**Crime Prediction Using Twitter Sentiments and Crime Data**

**Keywords:** XGBoost, Crime, SHAP, Machine learning, Sentiment Analysis

**Received:**

The incidence of crime is now of great concern globally. The culprits change their tactics on a regular basis. These crimes affect persons, groups, and the government to the extent a whole lot of budgets are allocated to serve as preventive measure to these crimes. The aim of this research is to predict crime based on Twitter hourly sentiments and crime data records. This is because it has been observed that existing crime prediction models that used Twitter data entail some drawbacks in predicting criminal incidents as a result of the unavailability of hourly sentiment polarity and demographic factors. Additionally, SHAP framework was used for the interpretability to rank the feature based on their importance. The xgboost algorithm was utilized with tuning to have an optimal model. The accuracy of 0.81 (81%) was obtained and an Area Under the Receiver Operating Curve (ROC AUC) score of 0.7079 was obtained. The result of this study indicated that crime could be predicted in real-time in contrast to earlier studies on this subject matter. Consequently, it is advised that this work be applied to real-world situations

Povzetek:

1. INTRODUCTION

In this recent time, crimes stand out amongst other social challenges that have an effect on the lifestyle, economy, as well as image of a country [1,2]. Crimes have affected people, establishments, and governments. Therefore, crimes influence a number of choices individuals must make on their own or as a group including moving to a new location, traveling at a proper time, evading unsafe areas, setting up security and safety policies, among others. As a result, this has significant impact on persons, establishments, and governments, a lot such as providing additional security troops and devices, plus court cases in order to keep the economy and repute on the rise.

Various reports have shown that crime rates are increasing. In a report, there is 13% increase in all crime documented by police across England and Wales. The report also shows the rise in violent crimes including weapon crime, sexual crimes, and violence against individual [3].

Consequently, persons, establishments and government are working towards the reduction and eradication of all forms of crimes in the society. In order for this to work there is need for a change of strategy; government must implement augmented strategy [1] and workable information system suitable for this purpose [2].

Crimes can be predicted as a result of the criminals’ actions and their mode of operation as there is a high tendency of repeating the crime under similar conditions. The occurrence of crime depends on several things which include the state of security of the neighbourhood, the intellect of the criminals and so on. According to previous research, there is likelihood for crime to occur again though this is not applicable to all crimes. Hence, making the crimes to be predictable [1]. Therefore, crime-solving is a painstaking work which entails human hard work and intellectual ability to analyze criminal data; therefore, the event of crimes is still the order of the day. It has also been discovered that for the past few years different crime data are being gathered [3], most especially for statistics purpose.

Recently, Machine learning has been active in predicting virtually most of human events and natural occurrences [4]. Machine learning, a subfield of Artificial Intelligence where computer uses systematic algorithms for carrying out task proficiently, without using explicit commands; but depends on patterns and inference in its place [5]. A whole lot of data have been collected; therefore, machine learning can help in crime identification challenges [6–8]. Various research works have been done using machine learning and data mining techniques in crime recognition and prediction [9].

A number of studies have proposed the use of decision trees for crime prediction [10–12] [13 -16]. Also, the research [10] has recommended the use of features such as population, the percentage of individuals above 16 years that are unemployed among others, to predict the level of violent crimes that are likely to occur in a particular area. The suggested methods did not put into consideration the type of crime that is likely to occur [10,11] [14-15]. In addition, these proposed methods used only decision trees classifiers.

In another research [1], the authors used dataset from UK police department [13] to visualize and predict crime using various machine learning algorithms. Also, a similar study [2], gathered data by crawling through various new archives such as The News, The Nation, Dunya News among others with a data miner tool. The data collected was then analyzed and visualized. The data mining techniques were used to gain more knowledge from the data by clustering the data and using various algorithms for crime prediction. Also, previous research [13] showed that GPS-tagged Twitter data can be utilized to predict future crimes in Chicago, Illinois, a major US city. However, current crime prediction models that incorporate Twitter data, have limitations in presenting criminal occurrences as a result of lack of hourly sentiment polarity and demographic factors. It is expected that adding sentiment polarity, crime data, and demographic factors to such models, will improve crime prediction. Furthermore, in crime prediction the interpretability of the machine or deep learning model is vital in order to know how the machine learning model has learnt and as well helped boost the reliability of such model as beneficiary especially law enforcement agencies cannot depend on “black box” system to forecast crime and influence their policies. Hence, affect the reliability of the existing system in crime prediction. Consequently, the lack of hourly sentiment polarity and demographic factors coupled with interpretability, pose a great problem; therefore, there is a need to alleviate this problem.

