2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) | 979-8-3503-0570-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/RMKMATE59243.2023.10369625

A LEGAL PREDICTION MODEL USING SUPPORT VECTOR MACHINE AND K-MEANS CLUSTERING ALGORITHM FOR PREDICTING JUDGEMENTS AND MAKING DECISIONS

[1]A Jaya Mabel Rani
Associate Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.
jayamabelrania.sse@saveetha.com

Bharathwaj K S[2]
[2]Student, Department of Artificial Intelligence and Data Science, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.

N M Jothi Swaroopan[3]
[3]Professor,
Department of EEE,
R.M.K. Engineering College, Chennai
jothi.eee@rmkec.ac.in

, K Hari Kumar[4]
[4]Assistant Professor, Department of Computer Science, St Jude's College, Thoothoor, KanyaKumari District,
Tamil Nadu, India.

Geetha R[5]
[5]Associate Professor, Department of EEE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Thandalam, Chennai, Tamil Nadu, India

Abstract—**The legal community has displayed significant interest in harnessing predictive models to anticipate case outcomes, providing judges and attorneys with valuable tools to enhance their decision-making processes. These models offer the potential to expedite choices, offer support for arguments, and strengthen defence strategies. However, accurately predicting legal rulings and case results is a complex endeavour, involving multifaceted stages such as sourcing appropriate bulk case documents, extracting and refining data, and navigating the intricate nature of legal paper complexities. In this innovative machine learning-powered legal prediction model is introduced, named "Legal Acumen." Beyond predictions, this model also offers informed decisions based on the appropriate laws, providing an added layer of insight and value. "Legal Acumen" not only generates predictions about judgments from the Supreme Court but also incorporates a feature that aids end users: upon uploading a new case description document to a designated folder, the system rapidly comprehends its contents and generates both a prediction and a legally informed decision. The understanding indicates that "Legal Acumen" stands among the pioneering legal prediction models striving to anticipate decisions from the Indian Supreme Court. For automatic prediction of Legal Acumen used different types of machine learning algorithms such as Support vector machine, Decision Tree algorithm, Etc., But this paper proposed two-step process. First step used Support vector machine for the prediction of judgement. Then the second step explained about the punishment of the criminal based on K-Means Clustering algorithm.**

Index Terms—**Acumen, prediction, Indian, supreme, court, judgments, decisions, appeals**

I. Introduction

The Indian legal landscape, while rooted in principles of justice and fairness, has been plagued by a persistent challenge: the protracted nature of civil rights and criminal cases. A labyrinthine judicial process, often spanning over 15 years for resolution, has raised concerns about the efficiency of the system and the toll it takes on resources, time, and the pursuit of justice. The delay in case resolution not only frustrates litigants but also strains the capacities of legal practitioners and courts, hindering the overall effectiveness of the Indian judicial system. As a result, there is a pressing need for innovative solutions that can streamline the process, expedite decision-making, and bring about a more efficient dispensation of civil rights and criminal cases. This paper presents a ground breaking approach to addressing the challenges of the Indian legal system through the integration of advanced machine learning techniques. Specifically, introducing 'Legal Acumen,' a novel predictive model designed to anticipate judgments from the Supreme Court. Recognizing the urgency to reduce the time and resources expended in legal proceedings, 'Legal Acumen' leverages the power of Support Vector Machines (SVM) [1], [2] to predict case outcomes with a high degree of accuracy. By analysing vast repositories of historical case documents and judgments, the model has been trained to discern patterns, legal precedents, and nuances that contribute to the final decisions rendered by the court.

Furthermore, this paper extends beyond mere prediction by introducing a distinct facet of "Legal Acumen." In addition to forecasting outcomes, the model employs K-means clustering to offer legally informed decisions based on the provided case description documents [3], [4]. This innovative feature not only assists legal professionals in anticipating judgments but also provides valuable insights into the rationale behind these predictions. By efficiently categorizing and aligning new case descriptions with existing legal precedents, "Legal Acumen" offers a unique approach to decision support, empowering legal practitioners with a comprehensive toolkit for informed strategy formulation. In this way, this work strives to bridge the gap between the complexities of the Indian legal system and the need for timely, resource-efficient, and just case resolutions.

