Machine Learning-based Decision Prediction in Turkish Legal Texts

Seher Solmaz[1] and Mahir Dursun[*2]

[1.] Grand National Assembly of Türkiye, Ombudsman Institution Ankara, Türkiye (S.S. ORCID: 0009-0007-8984-5841)
[2.] Department of Electrical and Electronics Engineering, Faculty of Technology, Gazi University Ankara 06560, Türkiye (M.D. ORCID: 0000-0003-0649-2627)
* Corresponding author: Mahir Dursun (mdursun@gazi.edu.tr)

Abstract: Artificial intelligence (AI) applications are becoming increasingly popular in the field of law and expanding its range of applications. In this study, machine learning supervised learning models are used to predict the decision type (Partial Recommendation, Partial Rejection, Recommendation, Rejection and Partial Recommendation) of the Ombudsman's Office (Ombudsman) decisions. Supervised learning models such as Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes, K-nearest neighbors (KKN), Logistic Regression (LR) and XGBoost were used in the study. During the training of the models, the "APPLICANT'S CLAIMS AND DEMANDS" section of the decision text was taken into account and the other parts were not shown to the models. Texts were transformed into TF-IDF vectors and the models were trained. The results obtained and the performance of different models are compared and analyzed in detail. It was observed that SVM was the most successful model in predicting the Ombudsman decisions on the test dataset. The SVM model achieved 61% F1 score and 62% accuracy in decision prediction.

Keywords: Machine learning, legal text, AI in law, decision predict, ombudsman

1. Introduction

As a result of research and applications in the field of artificial intelligence (AI), AI is improving its volume in our lives day by day. As different fields of study are affected by this development, the field of law is one of the affected fields of study. As a result of these developments, it is observed that studies on the use of AI in the field of law are increasing today in order to reduce the workload of legal professionals (Yılmaz, 2021). Although studies in the field of law are popular today, studies date back much further. Although Alan Turing's work (Turing, 1950) is not a direct study on the use of AI in the field of law, it has laid the foundations for its potential in legal decision-making.

According to MarketsandMarkets (2024a), the estimated value of the global legal AI software market for 2025 is US$2.7 billion. This represents an increase of 14.9% from the USD 2.35 billion estimated in 2024 (MarketsandMarkets, 2024b). As a result of the increasing market share of AI in the field of law, it is estimated that the work and transactions of lawyers, judges, prosecutors and legal professionals will become easier. In the legal field, AI can summarize texts and reports (Pike, 2018), present a persuasive argument to the court (Atkinson et al., 2020), follow current legislation and precedents (Von Lucke et al., 2022), match similar cases (Tong et al., 2024), and advise the judge (Aletras et al., 2016; Popple, 1990) without replacing the decision-making mechanism (Almuzaini & Azmi, 2023). At the stage of examining the files, it may be biased against the file (Smoliński & Brycz, 2024) and therefore, it can also prevent the situation where similar cases may be decided differently.

In the early stages of decision forecasting, studies focus on mathematical and statistical analysis of existing cases and do not provide any methodology on how to forecast (Long et al., 2019; Kort, 1957; Nagel, 1963).

There are many previous international studies on court decision prediction (Katz et al., 2017; Kowsrihawat et al., 2018; Ikram & Chakir, 2019; Chen et al., 2022; Lidén, 2024). There are limited studies in the literature on decision prediction with Turkish legal texts (Mumcuoğlu et al., 2021; Aras et al., 2022; Öztürk et al., 2022; Akça, 2023). The morphological structure of Turkish is different from other languages. Since it is a post-adjunctive language, the suffix added to the end changes the meaning of the root. For this reason, it causes some problems in the interpretation of texts with artificial intelligence. Another problem is that the terminology used in legal documents consists of semantically more complex and longer sentences than normal texts. This situation significantly affects the performance and success of the studies on legal texts using artificial intelligence.

There is a gap in the literature on decision prediction by converting Turkish legal texts into TF-IDF (Term Frequency-Inverse Document Frequency) vectors (Yassine et al., 2023). In this study, legal texts are transformed into TF-IDF vectors and machine learning supervised learning models are trained to predict the decisions of the Ombudsman's Office. All decisions were obtained from data available online on the Ombudsman's Office website. In this study, 10,600 case decisions available at the time of the study were used.

