I. J. Computer Network and Information Security, **2022, 1, 81-90**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2022.01.07

Ensem_SLDR: Classification of Cybercrime using Ensemble Learning Technique

Hemakshi Pandey[1]
[1]Department of Computer Science Engineering, Bhagwan Parshuram Institute of Technology, New Delhi-110089, India
E-mail: mahe173fas@gmail.com

Riya Goyal[2], Deepali Virmani[3] and Charu Gupta[4]
[2, 3, 4] Department of Computer Science Engineering, Bhagwan Parshuram Institute of Technology, New Delhi-110089, India
E-mail: riya.goyal2599@gmail.com, deepalivirmani@gmail.com, charugupta@bpitindia.com

Received: 08 June 2021; Revised: 21 August 2021; Accepted: 14 October 2021; Published: 08 February 2022

Abstract: With the advancement of technology, cybercrimes are surging at an alarming rate as miscreants pour into the world's modern reliance on the virtual platform. Due to the accumulation of an enormous quantity of cybercrime data, there is huge potential to analyze and segregate the data with the help of Machine Learning. The focus of this research is to construct a model, Ensem_SLDR which can predict the relevant sections of IT Act 2000 from the compliant text/subjects with the aid of Natural Language Processing, Machine Learning, and Ensemble Learning methods. The objective of this paper is to implement a robust technique to categorize cybercrime into two sections, 66 and 67 of IT Act 2000 with high precision using ensemble learning technique. In the proposed methodology, Bag of Words approach is applied for performing feature engineering where these features are given as input to the hybrid model Ensem_SLDR. The proposed model is implemented with the help of model stacking, comprising Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest and gave better performance by having 96.55 % accuracy, which is higher and reliable than the past models implemented using a single learning algorithm and some of the existing hybrid models. Ensemble learning techniques enhance model performance and robustness. This research is beneficial for cyber-crime cells in India, which have a repository of detailed information on cybercrime including complaints and investigations. Hence, there is a need for model and automation systems empowered by artificial intelligence technologies for the analysis of cybercrime and their classification of its sections.

Index Terms: Cybercrime, Bag of Words, Ensemble Learning, Machine Learning, Natural Language Processing.

1. Introduction

With the dynamic technological development, the dependency on cyberspace has increased [28]. Concepts and terminologies which seldom existed years ago have now been infused into our day-to-day life, as cyber-crime, computer-related crime, information crime, or internet crime. The crime which occurs with the aid of a computer, the internet, or any device is known as cyber-crime. Today, people all over the world are connected through social media networks which are vulnerable to cyber terrorism.

Due to the accumulation of a colossal amount of cybercrime data which may include complaint text or investigation description, there is huge potential to analyze the data with the help of Artificial Intelligence (AI), Machine Learning (ML), Ensemble Learning [29], and Natural Language Processing (NLP) [24, 25, 26, 27, 30]. With the culmination of extensive research in the field of NLP, there are multifarious applications in law enforcement like text summarization, relationship extraction, prediction of crime, criminal intelligence gathering, etc. [1].

In India, cybercrime is undertaken by cybercrime cells and the criminal acts committed may involve cyber terrorism, hacking, online stalking, online fraud, identity theft, or sending of offensive messages or circulation of obscene or toxic material. Under IT Act 2000, there are 94 sections originally however, two sections deal with the majority of these punishable offenses: section 66 involves computer-related offenses while section 67 involves punishment for transmitting obscene or toxic content [2]. The objective of this research is to implement a robust technique to categorize cybercrime into two sections, 66 and 67 of IT Act 2000 with high precision using ensemble learning techniques on the collected and processed data. To develop such classification frameworks, features may be extracted which is crucial in the identification of characteristics as defined in sections of various punishable offenses. For instance, the description containing the words like ‘fraud’, ‘terrorism’ will be classified under section 66 while the words like ‘child’ or ‘obscene’ will be classified under section 67.

In past, most of the research work deals with the classification of cybercrime offenses on the textual data which involves social media [3, 8]. However, there is a need for the extension of such methods to the law enforcement domain. At present, while registering a complaint or investigation, investigation officers may correlate the torts with the various penal codes and choose appropriate sections. This necessitates officers to have prior deep knowledge and a clear understanding of the criminal law definitions. Text classification can be very beneficial for cyber-crime cells for predicting and tagging the relevant sections to complaint text, description, and investigation reports.

Ensem_SLDR: The proposed dual-level framework is a combination of four algorithms which include SVM, Logistic Regression, Decision Tree, and Random Forest and the model is implemented using model stacking, one of the categories of ensemble learning techniques. Hence, we call the resulting framework as Ensem_SLDR.