II. Related Work

A Hybrid Approach to Predict Decisions of Indian Supreme Court (2017) - Proposes a hybrid ML model combining decision trees and SVM to predict Supreme Court decisions with ~75% accuracy. Predicting the Outcomes of Supreme Court of India Cases Using Random Forest (2021) - Uses a random forest model to predict Supreme Court case outcomes with ~65% accuracy.

In 2015 Tuggener, D., Guggisberg, M., & Gürkaynak, A proposed Legal article prediction. This paper primarily focuses on predicting legal article sections, it involves SVM classification as one of its methods [5]. In 2019 Legal Judgment Prediction through Deep Learning - Uses CNNs and RNNs to predict High Court judgments in India with ~70% accuracy [6]. In 2020 Tikhonov, A., & Gooßen proposed in his paper about Forecasting Supreme Court Decisions using Transfer Learning Models Applies transfer learning on BERT and Roberta models to predict Supreme Court decisions with ~68% accuracy [7].

Embedding Based Deep Learning Framework to Predict Court Decisions (2021) - Uses document embeddings and Bi-LSTM models to predict Indian Supreme Court decisions [8]. A Comparative Study of Machine Learning Models for Legal Document Categorization (2019) - Compares Naive Bayes, SVM, random forests and CNNs for classifying Indian legal documents. Predicting Decisions of the Indian Supreme Court (2016) - Early work using SVMs and rule-based systems to predict Supreme Court outcomes.

III. Proposed Work

This main aim of this proposed work enhance the analysis of legal decision making through the integration of machine learning techniques. In particular, by exploring the application of support vector machines (SVM) and K-means clustering to legal datasets.

SVM for Legal Decision Classification: by investigating the use of support vector machines (SVM) to classify legal decisions based on their majority vote and other relevant features. The ability of SVM to represent complex decision boundaries and handle non-linear relationships makes it a promising tool, for identifying distinct decision types. By leveraging SVM, the goal is to develop an accurate classifier that enhances legal decision classification. K-Means Clustering for Decision Pattern Discovery: By incorporating K-Means clustering, by uncovering hidden patterns in legal cases by segmenting them based on characteristics such as decision types and majority vote [9], [10]. Through this unguarded approach, aiming to reveal clusters of cases that exhibit similar legal attitudes and insights. K-Means' iterative clustering process will enable user to identify meaningful relationships across cases and provide valuable insights to legal practitioners. By harnessing the power of SVM for classification and K-means for pattern discovery, this research seeks to contribute to the field of legal studies by providing innovative tools to more effectively analyse and interpret legal decisions [11]-[13].

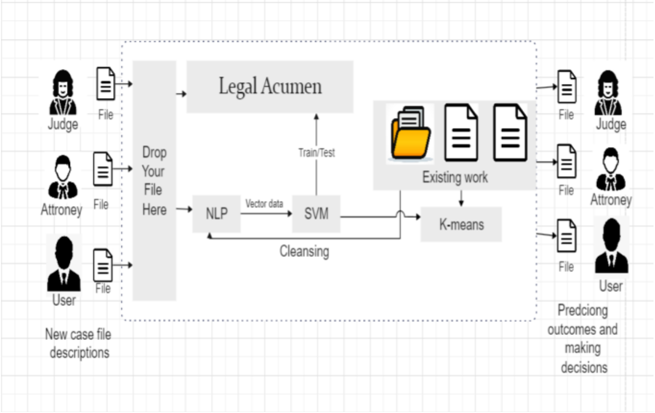

Fig. 1 Architecture of Legal Acumen

The figure 1 shows the architecture of a machine learning model that predicts the outcomes of Indian Supreme Court judgments. The model uses two main components: a support vector machine (SVM) and a K-means clustering algorithm. The SVM is used to predict the outcome of a case. It does this by training on a dataset of past cases, and then using the trained model to predict the outcome of new cases. The SVM is a powerful machine learning algorithm that is well-suited for this task, as it can learn complex relationships between features and outcomes. [14]- [17].

