ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse153862019.pdf
https://doi.org/10.30534/ijatcse/2019/153862019

# Data Mining on Prediction of Crime and Legal Judgements: A State of an Art

**Radha Mothukuri[1], Dr B Basaveswara Rao[2]**
[1]Research Scholar, Department of Computer Science& Engineering, Acharya Nagarjuna University, Guntur,, Andhra Pradesh, India.
[2]Professor, Department of Computer Science& Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.
radhahemanth12@gmail.com[1], bobbabrao62@gmail.com[2]

**ABSTRACT**

Information extraction and analysis on legal documents are the core research aspects lying in the stream of applying machine learning in jurisprudence. In the real world of lawsuits, the judgements consist of number of subtasks that are to be viewed in phenomenal form. In this paper a detailed analysis on data mining techniques that are applied on bother prediction of crime and decision-making process using the knowledge extracted from legal documents. The core intension of the paper is to list out the data mining mechanisms that were proposed for dealing the factor of judgement analysis.

**Key words:** Legal Documents, Data Mining, Crime Analysis

## 1. INTRODUCTION

Two basic terms Law and Society are the phenomenal terms which are bounded together in mankind of all time. Along with the advent of civilization of human there begins the formation of system of justice or judicial system. Legalization of Justice and Jurisprudence are the two basic significant terms that rises along with the society gloom. Law enforcement has become crucial when devising the communities in recent decade. Since the society has become more modernised and expanded technically, the revise version and the philosophy of law are also subject to face the changes. As far as our society is concerned, there is no inclusion of Jurisprudence in the primary education. Due to this non imposture, the human society remains unaware of the laws even for the basic needs.

Lawsuit is one among the most challenging legal proceeding for the civilians who are all unaware of legal processes that includes: interaction with lawyers technically, hiring the lawyer for precise companion, the court in which the case has to be filed, the proceeding decisions, the consequences of a law or decision, the impact of words in the case files, the time stamp of cases that are getting filed from different aspects. A civilian who is unaware of the lawsuits may get trapped with false claims and may lose his/ her ages along with time and funds. The intension of these days' lawsuit makers is to provide the proper guidance to such civilian to protect the law from getting trapped and to protect the civilians and provide them with the proper justice.

Meanwhile prevention of crime and detecting the crime are yet another two terms that becomes a trend in crime analysis in order to solve it. In the recent decade, researchers show an immense concentration in addressing the crime analysis scenario through different aspects. These studiesshow a drastic speeding up process in solving the crime cases since the models are computerized and data visualization is being applied to solve it. However even after addressing the crime cases using computerized format still there exists lacking in solving the crime issues and it is even still increasing rapidly. The data of the crime cases rapidly increases as the number of cases increases from all sources. Due to the increase in cases, the storage and analysis of these data remains as a challenging task.

Usually cybercrime requires a high-level specialization or organization which makes it to go nonidentically and it is hard to predict the structure of the crime group [25]. Due to the growth of digital technology among the individuals its every hard to restrict such individuals from being polluted. Irrespective of age and maturity many teenagers indulge in such activities which includes disabling of traffic signals, making them traffic jams, shutting down major power sources, manipulation of trades in stock exchange markets and so on. Apparently though not all, many organizations which indulge on criminal activities can engage the resources. The technology leads the individuals to make themselves equipped with all these proforma in a wide range of area. In such cases, an organized crime by an organization may indulge in a vast range od crimes including Mafia

with the IT professionals as their supporting streamers. The works that they undertake will almost be projects with low level of lifetime that has the potentiality to attack a huge organization or a individual who is wealthy. Apart from the institutional level of crimes, cybercrimes may also carry out by the individuals who have a link with macro level of group or network which can be stated as 'darknet'.

The criminals under cyber level security can be mostly working in loose type of networks, which has the potentiality to show different geographical locations. However, these intruders will be available within the proximity range of the victim where the attack takes place. Local networks, friends or relative home are some of the significance of locations where they stay altogether.Cybercrime hot spots with potential links to organized crime groups are found in countries of Eastern Europe and the former Soviet Union [26, 27].

