International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 5s
DOI: https://doi.org/10.17762/ijritcc.v11i5s.6650
***Article Received**: 01 March 2023 **Revised**: 10 April 2023 **Accepted**: 27 April 2023*

An Enhanced Scammer Detection Model for Online Social Network Frauds Using Machine Learning

Smita Bharne[1], Pawan Bhaladhare[2]
[1]School of Computer Sciences and Engineering, Sandip University/
Ramrao Adik Institute of Technology, D. Y. Patil Deemed to be University
Nashik, India
smita146@gmail.com
[2]School of Computer Sciences and Engineering
Sandip University
Nashik, India
pawan_bh1@yahoo.com

Abstract— The prevalence of online social networking increase in the risk of social network scams or fraud. Scammers often create fake profiles to trick unsuspecting users into fraudulent activities. Therefore, it is important to be able to identify these scammer profiles and prevent fraud such as dating scams, compromised accounts, and fake profiles. This study proposes an enhanced scammer detection model that utilizes user profile attributes and images to identify scammer profiles in online social networks. The approach involves preprocessing user profile data, extracting features, and machine learning algorithms for classification. The system was tested on a dataset created specifically for this study and was found to have an accuracy rate of 94.50% with low false-positive rates. The proposed approach aims to detect scammer profiles early on to prevent online social network fraud and ensure a safer environment for society and women's safety.

Keywords- Cyber security, Scammer profiles, Online social network frauds, Scammer detection model, Social threats, Compromised accounts, Fake profiles, Dating Fraud, Machine Learning.

I. Introduction

An online social network is medium that enables users to create personal profiles, connect with other users, and share information, content, and experiences with their network. Online social networks (OSN) can take various forms, including platforms (such as Facebook, Instagram, and Twitter,), dating platforms (such as Tumblr, Tinder, Bumble, Hinge, etc.), professional networking sites (such as LinkedIn), and online forums or discussion boards. They offer a range of features and tools that allow users to create and manage their online presence, build and maintain relationships with others, and access a wealth of information and resources [1]. Some of the key features of online social networks include: *a) Personal profiles*: users can create personal profiles that include information about themselves, such as their name, age, interests, and location. *b) Connections*: users can connect with other users on the platform, typically by sending friend or connection requests. *c) Sharing*: users can share various types of content, such as photos, videos, links, and status updates, with their network. *d) Communication*: Users can communicate with other users through messaging, commenting, and other forms of online communication. *e) Privacy*: most online social networks offer privacy settings that allow users to control who can see their content and interact with them on the platform. Overall, online social networks have transformed the way people connect and interact with each other, providing new opportunities for communication, collaboration, and socialization in the digital age. As the popularity of these OSN platforms increases, and fraudsters are taking advantage of the large number of users profiles to make OSN frauds. Online social network frauds happen for a variety of reasons, but most often they are motivated by financial gain or a desire to exploit other users for personal or professional gain. As fraudsters are creating the scam profile (false identity), it is typically for the purpose of deceiving or manipulating other users [2[[3]. Scam profiles may be used to perpetrate human-targeted frauds or other fraudulent activities, such as cyberbullying, dating fraud, compromised accounts, fake profiles, etc. It is important for users to be aware of these risks and to take steps to protect themselves and their personal information online. At present, social networking platforms do not offer any alerts or notifications to their users regarding the genuineness of profiles. This makes it difficult for inexperienced users to distinguish between genuine and fake profiles.

OSN sites like Twitter, Facebook, Instagram, Tinder, and others have similar user profile attributes like demographics, profile images, and short text descriptions worldwide. One of the challenges associated with these websites is that they enable users to generate numerous accounts, which gives rise to the issue of identifying replicated profiles. For instance, if a user already has an account on OSN websites or any dating websites, they can easily create another profile, making it challenging to detect duplicate profiles on these platforms. Besides the popularity of the OSN platforms, online dating websites have grown in popularity over the past decade, with more and more people turning to the internet to find romantic partners. Here are some key statistics that illustrate the popularity of online dating websites. According to a survey, 30% of people in the U.S. are using online dating services, and practice is more common among younger adults, with 48% of 18- to 29-year-olds reporting using online dating services [4]. The rise of mobile dating apps has also contributed to the popularity of online dating apps like Tinder, Bumble, and Hinge, which make it easier than ever to connect with potential partners from your phone. Thus, the rise of scammer profiles on these dating platforms has also increased over the past decade. Human-targeted frauds like online dating fraud, compromised accounts, and cloned profiles are types of fraud that involve the manipulation or exploitation of individual human beings through technological or systemic vulnerabilities. It is a form of social engineering fraud that relies on the natural tendencies of human beings to trust and cooperate with others [3][6][7]. Human-targeted frauds can have a significant impact on the safety and well-being of women, as they are often specifically targeted by scammers and fraudsters, where scammers create fake online profiles to lure victims into romantic relationships with monetary benefits. These OSN frauds can be emotionally devastating and result in significant financial losses. In this study, we aimed to analyze user profile attributes and develop an early-stage enhanced scammer detection model for detecting scammer profiles for OSN frauds. We extracted user profile attributes and utilized machine learning classification techniques to identify and detect scammer profiles.

