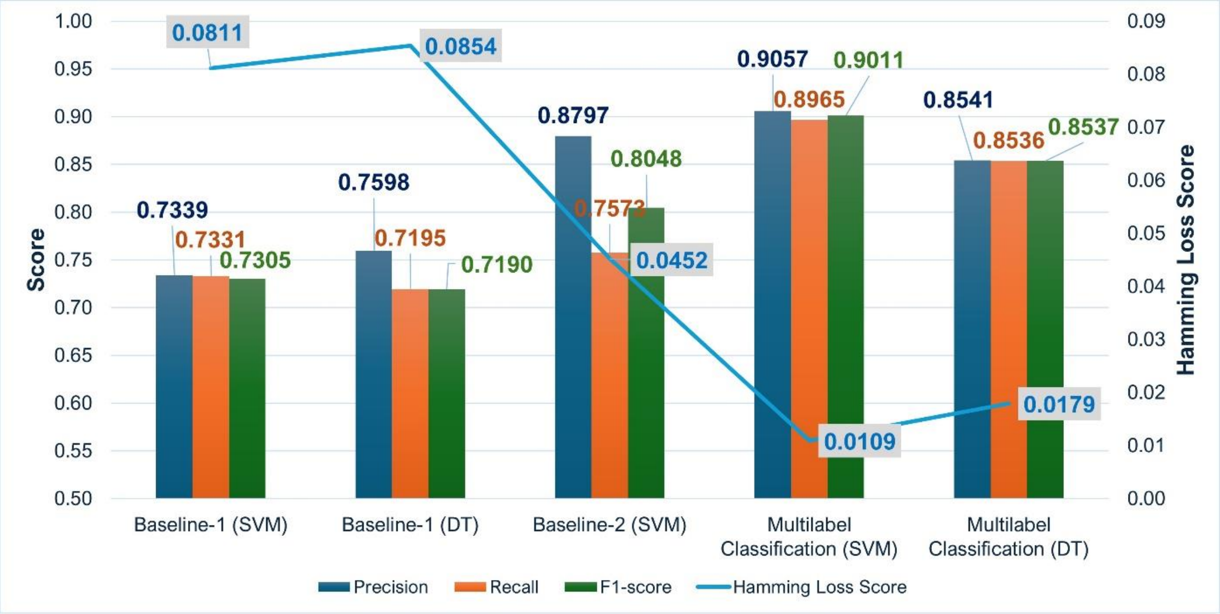
The experimental results demonstrate the effectiveness of our proposed model compared to the baseline models. A baseline implementation is used in this study to evaluate performance in comparison to different variations. Baseline implementation can refer to the initial version of a system or model, or a strategy for improving performance. A strategy for improving performance by identifying and addressing shortfalls. We achieved competitive performance in terms of classification performance, indicating that proposed model effectively addresses the challenges of multilabel legal text classification.

Table 4**:** Summary results of proposed model compared to baseline model

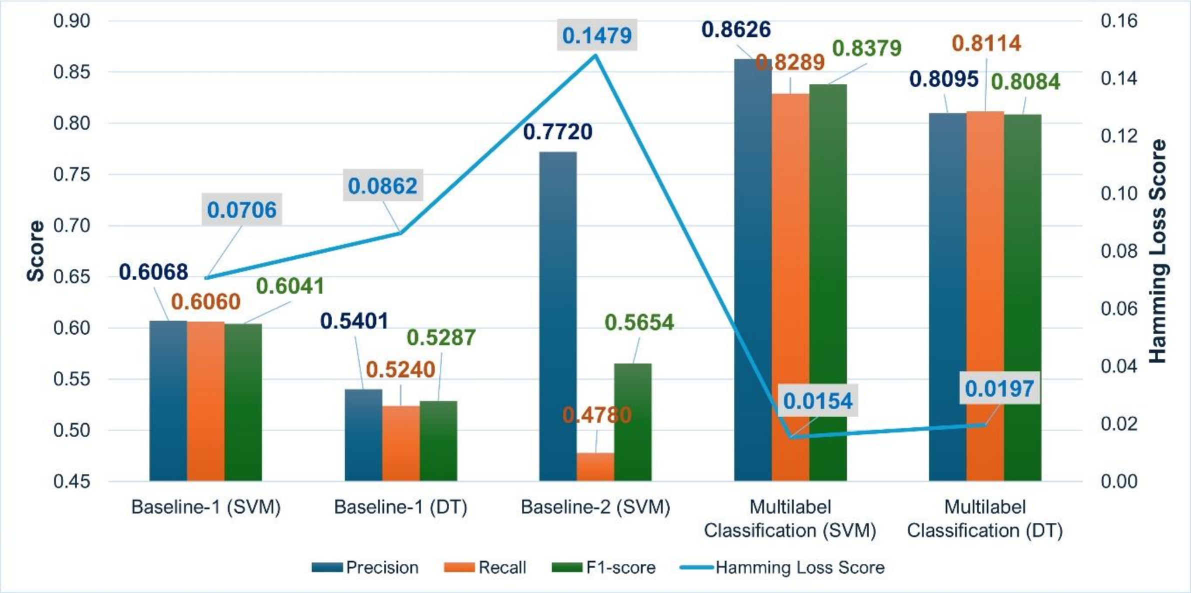
Datasets	Methods	Metrics			
		Precision	Recall	F1-score	Hamming Loss
EUR-Lex	Baseline-1 $_{\text{SVM}}$	0.7339	0.7331	0.7305	0.0811
	Baseline-1 $_{\text{DT}}$	0.7598	0.7195	0.7190	0.0854
	Baseline-2 $_{\text{SVM}}$	0.8797	0.7573	0.8048	0.0452
	Proposed model $_{\text{SVM}}$	**0.9057**	**0.8965**	**0.9011**	**0.0109**
	Proposed model $_{\text{DT}}$	**0.8541**	**0.8536**	**0.8537**	**0.0179**

ECHR	Baseline-1 $_{SVM}$	0.6068	0.6060	0.6041	0.0706
	Baseline-1 $_{DT}$	0.5401	0.5240	0.5287	0.0862
	Baseline-2 $_{SVM}$	0.7720	0.4780	0.5654	0.1479
	Proposed model $_{SVM}$	**0.8626**	**0.8289**	**0.8379**	**0.0154**
	Proposed model $_{DT}$	**0.8095**	**0.8114**	**0.8084**	**0.0197**

The comparison results are shown in Tables 4 and Fig. 4, where proposed models consistently achieve the best performance across all datasets. Compared to baseline models, proposed models further reduce Hamming Loss by at least 3.33% and increase the F1-score by at least 9.61%. The performance improvement of proposed model varies across different datasets due to the diverse characteristics of samples and labels. The average Hamming Loss from the baselines and proposed model is shown in Fig. 4. Prediction and the missing error are simultaneously considered by the Hamming Loss because of the ability to effectively represent the case when the output is partially correct. Hamming Loss is defined as the number of misclassified labels to the total number of labels. A low score of Hamming Loss denotes better performance of classification algorithm, with an optimal value of 0, showing there is no mistake. Compared to baselines, proposed models do not attain high recall. Despite maintaining a high recall, the precision is significantly lower. Since precision and recall are opposing traits, the F1-score is used to assess models showing the balance between the two metrics (precision and recall).