In the state of the art methods, only one or two classification or learning algorithms have been used. Although the accuracy of these methods is comparable to Ensem_SLDR, they suffer from substantial limitations of low accuracy [4, 8] and unreliability [5, 6]. The existing methods utilized single supervised learning algorithms like the Naive Bayes approach and Support Vector Machine algorithm on different cybercrime datasets, but more investigation is required using robust techniques like ensemble learning methods. Therefore, the ensemble learning approach solves the problem of low accuracy by improving model performance and has high reliability. One of the challenges concerning the classification of cybercrime in India is the availability of datasets related to complaints description and it is difficult to access such datasets due to their confidential nature. Hence, for our research, we have gathered the data from news articles for relevant sections of the IT Act 2000.

Approach: In this paper, we have proposed a hybrid model, Ensem_SLDR, for the classification of cybercrime offenses. The data has been collected with applicability to sections 66 and 67 of the IT Act 2000 for India through various news reports which involve the single line description of suitable incidences and the data had been processed by labeling them into these sections. A file was created with two attributes: definition and label. A total of 288 records were processed for our experiment and study. For text classification, we have proposed Ensem_SLDR which is implemented using NLP and ensemble learning. This hybrid framework is being developed using the technique of model stacking [7] with the combination of 4 ML algorithms and the model is trained with the relevant sections of 66 and 67 of the IT Act 2000. This framework helps in automating and categorizing cybercrime offenses and the adoption of such a method could also revolutionize many domains of law enforcement.

Outline*:* With the discussion of ML and NLP in the cybercrime domain in section 1, the literature review of extensive research work performed for text classification, ML, and ensemble learning in the area of cybercrime by previous researchers is provided in Section 2. The brief insight on the proposed work including NLP and ensemble learning techniques used for our research and study are depicted in Section 3. It also gives insights into the flow chart of the proposed model and involves three subsections: data collection, data preprocessing, and implementation. The outcome of the research on the performance of the model with other experimented models and their brief comparative analysis has been provided in Section 4. In the end, the deduced conclusion from our experiments with the future scope is presented in Section 5.

2. Related Work

This section discusses the existing methodologies implemented in the domain of cybercrime classification and text classification.

Kumari et al. [8] proposed a model which was trained using two datasets i.e., the online available dataset and the pure cybercrime data from Facebook and Twitter. In their model, they made a comparison between these two datasets to infer which cybercrime data gave better classification accuracy than the online datasets. To achieve this, they used Naive Bayes as a classifier and performed sentiment analysis with the help of NLTK. The sentiment analysis of the data extracted from social media can help mitigate cyber-terrorism with the method of text classification. The approach used in their research has limitations of low accuracy. Children and teenagers are also susceptible to cyber threats which may include cyberbullying [9], obscenity, or pornography. Studies and experiments have also been carried out to classify the online chat logs for such cyber threats using machine learning algorithms [3].

Sudha & Rupa [5] presented a sequential generalized model using the different machine learning algorithms such as K-means clustering algorithm, Naive Bayes classification, and prediction analysis. They transformed the collected data into structured data by using the TFID method. They applied the clustering algorithm to collect the identical kind of data in a single cluster. They employed a classification algorithm to check the classification accuracy and then applied the prediction analysis to take preventive measures against cybercrime or to reduce cybercrime. They carried out a study to categorize cybercrime offenses based on the characteristics like incident, location, year, and harm using text mining algorithms. Such frameworks are efficacious in analyzing and predicting cybercrime on parameters like location and year. This study has few limitations. It required comparative analysis of other classification algorithms to compare accuracy. Another research was conducted on a similar model where linear SVM improved the model performance [6] but this model is implemented only with categorization and clustering of patterns of cybercrime and hence it is restricted to this application only.

Cardoza & Wagh [10] collected cybercrime data from various news articles to study the text analysis framework. They used natural language processing (NLP), linguistic preprocessing, and Parts of Speech (POS) tagging to extract the information which was associated with cybercrime. In addition to this, the data mining algorithms were utilized to get the descriptive details of cybercrime data and perform the analysis.

Lekha & Prakasam [11] introduced a novel architecture using various data mining algorithms: the K-Means algorithm, J48 Prediction tree, and Influenced Association Classifier to predict the cybercrime data within the banking sectors and resolve the available harms. These techniques altogether improved and intensified the prediction precision. However, the efficacy of this model may not be feasible in a different scenario.

Fauzi & Yuniarti [12] developed a model to analyze whether a tweet was a hate speech or not by the application of ensemble ML methodology. Five diverse independent classification algorithms were employed to implement soft voting and hard voting: Naive Bayes, K-Nearest Neighbours, Random Forest, Maximum Entropy, and Support Vector Machine. The outcome of the ensemble method showed an improvement in classification accuracy compared to constituent algorithms. However, the model could be improved with the inclusion of certain features.

Ubing et al. [13] worked on the improvement of the classification accuracy for the detection of a phishing website. With the aid of an employed feature selection algorithm and combining the framework with the ensemble learning technique, it showed a considerable increase in forecasting precision.

