An Improved Traffic Crime Predictive System using Multinomial Naive Bayes Text Classification Algorithm

H. A. Akpughe
Department of Computer Science
University of Port Harcourt Nigeria

P. O. Asagba
Department of Computer Science
University of Port Harcourt Nigeria

C. Ugwu
Department of Computer Science
University of Port Harcourt Nigeria

ABSTRACT
Traffic law enforcement agencies in Nigeria have faced a huge setback as they do not have records of offenders or criminals that have been persecuted in the past. In this paper, a system was developed that can predict the possible class of traffic crime together with the penalty attached to that class of criminal offence that a known traffic criminal offender is most likely to commit next. The likelihood and frequency table will be constructed from a dataset of traffic crime data, the likelihood of a user falling under a particular class of traffic crime will also be established. Also, proposed to be designed and developed is a predictive system that uses object-oriented analysis and design methodology (OOADM), improved naïve bayes text classification algorithm to solve these problems. This will be achieved by implementing the stated model with python model-view-controller (MVC) framework known as Django Framework. This improved system is implemented using a real-time, cloud-hosted NOSQL database called FireBase which guarantees scalability. From the results, it was found out that the speed and predictability of probability of any user falling under a class 1 crime type was 81.42% and 10.39%, 8.19% for class 2 and class 3 respectively.

General Terms
Crime, improved, system, prediction, algorithm, online database, online, retrieval and storage.

Keywords
Predictive system, naïve bayes, classification algorithms, traffic crime, machine learning algorithm, NoSQL, Firebase, scalability.

1. INTRODUCTION
A traffic crime predictive system is a type of information filtering system that predict the probability of a user being categorized under a particular class of crime based on previous record. Predictive systems are sometime referred to as categorization system. Predictive system describes web applications that predicted response to options. According to [1], traffic crime predictive systems are targeted to offenders who have had an experience with the system and producing a potential result without overwhelming the officer with information that may or may not be readily available.

Some countries such as USA, UK and Germany are already are already utilizing the power of the predictive crime prediction systems in building their new and existing systems. Traffic crime predictive systems often provide personalized information of offenders showing their accurate bio information as well as previous details of offences committed.

According to [2] government's business mainly consisted of data processing and using information within its own departments as well as disseminating it in public for the benefit of the citizens. As examples, it is common to rely on inter agency documentations when profiling an individual for an offence committed; weather forecasters use predictive models in their prediction decisions; and football gamblers also use predictive models to determine outcome of matches.

In attempt to predict a crime class, the current predictive System uses manual processes to create and sort files based on a user detail such as a drivers licence which is unique to all active user. These pred8ictions were assigned for every crime class that similar offenders had committed.

In this paper, a novel architecture for the implementation of traffic crime predictive system with a classification algorithm which can be used to improve the information retrieval process for crime offenders was introduced.

1.1 Recent Application of Predictive Systems.
1. **Customer relationship management (CRM):** Predictive analysis applications are used to achieve CRM objectives such as marketing campaigns, sales, and customer services. Analytical customer relationship management can be applied throughout the customer's life cycle, right from acquisition, relationship growth, retention, and win back.
2. **Health Care:** Predictive analysis applications in health care can determine the patients who are at the risk of developing certain conditions such as diabetes, asthma and other lifetime illnesses. The clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care.
3. **Collection Analytics**: Predictive analytics applications optimize the allocation of collection resources by identifying the effective collection agencies, contact strategies, legal actions to increase the recovery and also reducing the collection costs.
4. **Cross Sell**: Predictive analytics applications analyse customers spending, usage and other behaviour, leading to efficient cross sales, or selling additional products to current customers for an organization that offers multiple products
5. **Fraud detection**: Predictive analytics applications can find inaccurate credit applications, fraudulent transactions both done offline and online, identity thefts and false insurance claims.
6. **Risk management:** A predictive analytics application predicts the best portfolio to maximize return in capital asset pricing model and probabilistic risk assessment to yield accurate forecasts.
7. **Direct Marketing**: Predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer.

8. **Underwriting:** Predictive analytics can help underwrite the quantities by predicting the chances of illness, default, bankruptcy. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data.

1.2 Definition of Terms

Database: This is a structured data set stored within a computer system for easy automatic retrieval and manipulation.

Incident Report: An incident report or accident report is a form that is filled out in order to record details of an unusual event that occurs at the facility or within a region such as a motor accident.