(a)



(b)

Fig. 4: Precision, Recall, F1-score in (a) EUR-Lex dataset; (b) ECHR dataset

Fig. 4 shows the comparison of Hamming Loss among different models across the two legal text datasets. It demonstrates that proposed model utilizing the Label Powerset problem transformation and label correlation has significant impact on improving model performance, proposed model reduces Hamming Loss by up to 3.33%-7.35% on the EUR-Lex dataset and reduces 6.52%-7.08% Hamming Loss on the ECHR dataset. Proposed model establishes correlations between labels and transforms multilabel problems using Label Powerset, enabling more accurate predictions and reducing the chance of missing labels. Fig. 4 shows the experiments on the EUR-Lex and ECHR datasets evaluating precision, recall, and F1-score. Proposed model achieves the highest F1-score. The improvements in these metrics align with the reductions observed in Hamming Loss, confirming the effectiveness of the approach in enhancing model performance. These results underscore the importance of incorporating label correlation information and problem transformation to achieve better performance across various evaluation metrics.

In the EUR-Lex dataset (Table 4), where text is relatively short but has more labels per sample, baseline models struggle to extract sufficient information for accurate label prediction. Proposed model addresses this challenge through the application of LBERT as semantic pre-trained model and transforms the problem using Label Powerset and label correlation, preventing the omission of true labels, resulting in 9.69%-17.04% improvement in F1-score and 3.33%-7.35% reduction in Hamming Loss compared to the baseline models.

In the ECHR dataset (Table 4), which has longer text and less labels per sample, models show improved performance. However, the number of legal cases from ECHR datasets is only 2,000 while EUR-Lex dataset has 57,000. Proposed model effectively utilizes the label correlation, enriching label distribution information and preventing overconfidence in predicted labels. This results in 23.38%-27.25% improvement in F1-score and 6.52%-7.08% reduction in Hamming Loss compared to baseline models.

Overall, our proposed model consistently outperforms baseline models across the evaluated legal text datasets, demonstrating its effectiveness in multilabel text classification tasks. The primary reason for baselines' underperformance lies in their model architectures, which are not fully optimized for multilabel classification. Baseline models did not transform charge prediction problems into multilabel problems, and they do not utilize label correlation to enhance the classification performance. Additionally, we compare four text embeddings which are BoW, TF-IDF, SBERT and LBERT. The baselines only utilized BoW and TF-IDF, so we employed BERT-based pre-trained model as text embedding: SBERT and LBERT. BERT is typically pre-trained on large-scale single-label data, focusing on tasks like text classification and sentence pair tasks but BERT-based text embedding does not adequately capture the dependencies and correlations between labels, resulting in performance that falls short compared to models specifically designed for multilabel tasks.

LBERT is a specialized pre-trained BERT model designed for legal NLP tasks, pre-trained on 12 GB of diverse English legal text [43]. It supports legal studies computational law, and legal technology applications, outperforming general BERT models on domain-specific tasks. The use of pre-trained BERT can solve the issue of small-size data without sacrificing the model's performance [46]. This is because pre-trained models reuse a data-driven model trained for one task to others. When data is insufficient, the pre-trained model aims to enhance performance in a target task by leveraging the knowledge learned from the original task [47]. Problem of small-sized datasets in existing legal studies is caused by the lack of lower court data in digital format [40] [48]. However, the number of legal documents has increased due to digitalization. This shows that the small-sized dataset might be sufficient for humans to start deciding what features can uniquely describe the target.

Another benefit of using pre-trained model is to reduce the data dependence on large-sample learning. Therefore, pre-trained model gains knowledge from a previous task to improve generalization about another. Generalization refers to the ability of a trained model to generate accurate predictions on unseen data. When the model does not generalize well, it will probably perform poorly in real-world scenarios despite showing high accuracy on the training set. Due to this restriction, the model is unreliable and impracticable for use in real-world scenarios.

Compared to TF-IDF and BoW which derive text embeddings based on the frequency of word occurrence in the trained data. LBERT captures the semantic meaning of words where similar meanings are closer together in the embedding space. This phenomenon facilitates easy classification of documents for the model based on content [42][49]. Therefore, the pre-trained LBERT is a more suitable option as the word embedding is applied for this study compared to TF-IDF and BoW. The experiments in Table 4 and 5 prove that proposed model gains higher performance when we employed LBERT than other text embeddings.

To evaluate the impact of label correlation and problem transformation on model performance, we conducted ablation experiments on the legal text datasets, isolating each type of text embedding. The experimental scenarios include label correlation and problem transformation with four different text embedding approaches and different values of Jaccard coefficient (*j*). The Jaccard coefficient is used to measure similarity between two distinct instances. We set Jaccard coefficient to a minimum of 0.06 to ensure an equal proportion of label similarity representations and multilabel vector information. When computing the Jaccard coefficient for pairs of labels, what one is essentially doing is quantifying a form of correlation based on co-occurrence.

To analyze the effect of Jaccard coefficient and prevent overfitting, we conducted experiments on the EUR-Lex and ECHR datasets. As shown in Table 5 and 6, smaller Jaccard coefficient values, indicating a larger proportion of similarity representation information, lead to faster model learning. Proposed model continues to improve, learning more effective information and enhancing performance.

Table 5: Proposed model's results in EUR-Lex dataset in different values of Jaccard coefficient (*j)*

Jaccard coefficient (*j*)	Metrics	Proposed model$_{SVM}$				Proposed model$_{DT}$			
		BoW	TF-IDF	SBERT	LBERT	BoW	TF-IDF	SBERT	LBERT
j >= 0.1	Precision	0.8081	0.8232	0.8781	0.9057	0.7660	0.7615	0.8395	0.8541

	Recall	0.7781	0.8071	0.8469	0.8965	0.7465	0.7718	0.8416	0.8536
	F1-score	0.7827	0.8142	0.8622	0.9011	0.7481	0.7673	0.8404	0.8537
	Hamming Loss	0.0237	0.0192	0.0140	0.0109	0.0244	0.0450	0.0287	0.0179
0. 08 <= *j* < 0.1	Precision	0.7928	0.7874	0.8642	0.8922	0.7507	0.7444	0.8357	0.8522
	Recall	0.7773	0.7969	0.8311	0.8654	0.7457	0.7421	0.8366	0.8525
	F1-score	0.7826	0.7931	0.8473	0.8786	0.7480	0.7412	0.8361	0.8523
	Hamming Loss	0.0266	0.0219	0.0189	0.0117	0.0247	0.0253	0.0241	0.0171
0.06 <= *j* < 0.08	Precision	0.7415	0.8061	0.8589	0.8902	0.7249	0.7348	0.8413	0.8425
	Recall	0.7484	0.7774	0.8198	0.8579	0.7288	0.7377	0.8391	0.8079
	F1-score	0.7511	0.7881	0.8389	0.8738	0.7274	0.7339	0.8401	0.8155
	Hamming Loss	0.0284	0.0227	0.0239	0.0124	0.0226	0.0230	0.0194	0.0254

