Augur Justice : A Supervised Machine Learning Technique To Predict Outcomes Of Divorce Court Cases

Somya Goel, Sanjana Roshan, Rishabh Tyagi, Sakshi Agarwal
Department of Computer Science & Engineering
Jaypee Institute of Information Technology
Noida, Uttar Pradesh, India
{ mailtogoelsomya,mailtoroshansanjana,r119d23 }@gmail.com, sakshi.agarwal@jiit.ac.in

***Abstract*—Machine Learning and Law are two disciplines that are rapidly gaining everyones attention due to Machines ability to understand, process, and learn the data. Legal data currently is present in tremendous amounts that are produced every day. If this legal data can be effectively classified and trained corresponding to a specific domain, it can prove to be a great help to the general public. This research paper focuses on one of the domains of law, which is of marriage and divorce belonging to three different religions, namely Hindu, Muslim, and Christian. The objective is to allow the user feed in information about their case relating to marriage and divorce domain of law. For the given user, the religion is detected, and for predicting the probability of winning or losing the case described by the user, the laws of the user's religion play a vital role. It is thus making laws on each of the three religions an important aspect of this work. Based on the pivotal features present in the trained data set of the previously fought court cases of the similar domain, the probability of losing or winning the case is determined through legitimate processes. Along with this tool which helps the user to determine the losing or winning probability of their case, this research paper also aims at attaining a comparative analysis of various Supervised Machine Learning algorithms on the domain of law and thus showing the proposed algorithms capability to predict the outcome better than that of commonly used Supervised Machine Learning Techniques.**

***Index Terms*—divorce, naive bayes, law, machine learning**

I. Introduction and Related Works

Legal information is produced in enormous amounts and is needed to be adequately classified in order to be reliably accessible. As inferred from the Census 2011, the Divorce stock rate is 2.0 and 3.7 for Hindus and Muslims, respectively. It shows that if we consider 1,000 married Muslims and 1,000 married Hindus, 2.4 and 2.0 are divorced, respectively. The conventional legal hard copies have a cost associated with them, and libraries cannot be accessed at all times by the general public. As a matter of fact, none of the existing legal platforms provide accountable histories of cases fought in the past. Also, there is no medium for the general public to get help in the serious legal issues that may arise in their lives. Nowadays, people are generally taken for a ride and are unnecessarily harassed and tortured at the hands of the prosecuting agencies. The United Nations Development Programme pondered over an urgent need to improve the living conditions of people all over the world. They listed 17 Sustainable Development goals, which are a universal call for action to protect the planet and ensure that all the people enjoy peace, prosperity, and a healthy lifestyle. One of the goals that inspired this paper was to attain peace, justice, and strong institutions. This paper Augur Justice aims at providing a tool to the general public for their legal problems relating to marriage and divorce with the help of intermixing of two disciplines, namely machine learning and law. The paper covers the following UNDP sub-targets 16.b promote and enforce non-discriminatory laws and policies for sustainable development, 16.3 promote the rule of law at the national and international levels,

978-1-7281-0899-5/19/$31.00 ©2019 IEEE

and ensure equal access to justice for all and 16.5 substantially reduce corruption and bribery in all its forms. A crucial section under the massive umbrella of Artificial Intelligence is Machine Learning. Machine learning is the scientific study of algorithms and mathematical, statistical models that help in efficiently performing tasks without giving instructions, relying on statistical models instead. In Machine Learning, computers use algorithms to analyze data, learn patterns to train the data for further usage in performing of orders. Machine Learning has started to emerge as a keen factor in driving the Law firms and the entire Legal sector towards the path of new technology. With the advancement in technology, innovation will be the key to transform the legal profession. Therefore, Machine Learning will prove to be a large factor shifting the way legal work is done. Lawyers are not programmers, and programmers are not lawyers. Machine learning algorithms have been used in the manner to train legal data to facilitate performing tasks involving decision making without providing explicit instructions. Natural Language Processing, an important component of Machine Learning, provides the ability for computers to understand and process human languages. There have been a variety of research done on machine learning and its applications. The preexisting machine learning algorithms have been applied in various fields to predict different kinds of things based on the training dataset. This variety includes the recognition of spam, and junk mails are done by applying the Naive Bayes Classifier [4]. Also, a document classifier has been built with the help of supervised learning techniques [6]. Machine Learning algorithms are also known for predicting the probability of occurrence of disease varying from heart ailment to something as crucial as cancer [5]. The filtering of vital facts from all the posts present on Facebook has also been achieved by the use of Machine Learning techniques [7]. The experimentation is also done on e-commerce websites by classifying the requirements of end-users [3]. Also, a supervised learning model has been made to classify posts on social media platforms. Despite its wide area of implementation, very little has been explored with the law domain. The little that has been achieved is the summarisation of legal documents [1,6,8].