To address all the problems that are descried earlier in this paper, data mining techniques are expected to be deployed to solve the problems such as extracting crime patterns from the previous case histories, extraction of key information from the text documents, discovering knowledge from the extracted documents are all the needs that are to be fulfilled using data mining techniques.

On the other hand, Machine learning has taken the stream of society to become smarter by interpreting the text documents into crisp form to extract the perfect content that the documents are intended for. Most of the judicial decisions become unnoticed since it is presented in the form of text documents and are hard to interpret all the documents by a single judge or through a cabinet of people. Analysis of the identified data is yet another interesting tasks in machine learning which can provide faster judgments where the lawsuits are defined to take the judge to make decisions after years of proceedings.

In this paper a detailed survey on the data mining techniques that are applied in the prediction of judgements and the prediction of crimes. The detailed survey includes almost all the components of the research contributed on the papers that are referred.

The rest of the paper is organized as follows: section 2 discusses the background knowledge needed for the crime pattern analysis and data mining techniques on legal judgement documents. Section 3 comprises of the detailed literature survey on data mining techniques on legal judgment documents and data mining techniques on crime prediction and analysis.

## 2. BACKGROUND

### 2.1. Data Mining Techniques on Legal Judgement Documents

Table 1 describes the related works made on the legal judgement documents using data mining techniques. The table comprises of the theme of each paper, the datasets used in which the data mining techniques are applied for relevant application, the way the information is processed for the representation on machine learning process, classification techniques applied on the paper for decision making, the pros and cons of the respective research works and the final remarks made by the author of this paper.

The theme of the papers discussed in this session includes, to predict the judgments based on the previous decisions made by the European Court under Human Rights [2], to speed up the estimation of judgements of slow judgements using Machine Learning [4], Predicting Legal Judgements based on the described facts in the case files using Artificial Intelligence [6], Prediction of charges for the cases using neural network model [10], Recognizing the DoS Denial of Service Attack in the sever using pattern recognition form in data mining [12], the Legal Information Retrieval and Focused Semantic Search (LIRFSS) is proposed which comprises of both information retrieval from the case documents and Extraction of Information are proposed [13], Similarities among the Legal Documents are measures and distinguished [24].

Unstructured information holds the hidden members of the family that include the links between exclusive documents known as citations. Expertise based legal records systems is need of the day. Citation analysis is a technique to discover the hidden relationships between the files and used to analyze the expertise switch from authors,articles and files from numerous domains. Consequently, the look at on citations of the legal documents becomes great. Citation from a patent to a scholarly article provides statistics of the industrial and business utility of that book. The evaluation of patent quotation offers an concept of the relationship among science and industry. Given that google patents do now not provide indexing for the instructional citations, the procedure of bibliometric look at on patent citations is time-eating. It additionally does not aid API searches and it becomes tough to search in a massive scale the use of its in-builtinterface. So, an automated bing seek the usage of the bing move slowly resolves the hassle. It's milesin mixture with the automated filtering of consequences for the duplicate information producedwith the aid of the bing seek.

Similarity seek and precedent seek is broadly used operation with the aid of felony specialists. There are many strategies like legal-term cosine similarity and all-time period cosine similaritythat gives a higher similarity end result. The felony-time period cosine method proves to be a highermanner to discover the similarity. Vector space version is used in this method for similarity computation. For a felony cosine similarity, handiest terms which are there inside the legal dictionary are taken into consideration to be the representative phrases and for each time period values are calculated.

The intension of this paper is to extract the core information proposed on the research articles related to data mining techniques applied on legal judgement documents. The core subject of each paper is given under the theme column which is given in precise form.

