IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 16, 2021, accepted August 2, 2021, date of publication August 9, 2021, date of current version August 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3103319

An Unsupervised Model for Identifying and Characterizing Dark Web Forums

SAIBA NAZAH[1], SHAMSUL HUDA[1], JEMAL H. ABAWAJY[1], AND MOHAMMAD MEHEDI HASSAN[2], (Senior Member, IEEE)

[1]School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia
[2]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Mohammad Mehedi Hassan (mmhassan@ksu.edu.sa)

This work was supported by the King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project, under Grant RSP-2021/18.

ABSTRACT Dark Web forums are significantly exploited to trade confidential information and illicit products by criminals. This paper addresses the problem of how to identify the cluster of discussion forums and their characteristics on the Dark Web. Exiting methods are mostly dependent on the continuous labeled contents, which are expensive and not feasible due to the nature of Dark Web data. Therefore, an approach that does not need a continuous availability of labeled forum and related knowledge is required. To this end, we propose an unsupervised model to identify and characterize Dark Web forums by combining clustering algorithm and decision tree algorithm. The proposed method presents the characteristics in an explainable form that can be used by the cyber threat intelligence system and law enforcement as scientific evidence to analyze any data breach or illicit activities in the Dark Web forums. To evaluate the performance of our model comprehensive experiments were conducted using real Dark Web forum data. The proposed approach achieves 98% accuracy and F1 score of 98% validating the efficacy of our proposed model to successfully characterize Dark Web forums. The experimental results suggest that the proposed model could be useful to the cyber threat intelligence and law enforcement community for building an intelligent source of knowledge that can be used for detecting data breach and illicit activities happening in the Dark Web forums.

INDEX TERMS Dark web, cyber security, data breach, cluster characteristics, decision rule.

I. INTRODUCTION

The World Wide Web is consisted of three different layers named Surface Web, Deep Web and Dark Web. The large sections of the Web that is unindexed and hidden is known as Deep Web which the normal search engines cannot crawl [41]. Approximately 96 percentage of the total Web is estimated to be made up with Deep Web [10]. A subset of the Deep Web that allows users and website operators to remain anonymous or untraceable, and only accessible through special software for instance the onion router (TOR) network is the Dark Web or Dark Net [8], [23] . The Dark Web is mostly popular for giving platform to a wide range of crimes. Illegal goods related to pornography, illicit finances, weapons, drugs, exotic animals, terrorist communication, stolen credentials and personal information are all traded on the Dark Web [3], [21]. There are much more things available on the Dark Web than the normal Web. A recent study found that 57% of the Dark Net is occupied by illegal activities [44]. However, the exact TOR traffic usage percentage at any particular time for sites serving the illicit markets in the Dark Web is uncertain [10].

Online forums have become the main place for discussing viral topics as well as illegal activities. Dark Web often serves as marketplaces for anonymous trading of illicit items or services. It also serves as a discussion forum for illicit activities [1], [8], [47]. These forums are an important platform for criminals looking to compromise or abuse personal financial details. Many forums focus on discussing financial frauds, identity thefts, data breach, stolen cards and accounts. However, the information available in the Dark Web is fragmented and unstructured. The sites also keep vanishing and disappearing. Analyzing these contents manually is almost impossible and erroneous. Moreover, all users of the Dark Web are anonymous. The intelligence data gathered from such forums can benefit the cyber threat intelligence in

The associate editor coordinating the review of this manuscript and approving it for publication was S. K. Hafizul Islam.

decision makings while investigating the threats. This can mitigate any breach and loss before attacks. Thus monitoring and analysis such type of discussion forum contents to gather the intelligence data can lead to one of the most convenient solutions to the law conformance agencies to identify financial frauds or data breach.

Despite the challenges and the rising threats, different studies have applied particular methods to monitor the Dark Web [8], [11], [48]. A study was proposed to detect threats in Dark Net data using cyber threat intelligence tool through Dark Net ecosystem network and cyber threat breach network [4]. The system has effectiveness in analyzing the networks yet the tool does not comply with the real time monitoring by crawling ongoing network as the tool is based on network analysis depending on different crawlers' specifications which is costly and time consuming. Another monitoring technique was proposed to identify underground threads those are related to data breaches [51]. They have also used different crawlers for collecting the data and the classification is based on labeled data only. A Dark Web monitoring tool has been developed [52]. They have used semantic analysis and used json files to do the classification to identify different categorizes of keywords. However, they used json files not real datasets for the analysis. In addition, sites without semantic contents cannot be analyzed by the tool. Although these studies have significant contributions in monitoring the Dark Web contents, they are mostly dependent on the continuous labeled contents. This could be time consuming and expensive to some extent. Thus, the motivation for our proposed method is identifying and characterizing the Dark Web contents without the direct dependency on labeled contents. This research attempts to understand how data in the underground forums traded on Dark Web forums can help in threat intelligence sphere.

In this paper, we propose an unsupervised model to identify and characterize the Dark Web forum. The main aim of the research is to monitor Dark Web discussion forums data to get the discussions and find out the most discussed topic or any data breach happening inside a particular page of the forum. Unsupervised clustering algorithm has been implemented to cluster the forum data. When the clusters are obtained, they are labeled with class names. Then a rule-based model using decision tree has been applied combined with unsupervised model to characterize the forum to identify a suspected forum which can lead cyber threat intelligence system and law enforcement to identify financial frauds or data breach. The cluster analysis has been validated using supervised classifications and K fold cross validation on the labeled dataset obtained from the cluster analysis. The significance of the proposed algorithm is that, the cyber threat intelligence and law enforcement can use the intelligence data for decision making or planning mitigation techniques for any data breach or illicit activities happening in the Dark Web forums. The novelties of the proposed approaches are listed in the following:

- Develop an unsupervised model using a combination of clustering and decision tree for an indicative system of data breach from Dark Web forums activities
- Develop an algorithm using a rule based and identify each forum discussions featured keywords by applying different feature selection algorithms
- Implementation of learning models for isolating significant information from noise and extracting discussion threads from huge amount of unstructured and raw contents
- Implementation of several classification algorithms to evaluate the accuracy results for the feature matrix with obtained characterized labels
- Evaluating the model performance with different performance metrics and K-fold cross validation

The remainder of this paper is organized as follows: Section II provides the literature review of the related works, Section III shows the research gaps and our research motivations, in Section IV we describe our proposed methodology and Section V and VI are covered with the results and conclusion respectively.

II. LITERATURE REVIEW

Monitoring Dark Web and Surface Web forums and retrieving data for cyber threat intelligence have been a great topic of interest in the research fields. Monitoring these forums can lead to some crucial findings that could be beneficial for the cyber threat intelligence. Different methods and analysis have been applied for mining and monitoring the Dark Web forum data.

A. DARK WEB FORUM DATA MINING

The Dark Web offers the platform for coordination, conversation and various actions through the forums. One of the most popular places to communicate about Dark Web was Reddit [10]. For the discussion of the Dark Net Markets, Deep Web, or Tor on diverse aspects Raddit was utilized by the users as a public platform. Links to different sites within the Dark Web can be found through these forums. Different strategies and methods have been developed to monitor different areas in the Deep Web by researchers. For monitoring the Internet's hidden portions various techniques have been discusses in a study [8]. Monitoring the Dark Net was proposed by researchers for identifying malicious threats and activities in the Dark Net [25], [38].

B. FOCUSED CRAWLING

Design of an application specific approach to hidden Web crawling was proposed in [29]. They have proposed a task-configurable hidden Web crawler that could automate the extraction of contents from hidden Web. They argue that human-assisted crawling of the hidden Web is feasible while accessing the contents than other techniques. This forum access technique was used in the study which developed a focused crawler in order to collect Dark Web forums [11].

They were succeeded in collecting 109 Dark Web forums from three regions in multiple languages. However, human assisted crawling could be less effective with the vastly growing forums. A Dark Web crawler was proposed along with network analysis of TOR network to identify terrorist groups [50]. A Web crawling technique was implemented to discover the latent topics from communities of extremists or terrorists in Dark Web [46].

C. ANALYSIS OF FORUM DATA CONTENTS

Analysis of specific contents of the forum can lead to targeted intelligence data. A Web mining approach was proposed for the monitoring and collection of online forums of the US extremist [48]. With the combination of expert knowledge and techniques of web mining their approach consisted of three steps including forum identification, forum collection and parsing, and forum analysis. As a result they were able to create storage of U.S. domestic extremist forums that contains 110 forums with more than 640,000 documents.

Searching and analysis of international Jihadist Dark Web forums with integrated approach was applied to develop a Web-based multilingual Dark Web Forums Portal [47]. The Dark Web Forums Portal consisted of seven Jihadist forums, where Arabic forums are six of them and one is an English forum. The system developed is functional with four options for operating. Several researches have been conducted for analyzing the usage and content of the Dark Web data by monitoring the extremist groups [1], [37], [49]. Their approaches identify targeted extremist forums from a wide range of resources; addresses practical issues in the extremist forum collection procedures faced by the researchers along with the analysis of performance in particular forums of the Dark Web.

D. ANALYSIS OF MARKET PLACE FORUMS

Dark Web marketplaces often serve as a pathway to trade illegally. A study on the online forum marketplace discussion was done for the detection of doping substances and suppliers of the online marketplaces [27]. Although the approach is good in finding doping related instance discussions the study is done on the surface web. In Dark Web the popularity measurement they have used may not be valid. Many studies analyzed the impacts of the Dark Net forums and online chat rooms serving as a marketplace for the illicit drug purchase [2], [6]. They identified that the discussion forums are encouraging the drug markets as the users from all around the world could get connected and share the knowledge on drug consumptions, purchase, legislation, drug manufacture and cultivation being totally anonymous [17], [43]. Analysis of one of the popular marketplace in Dark Web Evolution was done in a study to provide a comprehensive analysis on the illicit drug purchase process with the chemical profiling of the products [32]. They have used python scripts and orange canvas to extract specific html tags of the marketplace to extract drug listing contents. Microsoft Excel 2013 and R software have been used to analyze and visualize the information in a structured way. However, any clear research methodology was not addressed in the study rather hypothetical analysis was presented in this study.

E. TOPIC MODELING

Selecting appropriate keywords from a large volume of data is a complicated and challenging task. Latent Dirichlet Allocation (LDA) identifies latent topics from text corpus and works as a generative probabilistic model [20]. To determine the latent terms and topics in Dark Net Markets subreddit Latent Dirichlet Allocation (SLDA) unsupervised topic modeling techniques were applied [28]. Analysing the results they observed the topics are consistent all over the year of study. However, some topics emerged with the events related to real world incidents. Throughout the year crypto currency and security tools were consistent topics of conversation. PGP was also a popular topic that indicates the users demand on confidential means of communication and urge for authentication. A hybrid approach was applied to discover key members in virtual community of the Dark Web using topic models by Latent Dirichlet Allocation and social network analysis [16]. However, no specific patterns were identified as a result of their network analysis. Study and monitoring on Dark Web forum portal with the approach of combining network analysis and text mining technique based on topic modeling has been applied in to detect the overlapping communities in Dark Web portals [33]. Despite the wide applications of LDA, for complex datasets correlations analysis and dynamic topic evolution is not feasible with topic modeling.

F. MACHINE LEARNING APPROACHES TO DARKNET FORUM ANALYSIS

For the detection of data breach and financial frauds machine learning algorithms can have important implications. Both classification and clustering algorithms have been widely applied in the analysis of Dark Web data. Variations of fraud techniques were introduced with their detection and prevention techniques with different machine learning algorithms [34]. They concluded that in financial statement fraud, the probabilistic neural network performed the best with the accuracy of 98.09% following by Genetic algorithm with 95%. In credit card fraud with NSL-KDD dataset, Naives bays and Support Vector Machines (SVM) gives good results with 99.02% and 98.8% respectively. Discussion forums and marketplaces on the Dark Net have been analyzed to gather information related to hacking products and services with proposed operational system [26]. Supervised and semi-supervised machine learning algorithms have been implemented in the proposed study.