II. Related Work

Scammer profile detection has been a widely researched topic in recent years, with numerous studies proposing various techniques to detect fraudulent profiles on social networking sites. In the literature, several aspects of the scam profiles—cloned accounts, compromised accounts, fake profiles, fake dating profiles—are detected with the different OSNs like Facebook, Twitter, LinkedIn, dating sites, etc. The author in [8] proposed a technique based on Markov clustering (MCL) to detect scam profiles on online social networking websites. The detection is based on features like active friends, page likes, and URLs. The authors' analysis revealed that fake or spam accounts tend to share URLs more frequently than authentic users. The author [9] presented a model called "FakeBook" for detecting false identities from social media websites. They pointed out that most research on OSN websites has focused on privacy protection. The other users who do not have accounts on OSN are at risk due to scam profiles. To detect such fake profiles, the authors collected Facebook user data and developed the FakeBook model, which uses characteristics derived from real user accounts for real-time temporal evaluation. The author in [10] put forth a model for detecting scam profiles that employed a pattern-matching algorithm. By utilizing map-reducing techniques and a pattern recognition approach, they were able to identify a subset of fake user accounts that were highly reliable. The author in [11] utilized a clustering technique to distinguish between fake and genuine accounts in Twitter profile attributes. The system was able to successfully detect a subset of fake users, all of whom were later verified manually. A study by the author [12] focused on identifying fake users by analyzing their behavior in OSN. They employed various machine learning-based classifiers, such as support vector machines (SVM), random forests (RF), and adaptive boosting algorithms. The best results are given by the adaptive boosting classifier. The data on user nonverbal behavior was discovered to be valuable in detecting multiple accounts and false identity deception. Author [13] employed a model for identifying counterfeit accounts based on graph models along with weights feature. The weight values of the nodes in the graph were utilized to identify potential targets of the fake account. In [14], the author presented a model that utilizes regular expression (RE) and deterministic finite automaton (DFA) techniques for detecting scam profiles and verifying identities on social networking sites. Author [15] proposed a fake profile detection method called SybilWalk, in which the random walk technique was utilized to identify false accounts on OSN websites. The users and their relationships were represented as nodes and edges in the dataset. Additionally, the author [16] presented a neural network (NN)-based model for identifying fake profiles and bots on Twitter. Their method uses dimensionality reduction techniques to increase the performance of the system. An author in [17] developed a supervised machine learning-based approach to detect false identities on social media websites. Then, they validated the data with the social status of users with fabricated accounts. Authors in [18] proposed a method to detect duplicate profiles with profile comparison and communication matching methods. To detect duplicate profiles, they used node similarity communication matching. A study done by the author in [19] used data mining techniques to detect duplicate profiles in social networks. The dataset was obtained from the GitHub repository. The author utilized various machine learning algorithms and privacy-protected methods to identify false accounts. The author [20] proposed a model to detect false accounts on Facebook by studying users' emotions through a supervised machine-learning approach. Various emotion-based features like antagonism, desolation, fear, happiness, faith, expectation, etc. were used for the study. Authors in [21] introduced a Deep Learning (DL) model named "DeepProfile", which is specifically designed to detect false profiles on OSN. They made a significant modification to the pooling layer of the convolutional neural network (CNN), which was unique to their approach. In [22], the authors presented a model for identifying false accounts on Facebook and assessed the efficacy of Facebook's fake account detection algorithms along with the platform's artificial intelligence learning abilities in differentiating between genuine and false accounts. The author in [23] presented a model for detecting fake accounts on OSN websites using a blacklist method instead of a conventional spam word list. They created a blacklist by employing topic modeling and keyword extraction approaches. To detect automation on Twitter, the author [24] utilized natural language processing (NLP). Their model relied on natural-language text produced by humans as a standard for identifying accounts that use automated messages. The author in [25] proposed a model to distinguish between scammers, bloggers, and real experts on Twitter. Their approach utilized the Hyperlinking topic search algorithm to classify spam profiles and separate them from the actual specialists in a particular field. The model was designed to require less pre-classified data for accurate differentiation between bloggers and true experts. The author in [26] proposed a method to detect compromised accounts in OSN like Facebook and Twitter. However, their model has a limitation where attackers who are aware of the model's functionality can take steps to evade the detection of their compromised accounts. The study in [27] employed a technique to identify the compromised profiles on OSN by analyzing the history of a user. Their model considers seven features, including time, message source, message body text, message subject, message connection, direct user interaction, and proximity, to determine whether the account is compromised or not. The authors in [28] utilized the GAIN measure technique to assign weights to all attributes in the training dataset. These weights determine the significance of each attribute in the ML classification algorithm. RF gives the best performance in their study. An author in [29] developed a framework known as Social Profile Abuse Monitoring, which involves collecting data from 5000 Twitter users' profiles and their 200 most recent tweets. They analyzed the dataset using an SVM classifier and introduced a four-class classifier model using similarity profile calculations based on similarity interface characteristics. A study by the authors in [30] proposed a model for identifying spam profiles on the OSN application using an associative classification model. This model is slower and can improve performance based on given parameters. The author in [32] presented a methodology for classifying both spam messages and spam accounts on OSN websites. The model extracted 18 features, including behavior-content structures, to identify spammers using machine learning (ML) algorithms were used, and the best classification is done by the RF. The author in [33] proposed a hybrid technique for detecting spammer profiles that utilize user and graph-based features. Their model is able to discriminate between spam and non-spam by analyzing these features. These features are based on the relationships and properties of user accounts and are considered essential for an effective spam detection system. A comprehensive study done by the author in [43] for false profile detection in OSN summarized the identification of scam profiles using textual and non-textual features using ML and DL. In a study, authors [34] put forward a technique for identifying scammers on dating websites by utilizing a ML classifier. They employed ML-based classifiers such as SVM and decision trees to detect images associated with dating fraud. Furthermore, they compared the performance of various ML-based classifiers to obtain the most precise outcomes. Author [35] proposed the methods to detect scam profiles based on the user profiles attributes and profile descriptors for online dating platforms. The use of celebrity photos by scammers as profile pictures is a common tactic to mask their identity. To tackle this issue, a study in [36] focused on utilizing ML technology to scrape data from online dating sites and detect faces. The proposed model aims to assist users in identifying fake profiles based on two critical factors: whether the profile utilizes celebrity images and whether it includes any photos from external websites. However, very limited efforts have been made in the literature to identify scam profiles on dating websites.

