

Fake Profile Detection and Stalking Prediction on Twitter (X) using Convolutional Neural Network (CNN)

Christy Onoshokwue Isokpehi
Department of Computer Science
University of Port Harcourt
Nigeria

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

ABSTRACT

This research is directly focused on fake profile detection and stalking prediction in twitter using convolution neural network. Twitter is a real-time social media application that has gained global popularity, and the use of twitter is also raising serious issues in the form of cyber-stalking. Stalking is a serious cyber-attack in which the attacker uses digital media to harass the victim or group through personal attacks and the disclosure of false or confidential information among other persons. Stalking is categorized as email-stalking, internet-stalking, and computer-stalking. The main challenging problems in twitter social network security is to recognize fake profiles and stalking activities by followers. These Fake profiles are a preferred means for malicious users to commit various cybercrimes such as cyber-stalking, disseminating misinformation and fake news, stigmatize someone's personality, capturing subscribers' credentials and generating malicious communications, misleading users towards counterfeit sites and impacting notoriety. This study detect fake profiles and predict stalking activities in twitter. Convolution neural network is used to train with the dataset of twitter profiles which gives the accuracy of the model. Predicted results are produced after training and evaluation of the models, which is able to distinguish between fake and real twitter profiles based on attributes like follower and friend counts, status updates, and more. The adopted methodology is Object Oriented Analysis and Design Method (OOADM). Python programming language is used for implementation. The application results show that the proposed convolution neural network model performs better prediction accuracy than other considered algorithms and correctly classified over 95% of the accounts with a low error rate.

Keywords

Fake Profile; Stalking; Convolutional Neural Network; Twitter; Detection

1. INTRODUCTION

Fake profiles are a types of identity theft in which a real user is impersonated for a variety of malicious motives that violate OSN's terms of service, such as spamming [1]. Fake profile on social media accounts are profiles that are either not associated with a real person or are created with an actual person's personal information without their consent. Scammers, fraudsters, hackers, and just plain mean people create fake social media accounts every day to commit various cybercrimes. Cyber-stalking is defined as the act of stalking using the Internet, which can ultimately instigate threats, maltreatment, and/or harassment. Several of these cyberstalking acts can occur on open platforms such as Twitter or any other microblogging site as well as the membership only platforms / web sites like Facebook or Instagram to name a few. Twitter which is a form of a microblogging site is used

worldwide and is a platform in which users send, and read posts known as 'tweets' and interact [2]. Fake profiles can be generated by humans or computers (bots). One of the most widely used social networking sites is Twitter. Some people don't use these sites with good intent. Therefore they create fake accounts on social networking sites. Twitter is being targeted for many malicious activities such as creating fake profiles to stalk people, online impersonation, cyber harassment and other digital provocations which can harm the reputation and invade privacy in online social platform. A recent survey suggest that the number of accounts present in the social media is much greater than the users using it. This suggests that fake accounts have been increased in the recent years [3]. One of the challenging problems in social network security is to recognize these fake profiles. This has resulted in the need of cybersecurity measures and applications to forestall people from cyberbullying such as stalking from fake profiles. In this study, a framework to classify a Twitter profile as genuine or fake using machine learning techniques (Convolutional Neural Networks) is proposed and the same framework will be used for the prediction of stalking [4]. To overcome these issues, this research seeks to address the problems of recognizing the proliferation of fake profiles used by cybercriminals to commit various cybercrimes and cyber-stalking which cyber-stalkers used to instill fear in or gain control over a victim and also punish victim.

1.1 Review of Related Literatures

This study proposed a new model for fake profile detection and stalking prediction in Twitter using Convolution Neural Network (CNN). Over the past years, many researchers have investigated the problem of detecting malicious activities and spammers in social media using machine learning techniques. This section discusses and reviews some researchers' contributions that used the machine learning algorithm for fake profiles detection and cyberbullying detection on social media platforms and other internet applications. In the literature, researchers have applied machine learning from the supervised and unsupervised learning algorithm and natural language processing techniques for features extraction to better the detection model's performance. Saied in [5] conducted a research titled Fake profile identification on Instagram using Machine Learning techniques. The system was developed to detect fake profile accounts on Instagram using machine learning technique. Although, they achieved the developed framework but no evidence is available with regard to a unified framework that incorporates both fake profiles detection and stalking prediction. Balakrishnan et al. [6] implemented a framework utilizing Machine Learning algorithms for automated detection of cyberbullying in Twitter tweets. This method groups the tweets as bully tweets, aggressor tweets, spammer tweets, and valid tweets with psychological

