Journal of Construction Engineering, Management & Innovation
2025 8(1):64-88
DOI 10.31462/jcemi.2025.01064088

RESEARCH ARTICLE

When defendants speak: Quantifying the predictive value of defence arguments in construction litigation

Mahmut Sarı[1], Savaş Bayram[2], Emrah Aydemir[3]

[1] Kırşehir Ahi Evran University, Vocational School of Technical Sciences, Department of Construction, Kırşehir, Türkiye
[2] Erciyes University, Department of Civil Engineering, Kayseri, Türkiye
[3] Sakarya University, Sakarya Business School, Department of Management Information Systems, Sakarya, Türkiye

Article History

Received 04 March 2025
Revised 25 March 2025
Accepted 27 March 2025

Keywords

Court of cassation
Natural language processing
Judicial decision prediction
Text classification
Machine learning

Abstract

The global construction industry faces significant risks due to disputes. This study aims to predict outcomes in construction dispute judicial decisions by analyzing the linguistic interaction between plaintiff claims and defendant defenses in Turkish, addressing a methodological gap in the literature. The research examines 2,563 Court of Cassation decisions in Türkiye from 2011-2021 (from 15,667 cases), organized into three datasets: containing both plaintiff claims and defendant defenses (Dataset I), only plaintiff claims (Dataset II), and all decisions (Dataset III). Dataset I uniquely captures the impact of defendant voice, demonstrating how including counterarguments significantly enhances model performance. Standard preprocessing techniques were applied to address Turkish morphological challenges. Among various feature extraction methods, TF-IDF demonstrated superior performance. The HistGradientBoosting achieved optimal performance, with Dataset I reaching 87.38% accuracy compared to 84.53% for Dataset II, proving that modeling mutual arguments enhances prediction beyond using plaintiff claims alone, exceeding success rates in comparable literature. This study pioneers a framework for analyzing the dialectics of legal texts in construction disputes, with applications across different legal systems.

1. Introduction

Artificial Intelligence (AI) has catalysed a paradigm shift in legal analytics, enabling data-driven interrogation of judicial texts across jurisdictions [1, 2]. Yet, despite these advancements, the administration of justice remains mired in inefficiency — case backlogs, spiralling litigation durations, and eroding public trust plague courts globally [3, 4]. In Turkey, where the Civil Chambers of the Court of Cassation saw an 8.6% rise in average case duration (221 days in 2023 to 240 days in 2024) [5], the crisis underscores the unsustainability of traditional legal practices. Legal Judgment Prediction (LJP), which leverages Natural Language Processing (NLP) to automate outcome forecasting via case fact analysis [6], offers transformative potential. However, its application to agglutinative languages like Turkish and sector-specific disputes, such as construction, remains critically underexplored.

The Turkish judiciary's hierarchical structure — comprising Courts of First Instance, Regional Courts, and the precedent-setting Court of

Correspondence Savaş Bayram sbayram@erciyes.edu.tr

eISSN 2630-5771

Cassation [7] — faces unique challenges. Construction disputes, adjudicated finally by the Court of Cassation's Sixth Civil Chamber, are particularly prone to delays due to their reliance on complex contractual frameworks and technical evidence. This complexity exacerbates financial and reputational losses for parties [8, 9], yet alternative dispute resolution mechanisms remain underutilised, leaving litigation as the primary recourse.

NLP, an interdisciplinary field combining computer science and linguistics, focuses on the computational analysis of human language [10]. Modern NLP systems often treat words as atomic units for simplicity and robustness [11]; however, this approach proves inadequate for agglutinative languages such as Turkish, where words are formed by concatenating morphemes [12]. The growing volume of legal text data has amplified the need for automated classification, typically addressed through rule-based or data-driven methods [13, 14]. In the legal domain, LJP—the use of NLP to forecast judicial outcomes based on factual case descriptions [15] —has emerged as a critical tool for legal practitioners [16, 17]. However, existing LJP frameworks predominantly focus on non-agglutinative languages and prioritise outcome prediction over the linguistic complexity inherent in parties' arguments [18]. Furthermore, even studies emphasising linguistic complexity have not systematically examined how the interplay between claim and defence texts contributes to judgment prediction. This study addresses this gap by analysing 2,563 rulings from the Turkish Court of Cassation's Sixth Civil Chamber, the final authority in construction disputes. Unlike prior research, we evaluate the impact of claim-defence dynamics on prediction accuracy using three distinct datasets: claims-only texts, claims + defences texts, and full judgment texts. By applying NLP techniques for preprocessing and ML models for prediction, our findings demonstrate that incorporating defence texts significantly enhances model performance. For instance, jointly evaluating a plaintiff's "incomplete work" claim with a defendant's "force majeure" defence yielded higher prediction accuracy compared to datasets containing only claims or full judgments. These results highlight the critical role of defence arguments in judgment prediction.

Systematically assessing linguistic patterns in claim and defence texts enables legal professionals to identify focal points in litigation and optimise resource allocation. Additionally, modelling the relationship between legal arguments and judicial outcomes facilitates data-driven litigation strategy design. This approach offers a practical framework for addressing the 8.6% increase in case resolution times observed in Türkiye between 2023 and 2024. In the following, in Section II we present a literature review on predicting judicial decisions, in Section III we provide information on the legal process of construction disputes, in Section IV we detail the data collection, preprocessing, feature selection & prediction stages, in Section V we present the findings of the analyses, in Section VI we discuss the results, and in Section VII we present the conclusion of the study.

2. Literature Review

The prediction of judicial decisions started with the pioneering work of Lawlor [16] and has now become a global research area with the advancement of AI, Deep Learning (DA), Machine Learning (ML) and NLP techniques. These studies are carried out in a wide geography from developing countries such as Brazil, China, India, Thailand, Philippines and Turkey to developed countries such as the USA and the UK.

In pioneering studies in developed countries, Katz et al. [19] and Liu and Chen [20] evaluated algorithm performances on different judgement systems. Katz et al. [19] combined a novel feature engineering technique with the Random Forests method to predict two centuries of US Supreme Court decisions with 70-72% accuracy, while Liu and Chen demonstrated the superiority of Support Vector Classifier (SVM) over other algorithms in European Court of Human Rights decisions. Kowsrihawat et al. [21], who developed approaches to improve the effectiveness of SVM techniques in judgement prediction, stated that a Bag of Words

(BoW) approach in previous studies provides low accuracy due to the elimination of word order, and proposed a Bi-GRU model with attention mechanism for criminal cases of the Supreme Court of Thailand. Zhong et al. [17], on the other hand, developed the topological learning framework TOPJUDGE for Chinese criminal cases from a different perspective, modelling the hierarchical structure of the legal decision-making process and outperforming single-task baseline models. However, these studies focused on algorithm performance comparisons, did not address morphological challenges in adjacent languages, and did not focus on the dynamics of the parties' mutual arguments. Our research addresses this methodological gap by systematically analysing the linguistic interaction of claim and defence texts.

Examining data-driven modeling approaches in different disciplines, Koc [22] applied stochastic gradient boosting to model workers' susceptibility to accidents, while Mostofi et al. [23] combined Principal Component Analysis with Deep Neural Networks for housing price prediction. These studies enhanced model interpretability through visualization and methodological transparency. With a similar approach, our study focuses on integrating NLP and ML techniques in legal text analysis, emphasizing the visualization of linguistic patterns in construction disputes and making the results comprehensible for legal practitioners.

Analysing the impact of cultural and structural differences of judicial systems on algorithm selection, Virtucio et al. [24] and Long et al. [18] evaluate the effectiveness of different algorithms in Philippine Supreme Court and Chinese divorce cases, respectively. Focusing on methodological innovations in legal text analysis, Chalkidis et al. [25] proved the superiority of neural network-based language models in ECHR decisions with an F1-score of 82%, while Kaufman et al. [26] demonstrated the potential of decision tree methodology in US Supreme Court decisions. These studies failed to model the specific linguistic features of the case type and the interactional dynamics between the parties' arguments. Our approach with three different dataset configurations (only plaintiff claims, both plaintiff claims and defendant defences, all decisions) overcomes this methodological limitation and quantitatively measures the contribution of pleadings to predictive power.

Shaikh et al. [27], who examine the impact of specialised datasets in legal sub-fields on model performance, focus on murder cases in the Delhi District Court, while Medvedeva et al. [28] develop the JURI SAYS web platform for predicting ECtHR judgments. In contemporary research, Alrasheed et al. [29] analysed time-based disputes in the Kuwaiti construction industry, while Seo and Kang [30] developed an automatic text summarisation paradigm for construction disputes. Kalogeraki and Antoniou [31] conducted a taxonomic analysis of the recent dispute resolution literature. However, these studies have not analysed the linguistic structures and the interaction of party arguments in private law areas, especially in construction disputes, and have not performed decision prediction by using DBE and ML techniques. The approach we have developed integrates LDA and ML techniques in construction disputes and reveals the decisive role of linguistic factors in judicial processes.

Studies on the Turkish legal system gained momentum when Mumcuoglu et al. [32] emphasised the lack of applications of ML and LDA in the Turkish legal system. Akca et al. [10], Aras et al. [33], Ozturk et al. [34] and Sert et al. [35] examined the performance of various algorithms in the decisions of the Court of Cassation and the Constitutional Court. However, none of these studies focussed on a specific dispute, nor did they examine the effect of the linguistic interaction between the parties' arguments on the decision process. The model we present fills this disciplinary gap in the Turkish legal literature by analysing the decisions of the 6th Civil Chamber of the Court of Cassation on construction disputes.

Our study extends the scope of previous research and deals with construction disputes in a morphologically rich and contiguous language such as Turkish. Analysing 2,563 decisions from the 6th Civil Chamber of the Court of Cassation, our

research systematically examines the effect of claim-defence dynamics on prediction success over three different datasets (only plaintiff claims, both plaintiff claims and defendant defenses, all decisions). Our original contribution is the discovery that the inclusion of defence texts in the model significantly improves the prediction performance. Analysing the plaintiff's claim and the defendant's defence together resulted in higher accuracy compared to predictions based on the claim texts alone.

The findings of this study provide legal practitioners with an opportunity to strategically analyse the linguistic structure of the arguments of the prosecution and defence, and provide a practical framework for optimising the 8.6% increase in litigation times in Turkey between 2023 and 2024 [5]. Considering that in traditional legal proceedings, outcomes can only be determined by expert judgement [36], the potential of linguistic analysis-based decision prediction models in the construction industry, which requires large budgets and long periods of time, is significant in reducing financial and moral losses.