Index: This is a methodical arrangement of a list of items (such as names or terms) for easy search and retrieval.

Information Retrieval: This is the finding of materials of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [3]. This can also be said to be the activity of obtaining information resources relevant to an information need from a collection of information resources.

Information Storage: It is the part of a system that keeps data for easy access to the information processor.

Law Enforcement Agency: Any of a number of agencies chartered and empowered to enforce Nigerian laws within a certain jurisdiction. They are responsible for insuring obedience to the laws within the borders of a nation.

Offence: This is said to be the breach or transgression of an existing law or rule that has been placed within a given locality. This is also seen as an illegal act.

Offence Record: This is a record of a person's criminal history, generally used by law enforcement agencies, private organization, and others to assess his or her trustworthiness. It may include traffic offences such as speeding and drunk-driving, and in extreme cases records of actual convictions as declared by a qualified court.

Query: This can also be said in this context to be database queries. This is an inquiry into the database used to extract data from the database in a readable format according to the user's request.

2. RELATED WORKS

The effect of sequencing storage and retrieval requests on the performance of automated storage and retrieval systems (AS/RS) was studied by [4] where a storage request is given a predetermined storage location (known location). By taking advantage of this unique operating characteristic, they were able to present several optimum and heuristic sequencing methods under static and dynamic approaches. They also found out that the sequencing methods can significantly reduce travel time by a storage and retrieval machine, thereby, increasing throughput.

[5] Research purpose was to study and analyze whether information technologies contribute to the outcome of criminal investigations. The study is a sample one and its unit of investigation is U.S police departments. Mathematical analysis was carried out, clearance rates of crimes are used as the outcome variable and its independent variable is an index variable created by using survey items. Their study also shows that information technologies should be targeted to the policing fields where they would be most effective. In case of criminal investigations, primary investigation phase looks promising. It also argues that process evaluations should be conducted to ensure proper implementation and use of information technologies by police departments.

The volume of information being created, generated and stored is huge. Without adequate knowledge of Information Retrieval (IR) methods, the retrieval & predictive process for information would be cumbersome and frustrating [6]. They further studied the concept of existing information retrieval models, and the knowledge acquired was used to design a digital library information retrieval system. It was successfully implemented using a real life data. Noting that the major problem associated with the existing search engines is how to avoid irrelevant (unnecessary) information retrieval and to retrieve the relevant (necessary) ones.

[7] Described an algorithm which if executed by a group of nodes interconnected would provide a robust keyed indexed information storage and retrieval system with no element of central control or administration. This algorithm allows information to be made readily available to a large group of people in a similar manner to the World Wide Web. Several improvements made over the existing system include:

1. No central control or administration required
2. Anonymous information publication and retrieval
3. Dynamic duplication of popular information
4. Transfer of information location depending upon demand

There exist also several possibilities for this system to be used in a modified form as an information publication system within a large cooperation's whom may wish to utilize unused storage space which is allotted across the cooperation's. [7] report also describes different experiments designed to measure the efficiency and reliability of such a network. These experiments were performed upon a simulation of a working network written in the Java programming language.

[8] study investigates the use of databases in information storage and retrieval in some selected banks in Delta state, Nigeria. Key variables used within the study were reviewed under the following areas: concept of information, concept of databases, concept of information storage and retrieval, role of ICT in information storage and retrieval, challenges of effective information storage and retrieval. A descriptive survey research method was used and data were collected through the use of questionnaire. Research findings showed that parent bodies of banks are the sole source of funding the use of databases in the selected Banks in Delta State, that there are adequate skilled ICT personnel for rendering services through the storage and retrieval of information/data at the banks, there are ICT software and hardware facilities used for storage and retrieval of data/information in the banking industry which includes telephone, signature systems, Microsoft exchange, Microsoft excel, and credit card management. Staff training in the use of database for information storage and retrieval in Delta State banking industries is mostly quarterly, the normal duration of the maintenance of ICT facilities is mainly on weekly and monthly basis depending on the level of usage, there are problems militating against the services rendered by staff to users/customers in the use of database for information storage and retrieval in the selected banks in Delta state [8].