Anomaly detection in the Dark Web using unsupervised approach was proposed for identifying threats [15]. A data mining technique was developed for the anomaly detection from unlabeled data sets of the forums. As preliminary study they focused on Ansar1 data set. They implemented the unsupervised system with robust capability for anomaly detection

that does not require the user specification of normal and abnormal behavior. Fraud detection algorithm using the unsupervised learning has been implemented for detection financial fraud reporting [12]. Hidden information and incorrect information filling in the financial reports of corporate annual US Securities and Exchange Commission filings are detected using the quantitative method. SAS Enterprise Miner tool was used to develop the model. They have selected sixty-nine companies for their analysis and after preprocessing all the text documents were converted into singular value decomposition vector. Then clustering was applied with expectation maximization and hierarchical clustering algorithms. The clustering result could identify potential frauds in their dataset. The domain is limited to Management's Discussion and Analysis in 10-k fillings. Identification of user clusters in the Dark Web forums was proposed using temporal coherence analysis [45]. Identifying the clusters leads to analyze the user behavior and their interactions extremist forums. The time stamps of users messages in the Dark Web forums were used for measuring the activeness of users. The similarity matrix between forum users was used for the cluster discovery algorithm for the identification of user clusters of similar interests. Experimental Synthetic dataset and the Dark Web Ansar AIJihad Network were used as the data set for their experiment. Three clusters associated with particular theme were identified in the Dark Web forum data. However, while analysis only the user who is initiator of the discussion thread is considered. The language of the data set being Arabic, the interactions between the initiator and the users replying were not captured which could vary the cluster identification result.

III. RESEARCH GAPS AND INNOVATIONS

Corporate data breaches are a regular occurrence and growing problem. The Dark Web is utilized as a tool for cybercriminals to anonymously trade confidential information. Identifying and analyzing the contents of the Dark Web forums can play vital role for threat intelligence. Moreover, for the typical cyber security analyst using the Dark Net for gathering intelligence is less common. These factors have motivated us to study the contents of the Dark Web forums to gather intelligence data. This research attempts to characterize the Dark Web forum data to be applicable in the threat intelligence sphere. Specifically, this paper aims to examine the data on Dark Net for cyber security intelligence purposes, and analyse large amounts of unstructured data to derive emerging threat trends and cyber security risk factors.

From the literature studies we found that very little work has been done on analyzing the intelligence data from the Dark Web forums. Moreover, most of the existing monitoring approaches are dependent on labeled data and complex configurations. Based on the gaps found, the main aim of this research includes, monitoring and characterizing the Dark Web forums by analysing the cluster characteristics of the unlabeled data and find out the rules to obtain the discussions topic in a particular page of the forum which can lead cyber threat intelligence system and law enforcement to identify financial frauds or data breach. To the best of our knowledge, any model for the identification of potential data breach using the clusters analysis of the Dark Web forums contents has not yet been sufficiently explored. The proposed research outcome can answer the following questions for law enforcements:

- Can the Dark Net thread activities be indicative for potential data breach? If yes, what are the name of the organizations and the time of the data breach?
- What are the types of stolen data and how much information does the hacker have?
- Who are the victims and what immediate actions can they take to minimize the loss?

IV. METHODOLOGY

The proposed model has been presented in the Fig.1. The proposed method has several sub steps which are described as below.

First the system takes the unlabeled raw forum data. The forum data consists of several file formats of which the discussion pages are in html format. For extracting the texts from the html files HTML parser has been used. Once the texts are extracted regular expression is implemented to get each text word. After getting the words from the regular expression output filtering was applied to remove noise. The bag of word model was implemented for getting the vector of tokens. The word having a frequency greater than equal three were chosen for the further processing. After that, we created the document term matrix for our further implementation and as a pre-processing phase. Term Frequency (TF) -Inverse Document Frequency (IDF) ($TF - IDF$) has been used for

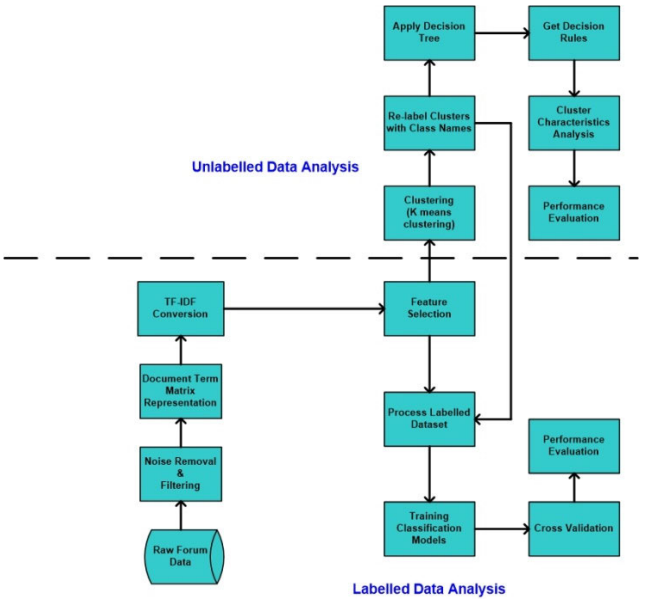

FIGURE 1. System overview of proposed model (a) Unlabeled data analysis (above dotted line) and (b) Labeled data analysis (below dotted line.

converting the text to features. For selecting the best features various feature selection algorithms have been implemented.

In the next step, we applied K-means clustering on the processed samples which clustered the data set. The dataset being complex, unknown and without explicit labels, clustering algorithm is needed to get the natural groupings among the forum data. We have applied K-means clustering algorithm into the processed dataset to get the clusters for each forum data examples. Once the cluster result is obtained, the clusters are assigned labels. These labels are class names based on the discussion forums analysis. After re labeling the clusters the decision tree was achieved with the most important features. The root node divides into child nodes and each child not iterates towards until leaf nodes are obtained. Each leaf node represents the clusters of the trees and the way the tree expands towards the end node or leaf node describes the cluster characteristics. Based on each cluster rules the characteristics are described.

For the final phase, we obtained labeled data from our cluster analysis and grouped the clusters according to the characteristics into binary and multiple classes for building supervised models and applying classification algorithms. To validate the labeling of the contents through clustering we have implemented the dataset into different supervised algorithms. This phase contributes in validating the obtained labeled data and evaluating the performance of the model. With the implementation of various classification models the proposed algorithm learns assigning class labels to the forum data examples. The accuracy of the model tells the model's performance based on the test data result. We have then implemented K fold cross validation to generate the classification reports with the performance metrics. We evaluated the model's performance with different performance parameters including precision, recall and f1-score. In this section we describe the phases of the overall system and the detailed results and outcome are discussed in section IV.

A. DATA SET

Dark Web forums are often used for anonymous trading of illicit items or services. For this research, we have used the scraped data collected from the Dark Net market archive crawled and created by researcher Gwern Branwen [5]. These dataset were scraped on a daily or weekly basis. The collections mostly contain raw HTML files, Jason files, images and php files. This database collection contains around 1.6 TB uncompressed Dark Net data of almost 89 marketplaces and 37 forums. Individual Dataset contains each forum dataset for example, Agora forum data [5], Agora data, Silk Road marketplace data and so on. The uncompressed files themselves are very large in size. For research purpose we have considered analysing one of the popular Dark Net forums named Agora forum for our experiment. For comparing our proposed algorithm with an existing work we have also implemented our model with Black Market Reloaded (BMR) forum dataset. This also shows the robustness of our proposed model in multiple datasets.

B. TEXT FILTERING AND TEXT CLEANING

Text cleaning and filtering the significant text is a prerequired step to progress further. The forum posts contain lots of non-meaningful words, digits, misspelling, slangs and non-English words. Every discussions text consists of many irrelevant words which occur only once those are not necessary to further processing. Popular English prepositions and punctuations also do not have any significance in further processing. From the documents English stop words and white-spaces are removed for further processing. For the English stop words removal nltk library has been used. Additional to these stop words some other irrelevant words were also removed by creating a list of the unwanted words and added by appending the new stop list in the stop words set. The occurrence of each word after removing stop words are counted and those occurred more than once are taken only as a part of text cleaning and filtering. Empty lists and all the digits in the list appear to be no significant use further so they were removed as well. Once the list of filtered words was achieved the next step involved the creation of dictionary and find unique token from the dictionary.

C. DOCUMENT TERM MATRIX

The Bag of words model breaks the text documents into statistics of individual word count. Data matrix with bag of words vectors is known as document term matrix [22]. Using the bag of word model the characteristics from the texts were extracted which gave us the vector of tokens. The number of occurrences for each unique token in the filtered dataset is represented by bag of word method. Data points with fixed length flat vectors represent the data matrix. To form a document-term-matrix, we represent each document's featured word per corpus in term of their occurrence in a matrix. In the document term matrix each row represents a unique document, where each column value is the featured keyword and each cell represents the number of occurrence within that document. Each cell of the matrix represents the frequency in that particular document.

D. TF-IDF

Term frequency-inverse document frequency or $TF - IDF$ is a popular method applied in text mining and information retrieval [30]. For our proposed algorithm $TF - IDF$ values are calculated for each keyword per document. $TF-IDF$ does the normalized count by dividing each word count in the data set by the number of documents where the word is present.

Term Frequency or TF measures the frequency of a term in a document. This calculation is obtained from the proportion of number of times a particular word is present in a document compared to the total number of words in that document. So, the value of TF goes higher with the frequency of that word within the document.

$$tf_{(t,d)} = \log(1 + freq_{(t,d)}) \tag{1}$$

Inverse term frequency or IDF is the measurement of a term's importance. Every term is considered equally

important when calculating TF. To upscale the weight of rare words in the documents of the whole text corpus IDF is computed. Equation (2) shows the formula to calculate the IDF value. Words occurring rarely across the corpus have. The words that occur rarely in the corpus have more IDF score.

$$idf_{(t,D)} = \log(\frac{N}{count(d \in D : t \in d}) \quad (2)$$

Equation (1) (2) and (3) show the formula of calculating $TF - IDF$ value of the term 't'.

$$tfidf_{(t,d,D)} = tf_{(t,d)} * idf_{(t,D)} \quad (3)$$

where t denotes the terms; d denotes each document; D denotes the collection of documents.

E. FEATURE SELECTION

In machine learning feature selection methods automatically select the most relevant attributes or variable to the dataset [7]. There are three different general classes for feature selection methods named embedded method, filter based method and wrapper method [13]. There are various feature selection method examples for each general class. For our model two different feature selection methods have been applied. At first feature importance method was implemented to get the scores of the most important features from the dataset then chi squared method has been implemented to get the ranking of the features and find the top features of our dataset. Feature importance of each feature of our dataset was obtained by applying the feature importance property of the Random Forest model applied in our dataset. This returns a score for each feature from the dataset. So the features with higher number of scores have more importance and are more relevant towards the output. These scores highlight which features are most relevant and least relevant to the target. The results of feature importance scores for our dataset are discussed in section V. We have implemented chi square method in our dataset for the top 10 features selection with the Random Forest classifier and getting the P value of each feature. The Recursive Feature Elimination or RFE module has been implemented with classifier to find the top k features. The chi square test returns a P value which is the test result of if a feature is significant neither of nor in our dataset. The higher p value indicates the less relevance of that feature for the model. The results of feature rankings and P values for our dataset are discussed in section V.

F. K-MEANS CLUSTERING

Clustering is the process of grouping data without labels [36]. Data having similar characteristics are grouped together and that group of similar data is called a cluster. There are many clustering algorithms of which K-means is one of the most popular one. We have implemented K-means algorithm as it can scale large unknown dataset with convergence assurance, is robust, fast and easy to understand. To define the clusters 'k' centroids are stored by K-means algorithm. An instance belong to a specific cluster if that is closer to that specific cluster's centroid point than any other centroids [9]. Once all the texts are processed K-means clustering is applied in our model. The featured keyword and their $TF - IDF$ values of frequency count were taken as the input of K-means. Each document is clustered with specific value for k. Clustered forum data within a particular cluster have high similarity among themselves than the data from the other clusters. The outcome and analysis of the clustering results are demonstrated in section V.