3. Legal Process of Construction Disputes

In construction works, the parties may disagree on different issues. These disputes may occur during the implementation phase of the construction contracts signed between the parties or during the contract phase. The court is often used as an official remedy to resolve disputes [37]. The court process varies according to each country. The Turkish Code of Obligations No. 6098, which determines the limits of the obligations and rights of the parties in the Turkish construction industry, is based on the Swiss Code of Obligations. In this respect, the relevant law is similar to the legal system of Switzerland, Germany, and France. When differences of interest between the parties turn into disputes and are brought to the judiciary, the judicial process varies between pre-contractual and post-contractual. Pre-contractual disputes fall under the jurisdiction of administrative jurisdiction, while post-contractual disputes fall under the jurisdiction of judicial jurisdiction. This study focuses on post-contractual disputes. In Türkiye, disputes are first referred to the courts of first instance and then to the courts of appeal. The final decision authority for ongoing appeals is the Council of State for pre-contractual disputes and the Court of Cassation for post-contractual disputes. The final decision authority for construction disputes is the 6th Civil Chamber of the Court of Cassation. The judgment process for post-contract disputes is shown in Fig. 1.

The 6th Civil Chamber of the Court of Cassation is the final decision-making authority in construction disputes. Its decisions are binding and have sanction power. The number of judgments and the time taken to reach decisions over the years, as presented in Fig. 2, shows a significant workload. Considering that the dynamics of each dispute are different, the importance of decision-support activities in judicial processes becomes apparent.

4. Methodology

ML is an effective technique for extracting valuable insights from extensive datasets. Text categorization is a crucial aspect of ML techniques and offers automated sorting of texts into specific groupings. Classification of legal texts is crucial for anticipating court rulings and examining legal proceedings. Precise categorization of these documents enables the anticipation of case results and the efficient management of legal proceedings. NLP methods involve computer-based methods to automatically interpret and assess texts created by humans. NLP studies language semantics, syntax, morphology, and phonology. NLP methods are used in the data preprocessing stage of ML approaches in this research. NLP plays a crucial role in preprocessing, enhancing the analysis of texts by ML algorithms and boosting the precision of prediction models. Taking actions like changing uppercase letters to lowercase, deleting punctuation, and eliminating unnecessary words aid in data cleaning and analysis preparation. Feature extraction and selection are crucial when using ML techniques to anticipate judgment decisions in the construction industry.

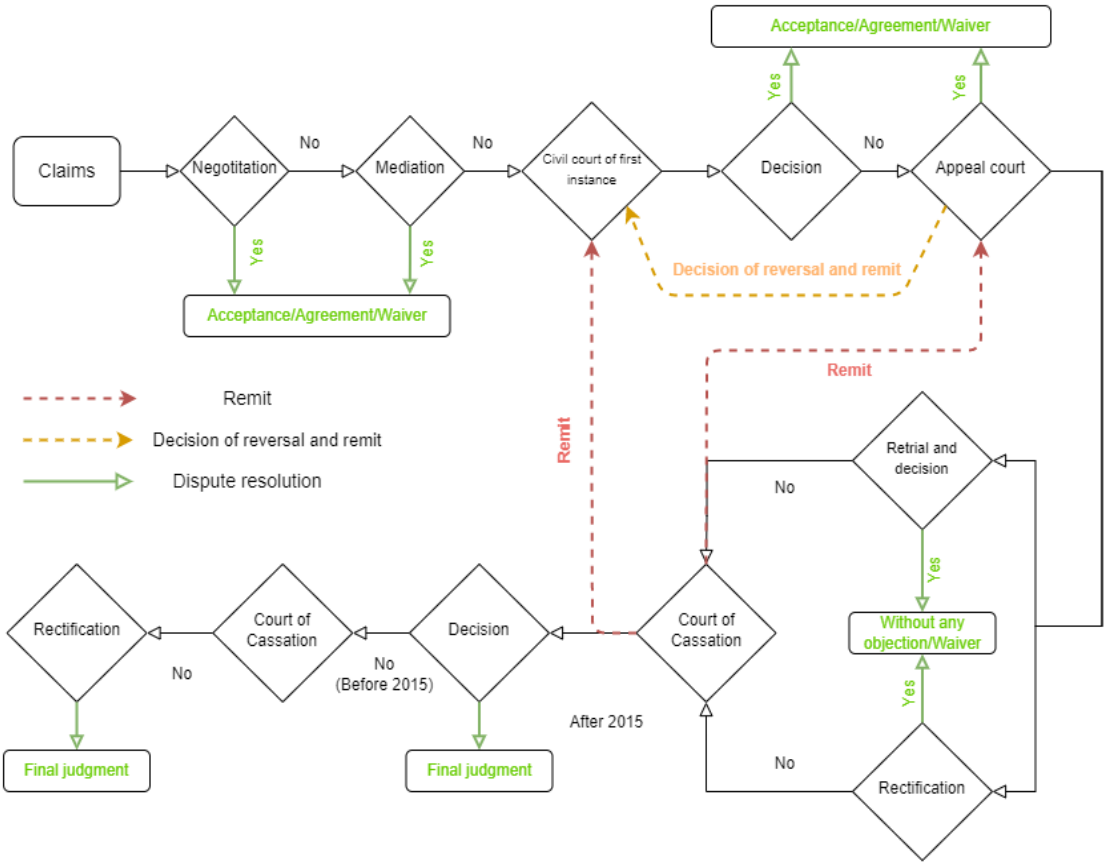

Fig. 1. The judicial process for disputes arising after the contract

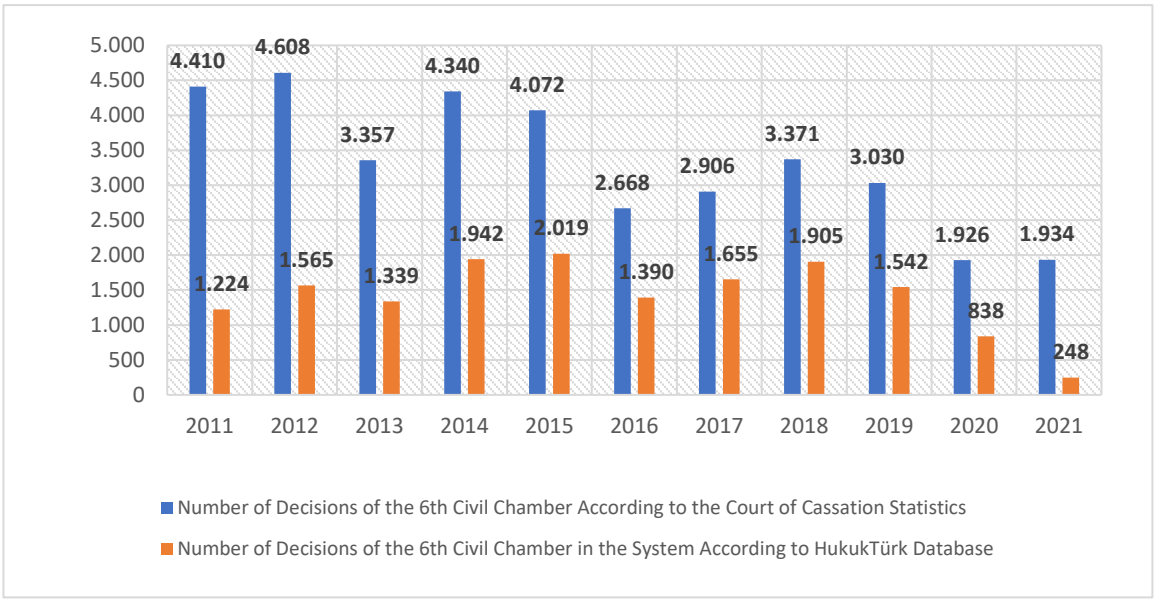

Fig. 2. Number of decisions of the 6th civil chamber of the court of cassation by years [38]

In this research, different techniques like Term Frequency-Inverse Document Frequency (TF-IDF), Word to Vector (Word2Vec), and FastText were employed to extract significant features from the text data. The top representative characteristics were chosen from the features that were extracted. Using different ML algorithms, models were developed with the chosen features. Fig. 3 illustrates the flow method employed in the study.