As a result, the goal of this research is to predict categories of crime. This is done by merging sentiment polarity resulting from lexicon-based sentiment analysis with historical crime data through the use of XGBoost machine learning algorithm and SHAP technique. They are employed to interpret the prediction in making the system reliable as it is accepted that machine learning methods have greatly enhanced crime prediction. However, the inability to interpret the predictions from these sophisticated models is still a limitation. With this, crime prediction will be done in real time which is more reliable. The second section of this study explains the method used in the study; after which the third section explains the results followed by the discussion of the results. Finally, the conclusion is drawn in the fourth section. The current state of the art lack the integration of social media information and crime data for crime prediction.

**Table 1 Summary of related works**

|  |  |  |
| --- | --- | --- |
| **References** | **Application** | **Model and result** |
| Umair et al. 2020 | Spatiotemporal Analysis of Web News Archives for Crime Prediction | Random Forest: Accuracy 62% |
| Ahishakiye et al. 2017 | Crime prediction using Decision Tree (J48) classification algorithm. | Decision tree: Accuracy 94% |
| Iqbal et al. 2013 | An experimental study of classification algorithms for crime prediction. | Naïve Bayes: Accuracy 83% |
| Chen et al. 2015 | Crime prediction using Twitter sentiment and weather | SVM: AUC 67% |

1. METHODOLOGY

This research entails three modules- data preprocessing, crime prediction and interpretability. Figure 1 showed the conceptual framework of the study.

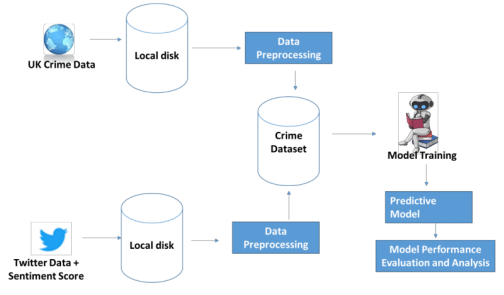


Figure 1: Conceptual framework of the study

**2.2 Data Description and Preprocessing**

The datasets for this research were gotten from the UK police department website [13] and Twitter. For the purpose of this research, the UK (stop and search) crime data was limited to Greater Manchester county and between January and June, 2019. The dataset entails records of crime with 12 attributes in which 5 attributes were taken into consideration for this research. The attributes taken into consideration are crime type, location, date, latitude and longitude. While the Tweets dataset entails GPS tagged tweets from Manchester between 2018 to 2019 were collected from Twitter using the Twitter streaming Application Programming Interface (API). Table 1 and Table 2 below respectively give the description of the dataset.

**Table 2 Description of the UK Stop and Search dataset**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Datatype** | **Description** |
| Type | Ordinal | Category of search carried out by the police officer:   * Vehicle Search * Person Search * Person and Vehicle Search |
| Date | DateTime | Date stamp when the search occurred. |
| Part of a policing operation | Boolean | Was the action part of police activity?   * True * False |
| Policing Operation | Ordinal | What part of the police activity occurred? |
| Latitude & Longitude | Float | Lattitude and Longittude of the location where the search took place. |
| Gender | Ordinal | Gender of the individual. |
| Age range | Ordinal | Age range of the suspect:   * Under 10 * 10-17 * 18-24 * 25-34 * Over 34 |
| Self-Defined Ethnicity | Ordinal | The ethnicity of the officer as stated by the person.   * Black/African/Caribbean/Black British – African * Asian/Asian British - Any Other Asian background * White - English/Welsh/Scottish/Northern Irish/British * Mixed/Multiple ethnic groups - White and Black Caribbean * Other ethnic group - Not stated. |
| Officer-Defined Ethnicity | Ordinal | Ethnicity group of the officer as stated by the officer. |
| Legislation | Ordinal | Law implemented on the person. Which includes:   * Misuse of Drugs Act 1971 (section 23) * Police and Criminal Evidence Act 1984 (section 1) * Firearms Act 1968 (section 47) |
| Object of search | Ordinal | The reason behind the search. Which may include:   1. Controlled drugs 2. Offensive weapons 3. Article for use in theft 4. Firearms 5. Stolen goods |
| Outcome | Ordinal | Outcome of the search or the action carried out by the officer.   * Caution (simple or conditional) * A no further action disposal * Arrest * Khat or Cannabis warning * Summons / charged by post * Community resolution |
| Outcome linked to object of search | Boolean | The outcome of the search linked to the object of search is stated in this attribute.   * True * False |
| Removal of more than just outer clothing | Boolean | Does the search method involve the clothing?   * True * False |