G. DECISION TREE

A decision tree is comparable to a flow chart in tree structure where an internal node denotes the attribute or feature, the branch denotes a decision rule and each leaf node denotes the outcome [18]. Once the cluster result is obtained with the implementation of K-means clustering, the clusters are assigned labels based on the specific classes from the discussion forums analysis. The results of the analysis have been discussed in the next section. Once the data are labeled with clusters, decision tree classifier is used to evaluate the model and test the prediction results. For a given keyword and its occurrence with the labeled train model predict the label. Data slicing is done by putting all the cell values of each featured columns and their cluster or label. There are different selection measures for the splitting criteria of the decision tree that performs the data portioning into best possible manner. Information Gain, Gain Ratio, and Gini Index are most popular examples [42]. We have measured our evaluation with Information Gain and Gini Index.

H. CLUSTER CHARACTERISTICS ANALYSIS

Analysis of the cluster characteristics is acquired from the result of the decision rules. The attributes of the generated decision tree are the most important features in the dataset. The leaf nodes of each branch of the tree contain the samples of forum data that belongs to a specific class those were relabeled according to the actual class names from our dataset analysis. These are the outcomes of the decision tee or the decision rules for each cluster named after the classes. Each decision rule was extracted from the characteristics of each cluster in the tree. The extracted decision rules and description of each cluster is explained in the result section.

I. CROSS VALIDATION WITH CLASSIFICATION MODELS

Once the cluster relabeling and characteristics have been analyzed with the decision tree, the forum data can be classified based on the analyzed cluster characteristics in a supervised way with classification algorithms. From the class labels obtained with relabeling the clusters we have multiple clusters against each class. To implement the supervised classification algorithms first we prepared the label dataset. In our proposed model, we have implemented both binary and multiclass classifications. In order to analyze the performance based on the binary classification results we assigned each forum data into two classes by grouping the classes into two

subsets of clusters and labeled our binary dataset for classification. For the multiclass classification we have used all the analyzed classes obtained from the cluster characteristics where we have one cluster per class.

After applying the classification algorithms, to validate and evaluate the model performance cross validation algorithm [14] has been implemented and the results of four different classification algorithms implemented into our dataset have been compared. Cross fold validation model is an effective approach to validate and compare any implemented model's performance with other classification models. Thus we have applied K fold cross validation in our model and also used it to find out which classification algorithm performs best. We have applied four supervised classification algorithms including Support Vector Machine [24], Logistic Regression [19] Naïve Bayes classification [31] and Random Forest classification [39] in the dataset with both multiclass and binary labels obtained from the cluster relabeling and analysis. The results of the cross validation with classification reports are discussed in section V.

V. EXPERIMENTAL RESULTS

In this section, the performance of our proposed model is evaluated. The results of the obtained features with feature selection algorithms, class labels analysis, characteristics of the cluster analysis and performance of the proposed model will be presented here. We have also evaluated our model performance with cross fold validation algorithm and compared the results of four different classification algorithms into our dataset. The classification reports along with the confusion matrix and Receiver Operating Characteristics (ROC) curves demonstrate that the proposed model can be effective with monitoring the forum data to analyze the contents when the labels of the dataset are uncertain. In addition, our model has been compared with an existing work and the results of each phase are shown. This comparison also demonstrates our model's performance with multiple dataset. Our analysis leads to the conclusion that with the cross fold validation and comparison results we can validate that our cluster analysis model is effective in describing the characteristics of Dark Web forum data which can lead to the particular discussion page that is indicative of any illicit activity or any data breach. We used different performance metrics which are presented in equations (4)-(7). In equations (4) to (7) ''TP'' is the total true positive and ''FP'' is the total false positive, ''TN'' is the total True negative, ''FN'' is the total False Negative.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{4}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{5}$$

$$F\ score = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

$$Total\ Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|} \tag{7}$$

TABLE 1. Features scores.

Sl	Feature	Score
1	categories	0.064339
2	information	0.059307
3	most	0.054655
4	last	0.047384
5	features	0.043436
6	view	0.042615
7	bugs	0.042081
8	post	0.040905
9	topics	0.040645
10	news	0.039275
11	online	0.03819
12	forum	0.038019
13	see	0.033674
14	for	0.033623
15	product	0.032971
16	ones	0.032899
17	board	0.032695
18	re	0.030491
19	market	0.029967

A. EXPERIMENTAL ANALYSIS ON AGORA FORUM DATASET

The obtained results implemented on the Agora forum dataset with our proposed model to monitor and analyze the discussion forums to identify any illicit activities happening inside the discussion pages are discussed in this sub section. The achieved results of each phase of our proposed model implemented on the Agora forum dataset are briefly described below.

B. FEATURES

The features of our experiment are the featured key words or tokens obtained from the processing of the data set. The feature extraction procedures are described in section IV.

The document term matrix implementation resulted 3926 tokens/words from the pre-processing phase. These tokens or words are filtered unique keywords from each document presented in a matrix form in terms of their occurrences. From the document term matrix, to convert the text words into feature Term Frequency- Inverse Document Frequency ($TF - IDF$) conversion has been applied. 50 featured words for 3736 forum data were selected from the conversion. In order to determine the most relevant attributes or features from the obtained feature, feature selection algorithms of feature importance and chi square method have been implemented. Features having very less score were omitted in the further processing when feeding the data to the model. In Table 1 the scores for all important features are summarized.

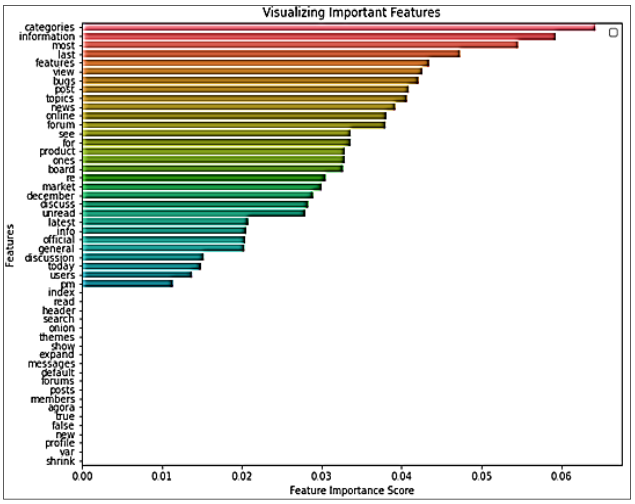

FIGURE 2. Visualization of features scorings based on feature importance.

The top 19 important features obtained from the 50 features of our dataset are shown in Fig.2. The colored bars show the importance of each feature in descending order. The features that do not show any bars have white bars and values 0 or less, except the *index* and *read* features. These two features have very low importance than the top 30 features but greater than 0 values than the below 18 features.

After the feature importance model applied we have implemented chi square method for getting the top 10 features and the p values of the features for ranking the features obtained. The feature rankings are shown in Table 2.

TABLE 2. Top 10 feature rankings.

Ranking	Features
1	today
2	topics
3	ones
4	online
5	discuss
6	users
7	last
8	latest
9	most
10	official

The Fig.3 shows the P values of each selected feature. The higher value indicates the less relevance of that feature for the model. For training the model we have discarded the features having P values greater than 0.9. According to the scoring *index* and *read* features have highest p values. Thus we choose top 30 features for the classification.

The overall summary of the final obtained features from keywords extraction to feature selection of the experimental Agora forum dataset is demonstrated in Table 3.

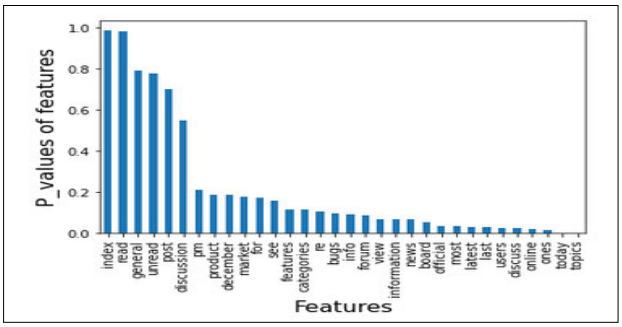

FIGURE 3. Scorings of features based on P values.

TABLE 3. Extracted features summary.

Attributes	Outcome
Total Tokens or keywords	3926
Converted Features ($TF-IDF$ Conversion)	50
Selected Features	32
Final Features (Feature Importance)	'categories', 'last', 'view', 'forum', 'latest', 'pm', 'online', 'users', 'topics', 'info', 'ones', 'news', 'official', 'for', 'bugs', 'board', 'discussion', 'post', 'discuss', 're', 'today', 'general', 'market', 'information', 'unread', 'most', 'features', 'see', 'product', 'december', 'read', 'index'
Top 10 Features (Chi Square Method)	'today', 'topics', 'ones', 'online', 'discussion', 'users', 'last', 'latest', 'most', 'official'

C. LABEL ANALYSIS

The Agora forum dataset consists of various forms of discussions on the forum pages. However, as mentioned earlier there is no direct labelling regarding the classes or discussion topics available on the dataset. For identifying the class of the forum data, a python script was run to find the topics on the discussions and the path of the files of that particular discussion file. Once the discussions files are obtained the pages were used for the analysis. Several random files were used for this label analysis. After analysing selected files and classifying the true labels of the files, total seven general classes were identified. To be noted the classes are not limited to what has been analysed in our study. The classes are discussed below.

i. Vendor: This class is related to discussion posts based on vendor reviews, asking on becoming vendors, discussing

various types of vendors and ways of vendor migration. All the posts related to this class are related to vendor issues.

ii. Breach: The most discussed topic found in Agora forum throughout the analysis was related to breach of information or data. Posts are mainly on scams, site links, hacking methods, security advice and issues, strategies on illicit activities, hiding keys, encryption breaking news, public key or PGP key discussions, sharing any old or illicit sites open, sharing ways to escape security and so on. These kinds of information related discussions were labeled as Breach.

iii. Financial: This class is based on discussion topics focused on bitcoin, bitcoin exchange related posts, escrow scams and discussions, money lost issues and so on. Posts related to finance and transactions are categorized into this class.

iv. Drug: Drugs are most highlighted and attractive topics on the forum discussions. This class is categorized on any advertisements on drugs, discussions on drug types, usage of drugs, availability of drug and some questions posts on the addiction and use level of the users.

v. Account: As Agora forum was a means of communication through discussions and marketplace, users must have account for it. So many posts were related to the user accounts. Such kinds of posts were labeled as account class. Posts on user account settings, inbox issues, updates on accounts, log in troubles and these related were into account

vi. Product: Posts and discussions on products and product related issues were labeled as product. Product listings, price issues, new products and ordering process or troubleshoot related posts were put into this class.

vii. Other: All the unspecified and hard to categorized posts were put into this class. Posts on personal questions and discussions such as colour or song choices, reading suggestions, guidelines on personal paths and these kinds of posts were there.

D. CLUSTERING AND RELABELLING

K-means clustering algorithm with 7 clusters has been applied on processed samples which clustered the data set. The random state was set as 1. The sample cluster outcome is shown at Fig.4 Here each document is clustered accordingly where each row represents each forum document, values of the features in each document are represented in each cell and the featured words or features of the dataset are presented in each column.

board	market	users	information	features	categories	Cluster
0.015129215	0.007535702	0.021194851	0.013666607	0.010314381	0.010322665	2
0.006142461	0.003059495	0.008605109	0.005548642	0.004187639	0.004191002	1
0.007142397	0.003557552	0.010005941	0.00645191	0.004869347	0.004873258	6
0.006662105	0.003318324	0.00933309	0.00601805	0.004541907	0.004545555	6
0.006576511	0.00327569	0.009213179	0.005940731	0.004483553	0.004487154	6
0.008876389	0.004421235	0.012435129	0.008018269	0.006051501	0.006056361	0
0.009812238	0.004887372	0.013746181	0.008863646	0.006689518	0.006694891	0
0.008800088	0.00438323	0.012328237	0.007949344	0.005999482	0.006004301	0

FIGURE 4. Sample output for K-means clustering.