### 2.2. Data Mining Techniques on Crime Prediction and Analysis

Table 2 describes the literature papers on the prediction of crime and the analysis of crime using the data mining techniques. The table comprises of the source cause that were taken for initial crime scene analysis, the dataset that were used for the prediction process, the actual mechanism of the paper, the features used for crime analysis or prediction, the pros and cons of the paper cited and the remarks.

A fraud is misdirecting or taking unfair gain of some other. A fraud carries any act, exclusion, or concealment, which includes a breach of prison or equitable responsibility or open up to, brings about the harm of different. One of a kind styles of frauds encompass takes a look at fraud, internet sale, coverage fraud and credit card fraud and so forth. Take a look at fraud manner issuance of a test whilst enough money is not present in account; net sale method selling faux gadgets; coverage fraud method faux coverage claimed for automobile damage, fitness care prices and other; credit score card fraud means acquiring credit score card facts from various way that's used for big amount of purchase without the permission of consumer. A violent crime is against the law wherein a responsible celebration threatens to utilize compel upon a casualty. This includes the two crime of hard act called goal, for example, killing or rape. Diverse forms of this crime are as follows: murdering of person with the aid of different.
Homicide: planned slaughtering of every other man or woman. 1st diploma murder: used to allude to a deliberate slaughtering. Second degree homicide: used to allude to kill accidently in which the executioner shows, outrageous detachment to lifestyles of human. Site visitors'violations appear while drivers harm legal guidelines that manage automobile operation on roads and highways. The increasing wide variety of cars in towns causes excessive extent of visitors and implies that visitors'violations end up extra important that may purpose extreme destruction of belongings and greater accidents which could endanger the lives of the people.

To remedy this trouble and save you such consequences, site visitors violation detection systems are wanted. Crook assault is the threat or endeavour to bodily strike a person, paying little recognize to whether touch is without a doubt made, insofar as the casualty is aware of about the peril included. Stage of sexual assaults encompass simple sexual assault: it includes constraining a person to participate in any sort of sexual motion without unequivocal assent. Sexual assault with a weapon: it carries the usage or hazard of the utilization of a weapon or damage to an interloper. Irritated sexual assault: it happens whilst the casualty is truly injured, mangled, fiercely beaten, or in hazard of passing on because of a rape.

Verbal assault: it's far a kind of non-bodily, oral ambush that results in a passionate, mental, and additionally intellectual damage to the casualty, instead of bodily massive damage way. Cyber-crime is the crime related to laptop. It incorporates of laptop and a network for crime to occur. Offenses which are perpetrated towards criminal process to harm the sufferers by way of current media transmission structures, for instance, internet and cell. Various types are web extortion, ATM misrepresentation, twine misrepresentation, record sharing and robbery, hacking, and so forth. Cyber-crime analysis is very critical responsibility of regulation enforcement machine in any. It consists of breakdown of protection, or harm to the computer framework residences, for example, documents, web site pages or programming.

The intension of this perspective in this paper is to analyse the data mining techniques that were used on prediction of crime from the legal documents.

**Table 1:** Data Mining on Legal Judgements

| Sl. No | Theme | Dataset | Information Processing | Classification | Pros | Cons. | Remarks |
|---|---|---|---|---|---|---|---|
| **1** | The theme of this research article is to predict the judgments based on the previous decisions made by the European Court under Human Rights. [2] | ECHR Dataset [3] 3 datasets: Article 3 Article 6 Article 8 | Text content are the basic input for the prediction and classification using N-gram and the topics | A binary classification is used to identify whether the judgement is violated or not | The classification accuracy is around 79%. | Case studies are less in this case | The access of case studies are significant barrier in this process. |
| **2** | To speed up the estimation of judgements of slow judgements using Machine Learning [4] | 25000 judgements from 800 people [5] | The input has been collected in text form and it is given to the Machine Learning for training the model | Different types of classifiers such as KNN, SVD, Linear Neural, Neural Hierarchal Linear Regression classifiers are used | Turning out slow judgments to fast judgements is a good initiative | Collection of datasets and variety of people that they belong to are not mentioned | Slow Judgements with fast ML prediction has high value provided that the human nature is effective in judging the sets. |
| **3** | Predicting Legal Judgements based on the described facts in the case files using Artificial Intelligence [6] | CJO [7] PKU [8] CAIL [9] | The subtasks are represented in the form of Graphs and the Directed Acyclic Graph format-based framework called Top-Judge | SVM classifier and Convolutional Neural Network (CNN) are used | Two different case studies are being done under this concept | Not effective for multiple defendant judgements | Temporary factors are still to be taken for consideration. |
| **4** | Prediction of charges for the cases using neural network model [10] | Prepared by 3 law students to annotate the charges for the criminal cases of more than 100 cases and from the news websites [11]. | The case document is encoded with AttentiveSequence Coder and Bi-GRU Sequence Encoder | Support Vector Machine and Neural Networks are imposed for classification model | Assuming the charges are most needed to reduce the human errors. | Achieving only 60% of success rate is not effective. | High pre-processing is required for classified document. |