An analysis of the literature review indicates that several techniques and models have been proposed to detect scammer profiles on OSN. Machine learning algorithms are still finding limits in recognizing scam profiles in OSN scams due to a lack of labeled data and a limited set of attributes. To prevent online social networking frauds (human-targeted frauds), we proposed an enhanced scammer detection model to detect a scammer and genuine user profile using user profile-based attributes and images at an early stage which can work on cross platforms also.

III. Proposed Work

We utilized an enhanced scammer detection model framework, as shown in Figure 1. An approach was developed to classify the scammer profiles using various user profile attributes and characteristics which can be effective for cross-platform like with OSN and dating websites. An integrated model was made by combining Naive Bayes, Support vector machines (SVM), Decision Trees (DT), and the Random Forest (RF) technique in order to produce the best classification results [42].

A. Data Collection

We constructed the datasets from scamdigger.com for scam user-profiles and datingnmore.com for real user profile data. The downloaded dataset is in raw format, which is then converted into the textual CSV format after pre-processing. The

DOI: https://doi.org/10.17762/ijritcc.v11i5s.6650
***Article Received**: 01 March 2023 **Revised**: 10 April 2023 **Accepted**: 27 April 2023*

dataset contains user profile data with more than 12000 unique user-profiles and the corresponding number of profile images. This textual data file contains the OSN user profile-centric information, which refers to the data and attributes related to the user's profile on OSN. The user profile textual data contains the demographic information such as age, gender, location, education, and occupation, as well as personal interests, hobbies, and relationship status in short profile descriptions. The dataset also contains scam and real user profile images. The constructed dataset contains the profile attributes, similar to the Facebook, Instagram, and other social networking sites, which also correspond to the profiles on dating websites. Thus, our aim is to make a scammer detection model that can identify the scammer profile based on these common profile attributes at an early stage in which can used in cross platform also. The dataset information is based on demographic attributes, profile images, and short profile descriptions.

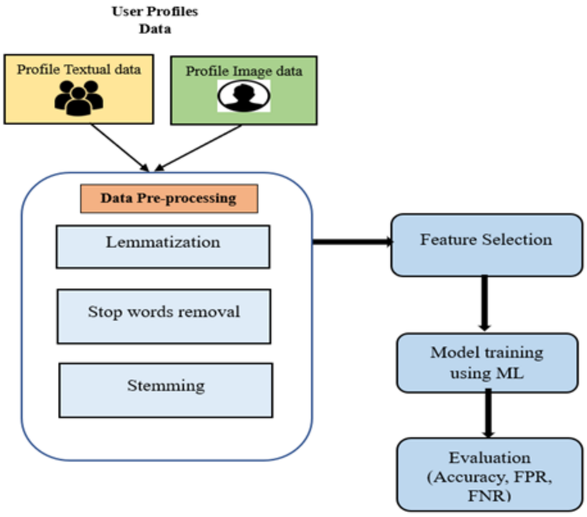

Figure 1. Proposed Model Workflow

B. *Data Pre-processing*

To detect scammers effectively, it is crucial to pre-process the data by cleaning, transforming, and structuring the raw text data for analysis. As a first step, the text files in the dataset need to be pre-processed before they can be used to train machine learning algorithms. One effective way to improve the performance of text classification is through text normalization techniques such as stemming, lemmatization, and stop word removal [37][38][39][40]. These techniques were applied to the data to investigate whether they could improve the detection model's performance. Below are the explanations of these techniques.

1) ***Stemming**:* is a text normalization technique used in the pre-processing of user profile data to improve the detection model's performance. Stemming is a natural language processing technique used to reduce words to their base or root form, known as a stem. The process involves removing any affixes, prefixes, or suffixes from a word in order to convert it into its simplest form. P represents the user profiles set and S represents is possible stems set. We can define stemming as a function that maps each word in a user profile to its corresponding stem given in equation 1.

$$\text{stem}(p)= [\text{stem}\ (p\ [1]\ \ldots\ldots\ldots\ \text{stem}(p[m])] \qquad (1)$$

where p is a profile of user, m is the words in p, and stem(p[i]) is the stem of the n-th word in p. A snowball stemmer is used with language used as English to process the text data to improve results.

2) ***Lemmatization**:* it reduces the number of unique words while preserving the intended meaning of the text. Additionally, lemmatization provides better results compared to stemming since it maintains the grammatical accuracy of words. P is a set of user profiles, and L is the set of possible lemmas. We can define lemmatization as a function that maps each word in a user profile to its corresponding lemma given in equation 2:

$$\text{lemma}(p)=[\text{lemma}(p[1]\ldots\ldots\ldots\text{lemma}(p[m])] \qquad (2)$$

where p is a user profile and m is the number of words in p, lemma(p[i]) is the lemma of the nth word in p.

3) ***Stop word removal:*** taking off stop words from text input can assist reduce noise and improve the detection model's performance. P is a collection of user profiles, while SW is a collection of stop words. Stop word removal is defined as a function that eliminates all stop words from a user profile as in equation 3.

$$\text{stop\_word\_removal}(p) = [\ w \mid w \text{ in } p \text{ if } w \text{ not in } SW] \qquad (3)$$

The stop_word_removal function works by iterating over each word in the user profile and checking if it is in the set of stop words. If the word is a stop word, it is excluded from the resulting list of words. Otherwise, the word is included in the output.

C. *Feature Extraction*

The bag-of-words, n-grams, TF-IDF, and word2vec methods are used to choose the demographic elements of user profiles, such as username, age, gender, location, marital status, and short descriptions. [39][40]. For user profile image features classification, the YOLOv4 (You only look once) algorithm is used [41]. The description is given below.