The organization of the paper will be as follows. Related works in the literature will be mentioned in Section II. The data preprocessing methods, machine learning models and techniques used in the study will be described in Section III. The results of the experiments will be analyzed in Section IV. Finally, the results of the study will be discussed in Section V.

2. Related Works

In his study, Lawlor (1963) predicted that computers will one day be able to analyze and predict the outcomes of judicial decisions (Aletras et al., 2016). As in this prediction, the idea that law can benefit from artificial intelligence was discussed in a 1970 study (Buchanan & Headrick, 1970). Nowadays, there are studies in the literature on judicial decision prediction, which is quite popular.

In a study to predict the behavior of the United States Supreme Court in a generalized, out-of-sample context, Katz et al. (2017) used a machine learning random forest classifier and found 71.9% accuracy from the model developed based on 28,000 case outcomes.

Kowsrihawat et al. (2018) proposed a prediction model for criminal cases in the Supreme Court of Thailand using End-to-End Deep Learning Neural Networks, where the model mimics the legal interpretation process. After performance testing, they found that the model can provide higher F1 than traditional text classification techniques such as Naive Bayes and SVM.

In a study (Buchanan & Headrick, 1970) on the automatic prediction of judgments of the European Court of Human Rights, Buchanan & Headrick, 1970 found that the average accuracy of predicting future judgments based on past cases ranged from 58% to 68%.

Lidén (2024), who used regression analysis to predict judges' decisions on new criminal case petitions, analyzed 3915 new criminal case petitions submitted to the Swedish Supreme Court and six Courts of Appeal in the period 2010–2020. These data formed the basis of a regression model that was then used to predict decisions on petitions in 2021. The regression model correctly predicted 100% of the decisions in 2021, based on access to legal representation and type of offense.

In a study on the prediction of high court judgments in Türkiye, Mumcuoğlu et al. (2021) used DT, Random Forests (RF), SVM and deep learning (DL) methods. They obtained accuracy values reaching 93% and, more importantly, F1 scores reaching 0.87.

In a study for the development of decision prediction methods using case texts of Turkish superior courts, Aras et al. (2022) trained models using Feed Forward Neural Networks (FFNN), word representations and features extracted from the texts by Principal Component Analysis (PCA) and achieved a Macro F1 score of 85.4% for decision prediction.

In the study by Öztürk et al. (2022), in which the results of the Supreme Court cases were predicted by using recurrent unit neural network (GRU), long and short-term neural network (LSTM) and bidirectional LSTM (BiLSTM) deep learning models, it was found that GRU was the most successful decision prediction model as a result of the experiments. They achieved an accuracy score of 96.8% in decision prediction with the GRU model.

In the study by Akça (2023), which included 90,000 files of high court decisions, various classifiers were experimented with. They found that BERT models outperformed the other classifiers by a large margin and led to an increase in the F1 score by about 2%.

3. Methodology

In this study, the Term Frequency-Inverse Document Frequency (TF-IDF) method was used to convert texts into numerical representations. TF-IDF measures word frequencies and their distinctive importance across documents, providing meaningful inputs for classification models (Rajaraman & Ullman, 2011).

In the study, software was developed to download decisions from the Ombudsman's web page. A data set was created with the downloaded decisions. The stages of the study are shown in Figure 1.
Empty and missing data were removed from the data set and preprocessing steps were applied. After the data set was divided into train, test and validation, the text data was converted into TF-IDF vectors.
The digitized text data was used to train the machine learning models. The models were subjected to performance tests to measure the success of the training process.
The process steps in the flowchart are explained in detail in the following sections.

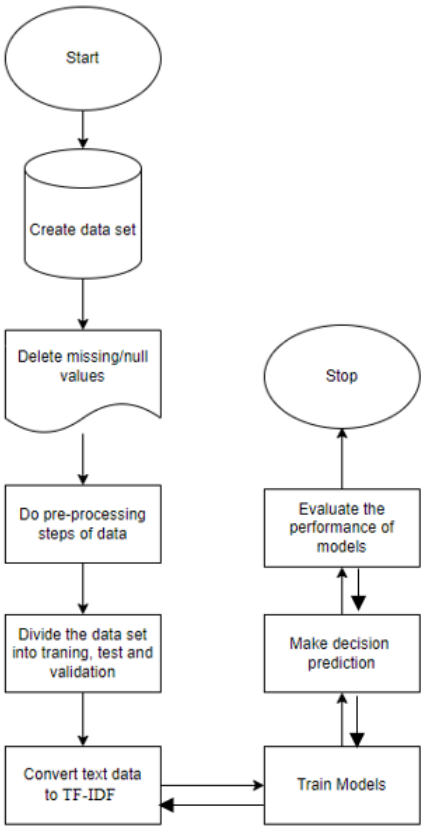

Figure 1. Flow chart.