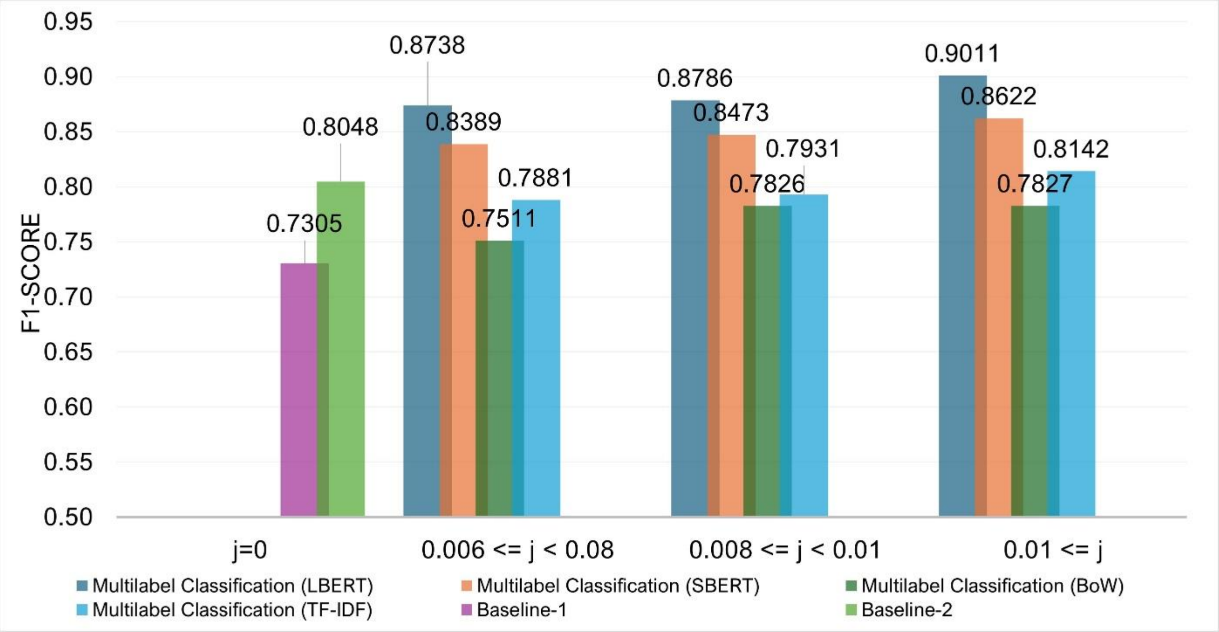
Table 6: Proposed model's results in ECHR dataset based in different values of Jaccard coefficient (*j)*

Jaccard coefficient (*j*)	Metrics	Proposed model$_{SVM}$				Proposed model$_{DT}$			
		BoW	TF-IDF	SBERT	LBERT	BoW	TF-IDF	SBERT	LBERT
j >= 0.1	Precision	0.6591	0.7247	0.7295	0.8626	0.5652	0.5760	0.8175	0.8095
	Recall	0.6173	0.6739	0.6899	0.8289	0.5478	0.5739	0.8017	0.8114
	F1-score	0.6346	0.6902	0.7017	0.8379	0.5548	0.5741	0.8032	0.8084
	Hamming Loss	0.0323	0.0256	0.0261	0.0154	0.0459	0.0416	0.0201	0.0197
0. 08 <= *j* < 0.1	Precision	0.6744	0.7076	0.7342	0.8522	0.5805	0.5589	0.8071	0.8048
	Recall	0.6181	0.6442	0.7110	0.8254	0.5486	0.5442	0.7982	0.7903
	F1-score	0.6347	0.6641	0.7156	0.8289	0.5549	0.5480	0.7942	0.7945
	Hamming Loss	0.0323	0.0272	0.0253	0.0157	0.0470	0.0456	0.0213	0.0215
0.06 <= *j* < 0.08	Precision	0.6078	0.6889	0.7312	0.8529	0.5394	0.5493	0.7974	0.8122
	Recall	0.5884	0.6637	0.6753	0.8114	0.5309	0.5398	0.7807	0.7982
	F1-score	0.6031	0.6691	0.6920	0.8245	0.5342	0.5407	0.7808	0.8018
	Hamming Loss	0.0347	0.0265	0.0264	0.0158	0.0452	0.0497	0.0210	0.0193

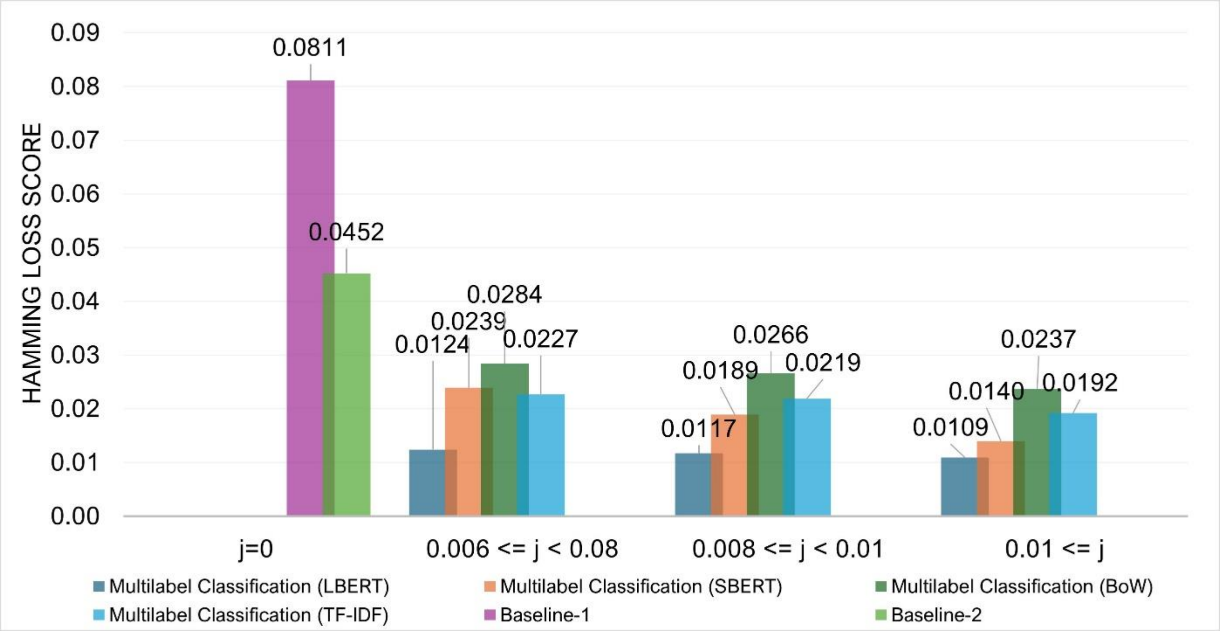
Table 5 and 6 show the experimental results of proposed model in terms of each evaluation metric. Each model is assigned three different Jaccard coefficient (*j*) as label correlation and four text embedding techniques. Figs. 5-8 illustrate the experimental results from Tables 5 and 6. The horizontal axis of each sub-figure indicates different values of the Jaccard coefficient, and the vertical axis represents the result of F1-score and Hamming Loss as evaluation metrics. Based on the experimental results, the following observations can be made:

- Positive label correlations will have a considerably smaller impact and multilabel classification performance will be worse if the Jaccard coefficient (*j*) is lower. The weaker label correlations would result in a decline in performance, strong label correlations are able to help classifiers perform better.
- In two distinct legal datasets, the multilabel classification performance varies depending on the text embedding techniques and the Jaccard coefficient (*j*). The classification performance appears to stay at a rather consistent level in the EUR-Lex dataset, which contains more instances and a larger label space but a shorter text sequence. As label correlation increases, performance also improves in the ECHR dataset, which has a longer text sequence but fewer instances and a smaller label space.
- In addition to applying the Jaccard coefficient (*j*) to the model, we used four text embedding techniques to transform the text sequence from both datasets into feature vectors. The findings indicate that BERT-based text esmbedding approaches (SBERT and LBERT) have outperformed frequency-based methods like BoW and TF-IDF. In contrast to BoW and TF-IDF, which disregard context, BERT is able to extract more context information from a sequence than the other two. The BoW and TF-IDF are unable to understand synonyms or alternative word forms since they only know how frequently a word appears and not its meaning. In both datasets, proposed model using the LBERT text embedding strategy from each experiment achieves a lower Hamming Loss score and a higher F1-score than alternative text embedding strategies.