| Sl. No | Theme | Dataset | Information Processing | Classification | Pros | Cons. | Remarks |
|---|---|---|---|---|---|---|---|
| **5.** | Recognizing the DoS Denial of Service Attack in the sever using pattern recognition form in data mining [12] | Dataset is not mentioned | Converts the text based log descriptions into Shortest Record Linkage Profiles for better computation | Not Applicable | Detecting DoS of a server and from where it comes. | The test is not made in real network. | Around 30% of server block can be presumed. |
| **6.** | The Legal Information Retrieval and Focused Semantic Search (LIRFSS) is proposed which comprises of both information retrieval from the case documents and Extraction of Information are proposed [13] | 20 PDF case documents collected from Supreme Court of India. | Date, Location and other preliminaries data will be extracted using Information Retrieval | Not Applicable | A detailed Sketch is given and the use is properly defined | Information Visualization technique is not discussed in detail | A better adaptation form of Information Visualization is expected to be incorporated |
| **7.** | Similarities among the Legal Documents are measures and distinguished [24] | 2430 legal documents from The Supreme Court of India Legal Website [14] | The information are classified into Case Citation, Out-Citation of the Judgement and In Citation of the Judgement. | The similarities of Documents are measured in four forms: All-term cosine similarity, Legal-term cosine similarity, Bibliographic coupling (BC) similarity, Co-citation (CC) similarity | Complete Level of Similarities are explained in details along with proper experimental views. | Not Applicable | Predefined level of investigation can be extended to Scalable form |