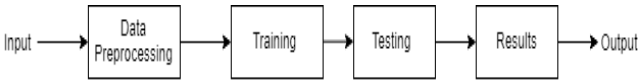

Fig. 1. Basic steps of Machine learning technique.

Therefore, this tool concerns about showcasing and guiding people on a legal path, which does not lead to disappointments in the domain of marriage and divorce laws. The aim is to allow general public feed descriptions about their case, thereby providing probability to whether or not to file their case after judging all aspects of the case relating to marriage and divorce. For predicting the outcome of the legal case belonging to marriage and divorce domain, various marriage and divorce laws of each religion plays a vital role in deciding the probability to win or lose a specific case. The religions focused on in this paper are Hindu, Muslim, and Christianity. So, in order to provide a better prediction, the name of the user is asked, and then the users religion is identified by the last name using datasets of surnames classified according to the three above mentioned religions. The case is taken as input from the user and is preprocessed. According to the pivotal features which are present in the trained dataset, the users probability of winning or losing the legal case is determined. Along with providing the user the probability of winning or losing the case through legitimate processes, the paper also aims at conducting a comparative analysis of various Supervised Machine Learning algorithms on the law domain.

II. BACKGROUND

This work is a unique amalgamation of law and machine learning. It has simplified the lives of people by gifting them suggestions on all domains. So, why not suggest them what to do when it comes to their rights that is the domain of law. The working of any ML technique follows a scheme, as shown in Figure 1. The paper presents a comparative study of several standard machine learning algorithms.

A. Data Preprocessing

Since the data had been collected by manually surfing various sites, it was then word tokenized,

and the keywords were searched to bring the data in the utilitarian form. To obtain processed data, the proposed algorithm follows a set of steps as shown in Figure 2. A CSV file of the segregated features based on the Hindu Marriage Law is made. The Hindu Marriage Act states these features as important in deciding the fact whether the divorce should be granted or not. These features include Cruelty, Violence, Forcible conversion of religion, Mutual Corporation, Sexual Disease to any of the partners, Physical or Mental Assault, Misbehaviour, Dowry, Unsoundness of Mind, Adultery by partner, Desertion for a long period, Wife forcing husband to Leave parents, Sexual Impotency of any of the partner. Another file has the features in accordance with the Dissolution of Muslim Marriage. The notable features of this act cover the parameters Cruelty, Physical Abuse, Desertion, Alimony, Imprisonment for 7 years and more, Verbal or Emotional abuse, Triple Talaq, Wife files case under Dissolution of Muslim Marriage. The last file is for the Christians, the act governing Christian Marriages in India is the Special Marriage Act and Christian Marriage Act. The Special Marriage Act also covers inter-caste marriages. The characteristics of this act are quite identical to the Hindu Marriage Act, but still there are few alterations. The parameters on which the divorce can be granted are Cruelty, Violence, Forcible conversion of religion, Mutual Corporation, Physical or Mental Assault, Misbehaviour, Adultery by partner, Desertion for a long period, Wife forcing husband to Leave parents. Based on these features, the data have been trained to give the prediction of the input case.

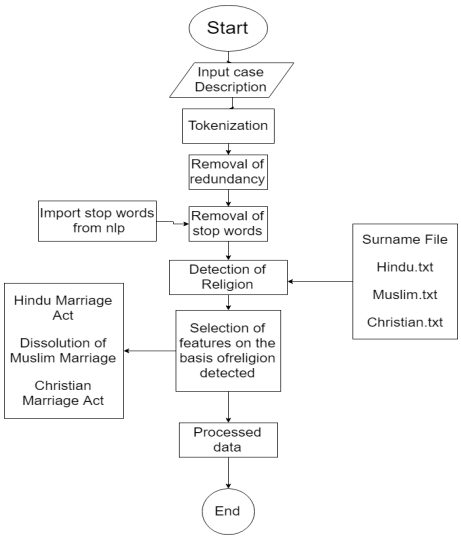

Fig. 2. The detailed layout for data preprocessing.