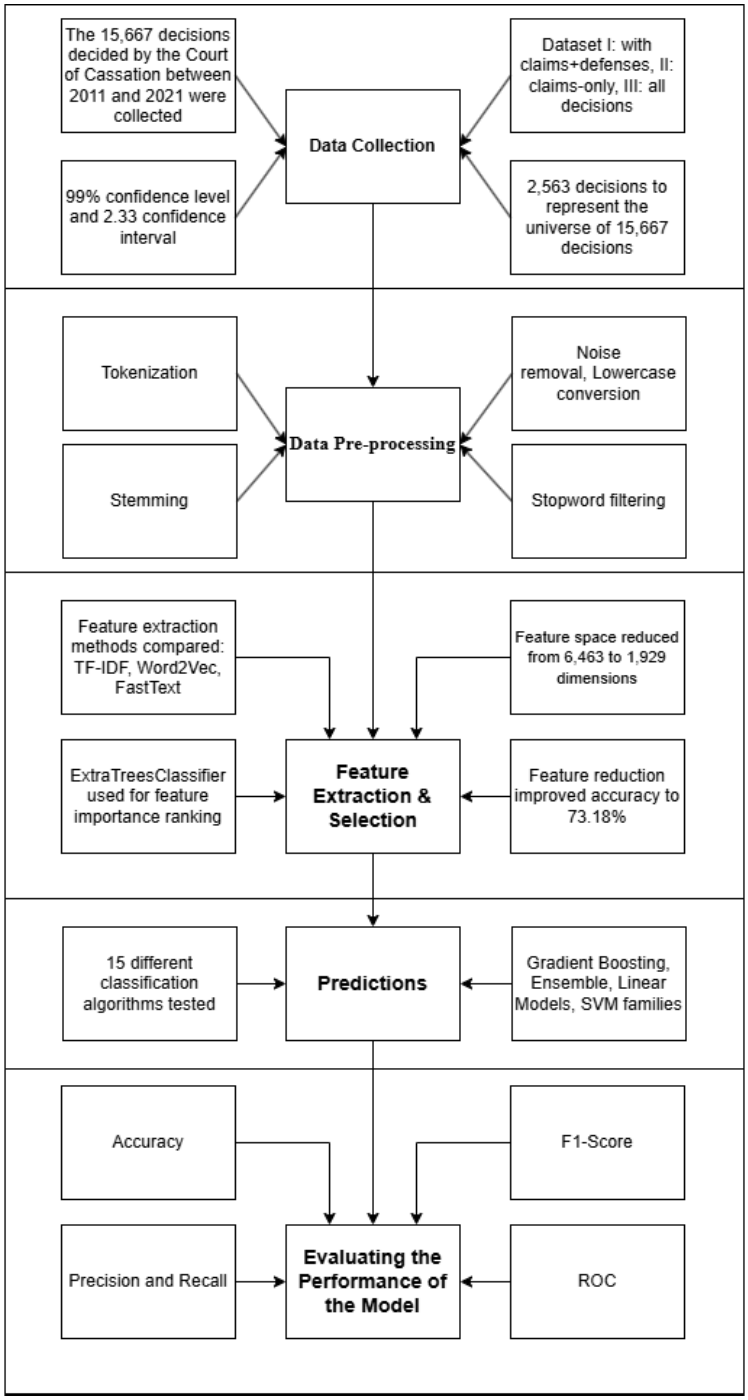

Fig. 3. Workflow of the research

4.1. Data collection

The 6th Civil Chamber of the Court of Cassation is responsible for handling construction disputes as well as disputes in the fields of co-operative law, works contracts and commercial law. Not only construction disputes constitute the workload of the Chamber. Between 2011 and 2021, the number of files decided by the Chamber as (i) reversal, (ii) approval, and (iii) partial decisions is 36,622 [39]. The Court of Cassation publishes a limited number of judgements to the public and legal databases by pre-processing the files it decides due to personal data and trade secret concerns. The published judgements include precedent-setting judgements as well as short and non-detailed judgements, which mostly only contain the decision of approval. In this context, the number of decisions of the department published to the public and legal databases between 2011 and 2021 is 15,667 [39].

The sample selection was designed with statistical rigour and a research-oriented approach. Out of 15,667 decisions published by the Court of Cassation between 2011 and 2021, 2,563 decisions were randomly selected using the Cochran formula (Fig. 4). While this calculation scientifically guarantees the representativeness of the universe with 99% confidence level and 2.33% margin of error, the assumption of p=0.5 has been adopted as the most common and reliable method for unknown distributions in the literature [40]. To ensure statistical representativeness, the target sample size (n) was calculated using Cochran's finite population correction formula for proportion estimation [41]:

$$n = \frac{Nx\ Z^2 x\ p\ x\ (1-p)}{(N-1)\ x\ E^2 + Z^2\ x\ p\ x\ (1-p)} \quad (1)$$

Where:

- N=15,667 (Total published decisions),
- Z=2.576 (Z-score for a 99% confidence level),
- p=0.5 (Conservative proportion maximizing variability),
- E=0.0233 (2.33% margin of error).

The selected decisions were optimised to create a dataset specific to construction disputes. In a comprehensive review process conducted by two independent researchers, only texts directly related to the research question (decisions detailing the parties' claims/defences) were retained, while short “Approval” and “Approval with Correction” decisions and texts not directly related to the research question were eliminated. This filtering increased the model's ability to capture meaningful language patterns and strengthened internal validity.

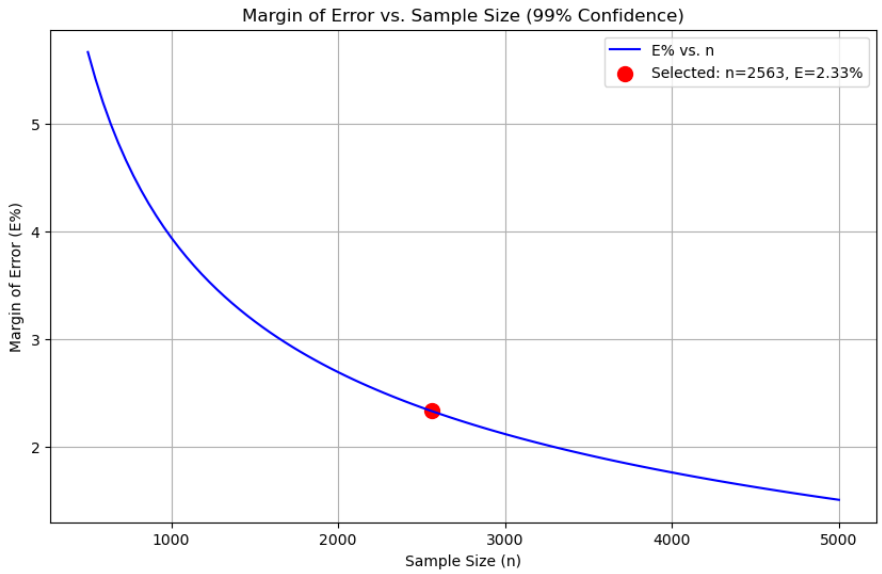

Fig. 4. Sample size and error margin at 99% confidence

The process was supported by multiple checks that maximised methodological reliability. Inter-researcher discrepancies were resolved through the refereeing of a third expert, thus minimising the risk of bias. Furthermore, the combination of randomisation and content filtering ensured that a domain-specific dataset was created while maintaining representativeness of the population.

This study represents a first in the literature by presenting a unique data set that systematically documents the claims and defences of the parties. The 2,563 construction dispute decisions rendered by the 6th Civil Chamber of the Court of Cassation between 2011 and 2021 naturally contain critical elements reflecting the linguistic and logical structure of the parties' legal struggle. This dataset is the first study to holistically analyse the claim-defence dynamics in Turkish legal texts. Unlike traditional LJP studies, it has made it possible to model not only the decision outcomes but also the linguistic complexity and logical context of the parties' legal arguments.

During the analysis, different decision writing styles of the judges in the construction dispute decisions drew attention. It is thought that the reason for this is that the judges who adjudicate the disputes do not adopt a general-universal decision writing style and that the judges' education in different periods may have an effect. As a result of the analyses, it was observed that some of the decisions were detailed decisions including the plaintiff's claim and the defence of the defendant, and some of the decisions included only the plaintiff's claim without the defence of the defendant (only the claim was rejected). According to these differences, the data set was classified as Dataset I (Plaintiff's claim-defendant's defence), Dataset II (Only plaintiff's claim) and Dataset III (All decisions). This detail captured constitutes another unique aspect of the study different from literature.

All the Court of Cassation decisions obtained consist of some generalised sections. These can be summarised as follows as shown in Fig. 5:

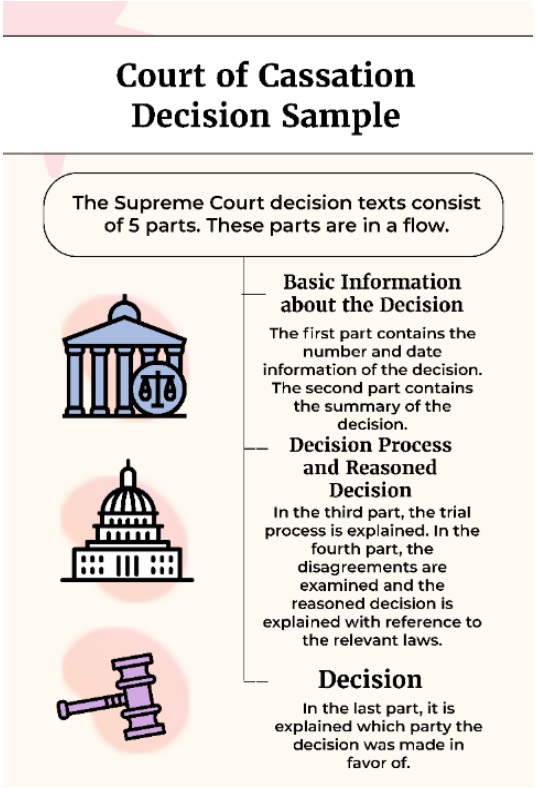

Fig. 5. Court of cassation decision sample [43]

i. The part containing the Court of Cassation decision number and date,
ii. The part containing the summary of the Court of Cassation judgment,
iii. The part that provides brief information about the case process,

The first four of the five sections of the Court of Cassation judgements described above were included in the study. On the contrary, the inclusion of the sections explaining the decision in the learning process may cause overfitting of the models and decrease the generalizability of the predictions. Since this situation is thought to negatively affect the prediction process, it is not included in the learning process in many studies [32, 42]. Since these sections contain explicit decision statements that the model can be directly associated with certain features, the ability of the model to make meaningful inferences about the process leading to the decision may be reduced. This may result in artificially high-performance metrics that do not accurately reflect the model's

ability to predict decisions based on the analysis and arguments in the text. For these reasons, the last part, where the decision is explained, is not included in the study. The last part is only used for labelling the decision result.

4.2. Data pre-processing

The agglutinative structure of Turkish is based on a linear arrangement of affixes, typical of the Ural-Altaic language family. This sequence provides a morphological flexibility not seen in analytical languages such as English but increases preprocessing complexity in NLP. Turkish poses significant challenges in text classification studies using ML methods due to its suffixed structure and increased variability of words. The various suffixes appended to word roots increase the number and variety of words that can be generated. This diversity and suffixation structure complicate text analysis. In studies aimed at improving Turkish text classification performance, morphological analysis techniques have been applied to reduce word diversity, resulting in consistent texts with word roots and enhancing text classification performance [44]. Turkish, as an agglutinative language, introduces significant challenges in text preprocessing due to its suffix-driven morphology. For example, a single root like yap- ("to do") can generate complex derivatives such as yapılamamıştı ("it could not have been done"), leading to high lexical variability [12]. Legal texts further amplify these challenges by incorporating domain-specific terminology (e.g., sorumluluk [liability], ibra [discharge]) and procedural phrases (e.g., mahkeme kararı [court decision]), which require careful handling to preserve legal relevance. The pre-processing pipeline, implemented using Python's NLTK library [45], included the following steps:

1. *Tokenization & Lowercasing:* The raw text was split into word-level tokens, and all characters were converted to lowercase to ensure consistency (e.g., İnşaat → inşaat).
2. *Noise Removal:* Punctuation marks, numbers, symbols, abbreviations, and extraneous whitespace were systematically removed. For instance, "§15'te belirtilen…" was simplified to "belirtilen".
3. *Stopword Filtering:* NLTK's default Turkish stopword list was applied to remove generic non-informative words (e.g., ve [and], için [for]).
4. *Stemming:* Words were reduced to their root forms using NLTK's Turkish stemmer to address agglutination. For example:
    - yüklenicinin ("contractor's") → yüklenici ("contractor")
    - sorumluluklar ("liabilities") → sorumluluk ("liability")

This workflow effectively balanced the reduction of morphological complexity with the retention of legally critical terms, ensuring robust input for ML models.