Table 3 Description of the Tweet Dataset

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| Row ID | Integer | ID for each tweet |
| Date | Date Time | The timestamp of when the tweet was tweeted |
| Tweet | String | The text of the tweet |

For the preprocessing,

During the pre-processing stage, the inconsistent data (including missing values, unnecessary information, etc.) were removed and the data was transformed to the format required for crime prediction in the following modules.

Manchester Crime Data

The cases where missing values exists were removed from the dataset. Then for the purpose of this research, the outcome of the stop and search crime data was classified into three groups viz.: Antisocial, Drugs and Criminal which will be the target class for prediction as shown in Table 3, while the fourth group was removed for the dataset as this work aims at predicting crime only. The latitude and longitude features were changed to specific location.

Table 4: Target/Label Classification

|  |  |
| --- | --- |
| Class | Category |
| Nothing found - no further action | NothingFound |
| A no further action disposal | Antisocial |
| Caution (simple or conditional) |
| Community resolution |
| Summons / charged by post |
| Local resolution |
| Offender cautioned |
| Offender given penalty notice |
| Penalty Notice for Disorder |
| Offender given drugs possession warning | Drugs |
| Khat or Cannabis warning |
| Suspect summonsed to court | Criminal |
| Suspect arrested |
| Arrest |

Tweet Data

Data Cleaning: The tweets collected entails different twitter handles (@user), that are being used for identification of users on twitter. These handles bear no tangible information hence, was removed. Additionally, special symbols including (&^123!), punctuation symbols, as well as numbers were replaced with blank spaces; thus, just characters and hashtags were the components of the tweets. Also, words with no meaning or information in the tweets including “oh”,” arg”,” hmm” and words with 3 letter words or lower were removed.

Finally, tokenization which entails splitting of individual strings to pieces referred to as tokens was carried out after which stemming, which entails the removal of suffixes in words including “ness”, “ly”, “ed”, ”s”, was performed then the tokens were joined again to form sentences.

Feature Extraction: this phase entails the extraction of meaningful features from the processed data. Due to the aim of this research, features need to be generated from the preprocessed data with the use of TF-IDF (Term Frequency – Inverse Document Frequency), which assigns lower weight to the most common word in a document, However, larger weight is assigned to words that are not common in the document.

Where M is given as the number of documents and m is the number of documents a term k is present.

TF is given as the frequency of a particular term k in a document

Therefore, TF-IDF is gotten by multiplying TF and IDF that is, TF\*IDF.

Sentiment Extraction: In this study, SentiWordNet [14] was used to categorize sentiment of tweets. It is referred to as approach utilized for opinion mining in a way that it applied lexicon by calculating sentiment terms found in document (tweet in this study) and determine sentiment based on the class with highest polarity score. It consists of positive and negative documents that analyze each document’s (tweets) words critically to classify that sentiment of such document (tweet). The sentiment categories ranges from 23 to -28 where tweets with zero sentiment score are neutral. Tweets with positive sentiment score are positive and tweets, while negative sentiment scores are negative.

Merging of Data and Feature Selection

From the date feature of the two datasets, hour, minutes and seconds were taken out so as to merge the data effortlessly. Based on the date and time components mined out, the two dataset were merged together. Finally, the attribute that are perceived to be useful to this task such as location, gender, age range, ethnicity, outcome, and sentiment score were extracted whereas the other attributes were discarded.

**2.2 Crime Prediction**

In order to have good prediction and high interpretability, XGBoost was used as the algorithm for predicting the crime. XGBoost is a common algorithm that has a good balance of accuracy, scalability, and efficiency [15]. Additionally, great performance by XGBoost was recorded in the previous works on this subject matter as compared to other algorithms including [13,16,17].

XGBoost is based on decision tree which utilizes an ensemble learning approach to develop diverse models such that each new model attempts to address the shortcomings of the previous models [18]. The given samples are classified using the decision rules in this tree model, and the prediction is carried out by computing the scores in the leaves following the cumulative classification.