B. Training

The data has been trained on four non-identical Machine Learning models, and the results have been compared. The models are mentioned below:

- Naive Bayes - This method gives the classification of a dataset by using probabilistic Naive Bayes mathematical theorem [9].
- Decision Trees - The idea of decision trees is really interesting, as the name suggests it an arboriform algorithm that is based on the concept that the constituent is dominant to segregate the parameters impacting the capability of the algorithm of making a decision [10].
- Random Forest - This is a favorable method for assigning the order of precedence to the variables used for training purposes. This method makes multiples trees by splitting the dataset into several randomly arranged chunks and then taking the majority vote of the independently made decision trees to make its final say [11].

The same dataset is trained on the algorithms mentioned above, and the performance of each is analyzed. Besides this the data trained is also used to predict the data given as input, which was preprocessed by tokenizing and extracting the keywords in accordance with the religion of the user detected.

C. Testing

Various applied techniques are tested by splitting the data in testing and training datasets on different ratios, and the accuracy is checked using the above mentioned three techniques. While carrying out this procedure, the ratios of training and testing datasets are varied, and later in the process. The accuracy is noted down for each process individually with help of confusion matrix. Further, cross-validation that to each of the techniques and the contrast in accuracy is observed. Each of the

Machine Learning algorithms is used to forecast the chances of winning the case of the user. For determining the winning and losing of the case, 50 percent is taken as the threshold criterion. After the completion of the procedure, if the percentage comes out to be higher than 50%, the user is most likely to win the case.

III. Methodology

The above-discussed comparison of the various commonly used supervised machine learning techniques works based on words only. The occurrence of a particular word has different sentiments subjected to the formation of the sentence. In order to resolve this issue, the use of Natural Language Processing (for Sentiment Analysis) along with Naive Bayes Classifier is suggested in this paper. The dataset had been trained on the basis of the divorce cases that have been filed earlier, and the results are known. The data was extracted sentence wise, and each sentence has been trained as either true or false based on its overall meaning. The true is assigned to the sentences which are in favor of the appellant (the one filing divorce) and false to those which cannot be used as an argument to win the case. The Naive Bayes Classifier of the Textblob Package has been used to train the data along with Naive Bayes from sklearn package, and the results have been tested on various ratios of test and train dataset. The classification is based more on the way how the sentence is presented. The working is explained in Figure 3. with the help of an example. The above scenario highlights the issue of training data only based on words. The sentence containing the feature Dowry meant that no dowry was taken, but this fact could not be understood by the word wise training algorithm, while the sentence wise training algorithm could figure out the actual sentiment of the word being used. The overall probability of winning the case was thus affected by this. In the proposed algorithm, a hybrid of word wise and sentence wise training algorithms is used to overcome the shortcomings of the standard algorithms.

IV. Analysis of Results

A. Dataset

The user giving case as an input needs to tell his/her last name as that would be the source