3.1. Data Set Preparation

In this study, 10.600 decisions existing in the ombudsman's web page decision database at the time of the study were downloaded with the software developed. The downloaded decisions were divided into columns of complaint number, complaint subject, complaint sub-subject and decision type and saved in .csv format.
Although the decision texts vary according to the type of decision and the content of the application, they generally consist of six (6) parts. These, in turn, consist of certain sections: summary of the applicant's claims and demands, the administration's explanations regarding the application, the relevant legislation, the Ombudsman's recommendation to the Ombudsman, the evaluation and justification, and the final decision. The section from the “APPLICANT'S CLAIMS AND DEMANDS” to the beginning of the next section was taken from the text of the decision and saved as a new field.
Since the decision type field contains text data, a new decision_id field was added and labeled {0: Partial Recommendation Partial Rejection Decision, 1: Recommendation Decision, 2: Rejection Decision, 3: Partial Recommendation Decision}. The columns with empty and incorrect fields were removed from the dataset. When empty and erroneous fields were removed from the dataset, 9674 decisions out of 10,600 decisions were made available. There were 2 decisions from the partial rejection decision and

they were added to the rejection decision in order not to create imbalance in the dataset.
After the arrangements made, the distribution of decision types in the data set is shown in Table 1. According to the data in the table, the distribution of decision types is numerically unbalanced. The decision type with the most data is the rejection decision, while the decision type with the least data is the partial recommendation decision.

Table 1. Distribution of Decision Type

No	*Number Distribution by Decision Type*	
	Decision Type	*Quantity*
1	Rejection Decision	4343
2	Recommendation Decision	3757
3	Partial Recommendation Partial Rejection Decision	1377
4	Partial Recommendation Decision	197

The section “APPLICANT'S CLAIMS AND DEMANDS” taken from the decision text was subjected to data preprocessing steps as it consists of textual data.
All text data were converted to lowercase. Turkish letters were taken into account during this process.
Punctuation and special characters have been removed from the text data.
Unnecessary spaces in the text data have been removed.
When the root of the word is analyzed, for example, in the word “judgment”, the root “judgment” is obtained. In this case, the semantic integrity of the sentence is disrupted and causes ambiguity. Considering these situations, stemming and lemmatization were not applied in the data preprocessing step.

3.2. Implementation of Machine Learning Models

The generated dataset is divided into training, testing and validation parts. It is shown in Figure 2. As can be seen in the figure, the data is divided into training, test and validation in the ratio of 70%-15%-15%.

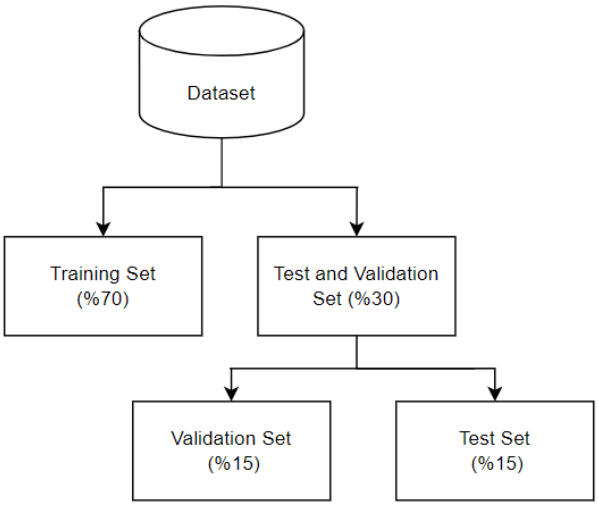

Figure 2. Splitting the dataset.

The text data was converted into a format suitable for machine learning models, i.e. a numerical data TF-IDF vector. With this conversion process, the importance of a word for a text is determined by looking at how often it occurs in the text (Term Frequency) and how often it occurs in all texts (Inverse Document Frequency).