characters, sentiment, and feelings. The experiment was performed utilizing Naïve Bayes and J48 Machine Learning algorithms on a dataset containing 5453 tweets. Asante et al. in [7] proposed a content-based technical solution for cyberstalking detection. The proposed model utilized a few modules: message identification, filtering, detection (content detection and profiling offender), and evidence modules. Authors utilized machine learning, data mining strategies, digital forensics, and profiling to investigate text, picture, and media substance, gather proof, and profile offenders accordingly. The authors did a good job. However, the analysis of their adopted methodology showed that they only simulated the implementation, and failed to deploy the work to a real smart environment. Mohammed and Asaad in [8] presented a new approach with dual functions, namely to identify and classify the twitter bots based on ontological engineering and Semantic Web Rule Language (SWRL) rules. The authors deployed Web Ontology Language (OWL), Semantic Web Rule Language (SWRL) rules, and reasoners to inductively learn the rules that distinguish a fake account (bot) from a real one, as well as to classify fake accounts into fake followers or spam bot. according to the authors, their approach could properly identify the false account with an accuracy of (97%) in the first stage, after which these fake accounts were classified into spam or fake follower bots with an accuracy rate of (94.9%). Furthermore, it has been found that the ontology classifier is a more interpretable model that offers straightforward and human-interpretable decision rules, as compared to other machine learning classifiers. Arun in [9] demonstrated the use of data hiding technique to hide information in profile picture or photo to detect fake profile and is associated with digital forms as cryptography, steganography and watermarking. They presented discrete wavelet transform algorithm for data hiding. This would prevent clone attack in social network. Also when the user uploading his/her profile photo it will be watermarked and then only it updated in social network. Java static watermarking algorithm was used for watermarking technique which help us avoid the clone attack in the social networks. Alhariri in [10] in his work titled early detection of similar fake accounts on twitter using the random forest algorithm analyze the early detection of similar fake accounts on Twitter using the following features based the confusion matrix: default_profile, default_profile_image, friends_count, statuses_count, followers_count, listed_count, listed_count, profile_background_image, verified, name, and id. Random Forest algorithm was used in the model which provides impressive results even in the validation phase. The Random Forest results depend upon the features selected to identify the similar fake accounts. The model produced impressive results in the early detection of similar fake accounts on Twitter. Priya in [11] proposed machine learning techniques such as Neural Networks and SVM for detecting the fake accounts on Facebook. Weka tool has been used for the

simulation of the algorithm and the obtained results were presented by the proposed plan. Weka is a data mining tool which allows quick user interaction with a simple tool for the identification of fake accounts from provided data. In this, we classify the data using above techniques, which identifies the fake accounts on the social networking sites. Furthermore, the proposed method compared to the five well-known machine-learning classifiers in terms of classification accuracy to better evaluate effectiveness of the method. The experimental results show that the proposed method performs better than other considered algorithms and correctly classified over 98% of the accounts with a low error rate.

2. MATERIALS AND METHODS

In the quest to developing detection of fake profile and predict stalking activities in Twitter using Convolution Neural Network (CNN) application, the authors employed an iterative object oriented design software engineering methodology known as Object-Oriented Analysis and Design Method (OOADM).

A RUP activity creates and maintains models and emphasizes the development and maintenance of models-semantically [12]. The object-oriented paradigm and concepts, such as visual modeling, are applied in the analysis and design of an application or system using the OOADM technical method. OOADM is adopted because it is a quicker, more reusable, dependable, efficient, and more effective method of system development. In addition, the Object-Oriented Analysis and Design Method (OOADM) for this research work involves the following phases:

- Analysis,
- Design,
- Implementation and
- Testing

We will follow the phases as detailed to get the research completed

2.1 Existing System

The existing system addressed by the study is a system for fake profile identification on Instagram using Machine Learning techniques and is illustrated in figure 1. The existing system is a fake account detector which uses very fewer factors to decide whether an account is Fake or not. The factors largely affect the way decision making occurs. When the number of factors is low, the accuracy of the decision making is reduced significantly. There is an exceptional improvement in fake account creation, which is unmatched by the software or application used to detect the fake account.