Once the cluster result is obtained, the clusters are assigned labels. These labels names are based on the actual labels discussed in section V.C which are assigned to each cluster. We have re labeled the clustered data set with the actual class names assigning to each cluster. Fig.5 demonstrates the relabeled cluster output.

board	market	users	information	features	categories	Cluster
0.015129	0.007535702	0.021194851	0.013666607	0.010314381	0.010322665	Financial
0.006142	0.003059495	0.008605109	0.005548642	0.004187639	0.004191002	Drug
0.007142	0.003557552	0.010005941	0.00645191	0.004869347	0.004873258	Account
0.006662	0.003318324	0.00933309	0.00601805	0.004541907	0.004545555	Account
0.006577	0.00327569	0.009213179	0.005940731	0.004483553	0.004487154	Account
0.008876	0.004421235	0.012435129	0.008018269	0.006051501	0.006056361	Vendor
0.009812	0.004887372	0.013746181	0.008863646	0.006689518	0.006694891	Vendor
0.0088	0.00438323	0.012328237	0.007949344	0.005999482	0.006004301	Vendor

FIGURE 5. Sample cluster output after Re label.

E. PERFORMANCE EVALUATION AND DECISION TREE

The decision tree achieved with the important features were generated with criterion of Gini index and Information gain with maximum depth of 20 and random state 25. For evaluating the performance of the model 20% of the data were sent to test data and the remaining to training data. The predictions on test data for Decision tree on Gini index and Information gain gave accuracy results of 89.30% and 97.5% respectively. Minimum leaves for both were put 5 and maximum depth of the tree as 3. The generated tree with Gini index criteria applied is presented in Fig.6.

The decision tree is generated with 2988 samples of training dataset. These samples are divided at the start or root of the tree where the most important attribute for a sample decision tree is *topics*. The root node divides into child nodes and each child node iterates towards until leaf nodes are obtained. Each leaf node represents the cluster outcomes of the dataset and the way the tree expands towards the end node or leaf node describes the cluster characteristics.

1) DECISION RULES AND CLUSTER CHARACTERISTICS ANALYSIS

The decision tree generates a value based on the $TF - IDF$ weights for the featured words used as attributes. The $TF - IDF$ weights of our dataset are proportional to the frequency of a word appearing in the dataset. According to the feature values generated from the $TF - IDF$ weights each cluster is generated with specific number of samples. When the Gini impurity reaches to 0.0, corresponding class is assigned to a specific cluster. According to the sample decision tree featured word obtained from our model having a specific $TF - IDF$ weight in the dataset refers to a class. Fig.6 visualizes the generated tree for the clusters. The extracted seven decision rules for seven clusters based on the sample decision tree are shown in Table 4.

Based on the extracted decision rules the characteristics of the clusters are described below in Table 5. To be mentioned the interpretation of each rule is not limited to the feature words mentioned.

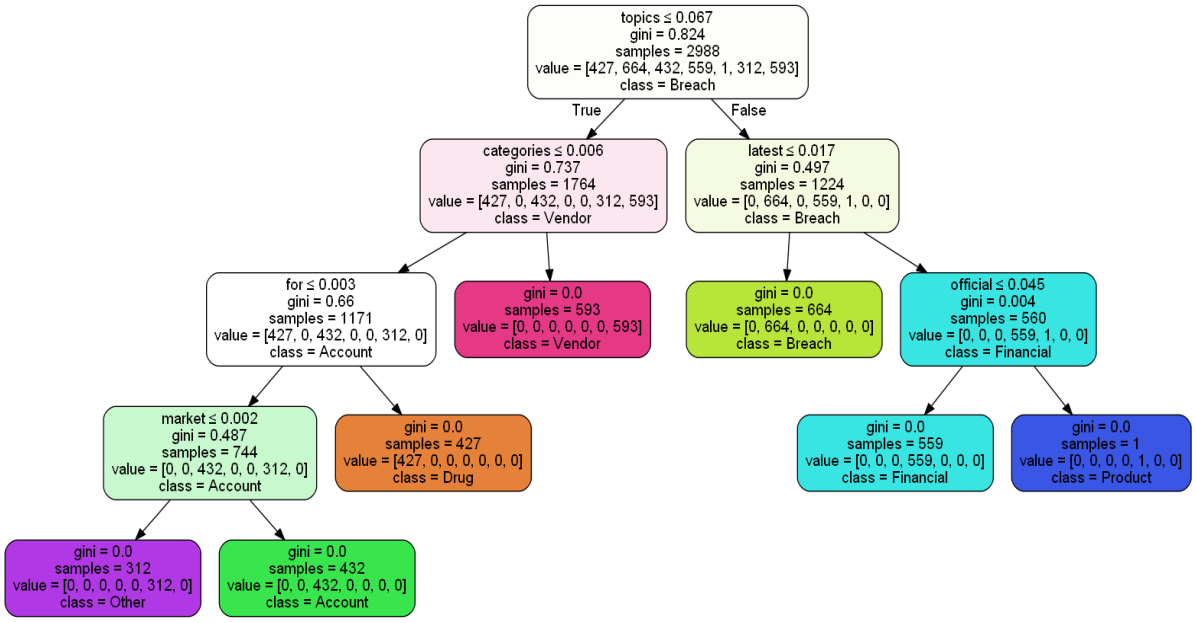

FIGURE 6. Generated sample decision tree.

TABLE 4. Extracted decision rules.

Cluster	Decision Rules
Breach	topics > 0.06662052497267723 & latest <= 0.01670150551944971 [[0. 664. 0. 0. 0. 0. 0.]] Breach
Financial	topics > 0.06662052497267723 & latest > 0.01670150551944971 & official <= 0.04527393355965614 [[0. 0. 0. 559. 0. 0. 0.]] Financial
Drug	topics <= 0.06662052497267723 & categories <= 0.00596160814166069 & for > 0.003210888826288283 [[427. 0. 0. 0. 0. 0. 0.]] Drug
Vendor	topics <= 0.06662052497267723 & categories > 0.00596160814166069 [[0. 0. 0. 0. 0. 0. 593.]] Vendor
Account	topics <= 0.06662052497267723 & categories <= 0.00596160814166069 & for <= 0.003210888826288283 & market > 0.0018587480881251395 [[0. 0. 432. 0. 0. 0. 0.]] Account
Other	topics <= 0.06662052497267723 & categories <= 0.00596160814166069 & for <= 0.003210888826288283 & market <= 0.0018587480881251395 [[0. 0. 0. 0. 0. 312. 0.]] Other
Product	topics > 0.06662052497267723 & latest > 0.01670150551944971 & official > 0.04527393355965614 [[0. 0. 0. 0. 1. 0. 0.]] Product

2) CLASSIFICATION RESULTS AND CROSS VALIDATION

To check the model accuracy, we have implemented different classification models into our dataset. Then K fold cross validation has been applied to choose the best model and generate the classification reports. The results of the classification prove that the cluster analysis model can be reliable in monitoring the forum data by characterizing the contents of the forums. The cross validation with classification models has been implemented for both binary classification and multiclass classification. In the following sub sections, the proposed model's performance based on the classification results and cross validation has been presented.

3) CLASS LABELING FOR CLASSIFICATION

The label dataset has been prepared for implementing the supervised classification algorithms. From the class labels obtained with relabeling the clusters we have seven classes for seven clusters. For multiclass classification all classes have been considered. In order to analyze the performance based on the binary classification results we divided all the classes into binary calculation to be more specific. So, we have analyzed the model performance for both binary and multiclass classification. Table 6 shows the class labeling and description for each label.

4) IMPLEMENTATION OF CLASSIFICATION ALGORITHMS

The results for the classification algorithms for binary and multiclass are presented in thus sub section. We have applied four supervised classification algorithms including Support Vector Machine, Logistic Regression, Naïve Bayes classification and Random Forest classification in the dataset with binary labels obtained from Table 6. For this task 20% dataset were given for testing and the rest for training. The results obtained from the models show that all four classification models perform well with the proposed model for binary classification. Among the four models Random Forest classifier

TABLE 5. Extracted cluster characteristics.

Rule	Interpretation
Cluster-Breach →Rule: *IF posts mention available hidden links, news revealed or users' security information THEN there is possible Data Breach discussion*	The highest number of samples (664) clustered is the Breach class. Forum data containing posts on hacking links, customers' hidden data, unpublished news and so on often refers to data breaches. Users often use the wording 'latest' 'news' 'most' 'information' in such posts. These feature words have high frequency occurrences with the posts in the dataset associated to breach.
Cluster-Financial→ Rule: *IF posts mention money lost, bitcoin index, transaction scam alerts THEN there is discussion on financial purpose*	The Financial cluster often belongs to forum data containing posts on bitcoin transaction help, transaction errors and so on . Users often use the wording 'index' 'board' 'discussion' 'latest' in such posts. These feature words have high occurrences with the posts in the dataset associated to financial.
Cluster-Vendor →Rule: *IF posts mention reviews on vendor types, becoming vendor THEN Vendor related discussions are present*	593 samples are into cluster Vendor class. Forum posts mostly ask for the ways to become vendor or giving review to a vendor. Users often use the wording 'categories' 'online' 'discussion' 'general', 'official' in such posts. These feature words have more $TF-IDF$ weights with the posts in the dataset associated to Vendor.
Cluster-Drug →Rule: *IF posts mention discussion on drugs, types of drugs, users of drugs THEN discussions are based on Drug class*	The Drug cluster mostly has the discussions topics on drugs in the forum data. Users often use the wording 'categories' 'product' 'market' 'discussion' 'users' in such posts. The $TF-IDF$ weights of these featured words associated with Drug class are more than any other class due to their occurrence.
Cluster-Account → Rule: *IF posts mention account settings, message read error, new user creation THEN discussion belong to class Account*	The Account cluster follows the wording related to discussions on account of the users. 432 samples were put into the cluster. The most frequent words used in such class are 'users' 'bugs' 'unread' 'read' 'view' but not limited to these.
Cluster-Other → Rule: *IF posts mention personal story sharing, general suggestion discussion THEN the discussions are classified as Other*	The Other cluster is the discussions matters those cannot be directly categorized into a specific class. The most frequent words used in such class are 'general' 'topics' 'view'.
Cluster-Product → Rule: *IF posts mention prices of products, buying new lists, ordering error THEN the discussions belong to Product cluster*	The remaining cluster is the Product class. Product listing, new buying, price settings are associated to these cluster. This cluster is often mixed with the rest of the classes while extracting the decision rules and classifying after relabeling.

TABLE 6. Class labeling.

Given Label	Cluster Description
Breach Related Data	• The cluster set for even clusters {0,2,4,6} for class labels "Vendor", "Financial", "Breach," "Account"
Non-Breach Related Data	• The cluster set for odd clusters {1,3,5} for class labels "Drug", "Product" and "Other"

gives the best accuracy of 98%. The accuracy results for the classification models are shown in Table 7.

The same four supervised classification with the same parameter settings of training and testing sample numbers

TABLE 7. Accuracy of classification models (binary class).

Model Name	Accuracy
Logistic Regression	0.798968423
Support Vector Machine	0.977590361
Random Forest Classifier	**0.989464525**
Gaussian Naïve Bayes	0.930937296

have been implemented in the dataset with the multiclass labels obtained from cluster labels described in Table 6. The results obtained from the models show that Logistic Regression model performs poorly but other three classification models perform well with the proposed model for multiclass classification. The accuracy results for the classification models are shown in Table 8.

TABLE 8. Accuracy of classification models (multi class).

Model Name	Accuracy
Logistic Regression	0.500267380
Support Vector Machine	0.979829192
Random Forest Classifier	**0.988930123**
Gaussian Naïve Bayes	0.976081223

With the comparative analysis for four different classification models on our proposed model for both binary and multiclass classifications the result shows that for multiclass classification Logistic Regression model performs poorly whereas for binary classification the all four classifiers generate good accuracy. The model accuracies for each model results comparing for binary and multiclass are summarized in Fig.7. It shows that for Random Forest and Support Vector Machine, the accuracy does not change much but Logistic Regression accuracy is reduced for multiclass and Naïve Bayes model increases the classification results a bit for the multiclass classification.