Ingole et al. [14] researched the tweets of Engineering Students to understand their problems and complications in their educational experiences. A hybrid approach and sequential architecture were presented, consisting of Naive Bayes and Support Vector Machine, and showed improvement in the precision of the categorization. However, the proposed methodology could be improved with the implementation of better NLP techniques before the input data is given to the hybrid model for classification and the proposed methodology is limited to Twitter data.

Kanakaraj & Guddeti [15] gathered data from Twitter's social networking platforms and incorporated NLP techniques to extract features from tweets. To improve the precision of forecasting, Word Sense Disambiguation was used along with diverse Ensemble classification methods. The entire methodology was found to outperform conventional classification techniques. Extremely randomized trees classification performed better than those of the ensemble approaches but this methodology is limited to sentiment analysis of Twitter posts.

3. Proposed Methodology

This section presents the proposed work with the explanation of different techniques and has been briefed by a block diagram shown in Fig. 1. The initial process for data pre-processing is covered in sub-section 3.1 and 3.2, followed by feature engineering and ensemble learning methods in subsections 3.3 and 3.4 respectively. Under ensemble learning methods, section 3.4.1 and section 3.4.2 give a brief explanation of model stacking and voting ensemble.

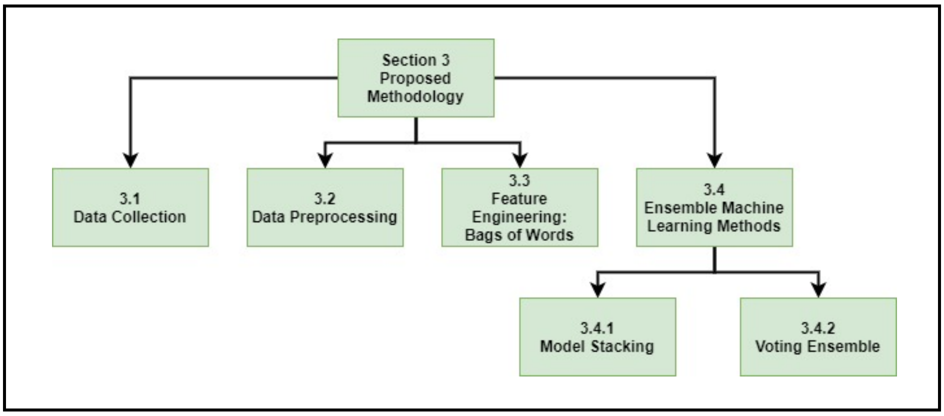

Fig.1. Detailed Structure of Section 3

The proposed methodology is represented with the help of a flow chart in Fig. 2. It shows the brief framework carried out to implement the model. The process initiates with the accumulation of suitably organized data (collected from news articles related to cybercrime), followed by data preprocessing which includes conversion of all the text to lowercase, sentence tokenization, removal of stop words (unnecessary words), and lemmatization. After the text cleaning process, feature engineering is applied to convert the text into a meaningful form that the classification algorithm could comprehend. The feature engineering was implemented with one of the elementary models of NLP, i.e. Bag of Words model. Then, sample data is divided into a suitable ratio, into a train and test set. The sample data inside the training set is given as input to all distinct models and is tuned for the intended accuracy to make the predictions and the performance was assessed using the test set. The proposed methodology is implemented using a hybrid approach comprising four classification algorithms with the technique of model stacking. In the stacking technique, the first level contains base-level classifiers and their output is supplied to the second level, which contains a meta-classifier. The final prediction is generated by the meta-classifier. In the final step, the model performance was evaluated on various metrics like F1-score, recall, accuracy, and precision. In this research, the performance metrics of all the ensemble learning models are compared to find the best model and all the programming for text analyzer and classifier is carried out with python programming language. The proposed model Ensem_SLDR is a robust and effective approach for the classification of cybercrime with the implementation of model stacking than the existing hybrid models and individual classifiers [8].