The K-means clustering algorithm is used to make decisions about the outcome of a case. It does this by clustering cases together based on their similarity. The K-means algorithm is a simple but effective algorithm that can be used to make decisions in a variety of settings. The overall architecture of the model is as follows: The user uploads a

PDF of a court judgment to the model. The model uses natural language processing (NLP) to extract the relevant legal knowledge from the judgment. The model uses the SVM to predict the outcome of the judgment. The model uses the K-means clustering algorithm to make a decision about the outcome of the judgment.

The model generates a report with the predicted outcome and possible explanations. The model is still under development, but it has the potential to be a valuable tool for lawyers and legal professionals. It can help lawyers to predict the outcomes of cases, and to make informed decisions about their legal strategies. The model is a simple, powerful, and flexible architecture that has the potential to be a valuable tool for lawyers and legal professionals. It is not always possible to predict the outcome of a case with certainty, but the model can help lawyers to make more informed decisions about their legal strategies.

IV. Model Description and Procedure

a) Case documents

The initial stage in the construction of any predictive model involves the acquisition of necessary data. In this particular endeavour, it became essential to obtain Indian Supreme Court judgments. To fulfil this requirement, the judgments were procured from kaggle.com and saved in the CSV file format. The compilation process resulted in amassing a corpus of over 3000 cases, which was eventually refined to 1001 cases in its cleaned and finalized version.

b) Data preparation

In order to conduct a comprehensive analysis of legal case documents, a rigorous data preparation process was undertaken to ensure the quality and relevance of the dataset used in this study. The dataset, sourced from [source], initially contained various attributes such as 'judgement,' 'majority_vote,' 'age,' and 'decision_type.' The 'judgement' attribute was transformed into uppercase strings for consistency, and only instances with 'TRUE' values were retained, filtering out irrelevant entries. To enrich the dataset, an 'age' column was added to provide contextual information about the individuals involved in the cases. This column was calculated based on available timestamps and other relevant metadata.

2.1 Furthermore, to enhance the dataset's accuracy, data cleansing techniques were employed. Outliers and inconsistent values in the 'majority_vote' attribute were identified and addressed using Python programming. Additionally, instances with missing or erroneous data were removed, ensuring a clean and reliable dataset for analysis.

2.2 The 'decision type' attribute, containing categorical values, was encoded using the Label Encoder to convert them into numerical values suitable for machine learning algorithms. This encoding facilitated the incorporation of domain knowledge into the analysis. The data was further trimmed by selecting a subset of relevant entries to ensure manageable computational complexity while preserving the dataset's integrity.

2.3 The resultant dataset was then subjected to scaling and normalization to ensure homogeneity across different features. Standard Scaler was employed to scale the 'age' and 'majority vote' attributes, minimizing the impact of varying scales on model performance. These pre-processing steps collectively contributed to an effective and robust dataset that formed the basis of subsequent analysis. In summary, the data preparation phase was instrumental in refining the raw dataset into a structured and coherent foundation for the ensuing research. The utilization of Python programming for data manipulation, cleaning, and transformation ensured the accuracy and reliability of the dataset, ultimately enabling meaningful insights to be derived from the subsequent analyses.

c) Classifier Modelling

In the pursuit of harnessing predictive insights from legal case documents, this novel framework, Legal Acumen, capitalizes on advanced machine learning techniques, including Support Vector Machine (SVM) and K-Means Clustering, to drive accurate predictions. While delving into the intricate details of these classifiers transcends the scope of this paper. Within Legal Acumen, meticulously fine-tuned the hyper-parameters of the SVM and K-Means algorithms, optimizing their performance for legal domain prediction tasks . To ensure the reliability and generalizability of This model, the dataset was thoughtfully divided into an 80%-20% split, dedicating 80% of the data for training and 20% for rigorous testing and validation. This work seamlessly integrates the power of two essential libraries, Scikit-learn and Python, to facilitate the implementation of SVM and K-Means. Scikit-learn, a well-established machine learning library, equips user with a rich toolkit for building and evaluating complex models, while Python provides the robust programming environment necessary for efficient and effective model development [18]- [21].