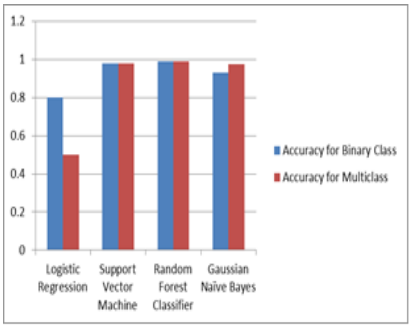

FIGURE 7. Model accuracy comparison of binary VS multiclass classification.

5) K FOLD CROSS VALIDATION

We have implemented K fold cross validation (k=5) to validate our model performance for both classifications. The performance metrics and the results are discussed and shown in this sub section. We have applied 5fold cross validation to compare the performance of each model and find the best model with the classification reports generated. The results of median accuracy for all models are shown in Table 9 and 10. From the tables we can see that, Logistic Regression performs poor on multiclass classification whereas for binary classification its performance average. The other three algorithms perform better with both binary and multi class. However, Random Forest classifier performs best for both cases followed by Support Vector Machine.

TABLE 9. Median accuracy of each classifier (binary class).

Model Name	Precision	Recall	F-Score
Logistic Regression	.92	.64	.67
Support Vector Machine	.97	**.98**	**.98**
Random Forest Classifier	**.98**	**.98**	**.98**
Gaussian Naïve Bayes	.91	.96	.94

TABLE 10. Median accuracy of each classifier (multi class).

Model Name	Precision	Recall	F-Score
Logistic Regression	.30	.37	.30
Support Vector Machine	**.85**	.84	.84
Random Forest Classifier	.84	**.86**	**.85**
Gaussian Naïve Bayes	.84	.85	.84

From the results obtained, we can find that Random Forest classifier gives the best results for our proposed model in predicting the classes for both multiclass and binary classification. Classification Reports of Performance Metrics for binary class labeling and multiclass labeling are shown in Table 11 and 12 respectively.

TABLE 11. Classification report of performance metrics (binary class).

	Precision	Recall	F-Score
Breach Related Data	**.98**	.94	**.97**
Non Breach Related Data	.83	**.97**	.90
Average	.91	.96	.93

From the above Tables, it can be seen that the overall classification report is better for binary classification. In multiclass classification Product class is giving poor prediction results and affecting the overall performance result.

6) CONFUSION MATRIX

We have evaluated the performance for both binary and multiclass classification with the generated confusion matrix

TABLE 12. Classification report of performance metrics (multi class).

	Precision	Recall	F-Score
Drug	0.97	0.97	**0.98**
Breach	**0.98**	**0.98**	**0.98**
Account	0.96	**0.98**	**0.98**
Financial	0.97	0.97	0.97
Other	0.91	**0.98**	0.95
Product	0.20	0.20	0.20
Vendor	**0.98**	0.97	**0.98**
Average	0.84	0.85	0.84

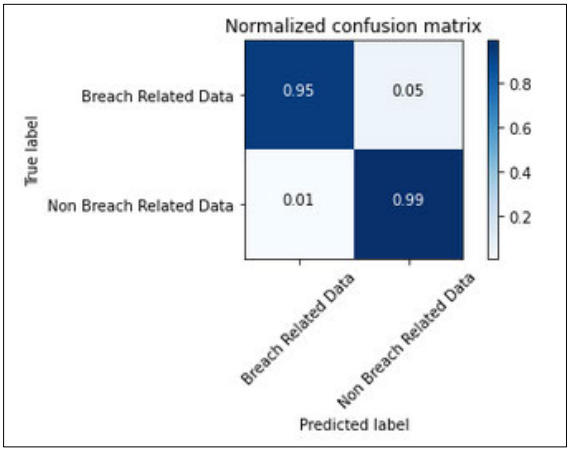

FIGURE 8. Normalized confusion matrix (binary classification).

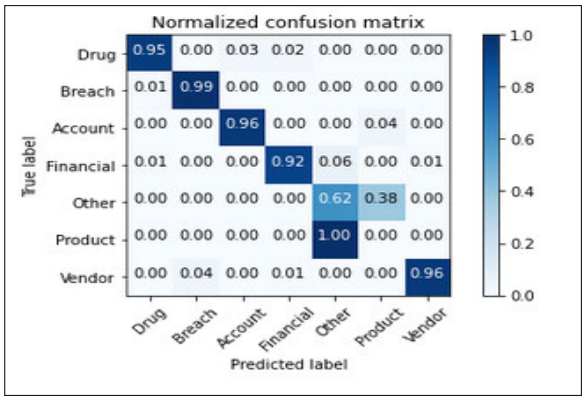

FIGURE 9. Normalized confusion matrix (multi class classification).

using Random Forest classifier as it performs best for both cases. Confusion matrix is a matrix form of table of errors used to evaluate the performance of classifier. The correct number of classified samples are presented in a diagonal way whereas the other cell values of the matrix represent misclassification [35]. The normalized confusion matrix of the classification results for binary classification and multiclass classification are presented in Fig.8 and Fig.9 respectively.

From the confusion matrix shown in Fig.8, it can be said the classification result of predicting the true labels for binary class is very accurate compared to misclassified labels. The diagonal values are representing the number of samples that predict the labels correctly against the true labels of the dataset.

The confusion matrix generated for the multiclass classification shows that the model performs very well classifying each class but performs badly at classifying the *product* class. The number of misclassified classes is very less for the rest of the classes. The reason behind this could be less number of instances in of the *product* class selected by the learning model. However, six classes out of seven classes are classified with high classification results as shown in the Fig.9.

7) AUC ROC CURVE

We have also evaluated the performance for both binary and multiclass classification with AUC ROC curves. ROC curve stands for a receiver operating characteristic curve which is a probabilistic curve defined by plotting the true positive and false positive rates at different thresholds. AUC stands for area under cover. The AUC ROC curve is useful in the performance measurement for multiclass classification where the AUC is the measurement of distinctions among the classes. The higher value of AUC represents the model can perform better in separating positive and negative classes [53]. The ROC AUC curves generated for four classification algorithms applied in our model for multiclass classification are presented in Fig.10.

From the curves shown in Fig.10, the results of ROC AUC scores for each classification models for multiclass classification are achieved. For all the four curves it can be visible that the models achieve poor score for product class with AUC score of 0.50. For Random Forest classifier except the product class other six classes achieve AUC score of 1 which is the best score thus all the lines of the curves are overlapped and only the last curve is visible. For Gaussian Naïve Bayes most of the AUC scores for all the classes expect product class are near 1. For SVM and

Logistic Regression models there are slight variations for each class but most of the classes have satisfactory AUC scores except the product and drug classes

a: COMPARISON WITH EXISTING ALGORITHM

We have compared our work with an existing work [40] on Dark Web forum dataset which revealed that our proposed model gives better accuracy than the comparing algorithm. We have compared our proposed model with the same dataset of the existing algorithm. This also demonstrates the robustness of our proposed model to perform on multiple datasets with great accuracy. The existing algorithm has implemented the classification with SVM and achieved 88% accuracy. With the implementation of SVM in our proposed model, our model gives 98% accuracy. From the cross-validation results in our proposed model the Random Forest classifier gives 99% accuracy. Beside this the precision and recall results have also higher accuracies for our model. With small dataset the comparing algorithm improves the accuracy but the recall

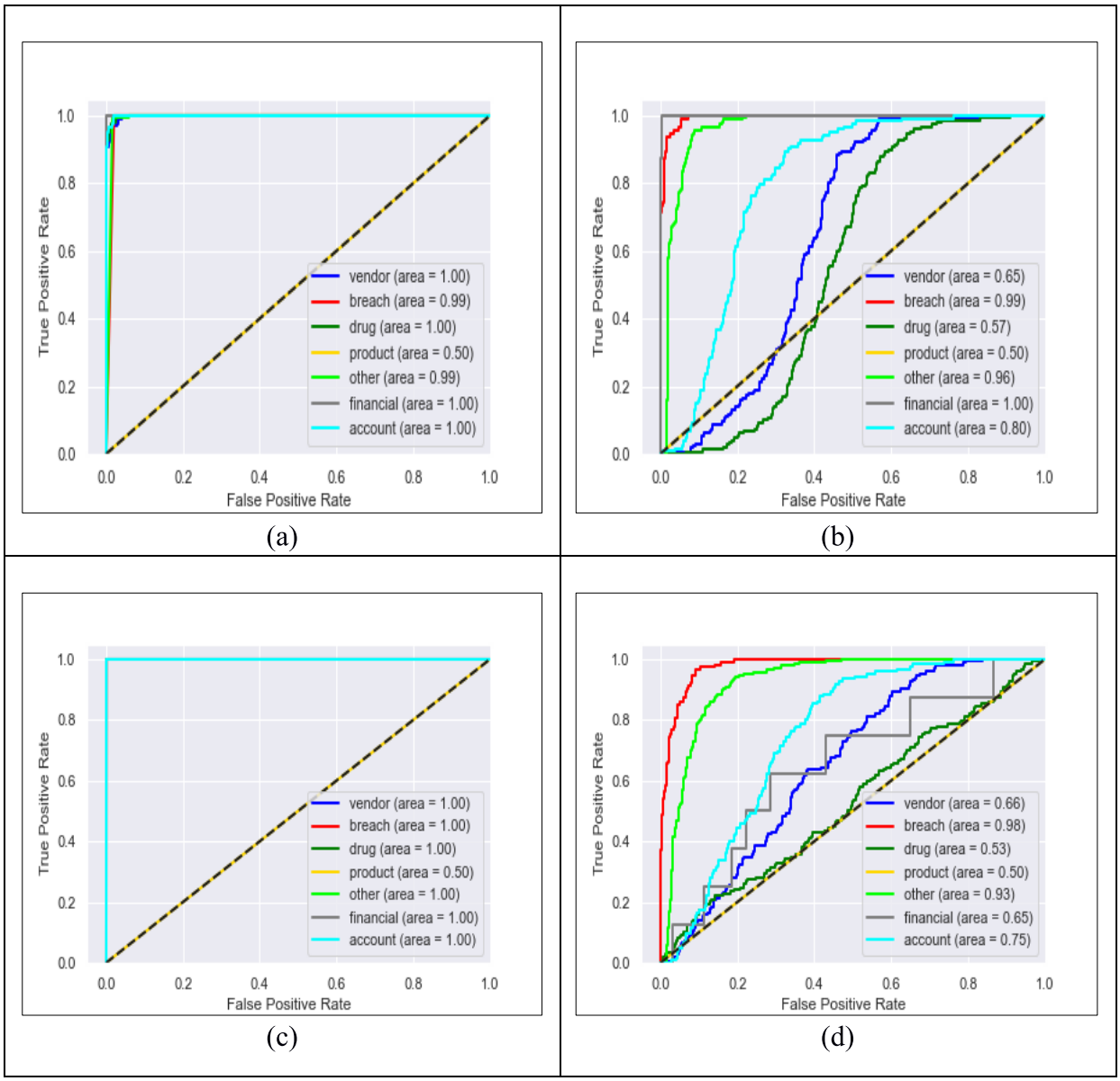

FIGURE 10. Illustration of AUC ROC curves for multiclass classification for a) Gaussian Naïve Bayes classifier b) SVM c) Random forest classifier and d) Logistic regression model.

is poor. However, with the huge Dark Web dataset the model must perform with great accuracy for large dataset. We have presented the comparison table in Table 13 that shows the dataset, features and methods used for our and the comparing article with their corresponding results.

From the table, we can see that for our proposed model, the overall accuracy including the precision recall results are higher than the existing algorithm for the same Dark Web forum dataset. With our proposed model, we have applied different feature selection algorithms in addition to $TF-IDF$ which ensures only the most relevant features are used in the model and also ranks the top keywords identified in the forum activities. Our cluster analysis model combined with classification algorithms can characterize the unlabeled forum data into meaningful clusters without the continuous availability of labeled forums and related knowledge. The achieved results of each phase of our model obtained from the Black-Market Reloaded forum dataset are briefly described in the below subsections.

b: PROCESSED FEATURES OF BMR DATASET

The Black Market Reloaded (BMR) forum [40] was another large marketplace forum in the Dark Net. The forum data contains thousands of documents, images, files and scripts and multi lingual discussion posts. The existing work [40] has proposed an authorship analysis technique as an de-anonymization process on the BMR forum dataset. They have used the discussion forums with

TABLE 13. Comparison with related research works.