The metric used to determine how often a word occurs in the text with Term Frequency (TF) is as follows;

$$TF(t,d) = n(t,d) / N(d) \quad (1)$$

In the definition in Eq. (1), TF(t,d) represents the term frequency of word t in document d, n(t,d) represents the total number of occurrences of word t in document d and N(d) represents the total number of words in document d.
The metric used to determine how often a word appears in all documents with Inverse Document Frequency (IDF) is as follows;

$$IDF(t) = \log(N / n(t)) \quad (2)$$

In the definition in Eq. (2), IDF(t) is the inverse document frequency of word t, N is the total number of all documents and n(t) is the number of times word t occurs in all documents. Rare words are considered more important than more frequent words.
The TF-IDF value is calculated by combining the above two metrics. The metric used is as follows;

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t) \quad (3)$$

In the definition in Eq. (3), TF-IDF(t,d) represents the TF-IDF value of word t in document d, TF(t,d) represents the term frequency of word t in document d and IDF(t) represents the inverse document frequency of word t.
In this study, the text data is transformed into TF-IDF vector and machine learning models are used as classification method. These models are DT, SVM, Naïve Bayes, KKN, LR and XGBoost, which are supervised learning models. The models were trained with training data.
Performance metrics of precision, recall, F1-score and accuracy were used to evaluate the models.
We calculated precision by measuring how many of the samples that the model predicts as positive are actually positive. The Precision metric is as follows.

$$Precision = TP /(TP+FP) \quad (4)$$

In the definition in Eq. (4), TP (True Positive): The number of true positive predictions and FP (False Positive): The number of false positive predictions. It refers to the number of samples that the model predicts as positive but are actually negative.
The ratio of true positives to all positives was calculated. That is, we calculated the recall value, which expresses how many of all instances that were actually positive the model correctly predicted. The recall metric is as follows.

$$Recall= TP / (TP+FN) \quad (5)$$

In the definition in Eq. (5); FN (False Negative): Refers to the number of false negative forecasts.
The F1 score is the harmonic mean of the precision and recall metrics. The F1 score provides the balance of these two metrics. The F1 score metric is as follows.

$$F1\ Skoru = 2 \times (precision \times recall) / (precision + recall) \quad (6)$$

In the definition in Eq. (6), precision refers to how precise the model is and recall refers to how comprehensive the model is.
Accuracy was calculated to determine how accurate the machine learning models are in making

predictions. The accuracy metric is as follows.

$$\text{Accuracy} = (\text{Accurate Predictions} / \text{Total Estimates}) * \%100$$

4. Experiments and Results

The results obtained with the test set of Ombudsman decisions using machine learning models are given in Table 2. SVM model has the highest accuracy and F1-score while DT model has the lowest value. It shows that the correct prediction rate of the SVM model is better than the other models. According to the Accuracy value, except for the DT model, the other models perform close to each other.

For all models, except for the DT model, precision values ranged between 57% and 63%. This indicates that the models are relatively more successful in positive estimates.

The recall values for all models ranged between 57% and 63%. This suggests that there is a possibility that the models may miss some of the samples that are actually positive.

F1-Score values for all models, except for the DT model, ranged between 58% and 60%. This supports the values we obtained from accuracy, precision and recall.

Table 2. Machine Learning Models Decision Prediction Results

	Decision Prediction of Models			
	Accuracy	***Precision***	***Recall***	***F1-Score***
DT	0.497933	0.496848	0.497933	0.497196
SVM	0.627410	0.635928	0.627410	0.615565
Naïve Bayes	0.592975	0.590627	0.592975	0.579366
KKN	0.584022	0.578942	0.584022	0.579207
LR	0.602617	0.615953	0.602617	0.594637
XCBoost	0.610881	0.616783	0.610881	0.604200

5. Conclusion

According to the decision prediction results of the machine learning models in Table II, SVM model is found to be more successful than the other models in decision prediction. The DT model, on the other hand, is found to be more unsuccessful in decision prediction.

During the experiments, it was observed that the high distribution of the data set according to the complaint subject affected the decision prediction. It has been observed that the “APPLICANT'S CLAIMS AND DEMANDS” field in the decision text has high differences according to the subject of the complaint and the applicant's demand, which affects the performance of the models.