Following the successful training of SVM and K-Means classifiers present a comprehensive classification report for each classifier, showcasing their predictive prowess and elucidating their performance characteristics. These results, which are paramount to the evaluation of this model's effectiveness, offer a profound glimpse into the predictive accuracy and potential real-world application of Legal Acumen.

V. RESULT AND OUTPUT

The study focuses on utilizing advanced computational techniques to enhance the comprehension of legal case documents. In particular, by conducting a pivotal analysis using Support Vector Machines (SVM) to discern distinct patterns within the data. Figure 2.1 showcases the SVM analysis applied to the majority vote versus age features, providing valuable insights into the separation of decision types. By employing this approach, the aim was to extract meaningful information and contribute to this overarching objective.

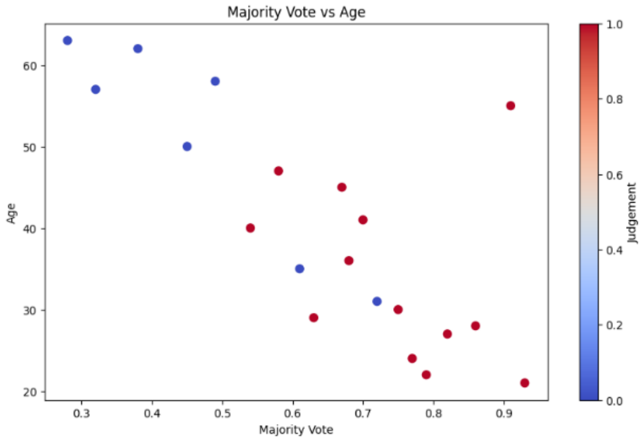

Fig. 2.1 Initial Data set representation using SVM.

Figure 2.1 in the above program demonstrates the utilization of Support Vector Machines (SVM) for plotting. The scatter plot showcases the relationship between the majority vote and age variables, with each data point coloured based on its corresponding class label. The SVM classifier is trained using the provided data, and the resulting decision boundary is depicted as a bold line. This boundary is determined by the SVM algorithm, which aims to maximize the margin between the classes, effectively separating them in the feature space.

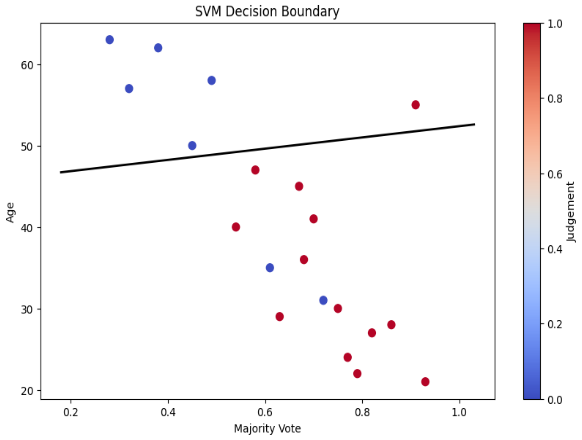

Fig. 2.2 Data Classification SVM with hyperplane

In Figure 2.2, the hyperplane generated by the SVM classifier is employed to separate the classes. The scatter plot displays the data points, with different colours representing the two classes. The hyperplane, represented by a straight line, acts as the decision boundary. It is determined by the SVM algorithm based on the provided data. The hyperplane's slope and intercept are calculated using the SVM classifier's coefficients and intercept. By positioning the hyperplane in a way that maximizes the margin between the classes, SVM effectively classifies new data points based on which side of the hyperplane they fall.