*1) **Bag of words** :* is a text representation technique in natural language processing. It involves converting a piece of text into a collection, or bag of its constituent words, disregarding the order in which they appear but retaining information on the frequency of each word. The system uses a set of user profiles (P and V) as a vocabulary of words. We can represent each user profile as a vector of word counts using the bag of words approach. Specifically, we define a function that takes a user profile as input and returns a vector of word counts given in equation 4.

$$\text{Bag\_of\_words}(p)=\{\text{count}(w,p) \mid w \text{ in } V] \qquad (4)$$

where p is a user profile, w is a word in the vocabulary V, and count (w, p) is the number of times the word w is found in user profile p. The function Bag_of_words operates by iterating through each word in the vocabulary V and tallying up the frequency of occurrences for each word in the user profile p. The resulting vector of word counts represents the user profile in terms of the frequency of each word in the vocabulary.

*2) **N-grams***: In the case of user profiles, an N-gram is a contiguous sequence of n words from a given model of text. The value of n determines the size of the N-gram. To use N-grams for user profile analysis, each profile p in a set P is tokenized into a list of words, and then N-grams of size n are created from the list of words in each profile. The feature selection process involved identifying the most frequently occurring n-grams in the trained dataset. These n-grams were then selected as the features for the model.

*3) **TF-IDF***: stands for Term Frequency-Inverse Document Frequency (TF-IDF). It represent textual data as vectors of weighted features, where each feature corresponds to a term and its weight is the TF-IDF score for that term in the textual data. P is a set of user profiles, and t is a term (i.e., a word or phrase) that occurs in one or more of the user profiles in P. The score of t in a particular user profile p is given in equation 5 as :

$$\text{TF-IDF}(t, p) = \text{TF}(t, p) \times \text{IDF}(t) \qquad (5)$$

where IDF(t) is the inverse document frequency of t over all user profiles in P, and TF (t, p) is the term frequency of t in p. The inverse document frequency of t, denoted IDF(t), is calculated as follows in equation 6:

$$\text{IDF}(t)=\log(N/n\_t) \qquad (6)$$

where N is the total no. of profiles in P, and n_t is the no. of user profiles in P that contain the term t. The TF-IDF score of t in p provides a measure of how important t is to p, relative to its importance in other user profiles in P. A high TF-IDF score indicates that t is both frequent in p and rare in other user profiles, which suggests that t may be a distinguishing feature of p.

*4) **Word2Vec**:* is a natural language processing technique used to represent words in space, where each word is mapped to a vector of real numbers. The technique is based on a neural network architecture that learns to predict the context in which a word appears. P is a set of user profiles, and w is a word in a user profile. The Word2Vec model aims to learn a vector representation, denoted v(w), for each word w in a given space, such that the vector representations of words that are semantically similar are closer together in the space. The representation of the Word2Vec model is as follows: The dimensionality of the vector space is D, and v(w) denotes a D-dimensional vector representing the word w. The objective of word2vec is to learn these vector representations by function in equation 7.

$$J = -1/N^* \sum\{i=1 \text{ to } N)\sum\{j=1\text{to } C) \log(p(w\_j|i) \qquad (7)$$

where N is the total number of user profiles in P, C is the size of the context window (i.e. the number of words on either side of the central word), and p(w_j|i) is the conditional probability of observing the word w_j in the context of the central word w_i

*5) **Image detection using YOLO v4 algorithm***: is a cutting-edge algorithm for detecting objects in images through the use of a deep learning network. The algorithm leverages multiple layers of CNN to perform the necessary transformations [41]. The network is trained on 12000 plus of datasets of images and learns to identify objects based on their visual features. The algorithm is designed to be fast and accurate, and it can detect multiple objects in an image in real-time[44]. We customised the framework for the YOLOv4 algorithm as per the images on our dataset. The layer-wise description is given below.

a) Convolutional Layers: x is the input image, and W is a set of convolutional filters. The output of a convolutional layer is given by the convolution of the input image with the set of filters as given in equation 8.

$$y=W^*x \qquad (8)$$

where * denotes the convolution operation.

b) Activation Functions: The output of a convolutional layer is passed through an activation function f(x) to introduce non-linearity into the network. The activation function

Rectified Linear Unit (ReLU) is used on the output of each convolutional layer is given in eqaution 9

$$y=\max(0, x) \tag{9}$$

where max (0, x) returns x if x is positive and 0 otherwise.

c) Pooling Layers: : utilized to decrase the spatial dimensions of feature maps produced by convolutional layer. The output is represent as in equation 10:

$$y=\text{max\_pool}(x, \text{pool\_size}, \text{stride}) \tag{10}$$

where max_pool is the max pooling operation, pool_size is the size of the pooling region, and stride is the stride of the pooling operation.

d) Fully Connected Layers: W is a set of weights for the fully connected layer, and x is the output of the final convolutional layer. The output of the fully connected layer represent in equation 11.

$$y=W*x \tag{11}$$

where * denotes the matrix multiplication operation.

e) Loss Function: Let y_pred be the predicted class probabilities and bounding box coordinates, and y_true be the true values for these variables. The loss function used in the YOLOv4 algorithm is given by the sum of squared errors (SSE) loss as given in equation 12:

$$L= (y\_\ pred- y\_\ true)\ \hat{}\ 2 \tag{12}$$

where ^2 denotes element-wise squaring.

We customize the darknet neural network on our image dataset with a learning rate is 0.01. The rate of momentum during the training of network is 0.9. Darknet neural network framework consist of different layers is shown in table I.

TABLE I. DARKNET NEURAL NETWORK

Type	Filters	Size	Output
Convolutional layer	32	3 x 3/ 1	180 x 180 x 3
Maxpool layer		2x 2/ 2	180 x 180 x 32
Convolutional layer	16	1 x 1/ 1	90 x 90 x 32
Convolutional layer	64	3 x 3/ 1	90 x 90 x 16
Maxpool layer		2x 2/ 2	90 x 90 x 64
Convolutional layer	32	1 x 1/ 1	45 x 45 x 64
Convolutional layer	128	3 x 3/ 1	45 x 45 x 32
Convolutional layer	64	1 x 1/ 1	45 x 45 x 128
Avgpool		45 x 45 x 2	
Softmax		45 x 45 x 2	

D. Model Training and Evaluation

There are several machine learning classifiers that are utilized to recognize the scammer profiles. On online social media platforms, there are a group of k people with P profiles: p1, p2, p3,... pk. User profiles p contain profile data such the user's name, gender, age, etc. The scammer detection model profile aims to determine if the user profile is a scam profile or a genuine profile based on the profile's textual attributes (pi) and image attributes (pt). A machine-learning classifier M is constructed to extract the set of m features F = f1, f2, f3,… fm from P.