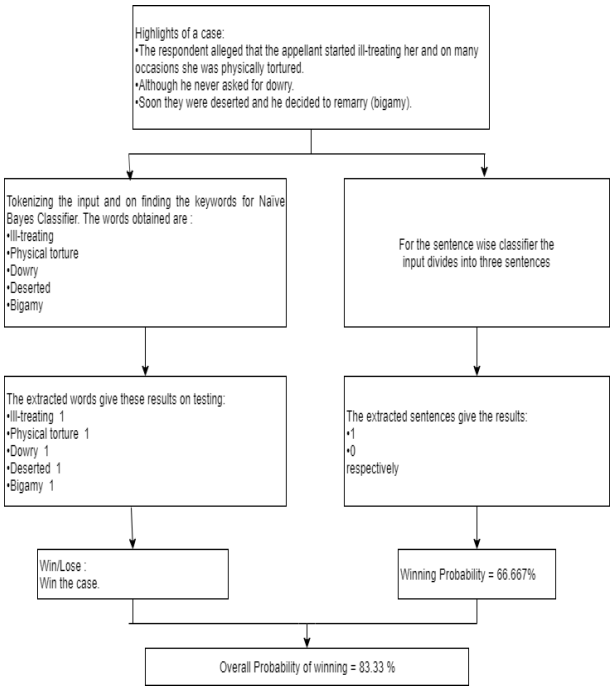

Fig. 3. Example showing working of the algorithm

of determining the religion for the selection of dataset to be used for training. So, the dataset of the surnames of different religions was collected. These datasets were taken separately for Hindu, Muslim, and Christian from family.com. Since, no former work has been done on this domain, finding the apt dataset was a tedious task. After scavenging numerous law-related books and websites, the data has been collected from the case studies available on IndianKanoon.com. Also, some of the cases have been picked from Wikipedia. The cases of each religion were collected and preprocessed separately. There are 26 Hindu, 6 Muslim, and 7 Christian cases. Each case was of around fifteen hundred lines. The relevant lines and features were extracted from these cases, thus creating extensive data.

B. Hardware Specifications

All the approaches were compiled and implemented using Anaconda-Spyder version Anaconda3. The programs were run on hardware configuration as Windows 64 bit operating sys-

tem with i5 6th Generation processor, 4gb DDR4 RAM, 1Tb hard drive.

C. Evaluation Parameter

The contemporary supervised machine algorithms have been compared with Augur Justice on the basis of its accuracy which is computed as shown in equation 1.

$$\frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

where, TP stands for true positive, that is the true instances classified as true. TN stands for true negative, that is the false instances classified as false. FP stands for false positive, that is the false instances classified as true. FN stands for false negative, that is the true instances classified as false.

D. Experimental Outcomes

The paper is focusing on the creation of a tool for enlightening people regarding the law-related issues, specifically in the divorce domain. The approach towards helping them is to apply the combination of Naive Bayes Algorithm (of Textblob package) along with the Sentiment Analysis of the given case. The comparison of the proposed algorithm with the existing Supervised Machine Learning Techniques is given in this section.

Table I. Comparison between various algorithms on Hindu dataset

Test-Train Ratio	Naive Bayes	Decision Trees	Random Forest	Augur Justice
20 : 80	75	75	75	80.25
33 : 67	70.85	70.85	52.94	78.95
46 : 54	92.3	69.23	69.23	74
60 : 40	70	70	70	73
87 : 13	75	50	50	84.09

The datasets of different religions have been segregated from each other. The accuracies of each algorithm applied at different percentages of testing and training datasets (divided based on the number of cases) have been displayed in Figure 4, Figure 5, and Figure 6 for the Hindu, Christian, and Muslim datasets, respectively. The graphs for training size with respect to accuracy also have been plotted for the given algorithm along with

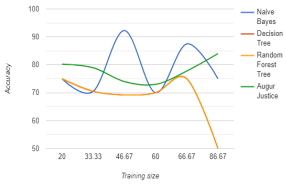

Fig. 4. Training size vs Accuracy for Hindu dataset.

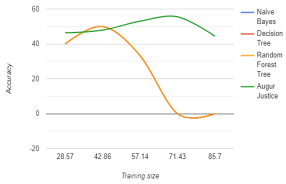

Fig. 5. Training size vs Accuracy for Christian dataset.

the other known supervised learning algorithms in Table I, Table II, and Table III for Hindu, Christian, and Muslim, respectively. The lines in the Muslim and Christian datasets for all three conventional ML techniques give the same percentages, so the same line denotes the accuracy of these algorithms in the graph. The accuracy increases and then begins to decrease at a point since the overfitting of data had taken place. Overfitting refers to the situation where the model learns too much that it starts predicting wrong values. In the tables mentioned above, it can be observed that the accuracy of the Hindu dataset begins to decrease at 46: 54 split of training testing. While in the Muslim dataset, it begins to reduce at 43:57 split ration. The Christian dataset shows this trait after the 50:50 train test split.