[9] noted that Today's decision support systems based on

predictive modeling are becoming more common, since organizations often collect more data than decision makers can handle manually. Predictive models are used to find potentially valuable patterns in the data, or to predict the outcome of some event. There are numerous predictive techniques, ranging from simple techniques such as linear regression, to complex powerful ones like artificial neural networks. Complex models usually obtain better predictive performance, but are opaque and thus cannot be used to explain predictions or discovered patterns. The design choice of which predictive technique to use becomes even harder since no technique outperforms all others over a large set of problems. It is even difficult to find the best parameter values for a specific technique, since these settings also are problem dependent. One way to simplify this vital decision is to combine several models, possibly created with different settings and techniques, into an ensemble. Ensembles are known to be more robust and powerful than individual models, and ensemble diversity can be used to estimate the uncertainty associated with each prediction.

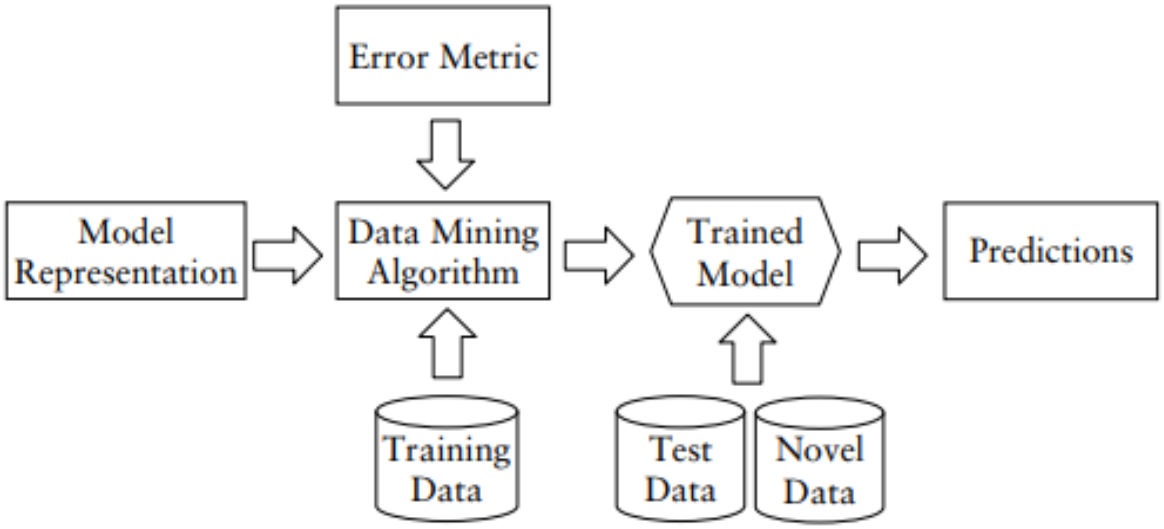

Fig 1. Predictive Model [9]

[9] Main contributions of their thesis were three novel techniques that enhance the performance of their purposed method. The first technique deals with ensemble uncertainty estimation and is based on a successful approach often used in weather forecasting. The other two are improvements of a rule extraction technique, resulting in increased comprehensibility and more accurate uncertainty estimations.