1) *Profile Classification Algorithm:*

D is the dataset of k user profiles, where each profile p has demographic features f_p and a profile image I_p.

X is the set of extracted features from the user profile attributes data and profile images for each profile p in D, where X_p = (x_p, pro, x_p, img) is a vector of demographic and image features for profile p.

Y is the set of labels for each profile p in D, where Y_p = 1 if profile p is a scam and Y_p = 0 if profile p is legitimate.

M be a binary classification model that takes as input X and outputs a probability that a profile is a scam, i.e., M(X) = P(Y=1|X).

Algorithm Steps:

Step 1. Input: D

Step 2: Pre-process demographic features from files and profile images I_p to obtain X_p = (x_p, pro, x_p, img) for each profile p in D.

Step 3: Extract relevant features from X_p to obtain X for all profiles in D.

Step 4: Train a binary classification model M on (X, Y) using an NB, SVM, DT and RF algorithm.

Step 5: Assess the performance of the best trained model M on a validation dataset using accuracy, false positive rate, and false negative rate metrics.

Step 6: Given a new profile p with demographic features f_p and profile image i_p, extract features X_p = (x_p, pro, x_p, img) and use M to predict the probability that p is a scam: M(X_p) = P(Y=1|X_p).

Step 7: Classify p as legitimate or a scam based on the predicted probability.

2) *Machine learning algorithms*

Machine learning algorithms are evaluated to discover the best outcomes. The feature vector X = x1…….. xn that represents a user profile, where each feature x_i is a binary variable that indicates the particular keyword or object in the profile text or image, and Y be a binary variable that indicates whether the profile is a scam or not. The aim is to calculate the P(Y=1|X), the probability that the profile is a scammer profile given the feature vector X.

a) *Naive Bayes*

The Naive Bayes algorithm operates on the assumption that the features are conditionally independent when given the class variable Y. This assumption simplifies the calculation of the posterior likelihood as given in equation 13.

$$P(Y=1|X) = P(Y=1) * P(X|Y=1)/P(X) \qquad (13)$$

where P(Y=1) is the prior probability of a scam profile, P(X|Y=1) is the likelihood of the feature vector given a scam profile, and P(X) is the marginal probability of the feature vector.

b) *Support Vector Machine*

By applying kernel functions to map the data in feature space, the SVM technique provides a robust and adaptable tool for identifying scam profiles based on their text and image properties. The goal is to find a decision boundary that separates the two classes in the feature space with maximum accuracy. The decision boundary is represented by a hyperplane in equation 14.

$$w\hat{}T\ x + b = 0 \qquad (14)$$

where w is representing weight and b is used as bias.

The SVM algorithm solves an optimization problem to find the weight vector and bias term that maximize the margin while ensuring that the data points are correctly classified as given in equation 15 and 16.

$$\text{minimize:} \qquad \|w\|\hat{}2/\ 2 \qquad (15)$$

$$\text{subject to:} \qquad y\_i\ (w\hat{}T\ x\_i + b) >= 1 \qquad (16)$$

where y_i is the class label (+1 for scam profiles, -1 for non-scam profiles), and the inequality constraint ensures that the data points are correctly classified.

c) *Decision tree*

The Decision Tree algorithm's goal is to construct a tree-like model that predicts the value of Y based on the values of X. It represents a tree-like model which recursively partitions the data into subsets based on feature values, then labels each subset based on the majority class. The algorithm selects the feature that best splits the data at each node of the tree depending on some set of criteria, such as information gain or Gini impurity. The split is chosen to maximize the homogeneity of the subsets with respect to the class label while minimizing the impurity or entropy of the subsets. The decision tree can be represented as a set of if-then rules where each internal node represents a split on a feature and each leaf node represents a class label as shown in equation 17.

$$\text{if } x\_i <= t\_1 \text{ then if } x\_j <= t\_2 \text{ then ... else ... else ...} \qquad (17)$$

where each internal node represents a split on a feature x_i and each leaf node represents a class label y_k. To classify a new profile, we traverse the decision tree by comparing the values of the features with the split thresholds until we reach a leaf node with a class label. The profile is then classified as a scam or non-scam based on the majority class of the leaf nodes in the subtree.

d) *Random Forest*

Multiple decision trees are combined in the RF algorithm, an ensemble learning technique, to increase the predictability and accuracy. The method involves constructing a forest of decision trees, where each tree is trained on a randomly selected subset of the training data and features. Based on criteria like information gain or Gini impurity, the algorithm chooses the feature at each node of each decision tree that best divides the data. The split is chosen to maximize the homogeneity of the subsets with respect to the class label while minimizing the impurity or entropy of the subset. The random forest can be represented as a set of decision trees in equation 18:

$$f(x)\backslash = 1/T * sum\_i=1\hat{}T\ (f\_i(x)) \qquad (18)$$

where f_i(x) is the prediction of the i-th decision tree, and T is the total number of trees in the forest. To classify a new profile, we apply each decision tree in the forest to the profile, and take the majority vote of the class labels across all trees. The profile is then classified as a scam or non-scam based on the majority vote.

IV. RESULTS AND DISCUSSIONS

The machine learning algorithm mentioned in this paper were trained utilizing profile feature sets, and their performance was evaluated using k-fold cross-validation to determine the best parameter choices. The optimal parameter setting was chosen based on accuracy, and the data was separated into 80% training sets and 20% testing sets for performance evaluation. The model evaluation parameters are the accuracy, false positive rate (FPR), and false negative rate (FNR). Scam profiles received positive labels, whereas legitimate profiles received negative marks. The effectiveness of the models is evaluated with feature sets and parameter settings.