4.3. Feature extraction and selection

After the preprocess of the raw text, highly representative and meaningful features need to be extracted and selected. The text remaining after pre-process contains large and intricate data. These are converted into more meaningful and smaller features, and vector space models are obtained. Word Embedding, based on the architecture of artificial neural networks, converts the words in the text into number vectors. TF-IDF is one of these models. TF-IDF is the product of the preponderance value (TF) of the most repetitive term in the text and the preponderance value (IDF) of the least repetitive term. This value explains the importance of the term in the text [46]. Word2Vec, another Word Embedding model, is a 2013 Google product library based on Cbow and skip-gram. The Word2Vec technique uses 2-layer (single hidden layer) trained neural networks that produce similar outputs from input data [47]. It develops a model by establishing the semantic relationship compared to the TF-IDF technique, which ignores the semantic relationship [48]. It is frequently used for unlabelled data [49].

Facebook AI Research developed a library called FastText in 2016 as a module of the Word2Vec technique. The technique performs text classification and performs better performance and speed than other techniques by converting texts into

vectors [50]. FastText includes Continuous Bag of Words (CBOW), which predicts the target word with its surrounding words, and skip-gram structures that predict the input and surrounding words. It uses and computes the n-gram structure within the skip-gram structure, which allows for a more specific word vector structure. High success is achieved with this structure for rare words in languages rich in word structure. After feature extraction, it is necessary to select between the features. Feature selection is essential for speed and controllability in high-dimensional models. It aims to increase the accuracy success by removing unnecessary and irrelevant data [51]. Following the extraction of features, the utilization of the Extra Trees Classifier represents a highly effective approach to enhance the performance of predictive models. Feature extraction represents a pivotal step, with the objective of transforming raw text data into numerical representations that can be employed by ML algorithms. Once features have been extracted, the Extra Trees Classifier, which is an ensemble method of decision trees, facilitates the identification of the most pertinent features by evaluating their importance based on the reduction of impurity or variance in the data. This method not only improves classification accuracy but also reduces computational complexity by eliminating irrelevant and redundant features. This process ensures that the most informative features are utilized for building robust predictive models [52, 53].

4.4. Predictions

Subsequent to feature extraction and selection, 15 ML algorithms were rigorously evaluated to predict judicial outcomes of construction dispute resolutions in the Court of Cassation, spanning diverse algorithmic families. Gradient Boosting algorithms (HistGradientBoosting, GradientBoosting, XGBoost) were prioritized for their scalability and robustness to missing data [54-56]. Ensemble methods (Voting, AdaBoost, Bagging) were employed to reduce variance and enhance generalization by combining multiple learners [57-59]. Linear models (RidgeClassifierCV, RidgeClassifier, SGDClassifier, LogisticRegressionCV, PassiveAggressive) provided stable baselines for high-dimensional TF-IDF vectors while addressing multicollinearity [60-62]. Support Vector Machines (LinearSVC, NuSVC, SVC) were selected for their ability to handle high-dimensional text data and class imbalance [63, 64]. A comprehensive rationale for each algorithm, including domain-specific justifications and references, is detailed in Table 1.

The default parameters provided by the respective libraries in Python for all ML methods were utilised in this study. No additional parameterisation was conducted, as the default settings were found to provide sufficient performance for the specified scope and objectives. The use of default parameters ensures the reproducibility of the study and facilitates its replication by other researchers.

In the classification of Turkish legal texts, the utilisation of a range of evaluation metrics is of paramount importance, as each metric assesses distinct performance aspects and provides a comprehensive evaluation of the classification models. While accuracy, which measures the ratio of correct predictions to total predictions, provides an overall indication of performance, it can be misleading in cases of class imbalance, which is a common occurrence in legal texts [68]. Therefore, the use of additional metrics in conjunction with accuracy is essential [69]. In the context of legal domains, precision, which measures the proportion of true positive predictions among all positive predictions, is of particular importance. This is because minimising false positives is vital in such domains [70]. In contrast, recall, which measures the proportion of actual positives correctly identified by the model, is crucial in minimising false negatives in legal text classification. This is because it ensures that significant legal cases are not overlooked [71]. F1-Scores, which are a harmonic mean of precision and recall, offer a balanced assessment, especially in scenarios with unbalanced classes [72].

Table 1. Algorithm selection reasons

No	Algorithm	Algorithmic Family	Selection Reason (Ref.)
1	HistGradientBoosting	Gradient Boosting	Fast, large data; robust to missing data [56]
2	GradientBoosting	Gradient Boosting	High accuracy via error correction [55]
3	XGBoost	Gradient Boosting	Scalable; L1/L2 prevents overfitting [54, 65]
4	Voting	Ensemble	Combines models; reduces variance [58]
5	RidgeClassifierCV	Linear Models	L2 reg.; handles multicollinearity [60]
6	RidgeClassifier	Linear Models	Stable via L2; baseline for high-dim [60]
7	SGDClassifier	Linear Models	For large data; supports online learning [62]
8	LinearSVC	Support Vector Machines	Fast for high-dim. text data [63]
9	NuSVC	Support Vector Machines	Flexible imbalance handling via Nu. [64]
10	LogisticRegressionCV	Linear Models	Auto hyperparameter tuning [66]
11	SVM	Support Vector Machines	Captures non-linear relations [63]
12	AdaBoost	Ensemble	Iteratively improves weak learners [59]
13	PassiveAggressive	Linear Models	Efficient for noisy/online data [67]
14	Bagging	Ensemble	Reduces variance via bootstrapping [57]
15	LogisticRegression	Linear Models	Interpretable, efficient baseline [61]

The receiver operating characteristic area under the curve (ROC-AUC) assesses the discriminative ability of the model across different threshold settings, providing an insight into the performance of the model at different decision boundaries [73]. Confusion matrices detail how the model performs, showing true positives, true negatives, false positives and false negatives. This is essential for understanding the specific areas in which the model performs well or needs improvement [69]. These metrics are crucial for capturing the complexity and linguistic nuances of legal texts, ensuring a robust and comprehensive evaluation of classification models used in legal text classification.

5. Findings

Between 2011 and 2021, 2,563 target datasets were determined with a 99% confidence interval and a margin of error of 2.33% from 15,667 decisions decided by the 6th Civil Chamber of the Court of Cassation and published in legal databases [39]. In contrast to the previous studies, 862 cases (33%) with the defendant's defence against the plaintiff's claim in the decision texts and 1,701 decision texts (67%) without the defendant's defence in the decision texts were classified separately. The data set designated as "Dataset I" comprises the information in the decision texts in which the defendant's defences counter the plaintiff's claims. In contrast, the dataset comprising the decisions in which the defendant's defences do not oppose the plaintiff's claims is designated as "Dataset II." Finally, the dataset comprising all the decisions is designated as "Dataset III," and all the methodology stages have been applied. Subsequently, the dataset underwent tokenization, the initial step of data cleansing. Subsequently, the large characters were transformed into smaller ones, and the punctuation, numerals, symbols, abbreviations, white spaces, and stop words, which are meaningless elements, were removed from the text. The removal of noisy entities and the implementation of stopword processing have been completed. To prevent the negative impact of stopwords on the accuracy of the prediction, the number of unique features in the corpus was reduced using the NLTK library, and the question words were removed from the text. In the final stage of the data processing, the words were classified according to their parts of speech, and the stems were identified through stemming. To ensure that the learning performance of documents with over 20,000 characters and less than 2,500 characters is not negatively affected, all datasets were filtered to exclude these documents. Consequently, the number of decisions in Dataset I was 730; in Dataset II, it was 1,482; and in Dataset III, it was 2,212. All instances of repeating words in

the decision-making text were eliminated. In this context, dataset I comprises 16,441 words, dataset II 19,185 words, and dataset III 24,395 words, extracted from the texts to present the most recent state of the datasets. The data from the three different datasets is presented in Table 2.

The word and character count statistics for different datasets and decision types are presented in detail in Table 2, as they are considered to have a significant impact on model performance, particularly in the context of feature extraction processes. The variability in the number of words and characters between different datasets and decision types demonstrates the differences in length and content of the texts. These statistics elucidate the context in which model performance is evaluated and furnish information on the generalisability of the results [74]. It is acknowledged that longer texts can provide more contextual information, thereby enabling more detailed analyses [75]. Furthermore, the uniformity in the number of characters indicates the homogeneity of the texts in terms of structure. This data is essential for optimising the feature extraction and model training processes and for understanding and interpreting the performance results of our text classification models [76].

Class imbalance is a critical issue that can significantly impact the performance of ML models. He and Garcia [77] documented that imbalanced datasets cause classifiers to be biased toward the majority class and exhibit poor sensitivity for minority classes. Similarly, Buda et al. [78], in their systematic study, examined the detrimental effects of class imbalance on classification performance and emphasized the importance of appropriate sampling strategies based on the degree of imbalance. Common mitigation strategies include synthetic data generation [79] and under sampling [80]. While data augmentation methods have the potential to reduce class imbalance and improve model performance in limited data scenarios [81], their application in the context of legal NLP raises critical concerns. Synthetic text generation through techniques such as synonym substitution or back-translation risks distorting domain-specific legal terminology, altering case law references, or misrepresenting the logical structure of judicial arguments [82, 83]. For example, Ishikawa et al. [84] showed that synthetic texts reduce model reliability by creating semantic inconsistencies. Similarly, Shorten et al. [85] warned that artificial examples cannot preserve the contextual integrity of decisions in languages such as Turkish, where morphological variations are semantically sensitive. In addition, Zhou et al. [86] showed that undersampling methods can provide effective results in imbalanced datasets by approaching the ideal classification boundary.

Table 2. Word and character count statistics for different datasets and decision types

Data Set	Decision Type	Nr. of Decision	Nr. of Reduced-Decision	Word Count			Character Count		
				Min	Avg	Max	Min	Avg	Max
I	In favor of defendant	497	365	240	653.87	2,176	1,981	5,368.45	17,828
	In favor of plaintiff	365	365	266	660.66	2,517	2,290	5,419.02	20,372
	Total	862	730						
II	In favor of defendant	960	741	165	538.72	1,828	1,433	4,411.19	15,132
	In favor of plaintiff	741	741	164	550.69	3,423	1,383	4,510.62	27,330
	Total	1,701	1,482						
III	In favor of defendant	1,457	1,106	165	578.01	2,176	1,433	4,737.73	17,828
	In favor of plaintiff	1,106	1,106	164	586.98	3,423	1,383	4,840.41	27,330
	Total	2,563	2,212						

In light of these findings, considering the terminological precision and contextual integrity of legal language, undersampling approaches that preserve the natural data distribution may be preferred to avoid the potential risks of synthetic data generation. To address class imbalance without compromising legal accuracy, we chose to create balanced datasets by under sampling rather than data augmentation (Table 2). This approach is in line with Mumcuoglu et al. [32], who obtained successful results on Turkish legal texts. By preserving the original linguistic and legal patterns, we have ensured that the model predictions remain consistent with the actual legal discourse.