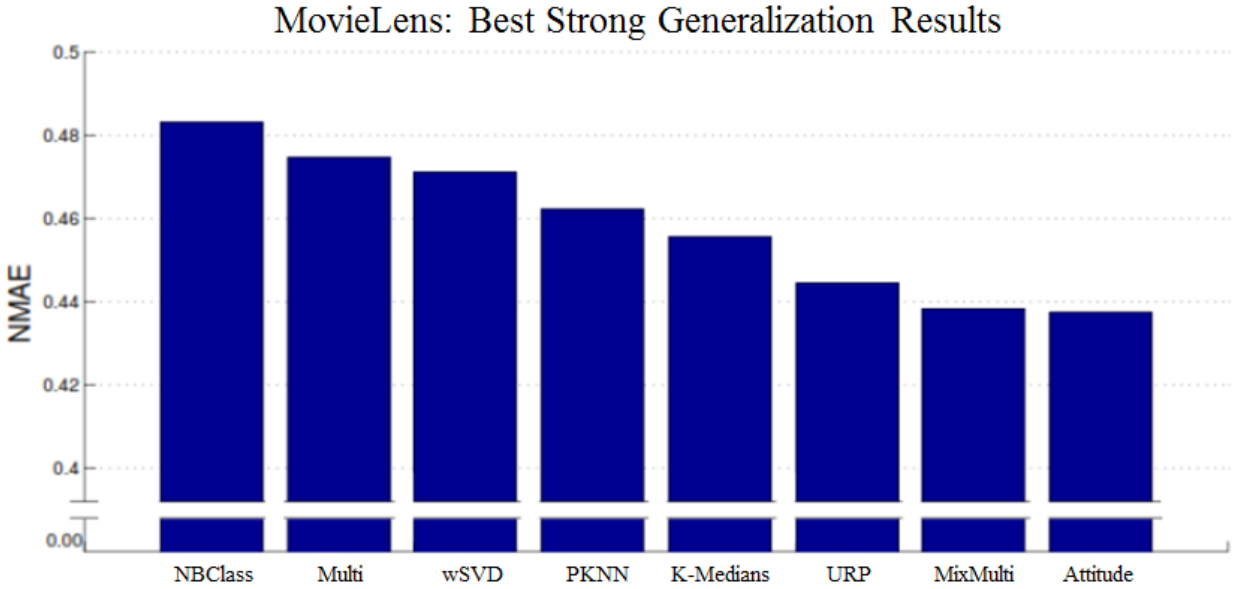

Fig 2. Best strong Generalization results from nine prediction models [10]

[10] conducted a study on the different machine learning approaches to collaborative filtering. In the study, the different K-NN classification and regression methods were applied to derive the class of neighborhoods from a recommenders system. A new method for prediction was introduced; this method was able to learn from a set of Naïve Bayes classifiers. The study also illustrated the application of K-medians clustering for prediction rating. Furthermore, dimensionality reduction techniques such as the weighted singular value decomposition, the principal component

analysis and the probabilistic principal component analysis were applied in rating predictions. A couple of density estimation methods in probabilistic models such as multinomial model, aspect model and user rating profile model were described and a new family of models known as Attitude model family was introduced. A total of nine (9) rating prediction models were implemented and compared as shown in Figure 2.

[11] in his paper presents the concept of accessibility from the field of transportation planning and embraces it within the context of Information Retrieval (IR). An analogy is drawn between the fields, which motivates the development of document accessibility measures for IR systems. Considering the accessibility of documents within a collection given an IR System provides a different perspective on the analysis and evaluation of such systems which could be used to inform the design, tuning and management of current and future IR systems.

3. MATERIALS AND METHODS

3.1 Architecture or the Exiting system.

This describes the existing system, explaining how modules and components integrate and communicate to bring about the working application of the existing system system.
The architecture of the existing system as shown in Fig. 3 comprises the following components;

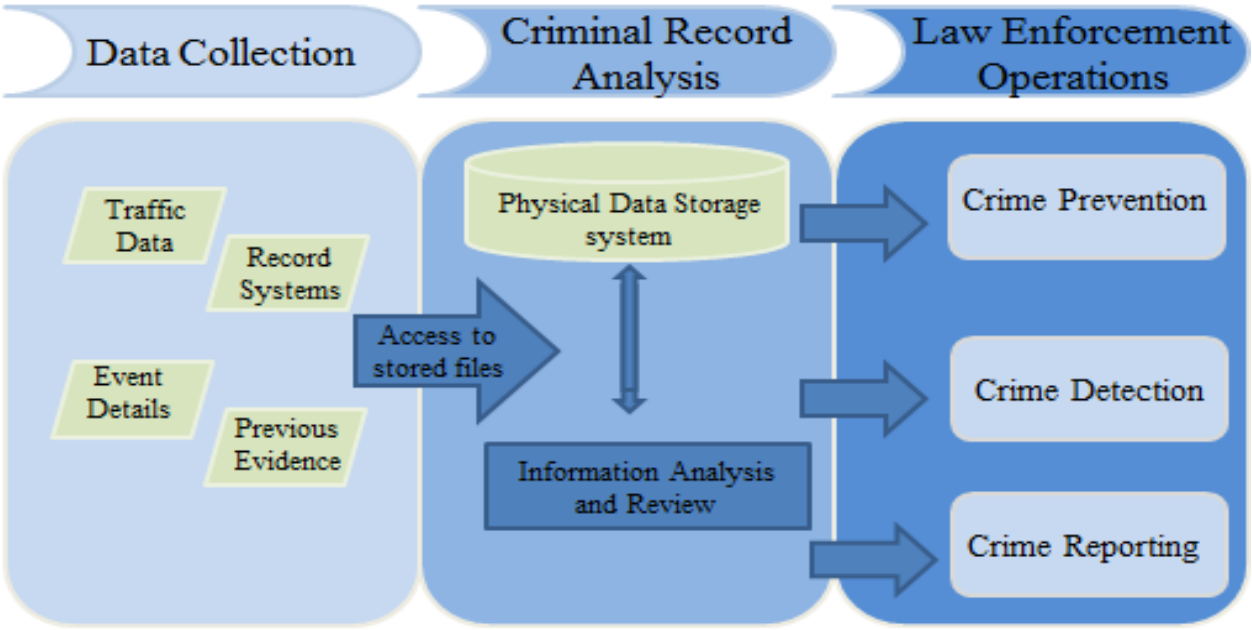

Fig 3. Architecture of the existing ISR for criminal data within the road safety in Nigeria