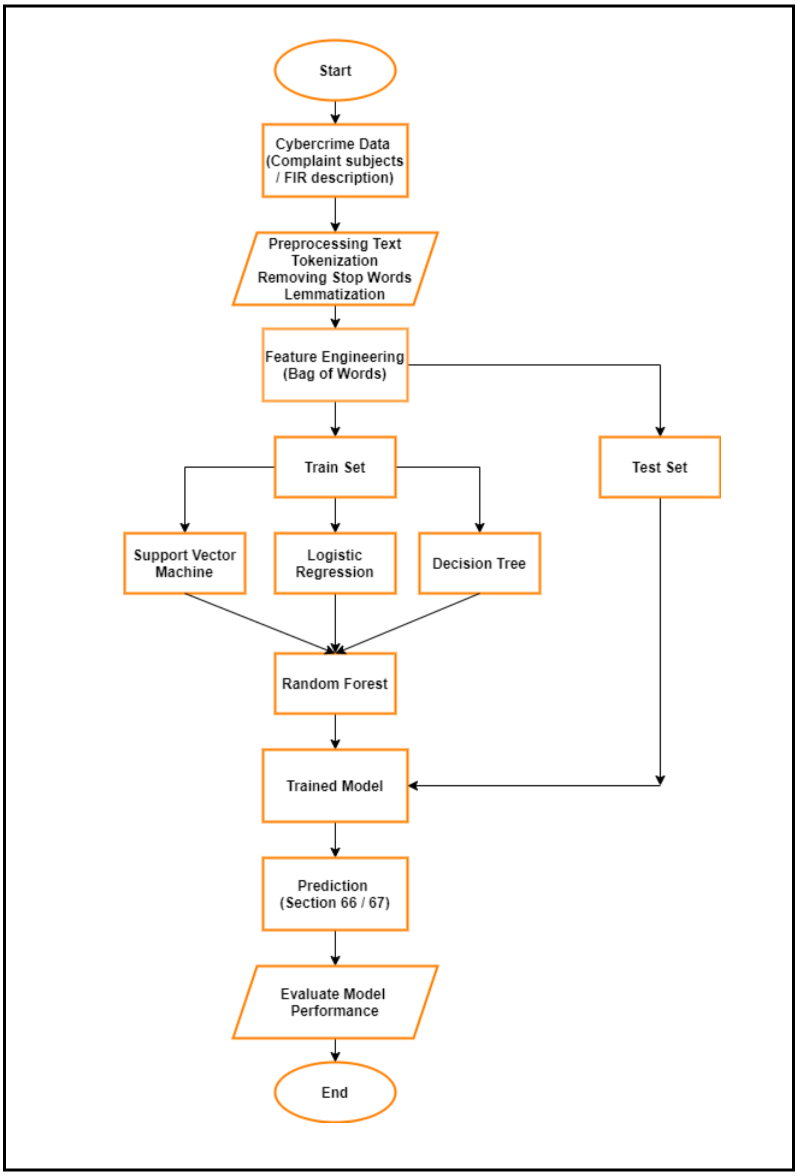

Fig.2. Flowchart of Proposed Hybrid Model (Model Stacking)

3.1. Data Collection

Since the actual complaint reports for the cybercrime are confidential in real-world scenarios, for the experiment and research, we have collected data through various news reports of Indian journals which involve the single line description of suitable incidences and we have organized and processed by labeling them into two sections, i.e., 66 and 67 of IT Act 2000. The dataset contains two attributes which are definition and label and two groups or classes; 66 and 67. There are 167 instances under the class of section 66 and 121 instances under the class of section 67. Therefore a total of 288 records in the dataset. Fig. 4. shows the view of sample data and Fig. 3. shows the distribution of records for two groups.

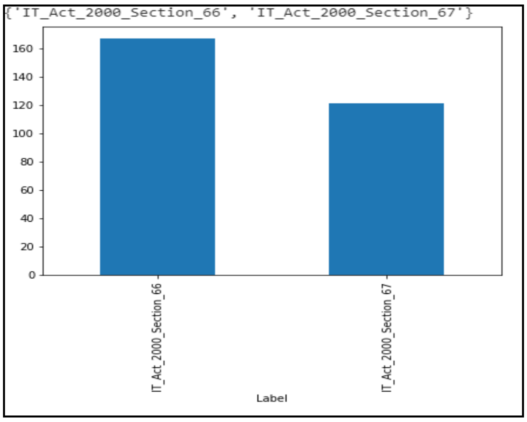

Fig.3. Distribution of Dataset

Fig.4. Cybercrime Dataset (Section 66 and 67)

3.2. Data preprocessing

Data cleaning is essential to highlight the attributes and filter them as the unstructured text is never directly given as the input to the machine learning algorithm. For this experiment, the process has been carried out using the NLTK library. The text is tokenized and transformed to lowercase to maintain uniformity. To extract distinguishable and unique features from the attribute, stop words (common words or useless words) are removed. Lemmatization is also performed on words to minimize inflectional endings and to compose the source form as in the dictionary. All these steps were constructive in polishing and normalization [16] for the text for the further process of feature extraction, training of the model and classification.

3.3. Feature Engineering: Bag of Words

The features were extracted from the preprocessed data using one of the chief models of natural language processing, known as Bag of words. This model is based on the computation of occurrences of words within any text data and it is also called the Vector Space Model. It is instrumental in reproducing text data as input feature vector into the machine learning algorithm by vectorizing and transforming the text into a matrix of numeric for which holds the count of words for each record for each of the unique features or words. The framework of this model is primitive and easy to comprehend and is suitable only for small-scale domain-specific and simple NLP applications. It also suffers from limitations like the dilemma of circumstantial perception, the issue of large dimensionality of feature space [17].