**2.3 SHapley Additive exPlanation (SHAP)**

The interpretability of the tree ensemble method is vital; however, tedious to accomplish. In some machine learning algorithms, when the weight of one significant attribute rises, the prominence of that attribute reduces, which leads to confusion [19] (Lundberg et al., 2018). Shapley additive explanation (SHAP) is a machine learning interpreter which can alleviate the challenge [20]. The aim of SHAP is to quantify the level of significance of attributes in machine learning models. Hence, FastSHAP package was used for visualizing the feature importance in this research.

1. RESULTS

The experiment was performed as stated previously. This section gives the details of the results of the experiment. XGBoost algorithm was fitted, and the grid search approach is utilized to optimize the model's parameters [21]. Then the system determines the most performing model on the basis of the evaluation metrics. The best combination of the parameters was chosen as the model after using cross-validation to assess how well each combination performed. The evaluation metrics used include accuracy, precision, recall, specificity, sensitivity and roc auc. performance evaluation. Details about the model’s performance is given in Table 4.

Table 5 Details of the Performance Evaluation of the Model

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Trees | Mtry | Min\_n | Tree\_depth | roc\_auc | Pr\_auc | Accuracy | Precision | Recall | specificity | sensitivity |
| 500 | 12 | 5 | 10 | 0.7049 | 0.733 | 0.8097 | 0. | 0. | 0.9472 | 0.4626 |
| 1000 | 10 | 20 | 20 | 0.7079 | 0.746 | 0.8130 | 0.7893 | 0.4651 | 0.9508 | 0.4651 |
| 2000 | 15 | 20 | 30 | 0.7079 | 0.746 | 0.8130 | 0.7893 | 0.4651 | 0.9508 | 0.4651 |

Mtry – is the number of variables randomly selected as candidates at each split

Trees – number of trees to grow

Min\_n – An integer for the minimum number of data points in node that is required in order for the node to be divided further.

Tree depth: The depth of each tree in the model

The model with an accuracy of 0.8130 in Table 4 demonstrated that it performed significantly better than the other models when the parameters were adjusted to 1000 trees, Mtry of 10, 20 min n, and Tree depth of 20. The hyperparameters were adjusted to have performance measures with better score and it was discovered that the achieved result in the last model converged as it was the same with the preceding hyperparameters that produced the higher evaluation metrics.

Figure 1. shows the ROC curve of the model.

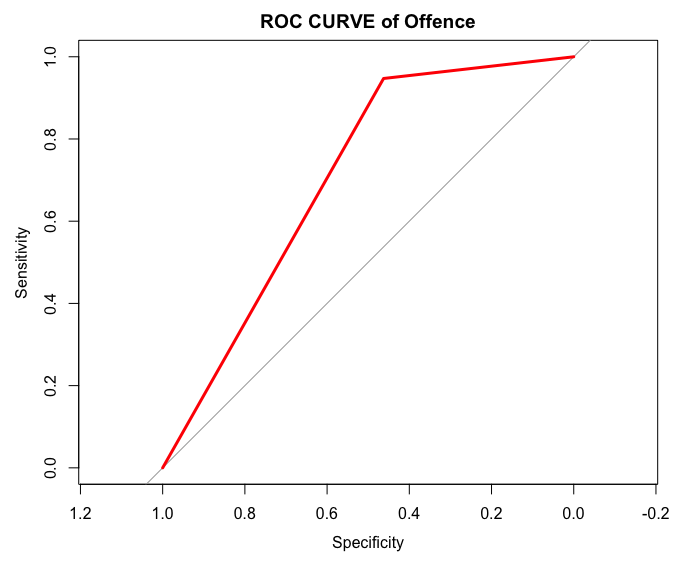


Figure 1. ROC Curve of the model

**3.1 Interpretability**

SHAP is generally used to get the description of the model. The importance of each feature to the prediction of the output is depicted with SHAP value which is weighted and summed over all conceivable feature value combinations. Figure 2 shows how each feature's mean absolute SHAP value is ranked from high to low. The features are arranged by their impact, with the most significant ones at the top. The age range (18-24) and sentiment score are the two most important features.