Algorithm	Data Set	Feature	Method	Results
Our Proposed Model	Black Market Reloaded forum	BOW, Doc-term-matrix, $TF-IDF$, Feature Importance, Chi-square Method	K means clustering with Decision tree Cross fold validation:LR,SVM,NB, RF	Classification:**98%** Precision: **98%** Recall:**99 %**
Comparing Article	Black Market Reloaded forum	Stylometric, character level n-gram, and Time	support vector machines (SVM)	Classification:88% Precision:91% Recall:25%

TABLE 14. Important features scores of BMR dataset.

Sl	Feature	Score
1	reply	0.049828
2	customers	0.047036
3	profile	0.046745
4	us	0.040853
5	topic	0.040766
6	items	0.037831
7	scamming	0.036228
8	gregory	0.034968
9	so	0.030775
10	post	0.030598
11	make	0.030504
12	messed	0.030389
13	stuff	0.029208
14	deliver	0.021737
15	possible	0.021517
16	report	0.019982
17	re	0.019911
18	vendors	0.019586
19	status	0.01951

English languages for their proposed algorithm and parsed the English forum posts of BMR dataset to analyze the contents. As features of their model they have used stylometric, character level n grams and time based features. For the classification results of alias and authorship attribution Support Vector Machine (SVM) has been implemented. For the alias classification using SVM they randomly created pseudo users of two with single user data and computed the mean of feature vectors using the similarity measurement of cosine angle. The classification result was obtained based on their given threshold value. The alias classification results for 177 users returned 91% precision and 25% recall. However, the recall increases to 45% with less number of users of 25. For the authorship attribution they achieved 88% classification accuracy. Table 13 summarizes the features, methods and results obtained from the existing work on BMR dataset. Although the existing work has utilized the unlabeled BMR forum data for authorship analysis, the proposed work has used character n gram features which cannot handle out of vocabulary words. Moreover, the stylometric and time based features can be intentionally altered and subject to uncertainty.

For our proposed algorithm we have processed the forum data containing discussions. The steps for extracting and processing the data were implemented following the section IV to obtain the features from the dataset. From the dataset the document term matrix implementation resulted 7429 tokens/words from the pre-processing phase. Term Frequency-Inverse Document Frequency ($TF-IDF$) have been implemented to convert the tokens into features which resulted 90 featured keywords for the dataset. Then the most important features and the top 10 features have been selected using feature selection algorithms. The results of the important features with scores and the summary of all the features of the targeted dataset have been presented in Table 14 and Table 15 respectively.

c: LABEL ANALYSIS FOR BMR DATASET

The Black Market Reloaded (BMR) forum dataset consists of various discussions on the forum pages. The classes identified are discussed below in brief.

i. Vendor: This class is related to discussion posts based on specific vendor reviews and asking on how a specific vendor dealt with them. These are common types of discussions related to vendor and this vendor class has been identified as a label.

ii. Drug: The most discussed topics in BMR forum throughout the analysis found, was related to information regarding drugs. Due to the popularity of the BMR marketplace being drug dealing, most posts found were related to drugs. The posts and discussions identified for this class are asking for suggestions on particular

TABLE 15. Extracted features summary of BMR dataset.

Attributes	Outcome
Total Tokens or keywords	7429
Converted Features ($TF - IDF$ Conversion)	90
Selected Features	46
Final Features	'reply', 'customers', 'profile', 'us', 'topic', 'items', 'scamming',
(Feature Importance)	'gregory', 'so', 'post', 'make', 'messed', 'stuff', 'deliver', 'possible', 'report', 're', 'vendors', 'status', 'ever', 'many', 'money', 'edited', 'offline', 'please', 'still', 'german', 'things', 'order', 'rite', 'prevent', 'works', 'back', 'put', 'wrote', 'good', 'time', 'even', 'quote', 'much', 'wanted', 'member', 'previous', 'know', 'where', 'very'
Top 10 Features (Chi Square Method)	'messed', 'rite', 'vendor', 'house', 'money', 'where', 'customers', 'limited', 'status', 'damn'

drugs, details of side effects of drugs, prescribed and non-prescribed drugs, facing troubles with drug dealing, escaping from police for dealing drugs and reviews on drugs.

iii. *Breach:* This class is based on discussion topics focused on creating illicit passports from different countries, credit card account selling, different ids information, TOR addresses, and vouchers of unusual titles. These kinds of information related discussions were classified as Breach. Unlike Agora forum this class is not the most popular one.

iv. *Financial:* Another popular discussion found in BMR forums from our analysis is discussion regarding finance and mostly on bitcoin. The posts of labeled into the financial class are about escrow models, bitcoin prices, bitcoin exchange sites, mediums and payment procedures through bitcoins.

v. *Account:* Users of the BMR forum must have accounts to post and reply for the discussions on the forum. As a result various account related queries and issues are posted for discussions in the forum. These posts are put into the account class throughput are analysis.

vi. *Product:* Posts and discussions on selling products and product related issues were classified as product. Various product listings, price issues, new products reviews and ordering process or shipping posts were put into this class.

vii. *Other:* BMR forum is multilingual and there are some posts on different languages such as Spanish. Both the comparing algorithm and our proposed model focus on English language only and thus all non-English posts have been classified as other class. In addition to that discussions posts those are hard to categorize were put into this class.

8) DECISION TREE OF BMR DATASET

With the obtained important features of the dataset and labeled clusters decision tree has been implemented with both Information gain and Gini index criterion. Decision tree classifier on Gini index and Information gain, the predictions on test data gave accuracy results of 98% and 95% respectively. A sample generated decision tree with Gini index criteria applied in the processed BMR dataset is presented in Fig.11. The decision tree is generated with 3997 samples of training dataset and the most populated cluster is the drug class following by breach class.

a: CLUSTER CHARACTERISTICS ANALYSIS OF DECISION TREE

We have extracted the decision rules from the generated decision tree of the BMR dataset that consists of featured keywords as attributes with values generated from their TF-IDF weights. Each cluster has specific number of samples in a tree based on the feature values of TF-IDF weights Fig. 14 visualizes the generated tree for the clusters. The extracted decision rules for clusters based on the sample decision tree are shown in Table 16.

b: CLASSIFICATION AND CROSS VALIDATION RESULTS OF BMR DATASET

The comparing algorithm has implemented the Support Vector Machine (SVM) for the binary classification results on the BMR dataset. Along with the SVM we have implemented the other three classification models used in our proposed model to check the accuracy of the model on the BMR dataset for both binary and multiclass classifications. The overall accuracy of our proposed model is 98% for the SVM classifier whereas the comparing algorithm has achieved an accuracy of 88%. Moreover, the classification reports, cross validation results of our proposed model demonstrate that our model performs superiorly than the existing algorithm for the same dataset. Our proposed model's performance based on the classification results and cross validation on BMR dataset has been presented below.

9) CLASS LABELING FOR CLASSIFICATION

According to the cluster characteristics analysis on the BMR dataset we have prepared the label dataset to implement the supervised classification algorithms. To analyze the performance of our model based on the multiclass classification we have implemented all the seven clusters as seven classes and thus there have been one cluster per class classification. For

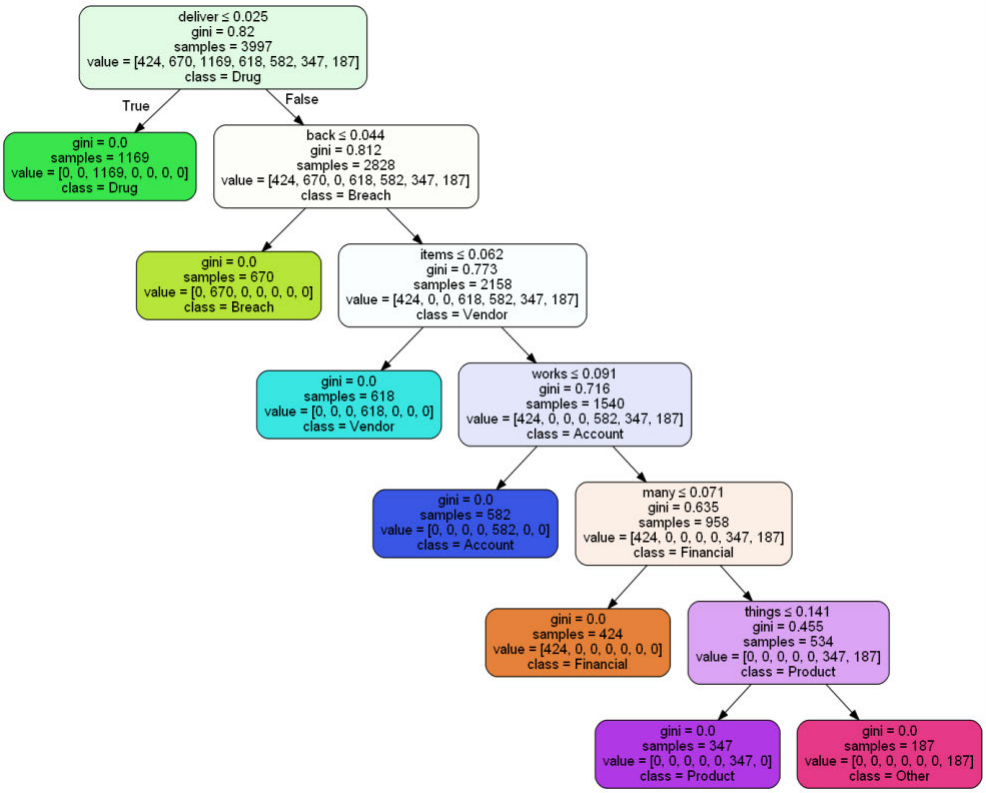

FIGURE 11. Sample decision tree of BMR dataset.

the binary classification we have grouped the classes into two subsets of clusters and labeled for binary classification based on the cluster characteristics. Table 17 represents the class labeling and description for each label.

10) CLASSIFICATION RESULTS

The classification results for both binary and multiclass classification implemented on the BMR dataset gives great performance based on our model. The results obtained for the classification algorithms including Support Vector Machine, Logistic Regression, Naïve Bayes classification and Random Forest classification in the dataset with binary and multiclass labels obtained from Table 17 are presented in Table 18 and Table 19 respectively.

From the above tables we can see that, for both binary and multiclass classifications our model performs very well (98%) for the SVM classifier compared to the existing algorithm on this dataset. Similar to our proposed model on Agora forum dataset, Random Forest classifier performs best on the BMR dataset as well. The other classifiers also present satisfactory results with our proposed model in the dataset. The model accuracies for each model results comparing for binary and multiclass are summarized in Fig.12. From the Figure, it can be seen that, the accuracy results do not vary much for Random Forest, Support Vector Machine and Naïve Bayes models but the accuracy of the Logistic Regression model is reduced for binary classification.

11) PERFORMANCE METRICS RESULTS

Comparing to the existing algorithm on BMR dataset our model also achieves great precision, recall and f scores. The performance metrics of all models have been evaluated with 5-fold cross validation for both binary and multiclass classification. The results of median accuracy for all models are shown in Table 20 and Table 21 for binary and multiclass classification respectively. Random Forest classifier performs best for both cases followed by Support Vector Machine. In the comparing article they have achieved 91% precision and 25% recall for binary classification with SVM classifier whereas for binary classification with SVM our proposed model shows 98% precision with 99% recall.

The classification reports from the generated cross validation results also show that our proposed model achieves great accuracy for both binary and multiclass classification. Classification reports of performance metrics for binary class labeling and multiclass labeling are shown in Table 22 and 23 respectively.

12) CONFUSION MATRIX

To evaluate our model performance for both binary and multiclass classification on the BMR dataset we have also the generated confusion matrix. The normalized confusion matrix of the classification results for binary classification and multiclass classification are presented in Fig.13 and Fig.14 respectively. The confusion matrix also shows that the

TABLE 16. Extracted cluster characteristics for BMR dataset.