3.4. Ensemble machine learning methods

To enhance the predictive effectiveness, ensemble learning methods function as a catalyst to optimize the model. Although major factors of error in training or learning of model are bias and variance, this technique leverages them and decreases the fluctuation of the model, and transforms the single weak classifiers into a strong learner called meta-classifier [18]. Thus, the reliability and stability of the model expand with high computation cost. Some of the ensemble machine learning techniques include voting, bagging, boosting, and stacking [19].

In this research, we have implemented a hybrid model *Ensem_SLDR* using model stacking and have obtained the highest precision. In addition to this, we have experimented with other models such as voting ensemble, Adaboost, XGBoost, Gradient Boosting techniques. The technique AdaBoost is iterative in operation. It constructs a strong classifier by merging multiplex weakly performing classifiers to increase the accuracy. XGBoost (Extreme Gradient Boosting) is a library and an algorithm that employs gradient boosted decision trees devised for fast computation and effective performance. Gradient boosting is a cluster of algorithms usually decision trees which associate many weak learners together to produce a strong predictive model.

A. Model Stacking

Model Stacking or blending is a popular mechanism that is implemented by associating and unifying diverse base classification algorithms followed by the meta-level classifier [20, 31, 32]. Stacking is the combination of distinct classification learning algorithms {Q1…Qn} on an individual sample dataset. The fundamental stage consists of a set of key or base-level classifiers {A1, A2…An}. The algorithm of model stacking [21] is presented in [23, Fig. 5.]. In the second stage, a meta-level classifier is generated by merging the key classifiers.

In this research, we have proposed the hybrid model *Ensem_SLDR* by combining four algorithms through stacking. In the beginning stage, three algorithms are used: Support Vector Machine (SVM), Logistic Regression, and Decision tree. The output of these algorithms is supplied as the input to the second stage which involves a meta-classifier; Random Forest. The output of the second stage gives the final prediction outcome. This dual learning architecture provides the best performance for the classification of cybercrime sections.

Input: Data set $D = \{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$;
First-level learning algorithm $L_1,\dots,L_T$;
Second level learning algorithm L;

Process:
For $t = 1,2,\dots,T$:
$h_t=L_t(D)$ % Train a first level individual learner h_t by applying the % first-level learning a algorithm L_t to the original data set D
end;
$D^{`} = \Phi$; % Generate a new data set
For $t = 1,2,\dots,m$:
For $t = 1,2,\dots,T$:
$z_{it} = h_t(x_i)$ % Use h_t to classify the training example x_i
end;
$D^{`} = D^{`} \cup \{((z_{i1},z_{i2},\dots,z_{iT}), y_i)\}$;
end ;
$h^{`} = L(D^{`})$ % Train the second level learner $h^{`}$ by applying the second-level % learning algorithm $L$ to the new data set $D^{`}$

Output: $H(x) = h^{`}(h_1(x),h_2(x),\dots,h_T(x))$

Fig.5. Steps Involved in Stacking Algorithm

B. Voting Ensemble

The voting ensemble is based on the principle of a collective decision of base learners through majority voting. This scheme includes different classifiers to be combined and the class or group allocated to a test sample will be the one implied by the majority of constituent classifiers. In the case of a 2-class problem, the groups which can be given as output by model: $h_t(x) \in \{+1,-1\}$. The weights w for a simple majority vote, for a weighted vote, can be uniform or non-uniform respectively. Hence the final output of the voting combiner is given by the [22, equation (1)]: -

$$H(x) = sign\left(\sum_{t=1}^{T} w_t h_t(x)\right) \tag{1}$$

Unlike stacking, the meta-level in voting is devoid of the learning algorithm. To perform this experiment, we used this technique using 3 machine learning classification algorithms: Logistic Regression, Support Vector Machine (SVM), and Decision tree. This architecture also improved performance and precision for classification.

4. Results and Discussion

In this research paper, we investigate five ensemble learning techniques for the classification of cybercrime sections 66 and 67. The performance metrics applied for evaluating our model are Recall, Precision, F1 Score, and Accuracy. Sensitivity or Recall can be defined as the ratio of precisely forecasted positive observations to the sum of all observations in the actual class. The ratio of precisely forecasted positive observations to the sum of all positive observations is called precision. F1 Score is the weighted mean of sensitivity and precision and the ratio of precisely forecasted observation to the sum of all observations is known as accuracy. The experimental result shows the proposed hybrid model Ensem_SLDR which was implemented through the stacking approach shows the best performance with the prediction accuracy of 96.55%, precision 100%, recall 92% and F1 score 96% which is higher than the model implemented using single learning algorithms and other existing models [4, 8]. This model could be effectively employed for the classification of cybercrime data and subjects in cybercrime cells. The accuracy of the voting strategy is 94.83%, which is also good compared to other techniques as shown in Table 1. Gradient Boosting and Adaboost have also been beneficial in improving the accuracy of the classification of the text-based cybercrime data with an accuracy of 93.10%, however, not better than models implemented through stacking and voting. However, the least accurate model is XGBoost with an accuracy of 87.93%. Fig. 6. represents a graph that shows the pictorial representation of the performance of all the experimented models.