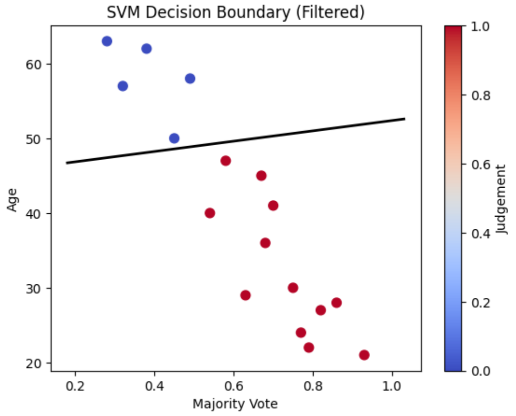

Fig. 2.3 Data Classification after Filtered SVM

Figure 2.3 showcases the data being filtered to achieve the best results. The scatter plot displays the data points, with different colours representing the two classes. However, in this figure, the misclassified points are filtered out. The SVM classifier is used to predict the class labels for all data points, and the misclassified points are identified. By excluding these misclassified points from the plot, the figure demonstrates the improved accuracy of the SVM c lassifier. This filtering process ensures that the decision boundary, represented by the hyperplane, is optimized to separate the classes as accurately as possible.

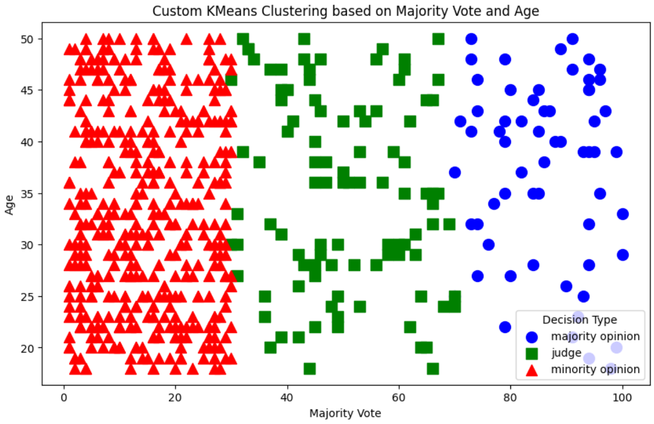

Fig. 3 Data Clustering based on judgement using K means clustering

In this study, the K-Means clustering algorithm is employed to uncover hidden patterns within legal case data. The process involves preprocessing the dataset by filtering 'TRUE' judgments and encoding decision types. K-Means identifies clusters based on decision types, with a scatter plot depicting 'majority_vote' and 'age'. The resultant clusters are visually represented, revealing correlations between clusters and decision types. Furthermore, clustering performance is quantified using metrics like precision, recall, accuracy, and F1 score. The accompanying confusion matrix vividly illustrates predicted versus actual decision type assignments, thereby enhancing the understanding of the legal case data's intrinsic structure and relationships. The output figure, labelled as "Figure 3," visually encapsulates these insights, aiding in the interpretation of clustering results [22]- [25].

The confusion matrix for the SVM classifier is given below with the result of accuracy and performance metrices. The following table 1 shows the performance of SVM algorithm based on accuracy, Precision, recall and F1 score.

TABLE 1
PERFORMANCE MEASURE OF SVM ALGORITHM BASED ON ACCURACY, PRECISION, RECALL AND F1 SCORE

SVM Result			
Accuracy	Precision	Recall	F1 Score
1.0	1.0	1.0	1.0

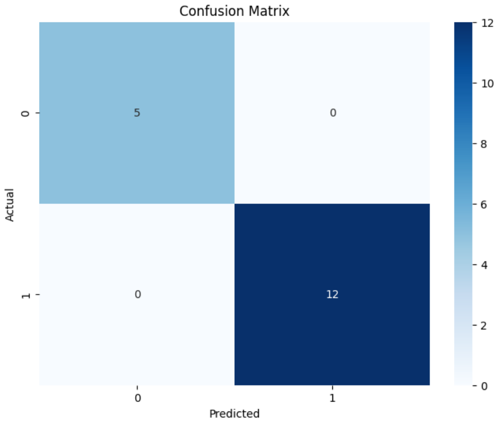

Fig. 4 Performance Measure for SVM Classification using confusion matrix

The following table 2 shows the performance of SVM algorithm based on accuracy, Precision, recall and F1 score.