**Table 2:** Data Mining on Crime Prediction

| Sl. No | Source | Dataset | Feature Extraction Phase | Mechanism | Pros | Cons. | Remarks |
|---|---|---|---|---|---|---|---|
| **1.** | e-mails[15] | Anticipated e-mails | A set of selection of keywords addressing the threatening concern such as "Kill", "Blast", "Guns", "Death", "Bomb", etc. | For classification of crime Decision Tree Algorithm "Enhanced $ID_3$" is imposed. | Importance on the attributes are considered in prior to the knowledge information from decision tree. | - | The results of this paper labels the mails such as suspicious or gives precaution measures in case of emergency. |
| **2** | History of Crime, Age of victim, # of similar arrests in History, Mode of Operation, Visited Countries, Birth Place, # of using Debit or Credit cards, Location of Debit or credit card accesses, mistakes of victim on the day [16] | Sensor devices, CCTV Footage, e-contents such as whatsapp, text messages, activities on Social Media. | Match of images via sensors and CCTV footages using sliding window. Sentimental Analysis on Texts and posts via lexical analysis framework and NLP. | A predefined trained Model of classification mechanism is used to map the input and the suspicious firm. | Current data such as precise location and data of mobiles and social media are put into action. | Clear flow of crime analysis if absence throughout the paper structure. | The level of suspect are defined in three classes namely High, Medium and Low. |
| **3.** | Attributes on Crime such as Theme of Crime Scene, the date and the day, offensive measures taken by victim, time, etc. [17] | Criminal Records of Colombo Jail | Stated Evidences are extracted | To Analyse the pattern of crime Clustering model has been imposed. To find the highest probability Naïve Bayes Classifier is used | For small datasets it works efficient | The clustering model is note crystal clear in this paper and the cross validation is absent. | 2 levels are described which first classifies the model to suspect and then based on the judgement it may be stated as criminal. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **4.** | Crimes on homicide and its similar references/ occurrences [18] | Dataset of Crime from England Police database from 1990-2012 | Based on the collected evidences the crime has been extracted using crime pattern | k-means clustering algorithm for pattern clustering | A year wise cluster is formed | Concerns are only on homicide crime clustering | A generic method is expected |
| **5.** | Burglary, Homicide and Robbery [19] | Dataset of Crime from England Police database from 1990-2011 | Detection of outlier is carried out by k-NN Classifier | GINI index is used as decision tree making mechanism for classification | Genetic algorithm is used for optimizing outlier detection | Cluster numbers are not optimized | The quality and effectiveness are deliberately earmarked. |
| **6.** | Crime Location, Date of crime, crime type from previous history via websites, blogs, social media, feeds. [20] | Websites, blogs, RSS feeds | Crime information are extracted based on the keywords "Vandalism", "murder", "robbery", "burglary", "abuse of sex", "arson", etc. | Naïve Bayes classifier and Support Vector Machine are used. | Prediction has been carried out using decision tree and cross validated with police. | Usage of decision tree will be highly efficient in spotting the crime | The location of crimes are geographically represented which gives multi-dimensional interpretation. |
| **7.** | Database of Crime and Information on Criminals [21] | National Level Crime Record Database | Frequency of crimes, nature of a crime, and its impact | Offensive crime profile has been created for every crime per year. | New distance measures metrics are defined | Overall execution time is heavy | Three Clusters are formed: Severe Criminal, One time criminal and minor criminal. |
| **8.** | Type of Crime and its location co-ordinates [22] | SNAP Gowalla, DataSF till Feb'15. | Geographical Co-ordinates, category of the location specified, Entropy of the neighbour, Density of Social Tightness, Venue from the road intersection | SVM, Random Forest and Linear Regression | Random split worked even. | - | Crime pattern has been devised with the given information precisely. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **9.** | All Criminal Personal details with criminal records along with Social Security Number (SSN) [23] | ACRJ and JACC | Crime History, Age, Employment History, “Assault”, etc. | Mental Strength of the prisoners are also considered | Identifying the illness of a prisoner. | Statistical analysis of the results are not present | Referred prisoners are given with long term jail |

## 3. CONCLUSION

In this paper a detailed analysis on the data mining techniques that are applied in the prediction of crime as well as the decision-making process that were automated from the information extracted from legal documents such as judicial documents. The theme of this paper is well defined and it is reflected in the Tables I and II. The extracted contents of every paper are very precise so that the contents of every paper are clearly described. From the details described in this paper it can be further extended to the approaches with higher level of machine learning approaches.