TABLE II. ACCURACY WITH SEPARATE UNI, BI, TRI GRAMS

Set of Feature [unigram, bi-gram and, tri-grams]	NB	SVM	DT	RF
[3,0,0]	0.794	0.867	0.883	0.919
[4,0,0]	0.845	0.887	0.883	0.912
[5,0,0]	0.869	**0.889**	0.842	0.912
[7,0,0]	0.881	0.884	0.869	0.912
[10,0,0]	0.896	0.885	0.895	0.915
[0,3,0]	0.818	0.815	0.834	0.831
[0,4,0]	0.816	0.839	0.821	0.831
[0,5,0]	0.862	0.837	0.867	0.832
[0,7,0]	0.873	0.842	0.856	0.837
[0,10,0]	0.883	0.853	0.842	0.894
[0,0,3]	0.783	0.777	0.73	0.826
[0,0,4]	0.796	0.805	0.805	0.829
[0,0,5]	0.807	0.809	0.832	0.835
[0,0,7]	0.793	0.818	0.819	0.834
[0,0,10]	0.802	0.83	0.835	0.851
TF-IDF [3]	0.826	0.863	0.893	0.913
TF-IDF [4]	0.864	0.851	**0.897**	0.917
TF-IDF [5]	0.875	0.818	0.892	0.918
TF-IDF [7]	0.915	0.772	0.893	0.927
TF-IDF [10]	**0.919**	0.761	0.893	**0.928**

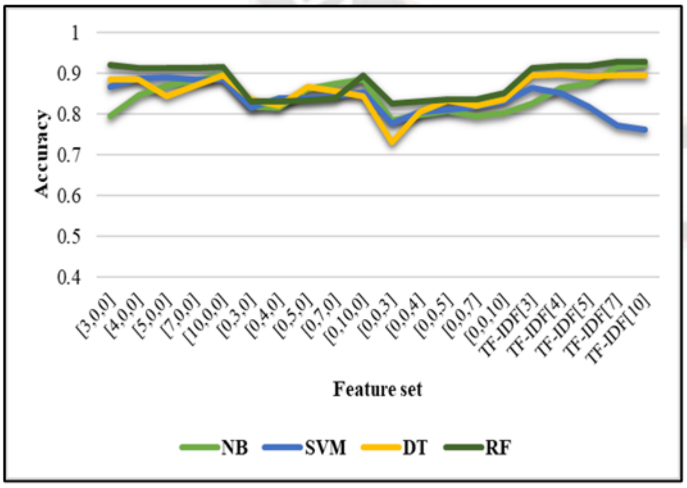

Figure 2. Comparison of accuracy by ML algorithm with separate feature set

1. Accuracy is the correctly identified d profiles within the total number of profiles as given in equation 19.

Accuracy= (TP+ TN)/ (TP+FN+TN+FN) (19)

2. FPR indicates no. of times real profile is considered as a scammer by the model as given in equation 20.

FPR= (FP/ actual negative N) (20)

3. FNR indicate no. of times a scam profile is not recognized by the model give in equation 21.

FNR= (FN/actual positive P) (21)

The overall accuracy for all different classes with separate unigrams, bigrams, and trigrams is shown in table II. By using the machine learning classifiers, the highest accuracy achieved by the random forest algorithm was 92.80%. The comparison of the accuracy achieved by the different ML classifiers with a separate feature set is shown in figure 2. This is the traditional approach used in the literature survey. This accuracy is improved by the feature sets of proposed methods, which combine the whole feature set as shown in table III. The comparison of accuracy by ML algorithm with combined feature set (word2Vec and n-grams) without profile image is shown in figure 3. Figure 4 shows the comparative analysis of the accuracy given by the ML algorithm with the proposed method features set (word2Ve, n-gram and profile image). In table IV is showing the performance s of the different ML algorithms along with the FPR and FNR.

TABLE III. MAXIMUM ACCURACY WITH PROPOSED METHOD

Set of Feature [n-grams, word2vec, profile image]	NB	SVM	DT	RF
n-gram [10]	0.925	0.922	0.92	0.933
word2vec [100]	0.926	0.916	0.928	0.938
n-gram [10]	0.928	0.935	0.933	**0.9455**

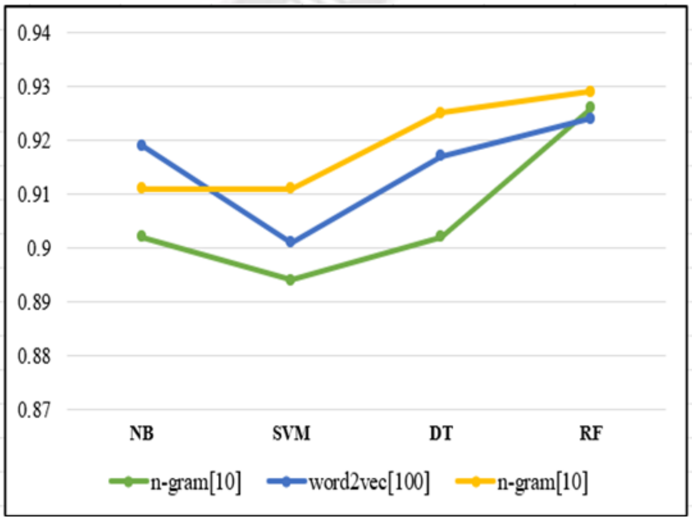

Figure 3. Comparison of accuracy by ML algorithm with combined features (word2Vec and n-gram) set