The system is grouped under three main phases:

1. Data collection phases.
2. Criminal record analysis phase.
3. Law Enforcement operation phase.

Data Collection Phase:
This talk about gathering and measuring information on variables of interest, in an established systematic fashion making all sources of data, information that the system needs to function how it should function. It can be said to either be training data or trained data depending on the stage the system is. Generally, the sources of traffic crime information are:

Traffic crime data: data gathered from the field either by the use of technology such as traffic cameras, traffic light information, collision and accident rates, congestions around certain locations and frequency of minor and major accident occurrences within these locations.

Record systems: findings that are in one way or another related to traffic crime within a certain location or tied to an individual which have been stored prior to this current need to access the information. Databases of other agencies that have similar aim of prosecuting traffic crime.

Event details: information leading up to the event of a crime or incident or those that occur after such as weather condition, health state of the drive, vehicle condition, number of casualties involved and time incident occurred as this would also add to forming a basis for future prediction.

Previous Evidence: those information relied on to get to a conclusion on what has transpired. Details already existing within the system, this might be offender's name, age, driver's license details, state of origin, his or her past criminal record and how they tend to relate to this current crime or offence.

Criminal Record Analysis Phase:
This phase details how all information and records are stored and analyzed within the system. The output of this stage show how effective the system has been in predicting the traffic crime and also show the complexity and ease of understanding of the system. The basic components of this phase are:

Physical storage system: for the existing system, files and folders alphabetically placed in storage cabinets are used. This cabinet is place under physical lock and key to prevent unauthorized access to the files. This means that access is given only when the key holder is available and grants permission.

Information analysis and review: this is a process of checks conducted by an officer in other to ascertain that information stored within the storage system is up to date or n its right format also this process ensures that the information store can

be analyzed and reliable conclusions drawn in other to assist the law enforcement officer effectively carry out their jobs.

Law Enforcement Operations Phase:
This phase talks about all the possible outcomes that could be carried out by the enforcement agencies once a successful analysis has been carried out and how these operations easily prevent crime and ensure speed while bring perpetrators to justice. The basis operations that are carried out in this phase are:

Crime Prevention: Crime can easily be detected and prevented even before they occur by studying trends and analyzing data, seen as the attempt to reduce and deter crime and criminals.

Crime Detection: this is matching techniques and patterns gotten from the analyses phase for crime detection, prevention and control. This is seen as a proactive measure of traffic crime combating.

Crime Reporting: This is a reactive measure or crime combating as it comes after the crime has been committed. It serves as input in future for analyses and it ensures that the details of that crime are not forgotten as offenders are prosecuted. It comes as a form which is designed to make it easier and more convenient for law enforcement agents and traffic officers to report certain traffic crimes and other incidents.

3.2 Architecture of Proposed System.

This describes the proposed system, explaining how modules and components integrate and communicate to bring about the working application of the proposed system. The systems design is developed to satisfy the requirement of modern classification and predictive systems architecture including computational structures and model training algorithms. The system design will also capture the major functional building blocks needed to understand the process of building traffic crime predictive system. The architectural design of the proposed system is illustrated in figure 5. The proposed system architecture is described in Fig. 4:

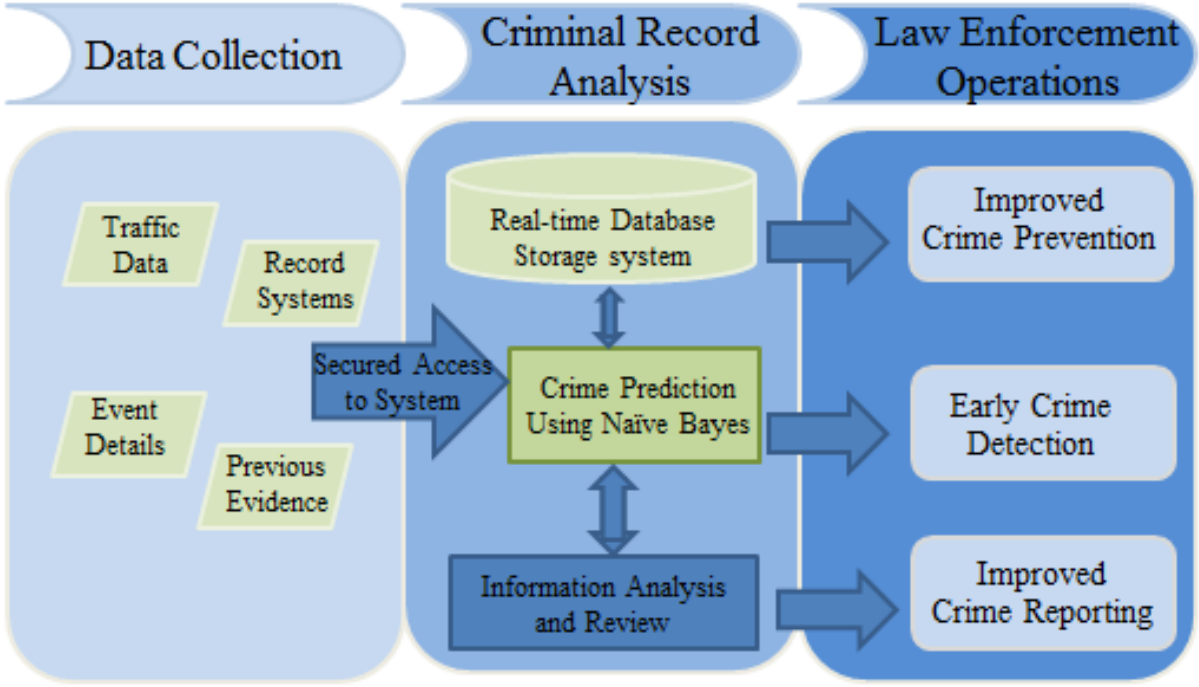

Fig 4. Architecture of the proposed system

The Naïve Bayesian Classifier is based on Bayes' theorem with independence assumptions between predictors. It assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors was used in developing the system. This assumption is called class conditional independence.

Naïve Bayes Classifier is often used to work out posterior probabilities given observations. For example, a driver may be observed to have a certain license number and using Bayes' theorem, the probability that a proposed crime class is correct, given the observation can be calculated.

In training for naïve bayes classification a set of predetermined license numbers have been picked out from a dataset containing various license numbers with different frequency matched to different crime class. The drivers licenses used for training are {UGH8574A, GHJ9856F, PHC3746G, LAG7645H, ABJ2008E, KAD4252W, JOS9080M, BEN7098U, WRR1109Z, DEL2222Q, GOM2008V, AHO5562B}. The frequency table as shown in table 1 was constructed. To eliminate zeros, add-one or Laplace smoothing was used, which simply adds one to each count.

Table 1 Likelihood Table for Training Dataset with Laplace Smoothing

License Number	**Class 1**	**Class 2**	**Class 3**	**Total**	Likelihood (Fraction)	Likelihood (Decimal)
UGH8574A	2	1	1	**4**	4/143	0.0279
GHJ9856F	3	2	2	**7**	7/143	0.0489
PHC3746G	3	2	4	**9**	9/143	0.0629
LAG7645H	3	4	3	**10**	10/143	0.0699
ABJ2008E	5	4	1	**10**	10/143	0.0699
KAD4252W	8	7	7	**22**	22/143	0.1538
JOS9080M	9	3	5	**17**	17/143	0.1188
BEN7098U	5	5	4	**14**	14/143	0.0979
WRR1109Z	2	6	2	**10**	10/143	0.0699
DEL2222Q	7	7	1	**15**	15/143	0.1048
GOM2008V	8	5	6	**19**	19/143	0.1328
AHO5562B	1	3	2	**6**	6/143	0.0419
Total	56	49	38	**143**		
Likelihood (Fraction)	56/143	49/143	38/143			
Likelihood (Decimal)	0.3916	0.3426	0.2657			

Multinomial Naïve Bayes Equation:

$P(H/E) = P(H) \prod_{i=1}^{n} P(E_i/H)$

Where

P(H) Is the probability of a crime class being true. This is known as the prior probability.

$P(E/H)$ Is the probability of a particular license number existing under a particular class being true.

$P(H/E)$ is the probability of the license number existing given that the class is true. It is also known as the posterior probability of class (target) given predictor (attribute).