Rule	Interpretation
Cluster-Drug →Rule: *IF posts mention suggestions on drugs, effects of drugs, reviews of drugs THEN discussions are based on Drug class*	The highest number of samples (1169) clustered is the Drug class. The Drug cluster mostly has the discussions topics on drugs in the BMR forum data. Users often use the wording 'deliver' 'prevent' 'wanted' 'where' 'messed' in such posts. The $TF-IDF$ weights of these featured words associated with Drug class is more than any other class due to their occurrence.
Cluster-Breach →Rule: *IF posts mention credit card selling, vouchers, security information THEN there is possible Data Breach discussion*	The next most populated sample is the cluster of the Breach class. Forum data containing posts on customers' hidden data, illicit id or passport buying, TOR links, unknown vouchers news and so on often refers to data breaches. Users often use the wording 'previous' 'items' 'possible' 'works' 'scamming' in such posts. These feature words have high frequency occurrences with the posts in the dataset associated to breach.
Cluster-Vendor →Rule: *IF posts mention reviews on vendor, vendor complains THEN Vendor related discussions are present*	638 samples are into cluster Vendor class. Forum posts mostly ask for giving review to a vendor or dealing with specific vendors on items . Users often use the wording 'wanted' 'items' 'stuff' 'rite' 'German', 'US' in such posts. These feature words have more $TF-IDF$ weights with the posts in the dataset associated to Vendor
Cluster Financial →Rule: *IF posts mention, bitcoin escrow models transaction procedures THEN there is discussion on financial purpose*	The Financial cluster often belongs to forum data containing posts on bitcoin transaction help, transaction mediums, escrow models and so on . Users often use the wording 'money' 'quote' 'things' 'many' in such posts. These feature words have high occurrences with the posts in the dataset associated to financial.

classification result of predicting the true labels for both binary and multiclass on BMR dataset are very accurate compared to misclassified labels.

TABLE 17. Class labeling for BMR dataset.

Given Label	Cluster Description
Non Breach Related Data	• The cluster set for clusters {0,1,3,5,6} for class labels "Drug", "Account", "Vendor", "Product" "Other"
Breach Related Data	• The cluster set for even clusters {2,4} for class labels "Financial" and "Breach"

TABLE 18. Accuracy of classification models on BMR dataset (binary class).

Model Name	Accuracy
Logistic Regression	0.670005205
Support Vector Machine	0.983195796
Random Forest Classifier	**0.998195796**
Gaussian Naïve Bayes	0.964179179

TABLE 19. Accuracy of classification models on BMR dataset (multi class).

Model Name	Accuracy
Logistic Regression	0.943773373
Support Vector Machine	0.987595395
Random Forest Classifier	**0.998195796**
Gaussian Naïve Bayes	0.984790591

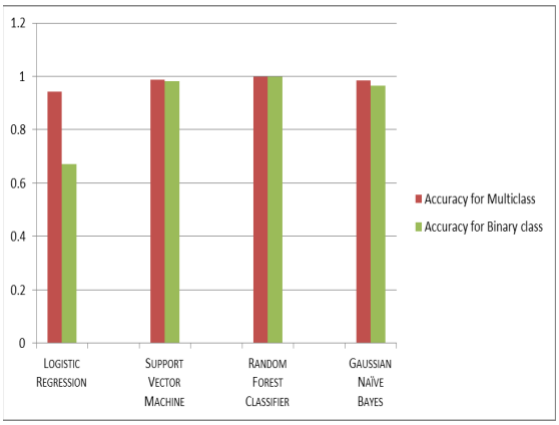

FIGURE 12. Comparison of binary VS multiclass classification on BMR dataset.

13) AUC ROC CURVE

To visualize the performances of each classifier with our proposed model the AUC ROC curves for all four classification algorithms for multiclass classification is presented in Fig.15.

TABLE 20. Median accuracy of each classifier on BMR dataset (binary class).

Model Name	Precision	Recall	F-Score
Logistic Regression	.34	.50	.40
Support Vector Machine	.98	**.99**	.98
Random Forest Classifier	**.99**	**.99**	**.99**
Gaussian Naïve Bayes	.96	.97	.96

TABLE 21. Median accuracy of each classifier on BMR dataset (multi class).

Model Name	Precision	Recall	F-Score
Logistic Regression	.95	.93	.93
Support Vector Machine	.98	.98	.97
Random Forest Classifier	**.99**	**.99**	**.99**
Gaussian Naïve Bayes	.98	.98	.98

TABLE 22. Classification report of performance metrics on BMR dataset (binary class).

	Precision	Recall	F-Score
Breach Related Data	.94	.96	.95
Non Breach Related Data	**.98**	**.97**	**.98**
Average	.96	.97	.96

TABLE 23. Classification report of performance metrics on BMR dataset (multi class).

	Precision	Recall	F-Score
Account	**0.99**	**0.99**	**0.99**
Breach	0.94	0.97	0.98
Drug	0.95	0.96	0.97
Financial	0.98	**0.99**	**0.99**
Other	**0.99**	**0.99**	**0.99**
Product	0.96	0.98	0.97
Vendor	0.96	**0.99**	0.98
Average	0.96	0.98	0.98

For all the four ROC curves it can be visible that the models achieve good AUC score for the majority of the classes. For Random Forest classifier all classes achieve AUC score of 1 which is the best score thus all the lines of the curves are overlapped and only the last curve is visible. For Gaussian Naïve Bayes most of the AUC scores for all the classes are 1

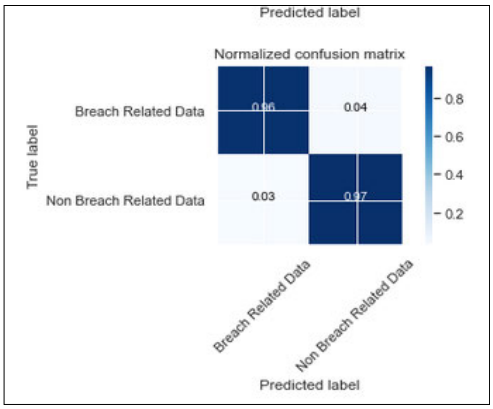

FIGURE 13. Normalized confusion matrix on BMR dataset (binary classification).

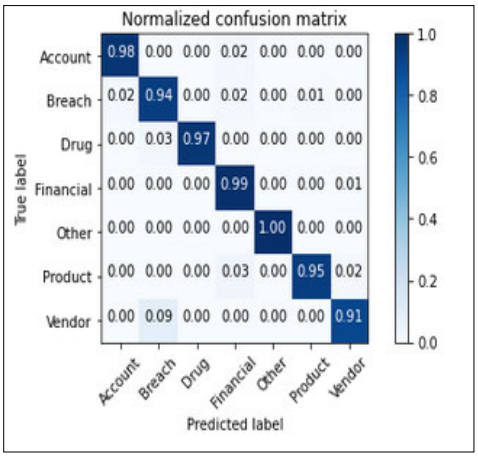

FIGURE 14. Normalized confusion matrix on BMR dataset (multiclass classification).

or near 1. For SVM and Logistic Regression models there are slight variations for each class but most of the classes have satisfactory AUC scores except the drug and other classes

F. DISCUSSION

We have evaluated the performance of our proposed model to characterize and understand the contents of the Dark Web forum data using cluster characteristics analysis. To validate the analysis of the model we have implemented and compared four different classification models into our dataset which gives the prediction results based on the class labels analyzed from the cluster characteristics. The analysis has been conducted in two ways one for binary and the other for multiclass classification. We have applied K fold cross validation to generate the classification reports and choose the best classification algorithm that performs best with our model. With the results obtained from the cross validation, it is proved that our model can effectively characterize the clusters with high classification accuracy. We have also demonstrated confusion matrix and ROC curve implementation and the

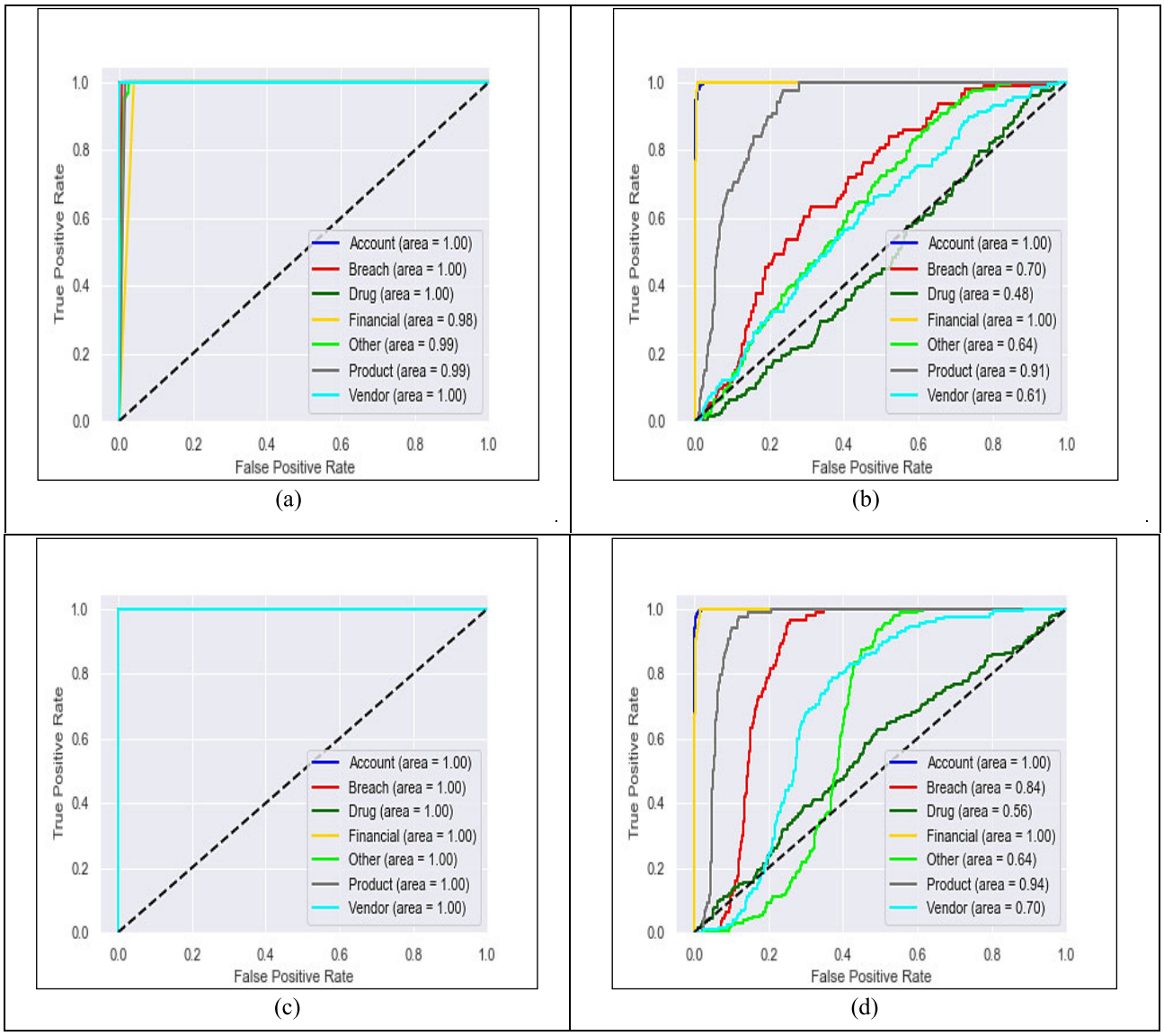

FIGURE 15. Illustration of AUC ROC curves for multiclass classification on BMR dataset for a) Gaussian Naïve Bayes classifier b) SVM c) random forest classifier and d) Logistic regression model.

results obtained for both classifications. Moreover, our proposed model has been compared with an existing algorithm on a different Dark Web forum dataset. This represents the robustness of our proposed model on multiple datasets. The experimental results also demonstrate the effectiveness of our model than the existing model on the same dataset.