After data preprocessing, the dataset was analysed comparatively with TF-IDF, Word2Vec and FastText methods for feature extraction. As a result of the analyses, TF-IDF was found to be more successful compared to other feature extraction methods (Table 3). There are various reasons for the prominence of the TF-IDF feature extraction method. The most prominent reason is that while Word2Vec and FastText methods are suitable for working with high computational capacity in large-scale data, TF-IDF method is more suitable for working with low computational capacity in small data sets as in the current study [87]. Another prominent reason is that the TF-IDF method allows for a more consistent and balanced analysis of language-specific complexities and semantic relations, since the TF-IDF method evaluates words with the assumption that they are independent of each other [75]. In addition, the TF-IDF method enables students to achieve success in comprehension and evaluation processes by determining the degree of importance of words and documents relative to each other [88]. For these reasons and due to its higher accuracy compared to other methods, the TF-IDF method was determined as the primary feature extraction method (Table 3).

The most successful approach to extracting features from the dataset was the TF-IDF method, which involved reducing the number of features. Although the TF-IDF method is effective in vectorising textual data, the resulting high-dimensional feature space (6,463 features) may adversely affect the computational efficiency and generalisation capability of the classification algorithm. Feature reduction provides notable advantages by helping to avoid the curse of dimensionality [89], enhancing model generalisation and computational efficiency [90], while reducing noise [91]. Among the widely used algorithms in text classification, ExtraTreesClassifier, developed by Geurts et al. [92], demonstrates superior performance particularly with high-dimensional textual data.

Unlike filter-based methods like Chi-square Yang and Pedersen [93] and Information Gain [94], ExtraTreesClassifier accounts for feature interactions and contextual dependencies inherent in agglutinative languages. In this context, the ExtraTreesClassifier ensemble method offers two significant advantages in feature selection: i. ExtraTreesClassifier provides a robust evaluation of feature importance in each tree structure, ii. As it calculates feature importance through multiple decision trees, it does not exhibit excessive dependency on individual features. This characteristic creates a feature selection mechanism that is more resistant to noise in the dataset. Additionally, the random splitting strategy of ExtraTreesClassifier minimises the effect of high correlations frequently observed among features obtained with TF-IDF, facilitating the identification of features that genuinely contribute to classification performance.

Table 3. Performance metrics of different methods

Method	Accuracy	Recall	Precision	F1-Score
FastText	0.6048	0.6050	0.6062	0.6035
Word2Vec	0.5931	0.5932	0.5937	0.5921
TF-IDF	0.6722	0.6724	0.6811	0.6673

Whilst Chi-square and Information Gain methods evaluate features independently, ExtraTreesClassifier can account for interactions between features, thus exhibiting superiority in capturing contextual meanings frequently observed in legal texts. Moreover, in the face of the limited dataset problem encountered in legal texts, ExtraTreesClassifier has been observed to be more resistant to overfitting due to its random sub-sampling and splitting strategy [92]. This methodological approach reduced the feature space from 6,463 to 1,929, both increasing computational efficiency and improving the classification performance of the Random Forest algorithm. The performance criteria presented in Tables 4 and 5 quantitatively demonstrate the improvement achieved after feature reduction.

Table 4. Performance metrics indicators obtained after feature reduction

Performance Indicator			
Accuracy	Recall	Precision	F1-Score
0.7318	0.7319	0.7457	0.7262

After reducing the number of features, only the RandomForest classification method was not used for dataset III at this stage, and the success of different classification methods was investigated. For this purpose, it was analyzed via 15 different classification methods, and the indicators in Table 5 were obtained.

The findings indicate that the HistGradientBoosting classification method provided most successful results. Since the best feature extraction method and the best classification method were determined, the classification was also performed for the remaining two datasets, I and II, and the final indicators were presented in Table 6. The final indica-tors are presented in Table 6. The confusion matrix tables for all three datasets are also presented in Fig. 5.

The examination of the performance metrics for the three datasets has revealed several significant trends and findings pertaining to the effectiveness of the various classification algorithms employed. With respect to Dataset I, the Ensemble Gradient Boosting Classifier has demonstrated the highest accuracy (0.8738) and precision (0.8758) rates. The classifier effectively identifies true positives and negatives, with 313 true negatives, 52 false positives, 40 false negatives, and 325 true positives in the confusion matrix (Fig. 5). The detailed documents, with an average word count of 653.87 for defendants and 660.66 for plaintiffs, emphasize the complexity that the classifier has to manage. The recall is 0.8904, indicating that the classifier is able to correctly identify a significant proportion of positive cases.

Table 5. Success results for different classifiers as a result of feature reduction of TF-IDF features for Dataset III

No	Algorithm	Accuracy	F1-Score	Recall	Precision
1	HistGradientBoosting	0.8598	0.8594	0.8598	0.8630
2	GradientBoosting	0.8557	0.8553	0.8558	0.8588
3	XGBoost	0.8490	0.8484	0.8490	0.8529
4	Voting	0.8485	0.8481	0.8485	0.8516
5	RidgeClassifierCV	0.8426	0.8424	0.8426	0.8446
6	RidgeClassifier	0.8426	0.8424	0.8426	0.8446
7	SGDClassifier	0.8422	0.8419	0.8421	0.8442
8	LinearSVC	0.8413	0.8411	0.8412	0.8425
9	NuSVC	0.8349	0.8345	0.8349	0.8378
10	LogisticRegressionCV	0.8336	0.8334	0.8335	0.8347
11	SVM	0.8286	0.8281	0.8286	0.8315
12	AdaBoost	0.8273	0.8264	0.8272	0.8317
13	PassiveAggressive	0.8241	0.8239	0.8241	0.8250
14	Bagging	0.8232	0.8226	0.8232	0.8265
15	LogisticRegression	0.8078	0.8073	0.8078	0.8102

Table 6. Final measures of performance metrics of algorithms on different datasets

Dataset	Algorithm	Accuracy	F1-Score	Recall	Precision
III	HistGradientBoosting	0.8598	0.8593	0.8598	0.8630
II	RidgeClassifierCV	0.8453	0.8451	0.8453	0.8470
I	GradientBoosting	0.8738	0.8736	0.8740	0.8758

Nevertheless, the majority of errors can be attributed to the ambiguity of legal terminology and the complexity of sentence structures. The F1-Score, which is a measure of the balance between precision and recall, is 0.8827, indicating a good overall performance.

The results of Dataset II, which was analyzed using a Linear Model Ridge Classifier CV, demonstrated lower accuracy (0.8453) and precision (0.8470) rates compared to those obtained for Dataset I. The confusion matrix indicated that there were higher misclassification rates, with 609 instances of true negatives, 132 instances of false positives, 97 instances of false negatives and 644 instances of true positives. The consistency of the data can be observed in the uniformity of the word and character count. The average word count for defendants was 538.72, and for plaintiffs, it was 550.69. Notwithstanding the classifier's balanced performance, the higher misclassification rates demonstrate the challenges inherent in managing the diversity of the dataset. The occurrence of misclassifications can be attributed to the inconsistent use of terminology. The recall rate and F1-Score of dataset II indicate that it has performed adequately in identifying and stabilising positive cases, although there is scope for improvement.

The HistGradientBoosting of Dataset III achieved competitive metrics with an accuracy of 0.8598 and precision of 0.8630. The confusion matrix records 953 instances of true negatives, 153 instances of false positives, 157 instances of false negatives, and 949 instances of true positives. This dataset contains the greatest number of decisions and the most extensive average word counts (578.01 for defendants and 586.98 for plaintiffs), indicating comprehensive and detailed documentation. Although the classifier performs well overall, it still struggles to correctly identify all positive cases due to the complexity and volume of the documents. In this dataset, errors exhibit a more complex structure due to the presence of detailed case histories, indicating that the processing of extensive cases and multiple types of evidence is challenging for the model. The recall rate and F1-Score of Dataset III are 0.8578 and 0.8604, respectively. This reflects a strong performance overall, albeit with a slight imbalance. This is slightly less balanced than Dataset I.

A more detailed examination of the confusion matrices reveals that Dataset I demonstrated superiority over the other datasets in the correct prediction of the positive classes (Fig. 6).

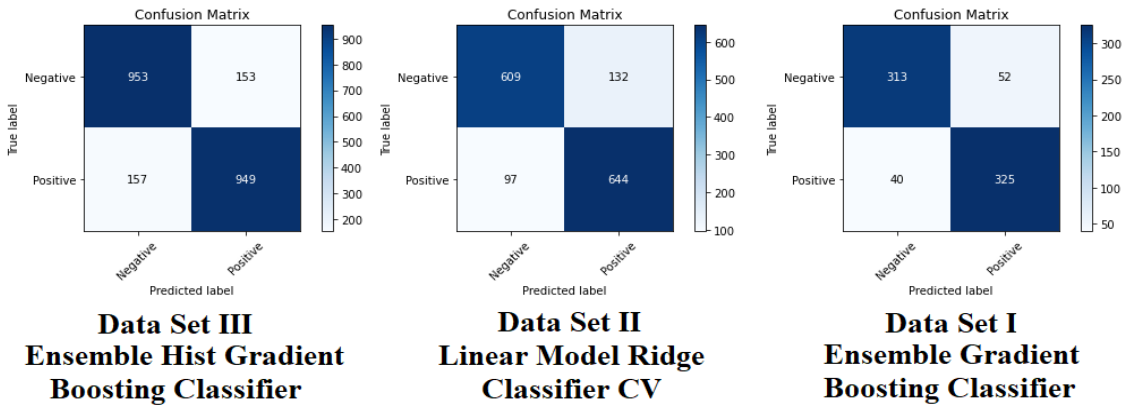

Fig. 6. Confusion matrices of three different datasets

However, the rate of positive misprediction of the negative classes remained high. Dataset II exhibited a reduction in classification accuracy due to inconsistencies in terminology. However, the rate of positive misprediction of negative classes was relatively lower. Dataset III exhibited the highest error rate, which was attributed to the presence of complex case information. Both negative and positive classes demonstrated elevated rates of misprediction.