Table 1 Shows that from the training data after Laplace smoothing, the frequency count of UGH8574A specified with a class 1 offence is 2 with a likelihood of 2/4 = 1/2. It also shows that the frequency count for UGH8574A specified with a class 2 is 1 with a likelihood of 1/4 also the frequency count of UGH8574A specified with a class 3 is 1 with a likelihood of 1/4. This is done for all training set.

Calculating the multinomial naïve bayes probability that a license picked falls under a class 1 crime category is shown below for all training dataset:

$$P(Class\ 1/License) = P(Class\ 1) \prod_{i=1}^{n} P(License_i/Class\ 1)$$

4. RESULTS AND DISCUSSIONS

We have proposed a generalised model for classifying items in classes which can help in reducing search time and the possible inclusion of redundant data using multinomial naïve based technique and weighting functions commonly used in IR models. Fig 5 shows the basic comparison of the proposed system with that of the existing system.

A cross section was sampled and analyzed using Naïve Bayes algorithm and figures was generalized across all data. Firebase has concurrent connections limit for up to 100,000 Concurrent connections. These are the number of users simultaneously connected to your database. Firebase issues the most recent copy of the database to a logged in user and attaches a time stamp to that user. This helps to ensure that the data integrity in the database is maintain as the user with the earliest time stamp updates a copy of that new database is sent out to other users still in connection to the database. It is hereby said that firebase handles a distributed database that is updated regularly.

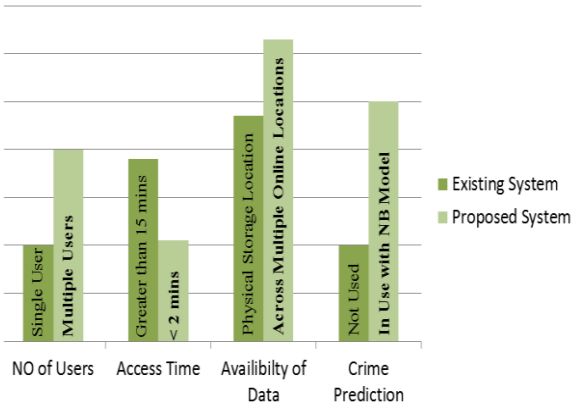

Fig 5. Comparison of existing system and proposed system.

Fig 6 shows all details about a user and his predicted crime and also the penalty for that class of crime. This was achieved using the naïve bayes algorithm embedded within the program. Also the system has been seen to be more secured with the three factor login criteria. The system is operated in real time as a secured connection must be established prior to log on as shown in Fig 7.

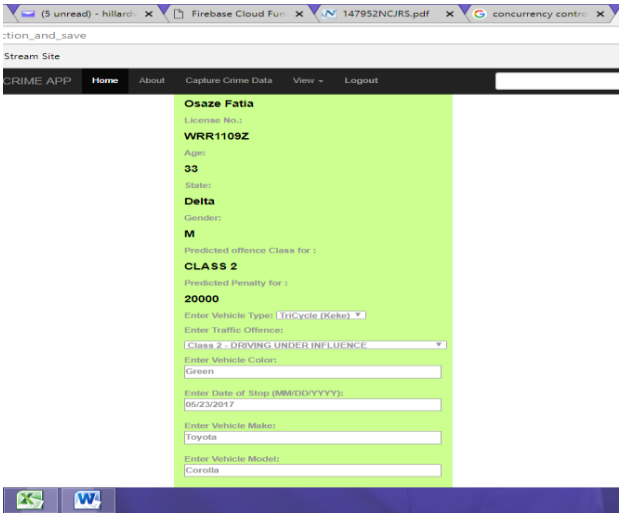

Fig 6. Offender Crime prediction.