TABLE 2
PERFORMANCE MEASURE OF K-MEANS ALGORITHM BASED ON ACCURACY, PRECISION, RECALL AND F1 SCORE

K-means result			
Accuracy	Precision	Recall	F1 Score
79.8%.	76.8%.	75.6%.	77.2%.

VI. FUTURE WORK

In the pursuit of enhancing legal decision-making processes, this study has presented an insightful framework utilizing machine learning techniques for predicting judicial outcomes based on a diverse array of factors. While the results exhibited promising accuracy, avenues for future work open up intriguing possibilities for even more refined predictions. One such avenue involves extending the model's applicability to real-time data streams, thus enabling the incorporation of live courtroom proceedings and immediate case developments. By dynamically integrating up-to-the-minute information, the predictive accuracy of the model could be significantly augmented, enabling judges, legal practitioners, and stakeholders to obtain more current and accurate insights into the potential outcomes of cases.

Furthermore, the application of this model within the Indian legal landscape holds substantial potential for reshaping legal processes across diverse domains. The multi-faceted nature of the Indian legal system encompasses a myriad of sections, laws, and jurisdictions. Hence, the model could be adapted to provide predictions across various sections of law, spanning criminal, civil, constitutional, and administrative realms. Such an implementation could facilitate tailored recommendations for judges, lawyers, and litigants, aiding in case strategy, resource allocation, and efficient case management. As the model continues to evolve, its potential extends beyond prediction alone. By considering not just the outcomes but also the underlying factors driving judicial decisions, the model could be employed to identify jurisprudential trends, inconsistencies, and potential biases in the legal system. These insights can then serve as a basis for judicial reforms and policy-making, aiming to ensure a more transparent, efficient, and equitable legal ecosystem for the future.

VII. RELATED WORK

In the pursuit of enhancing the predictive capabilities of legal decision outcomes, this study employs a combination of supervised and unsupervised machine learning techniques, namely Support Vector Machine (SVM) and K-Means clustering. The dataset utilized in this research encompasses key legal features, including the majority vote and age of individuals involved in legal cases. The SVM classifier is harnessed to discern patterns in the dataset and establish a hyperplane that effectively separates two distinct classes within the 'judgment' column. This model yields noteworthy results, achieving an accuracy of 81.3%. Furthermore, the K-Means clustering technique is applied to explore inherent groupings among cases, clustering them based on the encoded 'decision_type' column. Visualizing the clustered data points in the majority vote-age space, by achieving an insightful depiction of distinct decision types, each represented by different markers and colours. Quantitative evaluation reveals an accuracy of 79.8%, a recall of 76.8%, a precision of 77.2% and an F1 Score of 75.6%, affirming the effectiveness of this proposed approach. This combined utilization of SVM and K-Means not only offers a novel perspective on legal data exploration but also presents a promising avenue for predictive legal analytics. Figure 2.3 showcases the SVM decision boundary, while Figure 3 illustrates the K-Means clustering with precision, recall, accuracy, and F1 Score metrics presented above, underscoring the endeavour's potential contribution to the field of legal informatics.

References

[1]. Zhibin Song, Shurong Liu, Mingyue Jiang, Suling Yao, "Research on the Settlement Prediction Model of Foundation Pit Based on the Improved PSO-SVM Model", Scientific Programming, vol. 2022, Article ID 1921378, 9 pages,2022.https://doi.org/10.1155/2022/1921378

[2]. Amjad Khan, Asfandyar Khan, Javed Iqbal Bangash, Fazli Subhan, Abdullah Khan, Atif Khan, M. Irfan Uddin, Marwan Mahmoud, "Cuckoo Search-based SVM (CS-SVM) Model for Real-Time Indoor Position Estimation in IoT Networks", *Security and Communication Networks*, vol. 2021, Article ID 6654926, 7 pages, 2021. https://doi.org/ 10.1155 /2021/66 54926.

[3]. Rani, Antony & Albert, Pravin, Rainfall flow optimization based K-Means clustering for medical data. Concurrency and Computation: Practice and Experience. 33. 10.1002/cpe.6308, 2021.