6. Discussion

This study focuses on disputes in the implementation phase of contracts signed between parties in the public and private construction industry in Türkiye. It addresses a significant methodological gap in the literature by analyzing the linguistic interaction between claim and defense texts in Turkish, a morphologically rich and agglutinative language. It also performs decision prediction by text classification using ML methods on official judgement texts of construction disputes in the implementation phase. In the process of digitizing legal texts, a total of 15,667 decisions were collected from the public database of the Court of Cassation between the years 2011 and 2021. From this extensive dataset (Universe), a target data set representing 2,563 universes (sample) was determined with a 99% confidence interval and a 2.33% margin of error.

The texts were then tokenized for the application of NLP techniques and preprocessed by converting uppercase letters into lowercase letters, removing punctuation marks and meaningless words. Nevertheless, a number of difficulties were encountered during this process. One of the most significant challenges encountered during this process was the standardization of texts with varying formats and structures. It is important to note that judges do not write their decisions according to a uniform writing format. Furthermore, the fact that Turkish is a contiguous language necessitated morphological analysis and the reduction of words to their roots. As words are derived in adjoining languages through the addition of affixes, the accurate separation of the root and affix components of each word required a high degree of precision. To address this challenge, advanced morphological analysis tools and the Python NLTK library were employed to reduce words to their roots and extract meaningful features.

The judgement texts are divided into three different datasets: (i) containing the plaintiff's claims and the defendant's defences, (ii) containing only the plaintiff's claims, and (iii) all decisions. This different dataset approach distinguishes this study from previous studies by systematically measuring the effect of defense texts on prediction success, an aspect that has been overlooked in previous research. By applying ML techniques, a success rate of 87.38% was obtained for the prediction of decision texts containing the plaintiff's claims and the defendant's defences, 84.53% for the prediction of decision texts containing only the plaintiff's claims and 85.98% for the prediction of all decision texts. These findings demonstrate that the combined analysis of plaintiff claims and defendant defenses provides higher accuracy compared to predictions based solely on claim texts, proving that modeling the mutual arguments of the parties enhances prediction performance.

The relatively low number of decisions and high variability in text length indicate that Dataset I contains cases that are more complex and variable. This affects the performance of the classifier. Conversely, the higher number of decisions and more consistent text length indicate a dataset with more homogeneous cases, which leads to more efficient performance of the classifier. The improved performance metrics demonstrate better processing of legal documents with uniform features. The high number of judgments and the average length of texts in Dataset III indicate that it contains the most comprehensive and detailed cases. The classifier's performance demonstrates its ability to effectively process large volumes of complex legal documents. Additionally, the error analysis reveals that common errors in misclassified cases highlight specific challenges associated with each dataset. Dataset I addresses ambiguous legal terms and lengthy sentences, while

Dataset II confronts inconsistencies in terminology. Dataset III's errors are frequently attributed to the complexity of its case histories and the variety of evidence involved, exemplifying the intricacies of processing comprehensive cases. The success rates obtained exceed the success rates reported in the literature [4, 65], emphasizing the value of focusing on a specific legal discipline and analyzing the linguistic patterns in construction disputes with methodological transparency.

Our study addresses methodological gaps in the literature by developing decision prediction models for construction disputes in Turkish, a morphologically rich and agglutinative language. By analyzing 2,563 decisions from the 6th Civil Chamber of the Court of Cassation (representing a universe of 15,667 decisions with a 99% confidence interval and a 2.33% margin of error) through three different dataset configurations, we quantitatively measured the effect of defense texts on prediction success. Our findings demonstrate that the combined analysis of plaintiff claims and defendant defenses (Dataset I) provides higher accuracy (87.38% versus 84.53%) compared to predictions based solely on claim texts (Dataset II).

Mumcuoglu et al. [32], a pioneer in the application of NLP and ML techniques in Turkish judicial decisions, compiled many judicial decisions and achieved the highest rate of 91.80% decision prediction success. The corpus of the study includes the decisions of five different judicial units. The Civil Court of Appeal decisions, one of the five different judicial decisions, are similar in content to the dataset used in this study. Although 91.80% predicted success was achieved in the Court of Appeal on Taxation decisions, the highest success rate of 69% was achieved in Civil Court of Appeal decisions. Among the reasons for the variation in the success rates in different courts, as stated by the authors, it can be counted that the relevant court decisions are complex and contain different dispute issues. From this point of view, the high success rates obtained in the present study support the fact that the use of a data set specific to a particular legal discipline will increase the success rate, as suggested by related studies [32, 95]. This study, which focuses only on construction disputes, has achieved better success in terms of prediction of decision texts with similar content with the highest success rate of 87.38%. In addition, the BILSTM algorithm, which has the highest accuracy success in the related study, achieved an F1-Score value of 0.68. HistGradientBoosting algorithm outperformed the related study [32] with an F1-Score of 0.86 in the dataset containing all Supreme Court decisions. In terms of the number of data, the dataset containing 2,563 decisions (16%) representing 15,667 decisions showed high success despite having less data. In a similar study, Ozturk et al. [34], classified 59,822 Supreme Court of Appeals decisions and achieved the highest success rate of 96.80% in decision prediction. However, 92% of the decisions belonged to one prediction class and data augmentation was applied to the other class. It is thought that data augmentation in the field of law, which has its own domain-specific vocabulary, will cause the distribution of the augmented data to be different from the original data distribution [83]. In addition to this reason, in the present study, data augmentation was not applied since the number of decisions in the prediction classes were close enough to each other so as not to cause overlearning. The related study differs from our study in terms of including all Supreme Court of Appeals decisions and applying data augmentation process. For these reasons, the current study may have achieved a lower prediction success than this study.

Lage-Freitas et al. [65] achieved the highest accuracy of 81.35% in different data scenarios predicting Brazilian appellate decisions, including civil judgements. 81.35% prediction success was achieved via the XGBoost algorithm. The current study obtained 84.90% prediction success via the same algorithm, indicating that the two studies achieved close prediction success with the same algorithm. Although our study is similar in terms of the number of data, it has shown higher performance, 87.38%, in terms of prediction accuracy. Moreover, Zahir [4] predicted the Moroccan Supreme Court judgements with a success rate of 80.51% by using fewer judgment

texts and a data augmentation process. Our study differs from this study in terms of the amount of data and data augmentation and is ahead of this study in terms of prediction success.

The practical applications of our model can help improve efficiency and fairness within the Turkish legal system. Initially, it can help expedite legal proceedings by assisting judges and attorneys in the court process. For instance, it can help judges make decisions by predicting outcomes using similar case results. Additionally, by assisting parties to predict potential results early on, it could lead to the resolution of conflicts outside of the courtroom [96]. This would lessen the burden on the legal system and result in quicker delivery of justice [97]. Ultimately, the openness and responsibility of our model can help guarantee fairness in legal proceedings [98]. The incorporation of AI technologies can improve the consistency and fairness of decisions in the justice system.

Although using AI in legal decision-making has many benefits such as improving efficiency in the justice system and speeding up litigation processes, it is crucial to carefully address the ethical and bias concerns linked to this technology. The information used to train AI models can contain historical biases that could influence upcoming decisions [99, 100]. It is extremely important to be careful when choosing the datasets for model training in order to reduce the risk of bias. For instance, steps need to be implemented to guarantee the variety of data collections and to recognize and remove prejudices towards specific groups [101]. Furthermore, models should be held accountable and there should be transparency [102]. The use of AI in legal decision-making must follow the principles of human rights and justice [103]. In this study, the researcher endeavored to ensure that ethical considerations were adhered to, with particular attention paid to the careful selection of datasets, transparency and accountability of the model.

7. Conclusion

The aim of this study was to predict outcomes in construction dispute judicial decisions by text classification using ML methods, addressing a significant methodological gap in the literature by analyzing the linguistic interaction between claim and defense texts in Turkish, a morphologically rich and agglutinative language. Out of 15,667 judgments gathered from the Court of Cassation between 2011 and 2021, a representative sample of 2,563 judgments was selected with a 99% confidence level and a margin of error of 2.33%. The texts underwent comprehensive data preprocessing and feature extraction procedures to address the challenges of Turkish language structure. The decision texts were categorized into three distinct datasets: (i) decisions containing plaintiff claims and defendant defenses, (ii) decisions containing only plaintiff claims, and (iii) all decisions. Through the application of various ML algorithms, remarkable accuracy rates of 87.38%, 84.53%, and 85.98% were achieved respectively in these datasets, demonstrating that the combined analysis of plaintiff claims and defendant defenses provides significantly higher accuracy compared to predictions based solely on claim texts. These findings prove that modeling the mutual arguments of the parties enhances prediction performance, exceeding success rates reported in comparable literature.

Studies on construction disputes mostly focus on analysing the parties' obligations in contracts and the selection of standard forms used between the parties [104-107]. The decisions of the European Court of Human Rights, Courts of Appeal and Constitutional Courts have been analysed and decision predictions have often been made using ML and NLP techniques. Although some studies have used the decisions of the Court of Cassation, the decisions of the Court of Cassation have not been analysed specifically for construction cases. We believe that this study will pioneer future research in specialized legal domains with unique terminology and complex linguistic structures. The methodological innovations and significant contributions of this research can be summarized as follows;

- This study uniquely addresses the impact of defense texts on prediction performance by creating three distinct dataset configurations. While some of

the reasoned decision texts include the claims and defences of the parties together, some of them include only the statements of the claimant. Our approach systematically quantifies how the inclusion of defendant's arguments significantly improves prediction accuracy (87.38% versus 84.53%), demonstrating that the linguistic interaction between opposing parties' arguments contains valuable predictive patterns.

- Beyond merely applying ML methods to legal texts, this study offers a novel framework for analyzing linguistic patterns in construction disputes by examining how the mutual arguments of parties affect judicial outcomes. The use of three different datasets reflecting three different conditions related to construction issues provides a more nuanced understanding of how textual characteristics influence prediction performance, establishing a methodological template for future research in other specialized legal domains.

In the study, data augmentation was not performed by balancing the number of decisions in favour of the parties in the datasets according to the lower number. Our accuracy achievements indicate that high prediction success can be achieved in private law issues. In contrast to previous research on the Turkish legal system, the current study has successfully predicted outcomes using diverse and balanced datasets, despite limited data availability. Although it is compared with studies outside the Turkish legal system, studies conducted in different languages are not suitable for direct comparison due to the nature of the method. Although this aspect is in question, the unique terminology of our study will form the basis for future studies in this field.