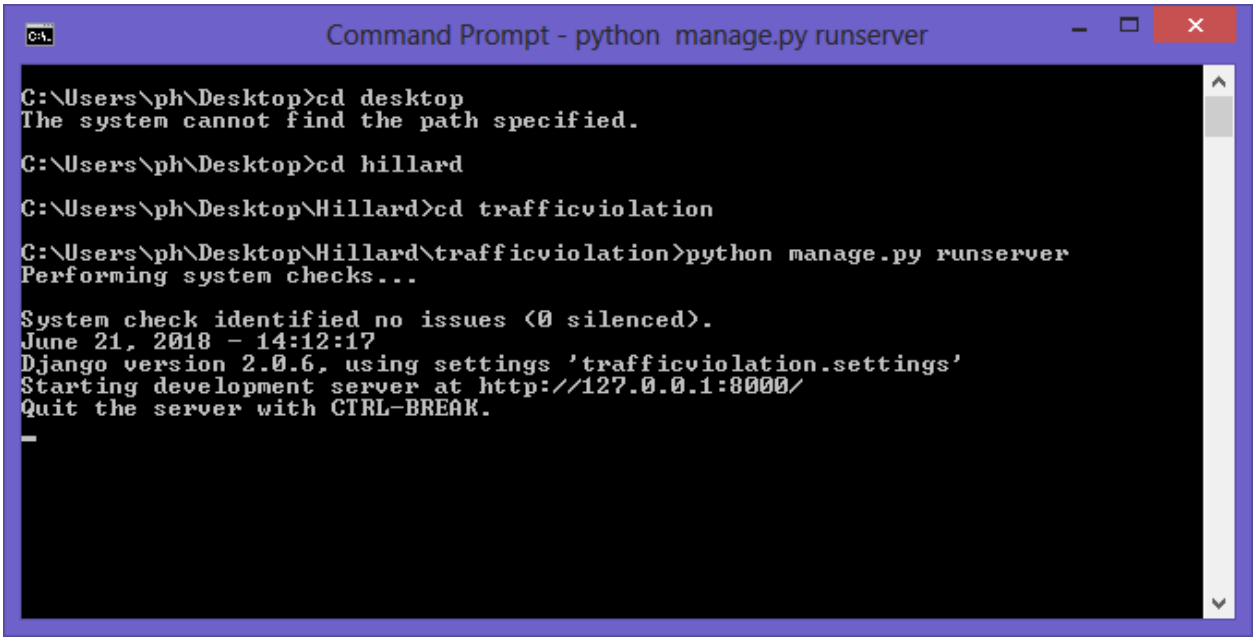

Figure 7: Realtime secured access connection prompt.

5. CONCLUSION

This paper was able to present a comprehensive review on researches previously targeted on improving classification and information retrieval systems. It also introduced a real-time database, an efficient classification algorithm and adjusted similarity in existing algorithm to improve on traffic crime predictive systems. From the results and visualizations, it can deduce that the accuracy of classification using the multinomial naïve bayes classification algorithm was more effective.

Hence, further research can also be done looking into the security of the various network layers to ensure data integrity protected when accessed at all times.

6. REFERENCES

[1] Adeyinka A. F., Ndako V. A., & Faith P. A, (2013). Design and Implementation of Crime Investigation System using Biometric Approach. The Pacific Journal of Science and Technology. 14(2)

[2] Olaniyan O, Mapayi T & Ibikunle F. A. (2012). An ICT-BAsed E-collaborative Application for law enforcement Agencies in Nigeria. Computing, Information Systems, Development Informatics & Allied Research, 13-18.

[3] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, (2009). Introduction to Information Retrieval. Cambridge University Press, Cambridge England.

[4] Lee H.F. & Schaefer S.K. (1997). Sequencing methods for automated storage and retrieval systems with dedicated storage. Computer and Industrial Engineering, 32(2).

[5] Hakan Hekim, Serdar K. Gul and Bahadir K, Akcam (2013). Police use of Information Technologies in

Criminal Investigations. European Scientific Journal, 9(4).

[6] Aruleba, K.D., Akomolafe, D.T. and Afeni, B. (2016) A Full Text Retrieval System in a Digital Library Environment. Intelligent Information Management, 8, 1-8.

[7] Ian Clarke and Chris Mellish, (1999). A Distributed Decentralized Information storage and Retrieval System. Division of Information, University of Edinburgh.

[8] Enakrire T. R., Olorunfemi Y. D. and Emmanuel O. A. (2013). The use of Databases for Information Storage and Retrieval in Selected Banks in Delta State, Nigeria. International Journal or Scientific and Technology Research 2(3), 11-24.

[9] Rikard K. (2012). Predictive Techniques and Methods for Decision Support in Situations with Poor Data Quality. University of Borås School of Business and Informatics,

[10] Marlin, B. (2004). Collaborative filtering: a machine learning perspective. Thes is research. Department of computer science, University of Toronto. 1-118. laborative filtering: a machine learning perspective. Thes is research. Department of computer science, University of Toronto. 1-118.

[11] Leif Azzopard and Vishwa Vinay, (2009). Accessibility in Information Retrieval. Department of Computing Science, University of Glasglow, Glasglow UK.