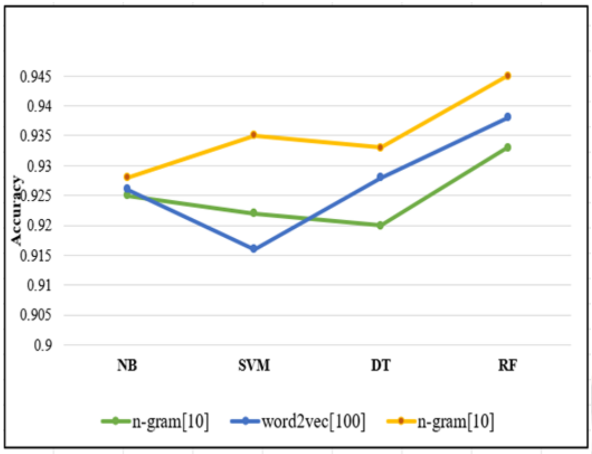

Figure 4. Comparison of accuracy by ML algorithm with proposed set of features

Our system's utmost accuracy is 94.55%, with a false positive rate of 0.01 and a false negative rate of 0.119. The top feature set combination is the feature set with n-grams and word2vec size 10 with stemming and stop word removal along with profile image.

TABLE IV. PERFORMANCE OF DIFFERENT ML ALGORITHM

Algorithm	Set of Feature [n-grams, word2vec, profile image]	Lemmatization Stemming	Stop word removal	Accuracy	FPR	FNR
Naive Bayes	n-gram[10]	lemmatization	yes	0.925	0	0.129
	word2vec[100]	lemmatization	yes	0.926	0.001	0.178
	n-gram[10]	stemming	yes	0.928	0	0.109
SVM	n-gram[10]	lemmatization	yes	0.922	0	0.109
	word2vec[100]	lemmatization	yes	0.916	0.001	0.133
	n-gram[10]	stemming	yes	0.935	0	0.161
Decision Tree	n-gram[10]	lemmatization	yes	0.92	0	0.123
	word2vec[100]	lemmatization	yes	0.928	0	0.19
	n-gram[10]	stemming	yes	0.933	0.001	0.118
Random Forest	n-gram[10]	lemmatization	yes	0.933	0	0.105
	word2vec[100]	lemmatization	yes	0.938	0	0.119
	n-gram[10]	stemming	yes	**0.9455**	0.001	0.119

V. CONCLUSION

This paper presents an enhanced scammer detection model for identifying scammer profiles based on the user profile features for online social network frauds such as online dating, false profiles, and compromised accounts. Identifying the scammer on OSN platforms is difficult due to the huge amount of OSN users data. The proposed approach is tested and evaluated on the dataset, which was created by us, using the evaluation parameters accuracy, FPR, and FNR. We use various techniques and algorithms, such as NB, SVM, DT, RF YOLOv4, to build effective classifiers that can detect scammer profiles based on user profile attributes and images. In order to increase the classifier's accuracy, pre-processing methods like stemming, lemmatization, stop word removal, bag of words, n-grams, and word2vec are also applied to the user profile characteristics. Using the proposed scammer detection model, we can efficiently identify and blacklist scammer profiles in early stages to prevent online social network frauds and provide a safety network for society. Future research can explore the use of multimodal techniques, such as fusion-based methods, to combine features from different modalities and build more effective scammer detection models.

REFERENCES

[1] Kayes, Imrul, and Adriana Iamnitchi. "Privacy and security in online social networks: A survey." Online Social Networks and Media 3 (2017): 1-21

[2] Guo, Zhen, Jin-Hee Cho, Ray Chen, Srijan Sengupta, Michin Hong, and Tanushree Mitra. "Online social deception and its countermeasures: A survey." IEEE Access 9 (2020): 1770-1806.

[3] Jain, Ankit Kumar, Somya Ranjan Sahoo, and Jyoti Kaubiyal. "Online social networks security and privacy: comprehensive review and analysis." Complex & Intelligent Systems 7, no. 5 (2021): 2157-2177.

[4] T. J. Holt and A. M. Bossler, The palgrave handbook of International cybercrime and cyberdeviance. 2020. doi: 10.1007/978-3-319-78440-3.

[5] Rathore, Shailendra, Pradip Kumar Sharma, Vincenzo Loia, Young-Sik Jeong, and Jong Hyuk Park. "Social network

security: Issues, challenges, threats, and solutions." Information sciences 421 (2017): 43-69.

[6] Apte, Manoj, Girish Keshav Palshikar, and Sriram Baskaran. "Frauds in online social net-works: A review." Social networks and surveillance for society (2019): 1-18.

[7] Cross C. (2020) Romance Fraud. In: Holt T., Bossler A. (eds) The Palgrave Handbook of International Cybercrime and Cyberdeviance. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-90307-1_41-1

[8] F. Ahmed and M. Abulaish, "An MCL-based approach for spam profile detection in online social networks," in Proc. IEEE 11th Int. Conf. Trust, Security. Privacy Comput. Commun., 2012, pp. 602–608

[9] M. Conti, R. Poovendran, and M. Secchiero, "Fakebook: Detecting fake profiles in on-line social networks," in Proc. IEEE/ACM Int. Conf. Adv. Social Network. Anal. Mining., 2012, pp. 1071–1078.

[10] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in Proc. Int. Conf. SocialMedia Soc., 2015, pp. 1–7

[11] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile characteristics of fake twitter accounts," Journal of Big Data, vol. 3, no. 2, 2016, 2053951716674236

[12] M. BalaAnand, S. Sankari, R. Sowmipriya, and S. Sivaranjani, "Identifying fake user's in social networks using non verbal behavior," International Journal Technol. Eng. Syst., vol. 7, no. 2, pp. 157–161, 2015.Technol. Eng. Syst., vol. 7, no. 2, pp. 157–161, 2015

[13] Y. Boshmaf et al., "Integro: Leveraging victim prediction for robust fake account detection in OSNs," in Proc.Netw.Distributed Syst. Secur. Symp., 2015, pp. 8–11

[14] M. Meligy, H. M. Ibrahim, and M. F. Torky, "Identity verification mechanism for detecting fake profiles in online social networks," Int. J. Comput. Netw. Inf. Secur., vol. 9, no. 1, pp. 31–39, 2017.