According to Jaccard coefficient (*j*), we learn that the optimal parameter settings, in particular the correlation for different data sets, are distinct. Thus, in the experiments, we searched for the best configuration for all parameters in each dataset by five-fold cross validation on the training data.

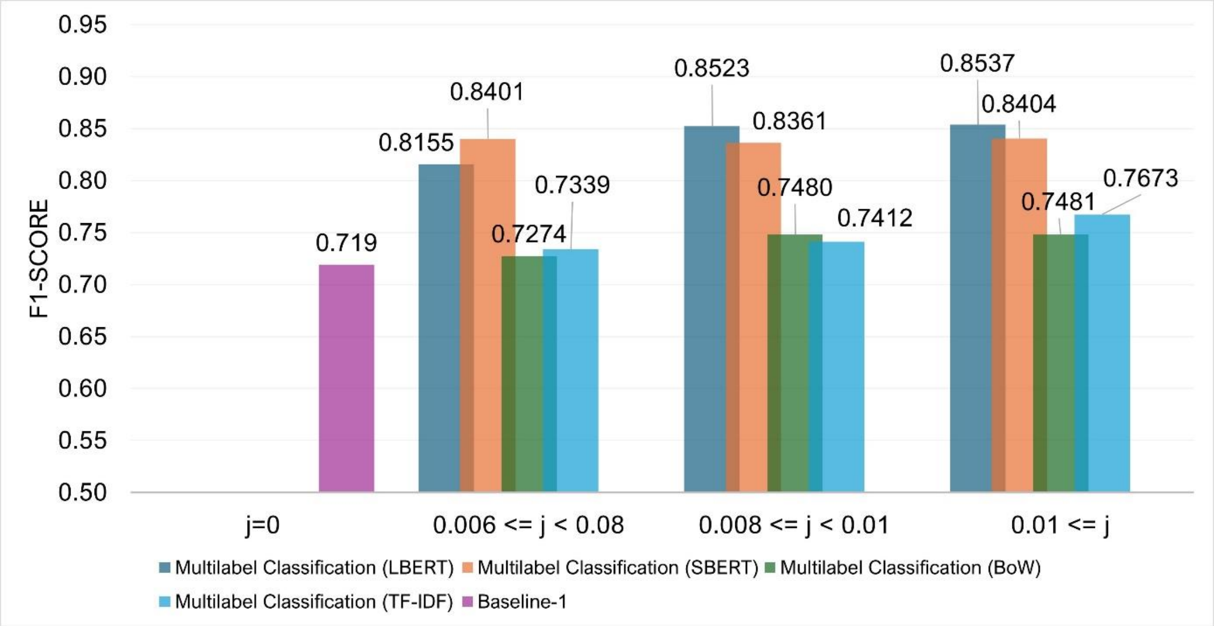


(a)

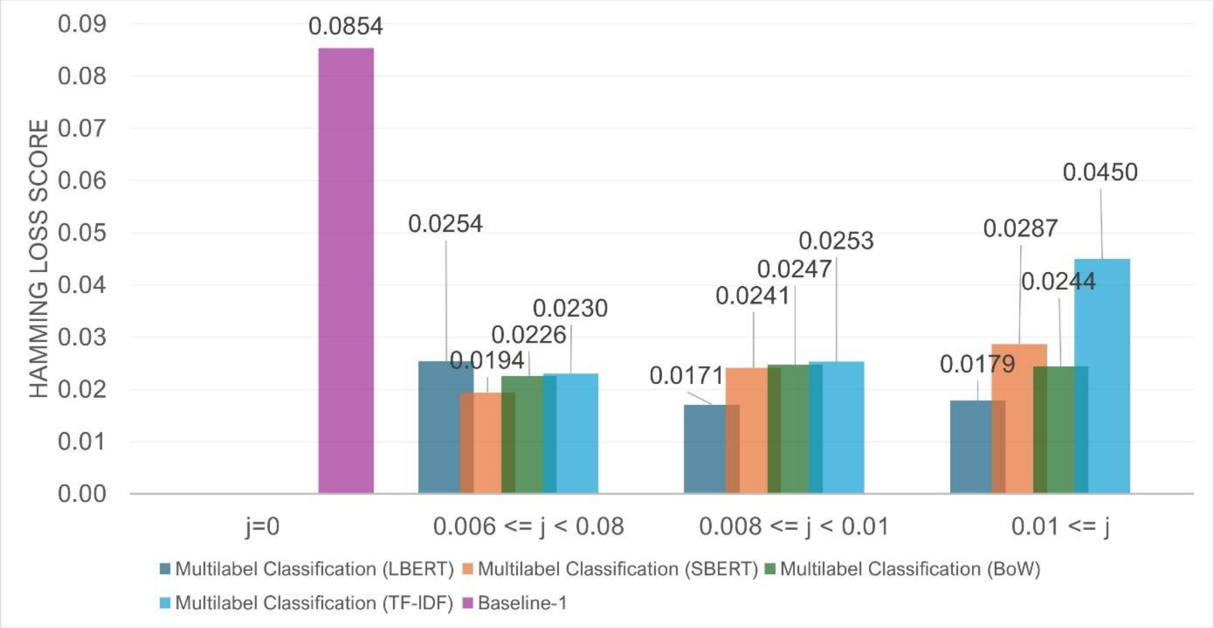


(b)

Fig. 5: Experimental Results SVM-based Classifier in comparation to different word embedding & Jaccard coefficient in EUR-Lex dataset (a) F1-score of the proposed model vs baseline per Jaccard coefficient; (b) Hamming loss score of the proposed model vs baseline per Jaccard coefficient