In terms of the utilisation of Turkish legal texts in the study, given that Turkish is an agglutinative language, there are challenges in text classification using ML methods studies due to the increased variability of words. This complexity of Turkish legal texts may impact the performance of the ML algorithms employed. Gradient Boosting algorithms (HistGradientBoosting, GradientBoosting, XGBoost) demonstrate high accuracy and robust performance on complex datasets. It is possible that these algorithms can be effective in capturing detailed and variable language structures in Turkish legal texts. However, it should be recognized that they do have limitations. One limitation is that there is a risk of overfitting [108]. Ensemble methods, including Voting, AdaBoost, and Bagging, can provide higher accuracy and a greater generalisation capability by combining different models. These methodologies are more effective at capturing various language structures and contexts in Turkish legal texts; however, they are limited by the necessity for high computational resources [33]. Linear models (RidgeClassifierCV, RidgeClassifier, LogisticRegression, PassiveAggressive) are more computationally efficient and can process data in a relatively fast manner. These models can be employed in large datasets, offering a basic level of accuracy and speed in Turkish legal texts. However, they may not fully capture complex relationships due to linear assumptions [34]. Support Vector Machines (LinearSVC, NuSVC, SVC) are effective in high-dimensional data sets and can capture subtle linguistic differences in Turkish legal texts. The accuracy of the results can be significantly enhanced by selecting the appropriate kernel functions. However, this approach may also lead to a notable increase in the computational costs [34]. Consequently, it is possible that a similar study conducted in different languages and legal systems may yield different outcomes. It is thought that the fact that judges do not form decision texts within a certain mould in judicial processes will affect the success of the model. The effects of judges having different education, background and experience on the formation of decision texts and thus on the success of the model can be addressed in future studies. In addition to the plaintiff's claim, the fact that the defence of the defendant is higher in the success of the model is seen as an important output. With the studies to be developed on this subject, systems that will make decision prediction without going to judgement can be developed based on the demands of the parties.

The incorporation of AI-based models in the Turkish legal system offers several benefits, such as

improved effectiveness and equity. The model helps speed up legal proceedings by assisting judges and attorneys in the court process. Moreover, by allowing the involved parties in the legal case to foresee possible results ahead of time, it can help in settling disputes before they reach the court, ultimately lessening the burden on the judicial system. Ultimately, the model's transparency and accountability can help enhance the delivery of justice, leading to more equitable and consistent outcomes in legal processes. These methods have the potential to improve both the overall effectiveness and equity of the Turkish judicial system.

There are significant benefits to be gained from using AI for predicting legal decisions, however, it is crucial to address potential biases and ethical concerns. To address these issues, it is crucial to meticulously choose the datasets utilized in training models and to guarantee transparency and accountability throughout the process. It is crucial to consider that AI should be employed for legal decision forecasting in alignment with human rights and justice principles in this situation.

The research has identified several future work directions within the scope of this study. It would be beneficial to explore and compare the performance of different ML models, with a particular focus on DA and transformer-based techniques. Furthermore, expanding the dataset and incorporating a wider range of legal decisions will enhance the model's generalisation capability. The results obtained through the use of data augmentation techniques can be evaluated by comparing them with the data without data augmentation. Finally, studies should be conducted on the practical ways of integrating the model into the justice system and real-world applications should be tested. These recommendations will increase the effectiveness and applicability of AI in legal decision prediction.

Declaration

Funding

This research received no external funding.

Author Contributions

M. Sarı: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Resources, Software, Validation, Visualization, Writing – Original Draft. S. Bayram: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing – Review & Editing. E. Aydemir: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Software, Validation, Supervision, Writing – Review & Editing.

Acknowledgments

This work was supported by the Scientific and Technological Research Council of Turkiye (TUBITAK) through the Innovative Solutions Research Projects Support Program in Social Sciences and Humanities (3005) under grant 122G126.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Ethics Committee Permission

Not applicable.

Conflict of Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Bhat PI (2020) Quantitative legal research. In: Bhat PI (ed) Idea and Methods of Legal Research. Oxford University Press. https://doi.org/10.1093/oso/9780199493098.003.0013.

[2] Camkerten AS (2023) Legal analytics. J Necmettin Erbakan Univ Fac Law 6(1): 234-254.

[3] Aydemir E (2023) Estimation of Turkish constitutional court decisions in terms of admissibility with NLP. In: 2023 IV Int Conf

Neural Networks and Neurotechnologies (NeuroNT).

[4] Zahir J (2023) Prediction of court decision from Arabic documents using deep learning. Expert Syst 40(6): e13236. https://doi.org/10.1111/exsy.13236.

[5] Cassation Co (2024) Court of Cassation Statistics. https://sgb.adalet.gov.tr/Resimler/Dergi/13052022095441T%C3%BCrk%20Adalet%20Sistemi.pdf.

[6] Cui J, Shen X, Wen S (2023) A survey on legal judgment prediction: Datasets, metrics, models and challenges. IEEE Access 11: 102050-102071. https://doi.org/10.1109/ACCESS.2023.3317083.

[7] Esen E (2022) Violation of the right to a fair trial in arbitration: Analysing the Turkish Court of cassation's decision of 10 february 2021. Ann Fac Droit Istanbul 71: 99-177. https://doi.org/10.26650/annales.2022.71.0003.

[8] Sarı M, Bayram S, Aydemir E (2024) Construction-related disputes: A comprehensive bibliometric investigation. In: Proc 8th Int Project Constr Manag Conf (IPCMC 2024). Istanbul, Türkiye.

[9] Sarı M, Sayın B, Akçay C (2021) Classification and resolution procedure for disputes in public construction projects. Rev Constr 20(2): 259-276. https://doi.org/10.7764/RDLC.20.2.259.

[10] Akca O, Bayrak G, Issifu AM, Ganiz MC (2022) Traditional machine learning and deep learning-based text classification for Turkish law documents using transformers and domain adaptation. In: 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA).

[11] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. https://doi.org/10.48550/arXiv.1301.3781.

[12] Oflazer K (2018) Morphological processing for Turkish. In: Oflazer K, Saraçlar M (eds) Turkish Natural Language Processing. Springer, pp. 21-52. https://doi.org/10.1007/978-3-319-90165-7_2.

[13] Minaee S, Cambria E, Gao J (2021) Deep learning based text classification: A comprehensive review. ACM Computing Surveys (CSUR) 54(3): 1-40. https://doi.org/10.1145/3439726.

[14] Sulea OM, Zampieri M, Vela M, Van Genabith J (2017) Predicting the law area and decisions of french supreme court cases. https://doi.org/10.48550/arXiv.1708.01681.

[15] Al-Alawi AI, Lmansouri AMA (2023) Artificial intelligence in the judiciary system of Saudi Arabia: A literature review. In: 2023 Int Conf Cyber Manag Eng (CyMaEn).

[16] Lawlor RC (1963) What computers can do: Analysis and prediction of judicial decisions. Am Bar Assoc J 49(4): 337-344. http://www.jstor.org/stable/25722338.

[17] Zhong H, Xiao C, Guo Z, Tu C, Liu Z, Sun M, Feng Y, Han X, Hu Z, Wang H (2018) Overview of CAIL2018: Legal judgment prediction competition. https://doi.org/10.48550/arXiv.1810.05851.

[18] Long S, Tu C, Liu Z, Sun M (2018) Automatic judgment prediction via legal reading comprehension. In: Chinese Computational Linguistics: 18th China National Conference.

[19] Katz DM, Bommarito MJ, Blackman J (2017) A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE 12(4): e0174698.

[20] Liu Z, Chen H (2017) A predictive performance comparison of machine learning models for judicial cases. In: 2017 IEEE Symp Series Comput Intell (SSCI).

[21] 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 Conf Defense Technol (ACDT).

[22] Koc K (2023) Determining the short-term susceptibility of construction workers to occupational accidents using stochastic gradient boosting. J Constr Eng Manag Innov (Online) 6(1).

[23] Mostofi F, Toğan V, Başağa HB (2021) House price prediction: A data-centric aspect approach on performance of combined principal component analysis with deep neural network model. J Constr Eng Manag Innov 4: 106-116.

[24] Virtucio MBL, Aborot JA, Abonita JKC, et al. (2018) Predicting decisions of the philippine supreme court using natural language processing and machine learning. In: 2018 IEEE 42nd Annu Comput Softw Appl Conf (COMPSAC).

[25] Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in English. https://doi.org/10.48550/arXiv.1906.02059.

[26] Kaufman AR, Kraft P, Sen M (2019) Improving supreme court forecasting using boosted decision trees. Political Anal 27(3): 381-387. https://doi.org/10.1017/pan.2018.59.

[27] Shaikh RA, Sahu TP, Anand V (2020) Predicting outcomes of legal cases based on legal factors using classifiers. Procedia Comput Sci 167: 2393-2402.

[28] Medvedeva M, Xiao X, Wieling M, Vols M (2020) JURI SAYS: An automatic judgement prediction system for the european court of human rights. In: Int Conf Legal Knowl Inf Syst.

[29] Alrasheed K, Soliman E, Albader H (2023) Delay dispute cases: Comparative analysis and claimed value prediction model. J Legal Aff Dispute Resolut Eng Constr 16. https://doi.org/10.1061/JLADAH.LADR-1038.

[30] Seo W, Kang Y (2024) Auto-summarization for the texts of construction dispute precedents. In: Construction Research Congress. https://doi.org/10.1061/9780784485286.018.

[31] Kalogeraki M, Antoniou F (2024) Claim management and dispute resolution in the construction industry: Current research trends using novel technologies. Buildings 14(4): 967. https://doi.org/10.20944/preprints202401.1195.v1.

[32] Mumcuoglu E, Öztürk CE, Ozaktas HM, Koc A (2021) Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. Inf Process Manag 58(5): 102684.

[33] Aras AC, Öztürk CE, Koç A (2022) Feedforward neural network based case prediction in Turkish higher courts. In: 2022 30th Signal Process Commun Appl Conf (SIU).

[34] Ozturk CE, Ozcelik SB, Koc A (2022) Predicting outcomes of the Court of Cassation of Turkey with Recurrent Neural Networks. In: 2022 30th Signal Process Commun Appl Conf (SIU).

[35] Sert MF, Yıldırım E, Haşlak İ (2022) Using artificial intelligence to predict decisions of the Turkish Constitutional Court. Soc Sci Comput Rev 40(6): 1416-1435. https://doi.org/10.1177/08944393211010398.

[36] Shang X (2022) A computational intelligence model for legal prediction and decision support. Comput Intell Neurosci 2022(1): 5795189.

[37] Tazelaar F, Snijders C (2010) Dispute resolution and litigation in the construction industry. J Purch Supply Manag 16(4): 221-229. https://doi.org/10.1016/j.pursup.2010.08.003.