VI. CONCLUSION

The proposed approach develops an unsupervised model to monitor and characterize the Dark Web forums. We proposed a rule discovery model by implementing unsupervised and supervised algorithms for the analysis of the cluster characteristics for each cluster obtained from the generated decision tree with most important featured keywords for multiple Dark Web datasets. The clustered samples and rules obtained from our experiment show that leaked information or breach is one of the most populated clusters for the Agora forum dataset whereas drug is the most populated one for the BMR dataset. The generated intelligent data can be utilized as scientific evidence of the crimes happening inside the Dark Web. Our implementation and analysis are limited to particular pre scrapped Agora forums dataset and BMR forum dataset which can be applied on any Dark Net forums. A possible enhancement of the paper can be analyzing the dataset by implementing deep learning with unstructured data. In future we also would like to propose a framework for semi-supervised approach by integrating unsupervised approach with supervised techniques. The overall performance of our proposed approach proves that our model could be effectively

implemented to characterize the Dark Web forums data to generate the intelligent data which can advantage the cyber security specialist and law enforcement.

REFERENCES

[1] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *Proc. IEEE Intell. Secur. Informat.*, May 2007, pp. 282–288.

[2] A. Afilipoaie and P. Shortis, *From Dealer to Doorstep-How Drugs are Sold on the Dark Net. GDPO Situation Analysis*. Swansea, Wales: Swansea Univ., Global Drugs Policy Observatory, 2015.

[3] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. de Paz, "Classifying illegal activities on TOR network based on web textual contents," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, Long Papers*, vol. 1, 2017, pp. 35–43.

[4] N. Arnold, M. Ebrahimi, N. Zhang, B. Lazarine, M. Patton, H. Chen, and S. Samtani, "Dark-Net ecosystem cyber-threat intelligence (CTI) tool," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 92–97.

[5] G. Branwen, N. Christin, D. Décary-Hétu, R. M. Andersen, P. E. StExo, and S. Goode. (Jul. 12, 2015). *Dark Net Market Archives 2011–2015*. [Online]. Available: https://www.gwern.net/DNM-archives

[6] J. Buxton and T. Bingham, "The rise and challenge of dark net drug markets," *Policy Brief*, vol. 7, pp. 1–24, Jan. 2015.

[7] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[8] V. Ciancaglini, M. Balduzzi, M. Goncharov, and R. McArdle, "Deepweb and Cybercrime," Trend Micro, Tokyo, Japan, Tech. Rep. 9, 2013.

[9] I. S. Dhillon, J. Fan, and Y. Guan, "Efficient clustering of very large document collections," in *Data Mining for Scientific and Engineering Applications*. Boston, MA, USA: Springer, 2001, pp. 357–381.

[10] K. M. Finklea, "Dark web," Congressional Res. Service, Washington, DC, USA, 2015.

[11] T. Fu, A. Abbasi, and H. Chen, "A focused crawler for dark web forums," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 6, pp. 1213–1231, 2010.

[12] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decis. Support Syst.*, vol. 50, no. 3, pp. 595–601, 2011.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[14] Y. Jung, "Multiple predicting K-fold cross-validation for model selection," *J. Nonparam. Statist.*, vol. 30, no. 1, pp. 197–215, 2018.

[15] S. Kramer, "Anomaly detection in extremist web forums using a dynamical systems approach," in *Proc. ACM SIGKDD Workshop Intell. Secur. Informat. (ISI-KDD)*, 2010, pp. 1–10.

[16] G. L'Huillier, H. Alvarez, S. A. Ríos, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 66–73, 2011.

[17] W. Lacson and B. Jones, "The 21st century DarkNet market: Lessons from the fall of silk road," *Int. J. Cyber Criminol.*, vol. 10, no. 1, p. 40, 2016.

[18] M. Li, H. Xu, and Y. Deng, "Evidential decision tree based on belief entropy," *Entropy*, vol. 21, no. 9, p. 897, Sep. 2019.

[19] D. Liu, T. Li, and D. Liang, "Incorporating logistic regression to decision-theoretic rough sets for classifications," *Int. J. Approx. Reasoning*, vol. 55, no. 1, pp. 197–210, Jan. 2014.

[20] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, p. 1608, Dec. 2016.

[21] S. Mancini and L. A. Tomei, "The dark web: Defined, discovered, exploited," *Int. J. Cyber Res. Educ.*, vol. 1, no. 1, pp. 1–12, Jan. 2019.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781

[23] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescape, "Anonymity services tor, I2P, JonDonym: Classifying in the dark," in *Proc. 29th Int. Teletraffic Congr. (ITC)*, Sep. 2017, pp. 81–89.

[24] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[25] H. Nishikaze, S. Ozawa, J. Kitazono, T. Ban, J. Nakazato, and J. Shimamura, "Large-scale monitoring for cyber attacks by using cluster information on darknet traffic features," *Procedia Comput. Sci.*, vol. 53, pp. 175–182, Jan. 2015.

[26] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 7–12.

[27] T. Pineau, A. Schopfer, L. Grossrieder, J. Broséus, P. Esseiva, and Q. Rossy, "The study of doping market: How to produce intelligence from internet forums," *Forensic Sci. Int.*, vol. 268, pp. 103–115, Nov. 2016.

[28] K. Porter, "Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling," *Digit. Invest.*, vol. 26, pp. S87–S97, Jul. 2018.

[29] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep., 2000.

[30] J. Ramos, "Using TF-IDF to determine word relevance in document queries," presented at the Proc. 1st Instructional Conf. Mach. Learn., Dec. 2003.

[31] Z. E. Rasjid and R. Setiawan, "Performance comparison and optimization of text document classification using k-NN and Naïve Bayes classification techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107–112, Jan. 2017.

[32] D. Rhumorbarbe, L. Staehli, J. Broséus, Q. Rossy, and P. Esseiva, "Buying drugs on a darknet market: A better deal? Studying the online illicit drug market through the analysis of digital, physical and chemical data," *Forensic Sci. Int.*, vol. 267, pp. 173–182, Oct. 2016.

[33] S. A. Ríos and R. Muñoz, "Dark web portal overlapping community detection based on topic models," in *Proc. ACM SIGKDD Workshop Intell. Secur. Informat. (ISI-KDD)*, 2012, pp. 1–7.

[34] I. Sadgali, N. Sael, and F. Benabbou, "Performance of machine learning techniques in the detection of financial frauds," *Procedia Comput. Sci.*, vol. 148, pp. 45–54, Jan. 2019.

[35] A. Santra and C. J. Christy, "Genetic algorithm and confusion matrix for document clustering," *Int. J. Comput. Sci. Issues*, vol. 9, no. 1, p. 322, 2012.

[36] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.

[37] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," *Secur. Informat.*, vol. 3, no. 1, pp. 1–10, Dec. 2014.

[38] A. Shimoda, T. Mori, and S. Goto, "Extended darknet: Multi-dimensional internet threat monitoring system," *IEICE Trans. Commun.*, vol. E95.B, no. 6, pp. 1915–1923, 2012.

[39] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, Nov. 2019.

[40] M. Spitters, F. Klaver, G. Koot, and M. van Staalduinen, "Authorship analysis on dark marketplace forums," in *Proc. Eur. Intell. Secur. Informat. Conf.*, Sep. 2015, pp. 1–8.

[41] D. Stupples, "ICITST-2013: Keynote speaker 2: Security challenge of TOR and the deep web," in *Proc. 8th Int. Conf. Internet Technol. Secured Trans. (ICITST-)*, Dec. 2013, p. 14.

[42] S. Tangirala, "Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020.

[43] M. C. Van Hout and T. Bingham, "Responsible vendors, intelligent consumers: Silk road, the online revolution in drug trading," *Int. J. Drug Policy*, vol. 25, no. 2, pp. 183–189, Mar. 2014.

[44] G. Weimann, "Going dark: Terrorism on the dark web," *Stud. Conflict Terrorism*, vol. 39, no. 3, pp. 195–206, Mar. 2016.

[45] C. C. Yang, X. Tang, and X. Gong, "Identifying dark web clusters with temporal coherence analysis," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jul. 2011, pp. 167–172.

[46] L. Yang, F. Liu, J. M. Kizza, and R. K. Ege, "Discovering topics from dark websites," in *Proc. IEEE Symp. Comput. Intell. Cyber Secur.*, Mar. 2009, pp. 175–179.

[47] Y. Zhang, S. Zeng, L. Fan, Y. Dang, C. A. Larson, and H. Chen, "Dark web forums portal: Searching and analyzing Jihadist forums," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jun. 2009, pp. 71–76.

[48] Y. Zhou, J. Qin, G. Lai, and H. Chen, "Collection of U.S. Extremist online forums: A web mining approach," in *Proc. 40th Annu. Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2007, p. 70.

[49] Y. Zhou, E. Reid, J. Qin, H. Chen, and G. Lai, "US domestic extremist groups on the web: Link and content analysis," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 44–51, Sep. 2005.

[50] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, "Surfacing collaborated networks in dark web to find illicit and criminal content," in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 109–114.

[51] Y. Fang, Y. Guo, C. Huang, and L. Liu, "Analyzing and identifying data breaches in underground forums," *IEEE Access*, vol. 7, pp. 48770–48777, 2019.
[52] N. Ferry, T. Hackenheimer, F. Herrmann, and A. Tourette, "Methodology of dark webmonitoring," presented at the 11th Int. Conf. Electron., Comput. Artif. Intell. (ECAI), 2019.
[53] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, 2013.

SAIBA NAZAH received the B.Sc. degree in computer science and engineering from Chittagong University of Engineering and Technology, Bangladesh. She is currently pursuing the master's degree with the School of Information Technology, Deakin University, Australia, with a focus on cyber security under the supervision of Prof. Dr. Jemal H. Abawajy. Her main research interests include cyber security, dark net, natural language processing, machine learning, the Internet of Things, and data analytics.

SHAMSUL HUDA received the Ph.D. degree in computer science, in 2010. He is currently a Lecturer with the School of Information Technology, Deakin University, Australia. Prior to join Deakin, he worked as an Academician with Federation University and as an Assistant Professor with Khulna University of Engineering and Technology (KUET), Bangladesh. He is a Certified Information System Security Professional (CISSP) by the International Information System Security Certification Consortium, $(ISC)^2$. He is also a member of the Cyber Security Research and Innovation Centre (CSRI), Deakin University. He is involved in many international cyber security projects, including cybersecurity capacity maturity for nations at Oceania Cyber Security Centre (OCSC), Melbourne with partnership of the Global Cyber Security Capacity Centre (GCSCC), University of Oxford. He has published more than 60 journal articles and conference papers in well reputed journals, including IEEE TRANSACTIONS. His research interests include communication and network security, strategies for secure operations for industrial control systems (SCADA) and critical infrastructure, intelligent counter measure for threats against mobile systems, detection of data breaches through the darknet, the IoT security, malware analysis and detection, reverse engineering for endpoint security, and malware analysis and detection for SCADA systems.

JEMAL H. ABAWAJY is currently a Full Professor with the Faculty of Science, Engineering and Built Environment, Deakin University, Australia. He has authored/coauthored over 250 refereed articles and supervised numerous Ph.D. students to completion. He has also served on the editorial board of numerous international journals, including the IEEE TRANSACTIONS ON CLOUD COMPUTING. He has delivered over 50 keynote and seminars worldwide and has been involved in the organization of over international conferences in various capacity, including the chair and the general co-chair.

MOHAMMAD MEHEDI HASSAN (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in February 2011. He is currently a Full Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He has authored or coauthored more than 260 publications, including refereed journals (more than 218 SCI/ISI-indexed journal articles, more than four ESI highly cited articles, and one hot article), conference papers, books, and book chapters. He is listed as one of the top 2% Scientists of the world in networking and telecommunication field. His research interests include cloud/edge computing, the Internet of Things, artificial intelligence, body sensor networks, big data, mobile computing, and cyber security. He has served as the chair and a technical program committee member in numerous reputed international conferences/workshops. He was a recipient of a number of awards, including the Distinguished Research Award from the College of Computer and Information Sciences, KSU 2020, Best Conference Paper Award from IEEE International Conference on Sustainable Technologies for Industry 4.0 (STI) 2020, Best Journal Paper Award from IEEE SYSTEMS JOURNAL, in 2018, the Best Conference Paper Award from CloudComp, in 2014 conference, and the Excellence in Research Award from the College of Computer and Information Sciences, KSU, in 2015 and 2016.