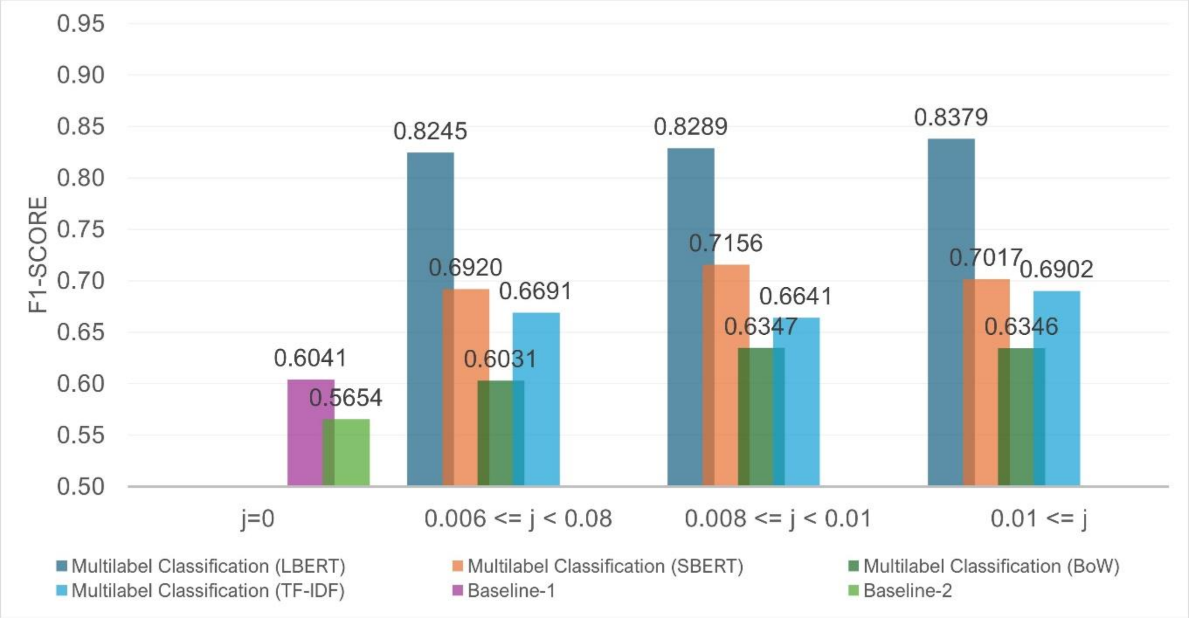


(a)

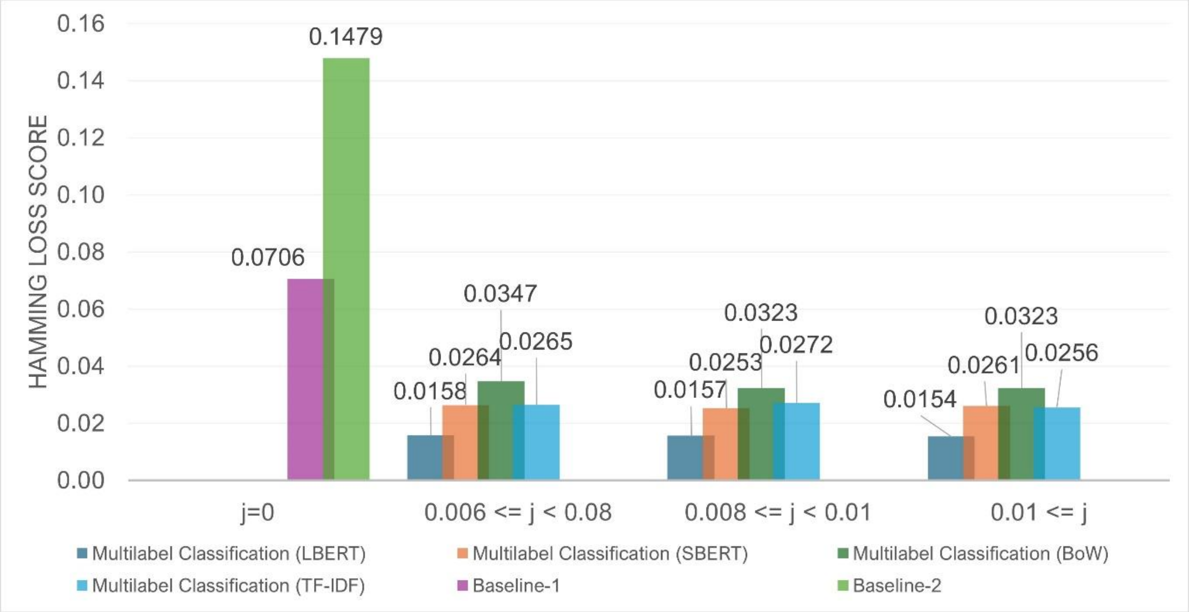


(b)

Fig. 6: Experimental Results DT-based Classifier in comparation to different word embedding & Jaccard coefficient in EUR-Lex dataset (a) F1-score of the proposed model vs baseline per Jaccard coefficient; (b) Hamming loss score of the proposed model vs baseline per Jaccard coefficient

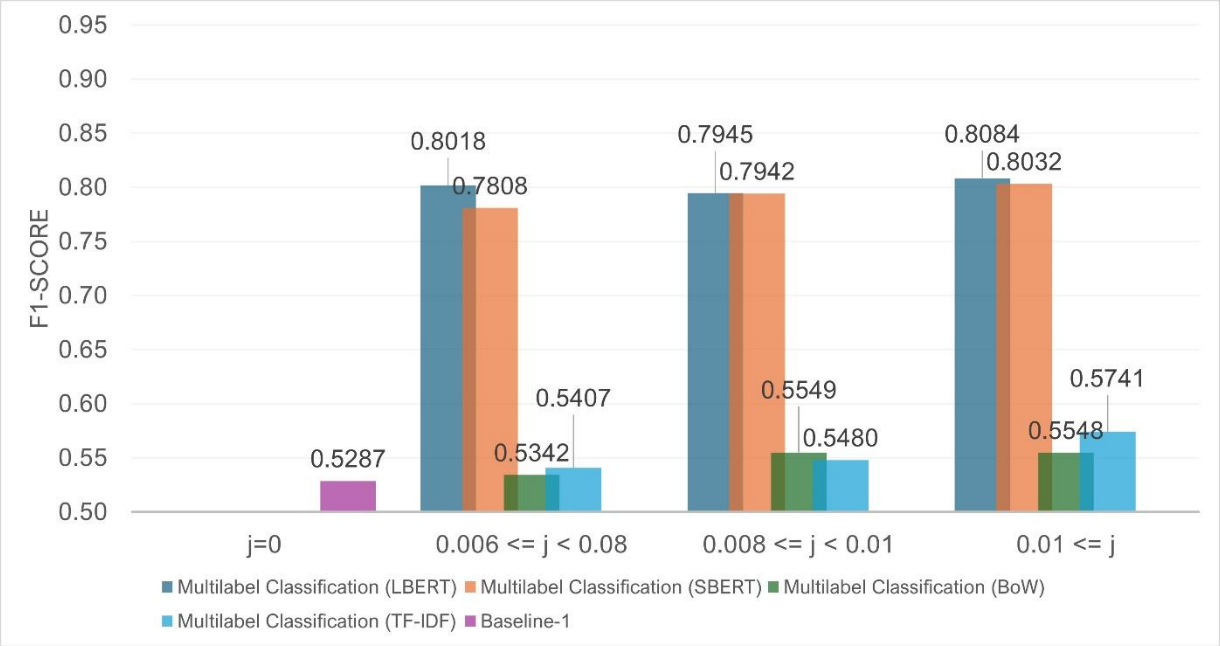


(a)