## REFERENCES


1. Agrawal, Ajay K., Joshua S. Gans, and Avi Goldfarb. *Prediction, Judgment and Complexity: A Theory of Decision Making and Artificial Intelligence*. No. w24243. National Bureau of Economic Research, 2018. https://doi.org/10.3386/w24243
2. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., &Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, e93. https://doi.org/10.7717/peerj-cs.93
3. https://figshare.com/s/6f7d9e7c375ff0822564
4. Evans, O., Stuhlmüller, A., Cundy, C., Carey, R., Kenton, Z., McGrath, T., & Schreiber, A. (2018). Predicting Human Deliberative Judgments with Machine Learning. Technical report, University of Oxford.
5. https://github.com/oughtinc/psj
6. Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3540-3549). https://doi.org/10.18653/v1/D18-1390
7. http://wenshu.court.gov.cn/
8. http://www.pkulaw.com/
9. http://cail.cipsc.org.cn/index.html
10. Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). **Learning to predict charges for criminal cases with legal basis**. arXiv preprint arXiv:1707.09168.
11. http://news.cn
12. Khan, Mohiuddin Ali, Sateesh Kumar Pradhan, and Huda Fatima. "**Applying data mining techniques in cyber crimes**." 2017 2nd International Conference on Anti-Cyber Crimes (ICACC). IEEE, 2017.
13. Gaur, Dhruv. "Data mining and visualization on legal documents." 2011 International Conference on Recent Trends in Information Systems. IEEE, 2011.
14. Supreme court of india judgments. http://www.commonlii.org/in/cases/INSC.
15. Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014.
16. Ehab Hamdy, Ammar Adl, Aboul Ella Hassanien, Osman Hegazy and Tai-Hoon Kim, "Criminal Act Detection and Identification Model", Proceedings of 7th International Conference on Advanced Communication and Networking, pp. 79-83, 2015. https://doi.org/10.1109/ACN.2015.30
17. Kaumalee Bogahawatte and ShalindaAdikari, "Intelligent Criminal Identification System", Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.
18. Jyoti Agarwal, Renuka Nagpal and Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications, Vol. 83, No. 4, pp. 1-4, 2013.
19. Rasoul Kiani, Siamak Mahdavi and Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 8, pp. 11-17, 2015.
20. Shiju Sathyadevan, M.S. Devan and S. Surya Gangadharan, "Crime Analysis and Prediction using Data Mining", Proceedings of IEEE 1st International Conference on Networks and Soft Computing, pp. 406-412, 2014.
21. Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", Proceedings of 6th IEEE International Conference on Data Mining, pp. 1-7, 2006.
22. Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, "Mining Location-based Social Networks for Criminal Activity Prediction", Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.
23. Kevin Sheehy et al., "Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System", Proceedings of IEEE Systems and Information Engineering Design Symposium, pp. 250-254, 2016.
24. Kumar, S., Reddy, P. K., Reddy, V. B., & Singh, A. (2011, March). Similarity analysis of legal judgments. In *Proceedings of the Fourth*

*Annual ACM Bangalore Conference* (p. 17). ACM.

25. Lusthaus, J. (2013). How organised is organisedcyber crime? Global Crime,

Kshetri, N. (2013a). Cybercrime and Cybersecurity in the Global South. New York: Palgrave Macmillan.

26. Jones, J. (2010, March 25). Profile of A Global Cybercrime Business – Innovative Marketing. Microsoft Security Blog Retrievedfrom http://blogs.technet.com/b/security/archive/2010/03/25/profile-of-a-global-cybercrime-business-innovative-marketing.aspx.
27. Machine Learning Techniques with IOT in Agriculture in International Journal of Advanced Trends in computer Science and engineering,8(3):742-747,June 2019. https://doi.org/10.30534/ijatcse/2019/63832019
28. Crime Trend Analysis Using Data Mining Technique in International Journal of Advanced Trends in Computer Science and Engineering 8(3):663-666,June 2019. https://doi.org/10.30534/ijatcse/2019/52832019
29. Predicting student's Performance by using classification Methods in International Journal of Advanced Trends in Computer Science and Engineering,8(4),2278-3091,July 2019 .
30. Radha Mothukuri,Cluster Analysis of Cyber Crime Data using R, International Journal of Computer Science and Mobile Applications, Vol.6 Issue. 2, February- 2018, pg. 62-70 ISSN: 2321-8363.
31. Radha Mothukuri,A Brief survey on classification,clustering and preprocessing techniques usage in text mining, International Journal of Research, ISSN NO:2236-6124,Volume7,Issue II,July/2018
32. Performance Prediction of Chronic Kidney Disease using various Data Mining Techniques, Radha Mothukuri, International Journal of Advanced in Management, Technology and Engineering Sciences,2017, ISSN NO : 2249-7455, Volume 7, Issue 12.