[15] J. Jia, B. Wang, and N. Z. Gong, "Random walk based fake account detection in online social networks," in Proc. 47th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw., 2017, pp. 273–284.

[16] S. Khaled, N. El-Tazi, and H. M. Mokhtar, "Detecting fake accounts on social media," in Proc. IEEE Int. Conf. Big Data., 2018, pp. 3672–3681

[17] N. Singh, T. Sharma, A. Thakral, and T. Choudhury, "Detection of fake profile in online social networks using machine learning," in Proc. IEEE Int. Conf. Adv. Comput. Commun. Eng., 2018, pp. 231–234.

[18] S. Revathi and M. Suriakala, "Profile similarity communication matching approaches for detection of duplicate profiles in online social network," in Proc. IEEE 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions, 2018, pp. 174–182.

[19] M. Suriakala and S. Revathi, "Privacy protected system for vulnerable users and cloning profile detection using data mining approaches," in Proc. IEEE 10th Int. Conf. Adv. Comput., 2018, pp. 124–132.

[20] M. A. Wani, N. Agarwal, S. Jabin, and S. Z. Hussain, "Analyzing real and fake users in Facebook network based on emotions," in Proc. IEEE 11th Int. Conf. Commun. Syst. Netw., 2019, pp. 110–117

[21] P.Wanda and H. J. Jie, "Deepprofile: Finding fake profile in online social network using dynamic CNN," Journal of Information Security, Appl., vol. 52, pp. 1–13, 2020

[22] P. Pourghomi, M. Dordevic, and F. Safieddine, "Facebook fake profile identification: Technical and ethical considerations," Int. Journal . Pervasive Comput. Commun., vol. 16, pp. 101–112, 2020.

[23] M. M. Swe and N. N. Myo, "Fake accounts detection on twitter using blacklist," in Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci., 2018, pp. 562–566. E. M. Clark, J. R.Williams, C. A. Jones, R. A. Galbraith, C.M. Danforth, and P. S. Dodds, "Sifting robotic from organic text: A natural language approach for detecting automation on twitter," J. Comput. Sci., vol. 16, pp. 1–7, 2016.

[24] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya "Segregating spammers and unsolicited bloggers from genuine experts on twitter," IEEE Trans. Dependable Secure Comput., vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018

[25] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, "Segregating spammers and unsolicited bloggers from genuine experts on twitter," IEEE Trans. Dependable Secure Comput., vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.

[26] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in Proc. Netw. Distrib. Syst. Secur. Symp., 2013

[27] P.V. Phad andM.Chavan, "Detecting compromised high-profile accounts on social networks," in Proc. IEEE 9th Int. Conf. Comput., Commun. Netw. Technol., 2018, pp. 1–4

[28] G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," Int. J. Inf. Technol. Knowl. Manage., vol. 2, no. 2, pp. 271–277, 2010.

[29] Chakraborty, J. Sundi, S. Satapathy, "SPAM: A framework for social profile abuse monitoring," Stony Brook Univ., Stony Brook, NY, USA, CSE508 Report, 2012.

[30] W. Hua and Y. Zhang, "Threshold and associative based classification for social spam profile detection on twitter," in Proc. 9th Int. Conf. Semantics, Knowl. Grids, 2013, pp. 113–120.

[31] K. S. Adewole, N. B. Anuar, A. Kamsin, and A. K. Sangaiah, "SMSAD: Aframework for spammessage and spam account detection," Multimedia Tools Appl., vol. 78, no. 4, pp. 3925–3960, 2019.

[32] E. V. D. Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," IEEE Access, vol. 6, pp. 6540–6549, 2018.

[33] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for Twitter," in Proc. 14th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2017, pp. 466_471

[34] Jong, Koen. "Detecting the online romance scam: Recognising images used in fraudulent dating profiles." Master's thesis, University of Twente, 2019.

[35] Suarez-Tangil, Guillermo, et al. "Automatically dismantling online dating fraud." IEEE Transactions on Information Forensics and Security 15 (2019): 1128-1137.'.

[36] S. Al-Rousan, A. Abuhussein, F. Alsubaei, O. Kahveci, H. Farra, and S. Shiva, "Social-Guard: Detecting Scammers in Online Dating", IEEE Int. Conf. Electro Inf. Technol., vol. 2020-July, no. August, pp. 416–422, 2020, doi: 10.1109/EIT48999.2020.9208268.

[37] Aytu_g Onan, Serdar Koruko_glu, and Hasan Bulut. Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57:232{247, 2016

[38] Liangxiao Jiang, Chaoqun Li, Shasha Wang, and Lungan Zhang. Deep feature weighting for Naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52:26{39, 2016.

[39] Bo Tang, Haibo He, Paul M Baggenstoss, and Steven Kay. "A Bayesian classification approach using class-specific features for text categorization", IEEE Transactions on Knowledge and Data Engineering, 28(6):1602{1606, 2016.

[40] Lungan Zhang, Liangxiao Jiang, Chaoqun Li, and Ganggang Kong. Two feature weighting approaches for naive bayes text classi_ers. Knowledge-Based Systems, 100:137{144, 2016.

[41] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).

[42] Mahesh, Batta. "Machine learning algorithms-a review." International Journal of Science and Research (IJSR), (381-386)2020.

[43] P. K. Roy and S. Chahar, "Fake Profile Detection on Social Networking Websites: A Comprehensive Review," in IEEE Transactions on Artificial Intelligence, vol. 1, no. 3, pp. 271-285, Dec. 2020, doi: 10.1109/TAI.2021.3064901.

[44] Jiang, Zicong, Liquan Zhao, Shuaiyang Li, and Yanfei Jia. "Real-time object detection method based on improved YOLOv4-tiny." arXiv preprint arXiv:2011.04244 (2020).