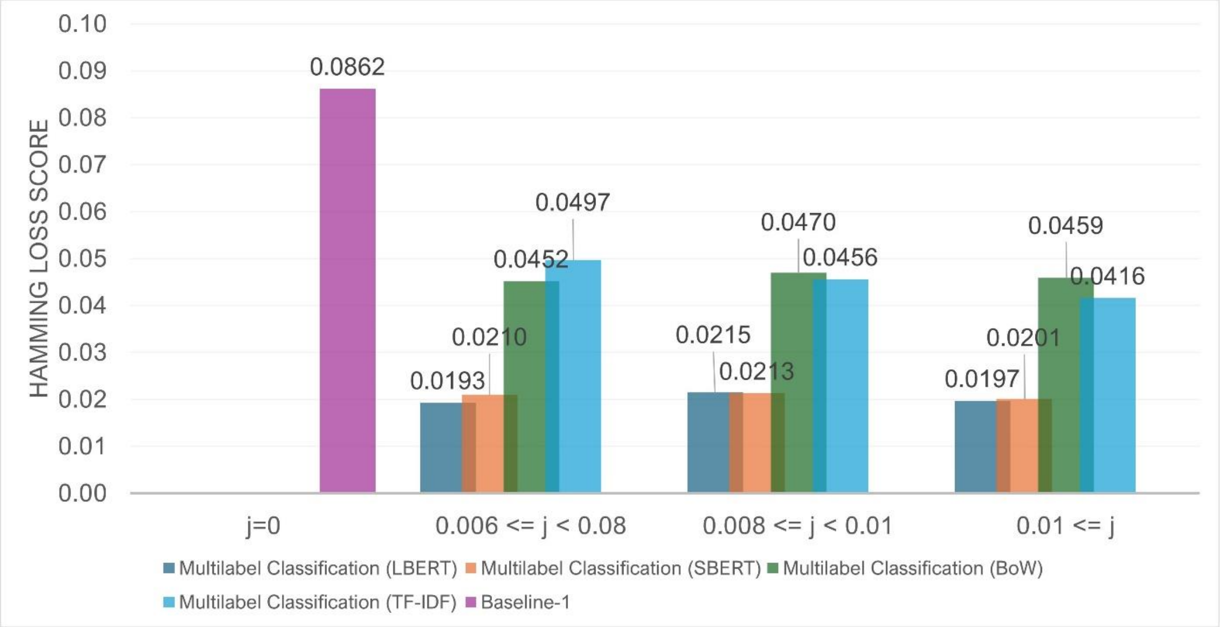


(b)

Fig. 7: Experimental Results SVM-based Classifier in comparation to different word embedding & Jaccard coefficient in ECHR dataset (a) F1-score of the proposed model vs baseline per Jaccard coefficient; (b) Hamming loss score of the proposed model vs baseline per Jaccard coefficient



(a)



(b)

Fig. 8: Experimental Results DT-based Classifier in comparation to different word embedding & Jaccard coefficient in ECHR dataset (a) F1-score of the proposed model vs baseline per Jaccard coefficient; (b) Hamming loss score of the proposed model vs baseline per Jaccard coefficient

To analyze the results from the experiments, a non-parametric statistical test is used to validate the performance of proposed model and baseline models. A Mann-Whitney U test is performed to compare proposed model that implemented multilabel classification with label correlation and problem transformation, and the baseline model. The Mann-Whitney U test, also known as the Wilcoxon Rank Sum Test, is performed to evaluate whether there are significant differences in the results. This non-parametric test is applied to rank the method over the datasets according to the *p*-value and *z*-score. The hypotheses for the Mann-Whitney U test for paired data are as follows:

H_0: There is no difference between the two groups (proposed model and baseline model) in the population.

H_1: There is a difference between the two groups (proposed model and baseline model) in the population.

There is a significant difference between proposed model and the baselines, with *z*-score= -2.50672, and *p*-value = 0.01208. The U-value is 0 and the critical value of U at p-value < 0.05 is 2, showing a significant at p-value < 0.05. For *Z*-ratio, the *z*-score is -2.50672 with *p*-value of 0.01208, showing that the result is significant (H_0 is rejected).

The application of label correlation and label powerset problem transformation is to resolve the limitations mentioned by [27] and [28]. The limitation of the current charge prediction model is the inability to predict violations of law articles simultaneously. This poses a significant challenge, as legal cases are mostly associated with more than one charge. However, existing studies only applied binary and multiclass prediction to address this challenge. Multilabel legal document classification has also been explored in previous studies [50], focusing on keyword-based legal search engines [51]. The majority of reports also use frequency-based word embedding as features without considering any possible label correlation between multilabel [50] [51]. Table 7 summarizes the topic by listing the studies, datasets, countries, the evaluation criteria, the classifiers with best results, the NLP, and ML methods, and the year of publication.

According to [52], multilabel classification problem can easily be solved by treating label individually. This method transforms multilabel problems into a series of binary classification tasks, where each label's existence in a sample is predicted independently. However, multilabel classification problem ignores the reality that label in multilabel data is not independent of one another but rather display certain associations. For example, certain labels may be mutually exclusive, while others appear frequently. Label correlation is the term used to describe the reliance between labels. To solve the limitation, this study used label correlation and powerset problem transformation for the model to predict charges simultaneously. Making good use of label correlation significantly improves multilabel classification procedure and aids in the development of more reliable as well as efficient classifiers [53]. This study measures correlations between labels using Jaccard. Consequently, a classifier can estimate the probability that an instance will have multiple highly correlated labels, in addition to predicting the presence of a specific label.

Table 7: Comparison of recent studies applying NLP and ML classifiers to relatively different datasets in law

Study	Year	NLP used	Dataset	Results	Classifier
Proposed model	2024	LBERT semantic embedding	EUR-Lex & ECHR	EUR-Lex: 86.75% precision 85.69% recall 86.18% F1-score 0.0359 Hamming Loss ECHR: 82.80% precision 79.44% recall 80.6% F1-score 0.0322 Hamming Loss	SVM
[32]	2021	TF-IDF & Tax2Vec	Drugs effect & Drugs side	47% f-measure drugs effect 52.3% drugs side	SVM
[28]	2020	TF-IDF	ECHR	75% accuracy	SVM
[33]	2017	Manual based on metadata	US Supreme Courts	70.2% accuracy 71.9% precision	Random Forest

Classification in charge prediction may speed up the decision support system [54], link similar legal cases, and prevent diverging judgments. Therefore, the use of classification to relate new legal cases with the previous judgments becomes essential for judges and other legal professionals. The practical implications of LJP include consideration of the potential impact of its use on legal practice and the ecosystem. For example, when courts base their decisions on LJP results, this can devalue legal reasoning, supporting the idea that computers are capable of legal reasoning, and degrade the skills of judges. The impact may be more significant when the LJP is used as a decision-making system. In this context, there can be serious repercussions when LJP fails to offer a legally valid response, as minor errors or biases have serious consequences. Based on these concerns, LJP remains an algorithmic investigation and will not be implemented directly in court. Rather than making final decisions without humans, the primary objective is to offer recommendations to courts. Human judges need to be the final option in practical applications for equality and justice.

4.0 CONCLUSION AND FUTURE WORK

In conclusion, extensive experiments were conducted on multilabel classification model to predict charges of legal documents utilizing label correlation and label powerset problem transformation. By evaluating two large-scale legal datasets, the proposed model substantially outperformed two baseline studies by attaining competitive results of 80.32%-90.09% F1-score and 0.0119-0.0210 Hamming Loss score, respectively. The generalizability of the model was proven by two large-scale datasets from legal domain. The implementation of label correlation in multilabel classification showed the broad range of actual problem-solving applications of the model. In addition to showing the significance of label correlation, this study suggested that future investigations should focus on effective management of multilabel in the absence of enough training data in several branches of knowledge with long texts. Our findings contribute to advancing the field of multilabel text classification, offering a powerful tool for various applications that require accurate and comprehensive text categorization. The scalability of the proposed model is limited to the legal domain and two different legal datasets, so for future work, it will be focused on improving the scalability of the model and exploring its application across different domains to realize its full potential.