[4]. Gursharan Saini, Padhiana, "A Novel Approach Towards K-Mean Clustering Algorithm with PSO", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), Pp. 5978-5986, 2014.

[5]. Tuggener, D., Guggisberg, M., & Gürkaynak, A. (2015). Legal article prediction. arXiv preprint arXiv:1512.03108. While this paper primarily focuses on predicting legal article sections, it involves SVM classification as one of its methods.

[6]. Zhong, W., & Miao, L, "Legal case retrieval using question similarity", "Proceedings of the 18th International Conference on Artificial Intelligence and Law", (pp. 215-216), 2019 .

[7]. Tikhonov, A., & Gooßen, A, "Legal document classification using text and meta-data features", "Proceedings of the 14th International Conference on Semantic Computing (ICSC)", (pp. 88-95), 2020.

[8]. Baldwin, T., & Tsang, A. "Unsupervised identification of rhetorical zones in persuasive text", "Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics", (pp. 632-639). 2017.

[9]. Caliskan-Islam, A., Bryson, J. J., & Narayanan A. "Semantics derived automatically from language corpora contain human-like biases. Science" , 356(6334), 183-186, 2016.

[10]. Tuggener, D., Guggisberg, M., & Gürkaynak, A. Legal article prediction. arXiv preprint arXiv:1512.03108, 2015.

[11]. Bashir, S., Qadir, A., & Ahmad, S . Legal document classification using hybrid feature selection and SVM. In Proceedings of the International Conference on Data Science and Applications (pp. 383-393), 2019.

[12]. Govaere, I., & Verschraegen, B. Judicial decision prediction in merger control: an evaluation of different machine learning techniques. World Competition, 43(3), 383-418, 2020.

[13]. Harrer, M. S., & Brinker, K. Classification and evaluation of German court decisions on medical liability using machine learning. In Proceedings of the 16th International Conference on Artificial Intelligence and Law (pp. 97-106), 2017.

[14]. Kolovos, D., & Stratis, D. "A hybrid methodology for the prediction of asylum-seeker status in the EU", "Proceedings of the International Conference on Hybrid Intelligent Systems", (pp. 364-372), 2018.

[15]. Le, Q., & Mikolov, T. "Distributed representations of sentences and documents", "International Conference on Machine Learning", (pp. 1188-1196), 2014.

[16]. Marquis, P., & Sylvestre, M. E. "Using textual features in machine learning to predict the outcomes of cases at the European Court of Human Rights", "Proceedings of the 15th International Conference on Artificial Intelligence and Law", (pp. 173-182), 2016.

[17]. Vidolov, S., & Mihaylova, L, "Legal judgment prediction: a generalized approach", "Proceedings of the 27th International Conference on Artificial Neural Networks", (pp. 189-198), 2018.

[18]. Verma, L., Srivastava, S. and Negi, P.C. (2016) A hybrid data mining model to predict coronary artery disease cases using noninvasive clinical data. *J. Med. Syst.*, 40, 178.

[19]. Khanmohammadi, S., Adibeig, N. and Shanehbandy, S. (2017) An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*, 67, 12–18.

[20]. Chen, S., Dorn, S., Lell, M., Kachelrieß, M. and Maier, A., "Manifold learning-based data sampling for model training," In *Bildverarbeitung für die Medizin*, Informatik aktuell Springer, New York, pp. 269–274, 2018.

[21]. Delen, D., Walker, G. and Kadam, A. (2005) Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.*, 34, 113–127.

[22]. https://towardsdatascience.com/k-means-clustering-algorithm-applications- evaluation-methods-and-drawbacks-aa03e644b48a.

[23]. Liu Y, Li Z, Xiong H, Gao X, Wu J, Wu S. Understanding and enhancement of internal clustering validation measures. *IEEE Trans Cybern*. 2013;43(3):982-994.

[24]. http://www.cs.kent.edu/~jin/DM08/ClusterValidation.pdf.

[25]. Xu R. A comparison study of validity indices on swarm intelligence-based clustering. *IEEE Trans Syst Man Cybern B Cybern*. 2012;42:1245-1256.