It was observed that the lack of publication of similar decisions on the Ombudsman's Office website and the numerical imbalance in the distribution of decisions affected the training of the models and the test results.

Ombudsman decisions are divided according to the subject of the complaint. It has been observed that differences in the subject matter of complaints affect the type of decision as well as the content of the text.

It is estimated that better results can be obtained in future studies when the number of samples in the data set is increased and a balanced distribution is provided according to the type of decision and the subject of the complaint.

Acknowledgment

We would like to thank The Ombudsman Institution for supporting this study.

Author Contributions: Conceptualization, S.S. and M.D.; Methodology, M.D.; Software, S.S.; Validation, Writing – Review & Editing, S.S. and M.D.; Supervision, M.D.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

Akça, O. (2023). *Natural Language Processings in Legal Domain: Classification of Turkish Legal Texts* [Master's thesis, Marmara University Institute of Science and Technology].

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science, 2*, e93.

Almuzaini, H. A., & Azmi, A. M. (2023). TaSbeeb: A judicial decision support system based on deep learning framework. *Journal of King Saud University - Computer and Information Sciences, 35*(8), 101695.

Aras, A. C., Öztürk, C. E., & Koç, A. (2022, May). Feedforward neural network based case prediction in Turkish higher courts. In *2022 30th Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE.

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence, 289*, 103387.

Buchanan, B. G., & Headrick, T. E. (1970). Some speculation about artificial intelligence and legal reasoning. *Stanford Law Review, 23*, 40–62.

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management, 59*(2), 102798.

G. Pike, G. H. (2018). AI in legal research: Casetext and LexisNexis battle it out. *Information Today, 35*(9), 16–17.

Ikram, A. Y., & Chakir, L. (2019). Arabic text classification in the legal domain. In *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)* (pp. 1–6). IEEE.

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE, 12*(4), e0174698.

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

Kowsrihawat, K., Vateekul, P., & Boonkwan, P. (2018). Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism. In *2018 5th Asian Conference on Defense Technology (ACDT)* (pp. 50–55).

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

Lidén, M. (2024). Can criminal justice be predicted? Using regression analysis to predict judges' decisions on petitions for new criminal trials. *Science & Justice, 64*(1), 43–49.

Long, S., Tu, C., Liu, Z., & Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019* (pp. 558–572). Springer.

MarketsandMarkets. (2020). *Legal AI software market – Global forecast to 2024*. https://www.marketsandmarkets.com/Market-Reports/legal-ai-software-market-88725278.html (Accessed: 20 March 2024).

Mumcuoğlu, E., Öztürk, C. E., Ozaktas, H. M., & Koç, A. (2021). Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Information Processing & Management, 58*(5), 102684.

Nagel, S. S. (1963). Applying correlation analysis to case prediction. *Texas Law Review, 42*, 1006.

Öztürk, C. E., Özçelik, Ş. B., & Koç, A. (2022, May). Predicting outcomes of the Court of Cassation of Turkey with recurrent neural networks. In *2022 30th Signal Processing and Communications Applications Conference (SIU)* (pp. 1–4). IEEE.

Popple, J. (1990). Legal expert systems: The inadequacy of a rule-based approach. In *Proceedings of the Thirteenth Australian Computer Science Conference (ACSC13)* (pp. 7–9). Monash University.

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Smoliński, P. R., & Brycz, H. (2024). Individual differences in inaccurate versus accurate economic judgment and decision making: Metacognitive approach. *Personality and Individual Differences, 219*, 112500.

Tong, S., Yuan, J., Zhang, P., & Li, L. (2024). Legal judgment prediction via graph boosting with constraints. *Information Processing & Management, 61*(3), 103663.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Von Lucke, J., Fitsilis, F., & Etscheid, J. (2022). Using artificial intelligence for legislation—Thinking about and selecting realistic topics. *EGOV-CeDEM-ePart*, 32–42.

Yassine, S., Esghir, M., & Ibrihich, O. (2023). Using artificial intelligence tools in the judicial domain and the evaluation of their impact on the prediction of judgments. *Procedia Computer Science, 220*, 1021–1026.

Yılmaz, O. G. (2021). Yargı uygulamasında yapay zekâ kullanımı – Yapay zekâ hâkim cübbesini giyebilecek mi? *Adalet Dergisi, 66*, 379–415.