[38] Rebuplic of Türkiye Court of Cassation: Reports, Statistics. https://www.yargitay.gov.tr/icerik/51/istatistikler. Accessed 14 Nov 2022.

[39] Hukukturk.com: Yargıtay Kararları. https://www.hukukturk.com/yargitay-kararlari. Accessed 22 March 2024.

[40] Krejcie RV, Morgan DW (1970) Determining sample size for research activities. Educ Psychol Meas 30(3): 607-610. https://doi.org/10.1177/001316447003000308.

[41] Cochran WG (1977) Sampling Techniques. John Wiley & Sons.

[42] Malik V, Sanjay R, Nigam SK, et al. (2021) ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. https://doi.org/10.48550/arXiv.2105.13562.

[43] Sarı M (2019) Investigation In The Framework Of Jurisdictions Of Conflicts In Public Construction. Master Dissertation, İstanbul University Cerrahpaşa.

[44] Sak H, Gungor T, Saraclar M (2011) Resources for Turkish morphological processing. Lang Resour Eval 45: 249-261. https://doi.org/10.1007/s10579-010-9128-6.

[45] Bird S, Klein E, Loper E (2009) Natural Language Processing With Python: Analyzing Text With The Natural Language Toolkit. O'Reilly Media, Inc.

[46] Robertson S (2004) Understanding inverse document frequency: On theoretical arguments for IDF. J Doc 60(5): 503-520.

[47] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proc Workshop at ICLR 2013.

[48] Goldberg Y, Levy O (2014) word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. https://doi.org/10.48550/arXiv.1402.3722.

[49] Enríquez F, Troyano JA, López-Solaz T (2016) An approach to the use of word embeddings in an opinion classification task. Expert Syst Appl 66: 1-6. https://doi.org/10.1016/j.eswa.2016.09.005.

[50] Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. https://doi.org/10.48550/arXiv.1607.01759.

[51] Prastyo PH, Ardiyanto I, Hidayat R (2020) A review of feature selection techniques in sentiment analysis using filter, wrapper, or hybrid methods. In: 2020 6th Int Conf Sci Technol (ICST).

[52] Berrouachedi A, Jaziri R, Bernard G (2022) Convolutional, extra-trees and multi-layer perceptron. In: 2022 IEEE/ACS 19th Int Conf Comput Syst Appl (AICCSA).

[53] Song SJ, Heo GE, Kim HJ, et al. (2014) Grounded feature selection for biomedical relation extraction by the combinative approach. In: Proc ACM 8th Int Workshop Data Text Min Bioinformatics. Shanghai, China. https://doi.org/10.1145/2665970.2665975.

[54] Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min.

[55] Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5): 1189-1232.

[56] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30.

[57] Breiman L (1996) Bagging predictors. Mach Learn 24: 123-140. https://doi.org/10.1007/BF00058655.

[58] Dietterich TG (2000) Ensemble methods in machine learning. In: Proccedings of First International Workshop, Multiple Classifier Systems. Cagliari, Italy.

[59] Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1): 119-139.

[60] Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1): 55-67.

[61] Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression. John Wiley & Sons.

[62] Zhang T (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proc 21st Int Conf Mach Learn.

[63] Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273-297.

[64] Scholkopf B (2000) New support vector algorithms. Neural Comput 12: 1083-1121.

[65] Lage-Freitas A, Allende-Cid H, Santana O, Oliveira-Lage L (2022) Predicting Brazilian court decisions. PeerJ Comput Sci 8: e904.

[66] Hosmer DW, Lemeshow S (2000) Assessing the fit of the model. In: Hosmer DW, Lemeshow S (eds) Applied Logistic Regression. John Wiley & Sons, Inc, pp. 143-202. https://doi.org/10.1002/0471722146.ch5.

[67] Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y (2006) Online passive-aggressive algorithms. J Mach Learn Res 7(Mar): 551-585.

[68] Tharwat A (2021) Classification assessment methods. Appl Comput Inform 17(1): 168-192. https://doi.org/10.1016/j.aci.2018.08.003.

[69] Naderalvojoud B, Akcapinar Sezer E (2020) Term evaluation metrics in imbalanced text categorization. Nat Lang Eng 26(1): 31-47. https://doi.org/10.1017/S1351324919000317.

[70] Imran AS, Hodnefjeld H, Kastrati Z, Fatima N, Daudpota SM, Wani MA (2023) Classifying european court of human rights cases using transformer-based techniques. IEEE Access 11: 55664-55676. https://doi.org/10.1109/ACCESS.2023.3279034.

[71] Keeling R, Chhatwal R, Huber-Fliflet N, Zhang J, Wei F, Zhao H, Shi Y, Qin H (2019) Empirical comparisons of CNN with other learning algorithms for text classification in legal document review. In: 2019 IEEE Int Conf Big Data.

[72] Howe JST, Khang LH, Chai I (2019) Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. https://doi.org/10.48550/arXiv.1904.06470.

[73] Riyanto S, Imas SS, Djatna T, Atikah TD (2023) Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. Int J Adv Comput Sci Appl 14(6).

[74] Wang D, Su J, Yu H (2020) Feature extraction and analysis of natural language processing for deep learning english language. IEEE Access 8: 46335-46345. https://doi.org/10.1109/ACCESS.2020.2974101.

[75] Deshwal D, Sangwan P, Kumar D (2019) Feature extraction methods in language identification: A survey. Wirel Pers Commun 107(4): 2071-2103. https://doi.org/10.1007/s11277-019-06373-3.

[76] Phan TTT, Ohkawa T, Yamamoto A (2017) Protein-protein interaction extraction from text by selecting linguistic features. In: 2017 IEEE 17th Int Conf Bioinf Bioeng (BIBE).

[77] He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9): 1263-1284. https://doi.org/10.1109/TKDE.2008.239.

[78] Buda M, Maki A, Mazurowski MA (2018) A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 106: 249-259. https://doi.org/10.1016/j.neunet.2018.07.011.

[79] Freitas LJG, Rodrigues T, Rodrigues G, Edokawa P, Farias A (2024) Text clustering applied to data augmentation in legal contexts. https://doi.org/10.48550/arXiv.2404.08683.

[80] Manjula B, Layaq S (2022) A new method for imbalanced data reduction using data based under sampling. In: Int Conf Soft Comput Signal Process.

[81] Chen J, Tam D, Raffel C, Bansal M, Yang D (2023) An empirical survey of data augmentation for

limited data learning in NLP. Trans Assoc Comput Linguist 11: 191-211. https://doi.org/10.1162/tacl_a_00542.

[82] Assunção GO, Izbicki R, Prates MO (2023) Is augmentation effective to improve prediction in imbalanced text datasets? https://doi.org/10.48550/arXiv.2304.10283.

[83] Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy E (2021) A survey of data augmentation approaches for NLP. https://doi.org/10.48550/arXiv.2105.03075.

[84] Ishikawa T, Yakoh T, Urushihara H (2022) An NLP-inspired data augmentation method for adverse event prediction using an imbalanced healthcare dataset. IEEE Access 10: 81166-81176. https://doi.org/10.1109/ACCESS.2022.3195212.

[85] Shorten C, Khoshgoftaar TM, Furht B (2021) Text data augmentation for deep learning. J Big Data 8(1): 101. https://doi.org/10.1186/s40537-021-00492-0.

[86] Zhou W, Liu C, Yuan P, Jiang L (2024) An undersampling method approaching the ideal classification boundary for imbalance problems. Appl Sci 14(13): 5421. https://www.mdpi.com/2076-3417/14/13/5421.

[87] Zhou W, Wang H, Sun H, Sun T (2019) A method of short text representation based on the feature probability embedded vector. Sensors 19(17): 3728. https://www.mdpi.com/1424-8220/19/17/3728.

[88] Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57: 232-247. https://doi.org/10.1016/j.eswa.2016.03.045.

[89] Bellman RE, Dreyfus SE (1962) Applied Dynamic Programming. Princeton Univ Press. https://doi.org/10.1515/9781400874651.

[90] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar): 1157-1182.

[91] Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507-2517.

[92] Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1): 3-42. https://doi.org/10.1007/s10994-006-6226-1.

[93] Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: Proc 14th Int Conf Mach Learn.

[94] Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3(Mar): 1289-1305.

[95] Erdoganyilmaz C, Mengunogul B (2022) An original natural language processing approach to language modeling in Turkish Legal Corpus. In: 2022 Innovations Intell Syst Appl Conf (ASYU).

[96] Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. Q J Econ 133(1): 237-293. https://doi.org/10.1093/qje/qjx032.

[97] Glaeser EL, Kominers SD, Luca M, Naik N (2015) Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. Harvard Business School NOM Unit Working Paper No. 16-065. http://dx.doi.org/10.2139/ssrn.2694723.

[98] Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/arXiv.1702.08608.

[99] Angwin J, Larson J, Mattu S, Kirchner L (2022) Machine bias. In: Martin K (ed) Ethics of Data and Analytics. Auerbach Publications, pp. 254-264.

[100] Chouldechova A (2016) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 5(2): 153-163. https://doi.org/10.1089/big.2016.0047.

[101] Binns R (2017) Fairness in machine learning: Lessons from political philosophy. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81: 149-159.

[102] Weller A (2019) Transparency: motivations and challenges. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (eds) Explainable AI: Interpreting, Explaining And Visualizing Deep Learning. Springer International Publishing, pp. 23-40. https://doi.org/10.1007/978-3-030-28954-6_2.

[103] Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104: 671.

[104] Candaş AB, Tokdemir OB (2022) Automated identification of vagueness in the fidic silver book conditions of contract. J Constr Eng Manag 148(4): 04022007. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002254.

[105] Hassan FU, Le T, Le C (2023) Automated approach for digitalizing scope of work requirements to support contract management. J Constr Eng Manag 149(4): 04023005. https://doi.org/10.1061/JCEMD4.COENG-12528.

[106] Ko T, Jeong HD, Lee G (2021) NLP-driven model to extract contract change reasons and altered work items for advanced retrieval of change orders. J Constr Eng Manag 147(11): 04021147. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002172.

[107] Moon S, Lee G, Chi S (2021) Semantic text-pairing for relevant provision identification in construction specification reviews. Autom Constr 128: 103780. https://doi.org/10.1016/j.autcon.2021.103780.

[108] Çetindağ C, Yazıcıoğlu B, Koç A (2023) Named-entity recognition in Turkish legal texts. Nat Lang Eng 29(3): 615-642. https://doi.org/10.1017/S1351324922000304.