ACKNOWLEDGEMENTS

This research was supported by the Universiti Malaya International Collaboration Grant (grant number ST080-2022).

REFERENCES

[1] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan. Advancements in natural language processing: Implications, challenges, and future directions. Telematics and Informatics Reports, vol. 16, 100173, 2024. https://doi.org/10.1016/j.teler.2024.100173

[2] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. Computer Science Review, vol. 54, 100664, 2024. https://doi.org/10.1016/j.cosrev.2024.100664

[3] B. Yang, J. T. Sun, T. Wang, and Z. Chen. "Effective Multi-Label Active Learning for Text Classification", in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, ACM, 2009, pp. 917-926. https://doi.org/10.1145/1557019.1557119

[4] T. D. Salma, G. A. P. Saptawati and Y. Rusmawati. "Text Classification Using XLNet with Infomap Automatic Labeling Process", in 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Bandung, IEEE, 2021, pp. 1-6. https://doi.org/10.1109/ICAICTA53211.2021.9640255

[5] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. Text Classification Algorithms: A Survey. Information, vol. 10, no. 4, 150, 2019. https://doi.org/10.3390/info10040150

[6] J. Zhang, Y. Li, F. Shen, Y. He, H. Tan, and Y. He. Hierarchical text classification with multi-label contrastive learning and KNN. Neurocomputing, vol. 577, 127323, 2024. https://doi.org/10.1016/j.neucom.2024.127323

[7] C. Meng, Y. Todo, C. Tang, L. Luan, and Z. Tang. MFLSCI: Multi-granularity fusion and label semantic correlation information for multi-label legal text classification. Engineering Applications of Artificial Intelligence, vol. 139, part B, 109604, 2025. https://doi.org/10.1016/j.engappai.2024.109604

[8] G. Y. Lin, Z. Y. Xiao, J. T. Liu, B. Z. Wang, K. H. Liu, and Q. Q. Wu. Feature space and label space selection based on Error-correcting output codes for partial label learning. Information Sciences, vol. 589, pp. 341-359, 2022. https://doi.org/10.1016/j.ins.2021.12.093

[9] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang. "Sequence Generation Model for Multi-label Classification", in Proceedings of the 27th International Conference on Computational Linguistics, New Mexico, Association for Computational Linguistics, 2018, pp. 3915-3926. https://aclanthology.org/C18-1330/

[10] J. Read, "Scalable multi-label classification," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 2010. Retrieved from https://hdl.handle.net/10289/4645

[11] N. Z. Dina, R. Triwastuti, and M. Silfiani. TF-IDF Decision Matrix to Measure Customers' Satisfaction of Ride Hailing Mobile Application Services: Multi-Criteria Decision-Making Approach. International Journal of Interactive Mobile Technologies, vol. 15, no. 17, pp. 104–118, 2021. https://doi.org/10.3991/ijim.v15i17.22509

[12] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus. Multilabel classification: problem analysis, metrics and techniques. Springer, 2016. https://doi.org/10.1007/978-3-319-41111-8

[13] M. A. U. H. Tahir, S. Asghar, A. Manzoor, and M. A. Noor. A classification model for class imbalance dataset using genetic programming. IEEE Access, vol. 7, pp. 71013–71037, 2019. https://doi.org/10.1109/ACCESS.2019.2915611

[14] M. R. Mustaffa, A. A. I. Mail, L. N. Abdullah, and N. A. Nasharuddin. Deep learning mango fruits recognition based on tensorflow lite. International Journal of Advances in Intelligent Informatics, vol. 9, no. 3, pp. 565-576, 2023. https://doi.org/10.26555/ijain.v9i3.1368

[15] K. Al-Rababah, M. R. Mustaffa, S. C. Doraisamy, F. Khalid, and L. F. de Pina Júnior. Hybrid discrete wavelet transform and texture analysis methods for feature extraction and classification of breast dynamic thermogram sequences. Malaysian Journal of Computer Science, pp. 116–131, 2021. https://doi.org/10.22452/mjcs.sp2021no2.8

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas. "Effective and efficient multilabel classification in domains with large number of labels", in Proceedings of ECML/PKDD 2008 workshop on mining multidimensional data, 2018, pp. 30–44. http://www.ecmlpkdd2008.org/files/pdf/workshops/mmd/4.pdf

[17] C. P. L. F. Carvalho, and A. A. Freitas. "A tutorial on multi-label classification techniques," in Studies in Computational Intelligence, 2009, pp. 177–195. https://doi.org/10.1007/978-3-642-01536-6_8

[18] N. Silla, and A. A. Freitas. A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery, vol. 22, no. 1, pp. 31–72, 2011. https://doi.org/10.1007/s10618-010-0175-9

[19] A. Shah, S. D. Ravana, S. Hamid and M. A. Ismail. Web credibility assessment: affecting factors and assessment techniques. Information Research, vol. 20, no. 1, 655, 2013. https://informationr.net/ir/20-1/paper663.html

[20] C. Ding, T. Pereira, R. Xiao, R. J. Lee, and X. Hu. Impact of label noise on the learning based models for a binary classification of physiological signal. Sensors, vol. 22, no. 19, 7166, 2022. https://doi.org/10.3390/s22197166

[21] X. Shang. A computational intelligence model for legal prediction and decision support. Computational Intelligence and Neuroscience, vol. 2022, no. 1, 5795189, 2022. https://doi.org/10.1155/2022/5795189

[22] O. M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. V. Genabith. "Exploring the use of text classification in the legal domain", in Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL 2017), London, 2017, pp. 1-5. https://ceur-ws.org/Vol-2143/paper5.pdf

[23] G. Boella, L. Di Caro, and V. Leone. Semi-automatic knowledge population in a legal document management system. Artificial intelligence and Law, vol. 27, pp. 227-251, 2019. https://doi.org/10.1007/s10506-018-9239-8

[24] S. Yang, S. Tong, G. Zhu, J. Cao, Y. Wang, Z. Xue, H. Sun, and Y. Wen. MVE-FLK: A multi-task legal judgment prediction via multi-view encoder fusing legal keywords. Knowledge-Based Systems, vol. 239, 107960, 2022. https://doi.org/10.1016/j.knosys.2021.107960

[25] R. A. Shaikh, T. P. Sahu, and V. Anand. Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers. Procedia Computer Science, vol. 167, pp. 2393-2402, 2020. https://doi.org/10.1016/j.procs.2020.03.292

[26] S. I. Moghadasi, S. D. Ravana, and S. N. Raman. Low-cost evaluation techniques for information retrieval systems: A review. Journal of Informetrics, vol. 7, no. 2, pp. 301-312, 2013. https://doi.org/10.1016/j.joi.2012.12.001

[27] A. S. Imran, H. Hodnefjeld, Z. Kastrati, N. Fatima, S. M. Daudpota, and M. A. Wani. Classifying European Court of Human Rights Cases Using Transformer-Based Techniques. IEEE Access, vol. 11, pp. 55664–55676, 2023.