Table II. Comparison between various algorithms on Christian dataset

Test-Train Ratio	Naive Bayes	Decision Trees	Random Forest	Augur Justice
29 : 71	40	40	40	46.31
43 : 57	50	50	50	47.8
57 : 43	33.33	33.33	33.33	53.06
70 : 30	0	0	0	55.56
86 : 14	0	0	0	44.44

The vast differences, in many instances, can be

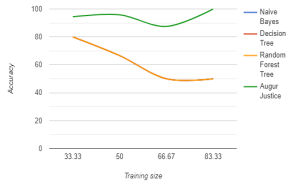

Fig. 6. Training size vs Accuracy for Muslim dataset.

jotted, as shown in the tables and graphs below. The Muslim dataset gave 94.44% as compared to 80%, 95.83% and 87.5% as compared to 66.66% and 50% respectively accurate results and later with a good amount of training dataset (5 cases in training and 1 in testing) it was 100 percent accurate. The Christian dataset, the conventional techniques show 40,50 and 33.33 percent accuracy in comparison to 46.31, 47.8, 53 percent accuracy and at 70 percent the result was not at all predictable by conventional methods, but Augur Justice, was able to give results with an accuracy of 55.56 and 44.44 percents which is far more better and even on the other instances the accuracy was more that is, 46.31, 53.06 in place of 40 and 33.33. In the case of Hindu divorce cases, the Naive Bayes has performed better at two instances, but otherwise the Augur Justice algorithm was quite successful overall as it outshone 50 percent with 84.09 percent of prediction accuracy. It can be observed that Augur Justice has outshone the conventional supervised machine learning techniques. This algorithm is giving consistent results on the various splits of testing and training dataset, which is itself an achievement.

Table III. Comparison between various algorithms on Muslim dataset

Test-Train Ratio	Naive Bayes	Decision Trees	Random Forest	Augur Justice
33 : 67	80	80	80	94.44
50 : 50	66.67	66.67	66.67	95.83
67 : 33	50	50	50	87.5
70 : 30	0	0	0	55.56
85 : 15	50	50	50	100

V. Conclusion

In this paper, an algorithm is proposed, Augur Justice, which has outshone the performance of the contemporary classification techniques. The paper supports the sustainable development goal of peace, equality, and justice. The prediction of the outcome of divorce cases using an efficient algorithm has been achieved. The historical data has been collected separately for Hindu, Muslim, and Christian religions since the divorce laws vary from one religion to another. The sentences of these cases have been classified either as true or false manually. The cases were pre-processed by using various functions of the NLTK before applying the techniques of supervised machine learning and the Augur Justice algorithm.

References

[1] 1. A. Farzindar, G. Lapalme, LetSum, an Automatic Legal Text Summarising system.,Legal knowledge and information system. JURIX, 2004.
[2] 2. H. Surden , Machine Learning and Law. Wash. L. Rev., 89, 87., 2014
[3] 3. J.L.Hellerstein,T.S.Jayram,I.Rish, Recognizing End-User Transactions in Performance Management. IBM Thomas J. Watson Research Division.Hawthorne, NY , 2000.
[4] M. Sahami, S. Dumais, D. Heckerman,E. Horvitz,." A Bayesian approach to filtering junk e-mail". In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105). 1998
[5] M. Langarizadeh, M. Fateme. "Applying naive bayesian networks to disease prediction: a systematic review." Acta Informatica Medica 24.5 ,2016
[6] D. Kalita. "Supervised and Unsupervised Document Classification: A Survey." International Journal of Computer Science and Information Technologies.2015.
[7] R. Benkhelifa, FZ Laallam. "Facebook Posts Text Classification to Improve Information Filtering." WEBIST (1). 2016.
[8] C. Grover, B. Hachey, and C. Korycinski. "Summarising legal texts: Sentential tense and argumentative roles." Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5. Association for Computational Linguistics, 2003.
[9] C.R. Stephens, H.F. Huerta, and A.R. Linares. "When is the Naive Bayes approximation not so naive?." Machine Learning 107, no. 2 (2018): 397-441.
[10] Abdallah, Imad, V. Dertimanis, H. Mylonas, K. Tatsis, E. Chatzi, N. Dervilis, K. Worden, E. Maguire, "Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data." Safety and ReliabilitySafe Societies in a Changing World (2018): 3053-3061.
[11] A. Parmar, R. Katariya, V. Patel, "A Review on Random Forest: An Ensemble Classifier." International Conference on Intelligent Data Communication Technologies and Internet of Things. Springer, Cham, 2018.