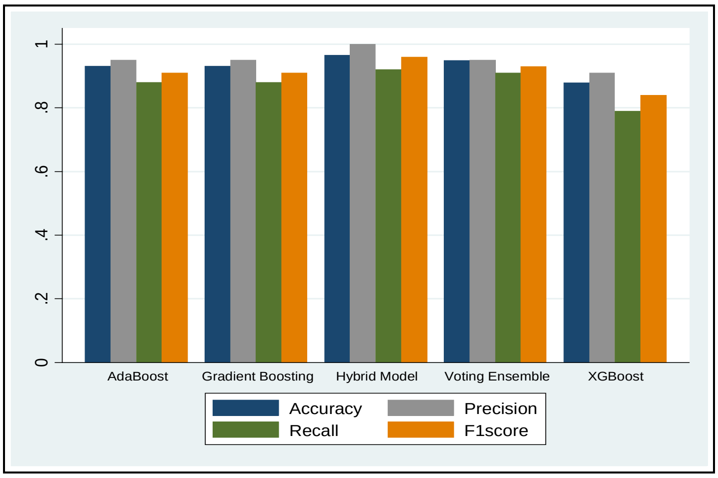

Fig.6. Comparative Analysis of Ensemble Learning Algorithms

Table 1. Performance of ensemble machine learning algorithms

Algorithm	Accuracy (%)	Precision	Recall	F1 score
Hybrid Model	96.55	1.00	0.92	0.96
Voting Ensemble	94.83	0.95	0.91	0.93
XGBoost	87.93	0.91	0.79	0.84
Gradient Boosting	93.10	0.95	0.88	0.91
AdaBoost	93.10	0.95	0.88	0.91

5. Conclusion and Future Work

The paper analyses the results of the proposed model using machine learning and NLP technologies which is beneficial in tackling the text-based complaints about cybercrime and the corresponding sections regarding the law (section 66 and 67 of IT Act 2000). The unseen text was classified and forecasted according to unique characteristics or features of the data. We collected and processed the data for the research from the news articles and extracted the features using the Bag of Words model. We have applied various ensemble learning techniques for the classification of cybercrime data. The outcome shows that the proposed hybrid model Ensem_SLDR implemented through model stacking technique, gives excellent performance by having an accuracy of 96.55%, precision 100%, recall 92%, and F1 score 96% which is a better and robust approach than some existing models [8]. Further, it is concluded that the ensemble of individual machine learning algorithms and model stacking helps in increasing the accuracy of the model and is effective than an individual algorithm. The idea of the proposed model could be applied for the classification of different sections of the IT Act 2000 depending upon the size and nature of the available data in data repositories of cybercrime cells. In addition to this, the proposed model and its approach could be applied to different domains apart from cybercrime taking into account the type of scenario and problem statement.

Our future work would be focused on investigating the proposed work more deeply and using a combination of other better features engineering models with more data samples and deep learning techniques. However, other approaches like NER (Named Entity Recognition) may be beneficial in the extraction of informative data which can be applied to IPC (Indian Penal Code) offenses as well.

References

[1] van Banerveld M, Kechadi M-T, Le-Khac N-A (2016) A Natural Language Processing Tool for White Collar Crime Investigation. In: Hameurlain A, Küng J, Wagner R, et al (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIII. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–22

[2] Deshpande DrAdvMrsN, V. P. Institute of Management Studies and Research, Sangli, Affiliated to Shivaji University, Kolhapur, Maharashtra, India (2018) A Brief Study on Cyber Crimes and IT Act in India. Int J Trend Sci Res Dev Special Issue:141–149. https://doi.org/10.31142/ijtsrd18693

[3] Ngejane CH, Mabuza-Hocquet G, Eloff JHP, Lefophane S (2018) Mitigating Online Sexual Grooming Cybercrime on Social Media Using Machine Learning: A Desktop Survey. In: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). IEEE, Durban, South Africa, pp 1–6

[4] Haidar B, Chamoun M, Serhrouchni A (2017) A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. Adv Sci Technol Eng Syst J 2:275–284. https://doi.org/10.25046/aj020634

[5] Sudha TS, Rupa C (2019) Analysis and Evaluation of Integrated Cyber Crime Offences. In: 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). pp 1–6

[6] Ch R, Gadekallu TR, Abidi MH, Al-Ahmari A (2020) Computational System to Classify Cyber Crime Offenses using Machine Learning. Sustainability 12:4087. https://doi.org/10.3390/su12104087

[7] Džeroski S, Ženko B (2004) Is Combining Classifiers with Stacking Better than Selecting the Best One? Mach Learn 54:255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