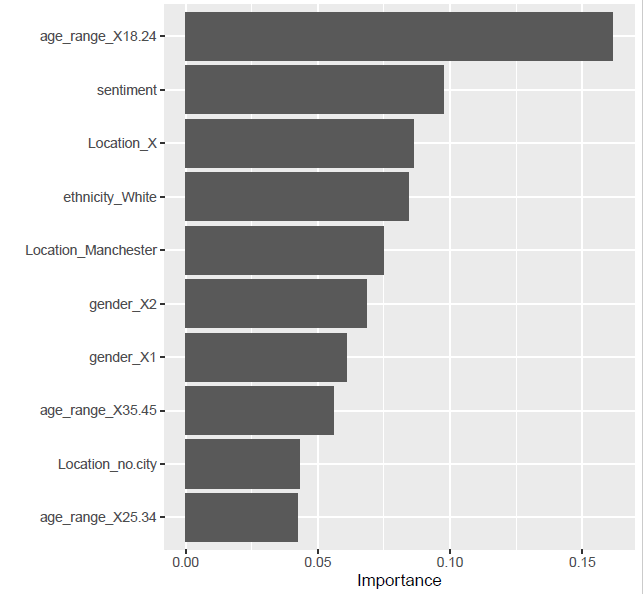


Figure 2. Ranking of the absolute value of SHAP value of all features

1. Discussion

The previous works on crime prediction that used machine learning approaches such as (Qi, 2020; Zhang et al., 2022) study is referred to as “black box” because it cannot determine what really occurs during the process. In essence, it is only to provide the data and obtains the outcome. Hence, lacks interpretability which may affect people’s trust in the prediction models. (Umar et al. 2020; Chen et al., 2015) used spatiotemporal news and crime data only respectively for modeling algorithms prediction and achieved an accuracy of 62% and AUC of 67% respectively, while (Iqbal 2015; Ahishakiye 2017) used Twitter and weather information as source of data for crime and achieved accuracy score of 83% and 94% respectively, hence limitation in source of data used for modeling by these studies could affect the result obtained and the use of accuracy or AUC could not provide enough information to evaluate the classification property of used models . However, in this study, the challenge has been alleviated not only by improving the performance of the model with an accuracy score of 81% and AUC of 71% through integration of crime data records and Twitter hourly sentiment data, also bringing about interpretability of the model through visualizing the features based on their level of importance through the SHAP method through nsim parameter responsible for setting number of repetition to calculate shapley values set to 150 to tradeoff between computation time and accuracy, pred wrapper to select the multiclass classification and test data. While parameter such as mtry the number of predictors that would be sampled, tree depth as the maximum depth in the tree, minimum number of data point required in a node for further split, and tree as the number of trees in the model were set with boost\_tree to achieve an optimal classification result. The result of this research highlights that real time crime can be predicted through the merging of social media and historical crime data in which the reason can be drawn from the sentiment polarity of the social media data. Additionally, it was discovered that this study will be supplementary to existing works on crime prediction that have used a variety of data, including socioeconomic, spatiotemporal, and criminal data, causing the earlier research models to perform poorly in real-time. Hence, this study possessed an advantage over previous studies because it can be used on a real time for crime prediction. The relationship that exist between crime and Twitter data is such that on several cases when crime happened people talk about it on social media, while on occasional cases there are situations where crime perpetrators would first talk about executing a crime, hence the need for this study to capture real time crime before it happens or during the crime execution. The accuracy score of 81% showed that the XGboost model possessed the capacity to correctly classify crime cases from antisocial and drug 81 times out of 100. While the AUC of 71% showed that the classification of these cases (drug, antisocial and crime) is being discriminated with a score of 71. Hence, it is actually the classification beyond chances.

Also, it was discovered that attributes such as age range (especially, 18 -24), location, and sentiment are the crucial factors in predicting crime. Based on the Telephone-operated Crime Survey for England and Wales (TCSEW), which was conducted in 2021 (ONS, 2022), the office of national statistics also validated this. The survey showed that those between the ages of 18 and 34 are the most likely to commit crimes. Furthermore, it can be said from the attribute importance plot that sentiment score also plays a major role in predicting crime. This study could be embraced by major stakeholders such as law enforcement agencies to strategically plan on mobilization of officers to detected crime cases and would help policymakers to plan and mobilize resources based on where crime is prevalent, hence assists in detection of crime cases on a real time or foresee crime happening, hence reduce crime in the society.

The major challenge and limitation of this study is the lack of location of crime tweets and use of other social media platforms to generate crime information, this would have help improve the performance of the model by complimenting the precise location of crime data collected offline. Hence future study should consider the integration of other social media data with their precise location to corroborate the location of crime data from the authority.