[28] M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law, vol. 28, pp. 237–266, 2020. https://doi.org/10.1007/s10506-019-09255-y

[29] F.de-. Arriba-Pérez, S. García-Méndez, F. J. González-Castaño, and J. González-González. Explainable machine learning multi-label classification of Spanish legal judgements. Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 10, pp. 10180-10192, 2022. https://doi.org/10.1016/j.jksuci.2022.10.015

[30] L. Chen, Y. Wang, and H. Li. Enhancement of DNN-based multilabel classification by grouping labels based on data imbalance and label correlation. Pattern Recognition, vol. 132, 108964, 2022. https://doi.org/10.1016/j.patcog.2022.108964

[31] C. Wang and X. Jin. "Study on the Multi-Task Model for Legal Judgment Prediction", in 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, IEEE, 2020, pp. 309-313. https://doi.org/10.1109/ICAICA50127.2020.9182565

[32] Skrlj, M. Martinc, J. Kralj, N. Lavrac, and S. Pollak. Tax2vec: Constructing interpretable features from taxonomies for short text classification. Computer Speech & Language, vol. 65, 101104, 2021. https://doi.org/10.1016/j.csl.2020.101104

[33] D.M. Katz, M. J. Bommarito, and J. Blackman. A general approach for predicting the behavior of the supreme court of the United States. PLOS ONE, vol. 12, no. 4, e0174698, 2017. https://doi.org/10.1371/journal.pone.0174698

[34] P. Prajapati, and A. Thakkar. Performance improvement of extreme multi-label classification using K-way tree construction with parallel clustering algorithm. Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 8, pp. 6354-6364, 2022. https://doi.org/10.1016/j.jksuci.2021.02.014

[35] W. Siblini, P. Kuntz, and P. Meyer. A review on dimensionality reduction for multi-label classification. IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 3, pp. 839-857, 2019. https://doi.org/10.1109/TKDE.2019.2940014

[36] P. Li, P., H. Zhang, X. Hu, and X. Wu. High-Dimensional Multi-Label Data Stream Classification With Concept Drifting Detection. IEEE Transactions on Knowledge & Data Engineering, vol. 35, no. 8, pp. 8085-8099, 2023. https://10.1109/TKDE.2022.3200068

[37] S. Huang, W. Hu, B. Lu, Q. Fan, X. Xu, X. Zhou, and H. Yan. Application of Label Correlation in Multi-Label Classification: A Survey. Applied Sciences, vol. 14, no. 19, 9034, 2024. https://doi.org/10.3390/app14199034

[38] W. Ferreira, and A. Vlachos. “Incorporating Label Dependencies in Multilabel Stance Detection”, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, Association for Computational Linguistics, 2019, pp. 6350–6354. https://doi.org/ 10.18653/v1/D19-1665

[39] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, vol. 2, e93, 2016. https://doi.org/10.7717/peerj-cs.93

[40] M. B. L. Virtucio, J. A. Aborot, J. K. C. Abonita, R. S. Avinante, R. J. B. Copino, M. P. Neverida, V. O. Osiana, E. C. Peramo, J. G. Syjuco, and G. B. A. Tan. “Predicting decisions of the Philippine supreme court using natural language processing and machine learning”, in Proceedings of the 2018 IEEE 42nd annual computer software and applications conference (COMPSAC), Tokyo, IEEE, 2018, pp. 130–135. https://doi.org/10.1109/COMPSAC.2018.10348

[41] A. R. Kaufman, P. Kraft, and M. Sen. Improving supreme court forecasting using boosted decision trees. Political Analysis, vol. 27, no. 3, pp. 381-387, 2019. https://doi.org/10.1017/pan.2018.59

[42] S. K. W. Chu, S. D. Ravana, S. S. W. Mok, R. C. H. Chan. Behavior, perceptions and learning experience of undergraduates using social technologies during internship. Educational Technology Research and Development, vol. 67, no. 4, pp. 881–906, 2019. https://doi.org/10.1007/s11423-018-9638-2

[43] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. “LEGAL-BERT: The muppets straight out of law school,” in Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, 2020, pp. 2898-2904. https://doi.org/ 10.18653/v1/2020.findings-emnlp.261

[44] N. Z. Dina, and N. Juniarta. Deriving customers preferences for hotels from unstructured data. Geojournal of Tourism and Geosites, vol. 43, no. 3, pp. 872–877, 2022. https://doi.org/10.30892/gtg.43305-899

[45] H. Dong, F. Yang, and X. Wang. Multi-label charge predictions leveraging label co-occurrence in imbalanced data scenario. Soft Computing, vol. 24, pp. 17821–17846, 2020. https://doi.org/10.1007/s00500-020-05029-w

[46] R. Keshari, S. Ghosh, S. Chhabra, M. Vatsa, and R. Singh. “Unravelling Small Sample Size Problems in the Deep Learning World”, in IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, IEEE, 2020. pp. 134-143. https://doi.org/10.1109/BigMM50055.2020.00028

[47] V. Suhag, S. K. Dubey, and B. K. Sharma. Transfer Learning based Low Shot Classifier for Software Defect Prediction. Journal of Information Systems Engineering and Business Intelligence, vol. 9, no. 2, pp. 228–238, 2023. https://doi.org/10.20473/jisebi.9.2.228-238

[48] S. Thammaboosadee, B. Watanapa, J. H. Chan, and U. Silparcha. A two-stage classifier that identifies charge and punishment under criminal law of civil law system. IEICE Transactions on Information and Systems, vol. 97, pp. 864-875, 2014. https://doi.org/10.1587/transinf.E97.D.864

[49] S. Hamzah, M. Mohd, and L. Q. Zakaria. “Exploring the Hybrid Neural Network and Attention Mechanism for Classification of Social Bias”, in Proceedings - International Conference on Knowledge and Systems Engineering, Hanoi, IEEE, 2023, pp. 1-4. https://doi.org/10.1109/KSE59128.2023.10298845

[50] D. Song, A. Vold, K. Madan, and F. Schilder. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Information Systems, vol. 106, 101718, 2022. https://doi.org/10.1016/j.is.2021.101718 101718

[51] G. M. Csányi, R. Vági, D. Nagy, I. Üveges, J. P. Vadász, A. Megyeri, and T. Orosz. Building a Production-Ready Multi-Label Classifier for Legal Documents with Digital-Twin-Distiller. Applied Science, vol. 12, no. 3, 1470, 2022. https://doi.org/10.3390/app12031470

[52] M. -L. Zhang, Y. -K. Li, X. -Y. Liu, and X. Geng. Binary relevance for multi-label learning: An overview. Frontiers of Computer Science, vol. 12, pp. 191–202, 2018. https://doi.org/10.1007/s11704-017-7031-7

[53] Bao, Y. Wang, and Y. Cheng. Asymmetry label correlation for multi-label learning. Applied Intelligence, vol. 52, pp. 6093–6105, 2022. https://doi.org/10.1007/s10489-021-02725-4

[54] W. N. H. W. Ali, M. Mohd, F. Fauzi, K. Shirai, and M. J. M. Noor. Implementation of hyperparameter optimisation and over-sampling in detecting cyberbullying using machine learning approach. Malaysian Journal of Computer Science, pp. 78–100, 2021. https://doi.org/10.22452/mjcs.sp2021no2.6