[8] Kumari S, Saquib Z, Pawar S (2018) Machine Learning Approach for Text Classification in Cybercrime. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, Pune, India, pp 1–6

[9] Andleeb S, Ahmed R, Ahmed Z, Kanwal M (2019) Identification and Classification of Cybercrimes using Text Mining Technique. In: 2019 International Conference on Frontiers of Information Technology (FIT). IEEE, Islamabad, Pakistan, pp 227–2275

[10] Department of Computer Science, Christ University, Bengaluru-560029, India, Cardoza C, Wagh R, Department of Computer Science, Christ University, Bengaluru-560029, India (2017) Text analysis framework for understanding cyber-crimes. Int J Adv Appl Sci 4:58–63. https://doi.org/10.21833/ijaas.2017.010.010

[11] Lekha KC, Prakasam S (2017) Data mining techniques in detecting and predicting cyber crimes in banking sector. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). IEEE, Chennai, pp 1639–1643

[12] Fauzi MA, Yuniarti A (2018) Ensemble Method for Indonesian Twitter Hate Speech Detection. Indones J Electr Eng Comput Sci 11:294. https://doi.org/10.11591/ijeecs.v11.i1.pp294-299

[13] Ubing AA, Kamilia S, Abdullah A, et al (2019) Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning. Int J Adv Comput Sci Appl 10:. https://doi.org/10.14569/IJACSA.2019.0100133

[14] Ingole P, Bhoir S, Vidhate AV (2018) Hybrid Model for Text Classification. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, Coimbatore, pp 450–458

[15] Kanakaraj M, Guddeti RMR (2015) Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques. 2

[16] Han P, Shen S, Wang D, Liu Y The Influence of Word Normalization in English Document Clustering. 5

[17] Wang F, Wang Z, Li Z, Wen J-R (2014) Concept-based Short Text Classification and Ranking. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM ’14. ACM Press, Shanghai, China, pp 1069–1078

[18] Bian W, Wang C, Ye Z, Yan L (2019) Emotional Text Analysis Based on Ensemble Learning of Three Different Classification Algorithms. In: 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). IEEE, Metz, France, pp 938–941

[19] Khanday AMUD, Rabani ST, Khan QR, et al (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12:731–739. https://doi.org/10.1007/s41870-020-00495-9

[20] Li Y, Chen W (2020) A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. Mathematics 8:1756. https://doi.org/10.3390/math8101756

[21] Wang G, Hao J, Ma J, Jiang H (2011) A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl 38:223–230. https://doi.org/10.1016/j.eswa.2010.06.048

[22] Brown G (2010) Ensemble Learning. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer US, Boston, MA, pp 312–320

[23] O. O. Olasehinde, O. V. Johnson and O. C. Olayemi, "Evaluation Of Selected Meta Learning Algorithms For The Prediction Improvement Of Network Intrusion Detection System," 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), Ayobo, Nigeria, 2020, pp. 1-7, doi: 10.1109/ICMCECS47690.2020.240893.

[24] Amit Pandey, Achin Jain,"Comparative Analysis of KNN Algorithm using Various Normalization Techniques", International Journal of Computer Network and Information Security, Vol.9, No.11, pp.36-42, 2017.

[25] Shubham Bauskar, Vijay Badole, Prajal Jain, Meenu Chawla, " Natural Language Processing based Hybrid Model for Detecting Fake News Using Content-Based Features and Social Features", International Journal of Information Engineering and Electronic Business, Vol.11, No.4, pp. 1-10, 2019.
[26] Volodymyr Tolubko, Viktor Vyshnivskyi, Vadym Mukhin, Halyna Haidur, Nadiia Dovzhenko, Oleh Ilin, Volodymyr Vasylenko, "Method for Determination of Cyber Threats Based on Machine Learning for Real-Time Information System", International Journal of Intelligent Systems and Applications, Vol.10, No.8, pp.11-18, 2018.
[27] Semih Sevim, Sevinç İlhan Omurca, Ekin Ekinci, "An Ensemble Model using a BabelNet Enriched Document Space for Twitter Sentiment Classification", International Journal of Information Technology and Computer Science, Vol.10, No.1, pp.24-31, 2018.
[28] Raghad Khweiled, Mahmoud Jazzar, Derar Eleyan, "Cybercrimes during COVID -19 Pandemic", International Journal of Information Engineering and Electronic Business, Vol.13, No.2, pp. 1-10, 2021.
[29] Dimple Tiwari, Nanhay Singh, "Ensemble Approach for Twitter Sentiment Analysis", International Journal of Information Technology and Computer Science, Vol.11, No.8, pp.20-26, 2019.
[30] Mohammad Mojaveriyan, Hossein Ebrahimpour-komleh, Seyed jalaleddin Mousavirad,"IGICA: A Hybrid Feature Selection Approach in Text Categorization", International Journal of Intelligent Systems and Applications, Vol.8, No.3, pp.42-47, 2016.
[31] Bodunde Akinyemi, Oluwakemi Adewusi, Adedoyin Oyebade, "An Improved Classification Model for Fake News Detection in Social Media", International Journal of Information Technology and Computer Science, Vol.12, No.1, pp.34-43, 2020.
[32] Yasin Görmez, Yunus E. Işık, Mustafa Temiz, Zafer Aydın, "FBSEM: A Novel Feature-Based Stacked Ensemble Method for Sentiment Analysis' Comments in E-Government", International Journal of Information Technology and Computer Science, Vol.12, No.6, pp.11-22, 2020.