1. Conclusion

In this research, historical crime data and twitter (sentiment scores) were utilized together with the use of XGBoost and SHAP. The model's performance during training was improved by adjusting the model hyperparameter. Besides, this work has been able to produce an interpretable model to predict crime. Area Under the Receiver Operating Curve (ROC AUC) of 0.7079 and Accuracy of 0.81 (81%) were both achieved.

It will be fascinating to see how sentiment analysis is improved in the future because social networks frequently utilize slang and other languages that the Natural Language Processing (NLP) system can understand, which in some way affects the model's effectiveness.

**References**

1. ToppiReddy HKR, Saini B, Mahajan G. Crime Prediction & Monitoring Framework Based on Spatial Analysis. Procedia Comput Sci [Internet]. 2018;132(Iccids):696–705. Available from: https://doi.org/10.1016/j.procs.2018.05.075

2. Umair A, Sarfraz MS, Ahmad M, Habib U, Ullah MH, Mazzara M. Spatiotemporal Analysis of Web News Archives for Crime Prediction. Appl Sci. 2020;10.

3. Tompson L, Johnson S, Ashby M, Perkins C, Edwards P. UK open source crime data: Accuracy and possibilities for research. Cartogr Geogr Inf Sci. 2015;42(2):97–111.

4. Oladimeji OO, Oladimeji A, Oladimeji O. Classification models for likelihood prediction of diabetes at early stage using feature selection. Appl Comput Informatics. 2021;

5. Oladimeji OO, Oladimeji O. Predicting Survival of Heart Failure Patients Using Classification Algorithms. JITCE (Journal Inf Technol Comput Eng [Internet]. 2020 Sep 30;4(02):90–4. Available from: http://jitce.fti.unand.ac.id/index.php/JITCE/article/view/75

6. Malathi A, Baboo SS. Enhanced Algorithms to Identify Change in Crime Patterns. Int J Comb Optim Probl Informatics. 2011;2(3):32–8.

7. Brayne S, Christin A. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. Soc Probl. 2021;68(3):608–24.

8. Manzanares MCS, Diez JJR, Sánchez RM, Yáñez MJZ, Menéndez RC. Lifelong learning from sustainable education: An analysis with eye tracking and data mining techniques. Sustain. 2020;12(5).

9. Kotevska O, Kusne AG, Samarov D V., Lbath A, Battou A. Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics. IEEE Access. 2017;5:20524–35.

10. Ahishakiye E, Omulo EO, Taremwa D, Niyonzima I. Crime prediction using Decision Tree (J48) classification algorithm. Int J Comput Inf Technol. 2017;06(03):188–95.

11. Nasridinov A, Ihm SY, Park YH. A decision tree-based classification model for crime prediction. Lect Notes Electr Eng. 2013;253 LNEE:531–8.

12. Iqbal R, Murad MAA, Mustapha A, Panahy PHS, Khanahmadliravi N. An experimental study of classification algorithms for crime prediction. Indian J Sci Technol. 2013;6(3):4219–25.

13. Chen X, Cho Y, Jang SY. Crime prediction using Twitter sentiment and weather. 2015 Syst Inf Eng Des Symp SIEDS 2015. 2015;(c):63–8.

14. Ohana B, Tierney B. Sentiment classification of reviews using SentiWordNet. 9th IT T Conf. 2009;

15. Mousa SR, Bakhit PR, Osman OA, Ishak S. A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments. Transp Res Rec. 2018;2672(42):268–79.

16. Zhang X, Liu L, Lan M, Song G, Xiao L, Chen J. Interpretable machine learning models for crime prediction. Comput Environ Urban Syst [Internet]. 2022;94(November 2021):101789. Available from: https://doi.org/10.1016/j.compenvurbsys.2022.101789

17. Qi Z. The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model. Proc 2020 IEEE Int Conf Artif Intell Comput Appl ICAICA 2020. 2020;1241–6.

18. Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. PeerJ Comput Sci. 2017;2017(7).

19. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng [Internet]. 2018;2(10):749–60. Available from: http://dx.doi.org/10.1038/s41551-018-0304-0

20. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology. 2019;126(4):552–64.

21. Putatunda S, Rama K. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. ACM Int Conf Proceeding Ser. 2018;6–10.