Authors' Profiles

Hemakshi Pandey is a student at Bhagwan Parshuram Institute of Technology, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi) pursuing the Bachelor of Technology in Computer Science Engineering. Her area of interest includes Data Structures, Machine Learning and Natural Language Processing.

Riya Goyal is a student at Bhagwan Parshuram Institute of Technology, New Delhi (Affiliated to Guru Gobind Singh Indraprastha University, Delhi) pursuing the Bachelor of Technology in Computer Science Engineering. Her area of interest is Artificial Intelligence.

Dr. Deepali Virmani is Head of the Department and Professor in the Department of Computer Science and Engineering at Bhagwan Parshuram Institute of Technology affiliated to GGSIPU, New Delhi. Dr. Deepali Virmani has received her B.Tech. degree in Computer Science from MDU, Rohtak, M.Tech. degree in Information Technology, from GGSIPU, and the Ph.D. degree in Computer Science from Delhi University,India. She has an innovative work experience of more than 19 years in both research and academics. She has published more than 90 research papers in international journals/National journals/International conferences of repute. She works in a multi-disciplinary environment involving sensor networks, web intelligence, data mining and intelligent information retrieval systems and machine learning applied to various real-world problems. Dr. Virmani has more than 500+ academic citations index as per Google Scholar. She has guided more than 70 B.Tech Projects. Presently, she is guiding many Ph.D. scholars registered with reputed universities like GGSIPU and UPTU. She is branch counselor of BPIT-IEEE student chapter and BPIT- CSI student branch. She is the remote center coordinator for IIT Bombay, Spoken Tutorial IIT Bombay, IIT Kharagpur, and Virtual Labs IIT Delhi. She is the associate editor of journal Open Computer Science De Grutyer and also on the reviewer board of various IEEE transactions, Elsevier and Springer journals. She has organized many professional activities like FDPs, workshops, expert lectures and conferences. She has been the session chair in National/International conferences. Her papers have won Best Paper Award at various International Conferences. She has won Best Researcher Award and Best Faculty Award at BPIT. She has been awarded Excellence in Research Award by International Research Awards on New Science Inventions.

Dr. Charu Gupta graduated B.E. in Computer Science and Engineering and completed M.Tech from JSS Academy of Technical Education, Noida in Computer Science and Engineering with Honors. She received her Doctoral degree from the Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. Presently serving as assistant professor at Bhagwan Parshuram Institute of Technology (Affiliated to GGSIPU, Dwarka), Delhi with a teaching experience of 10 years. She has to her credit research publications in various National and International Journals/Conferences of repute (SCIE/ESCI/SCOPUS). She is a reviewer, section editor and editorial member of many international journals/conferences. She is also a member of Encyclopedia of Neutrosophic Researchers, Vol. 3, Book edited by Prof. Florentin Smarandache, University of New Mexico. She is a subject matter expert in VIDWAN, an MHRD project on Expert Database and National Researcher's Network. Vidwan-ID: 136697. She is a Research Associate with Nokia in "Invent with Nokia" venture. She has been the co-convenor of various webinar series and faculty coordinator of National and International Conferences at BPIT. She is the Lead Editor and Co-Editor in various book series by Wiley, De Gruyter, Nova publishing and many more. She has to her credit various sessions in National and International Conferences. She is a faculty coordinator (Delhi Section) of Free and Open Source Cell (FOSS cell) from International Centre for Free and Open Source Software (ICFOSS), Govt. of Kerala, India. She is also the faculty co-ordinator of e-Yantra Lab setup initiative (eLSI) in collaboration with IIT Bombay & Anusandhan Research Lab in Department of Computer Science and Engineering, BPIT. She has also served as a Coordinator for Software Engineering for GGS Indraprastha University (IPU) B.Tech syllabus designing committee. Her areas of interest are Natural Language Processing, Neutrosophic logic and its applications, Time Series Analysis and Forecasting, Evolutionary Computation.

How to cite this paper: Hemakshi Pandey, Riya Goyal, Deepali Virmani, Charu Gupta, "Ensem_SLDR: Classification of Cybercrime using Ensemble Learning Technique", International Journal of Computer Network and Information Security(IJCNIS), Vol.14, No.1, pp.81-90, 2022. DOI: 10.5815/ijcnis.2022.01.07