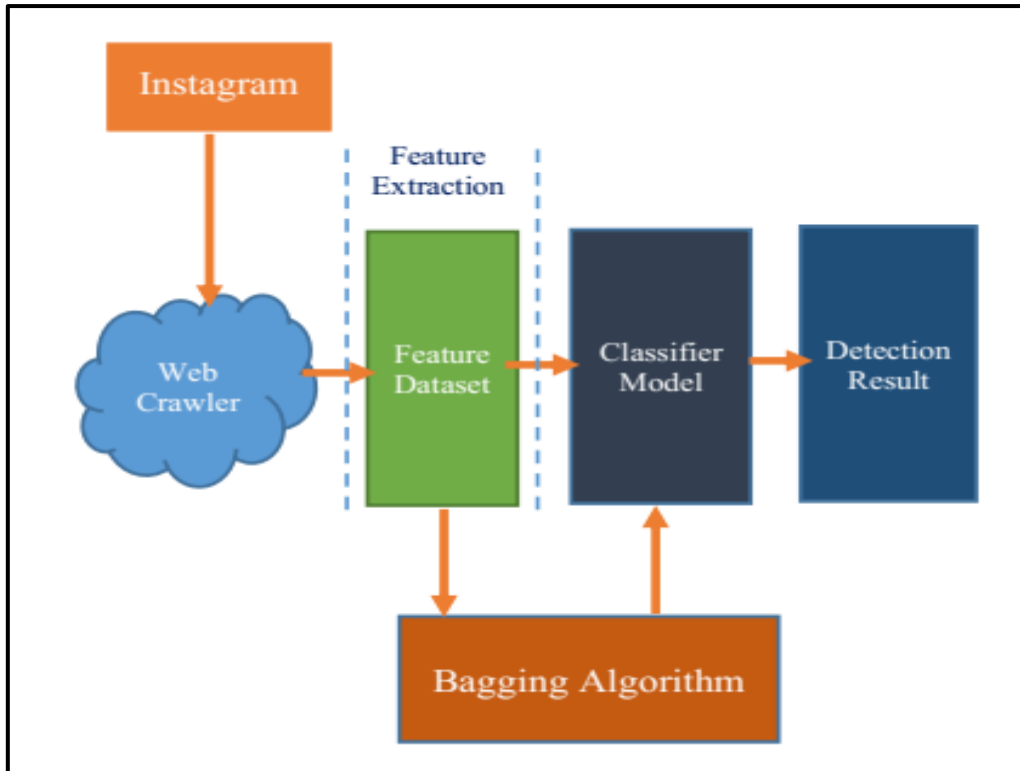


Figure 1: Shows the Existing Architecture (Source: Saeid, 2020)

Algorithm 1: Fake profile detection Model

Procedure: Gradient Boosting Machine (GBM), Bagging Method (BM) classifiers were trained and validated with training and validation sets after feature selection and then accuracy was tested.

Input:

TrainData = The labeled training set (70%)
ValidationData = The validation dataset (10%)
TestData = Unlabeled dataset (20%)

Output: Predictions = prediction from classifiers used.

Step 1: Load TrainData

Step 2: For all instances in TrainData

Step 3: For each feature matrix fed to the Classifier (GBM, BM)

Step 4: Accuracy, precision = Prediction.metrics

Step 5: Result Comparison

Step6: Stop

2.2 Disadvantages of the Existing System

The following are challenges of the existing system under investigation;

- i. The existing system use very fewer factors to decide whether an account is Fake or not. The factors largely affect the way decision making occurs. When the number of factors is low, the accuracy of the decision making is reduced significantly.
- ii. **Loss of interpretability:** It is difficult to draw very precise insights through bagging due to the averaging involved across predictions.

2.3 Fake Profile Detection and Stalking Prediction on Twitter

The proposed system provides a justification and means for overcoming the drawbacks of the existing model while accounting for recent developments in the state-of-the-art. The new system is an improvement of the existing system and it is an enhanced fake profile detection. Architectural design for the proposed fake profile detection and stalking prediction on twitter (X) using convolution neural network (CNN) is illustrate in figure 2. This study proposed an analytical model for fake profiles detection and stalking prediction. The analytical system architecture is designed for the detection of fake profiles and stalking prediction on Twitter platform. This study uses Convolutional Neural Networks. The dataset is classified based on its performance and accuracy in detecting fake profiles and cyberstalking features from the dataset is determined. Thus, the training dataset and testing dataset are loaded into classifiers to acquire the output. The proposed work is divided into two steps; Fake profile detection and the Stalking prediction. The detection process starts with the selection of the profile that needs to be tested. After selection of the profile, the suitable attributes (i.e. features) are selected on which the classification algorithm is implemented. The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is fed into the classifier. The classifier determines whether the profile is fake or real. The classifier may not be 100% accurate in classifying the profile so; the feedback of the result is given back to the classifier. This process repeats and as the time proceeds, the number of training data increases and the classifier becomes more and more accurate in predicting the fake profiles.

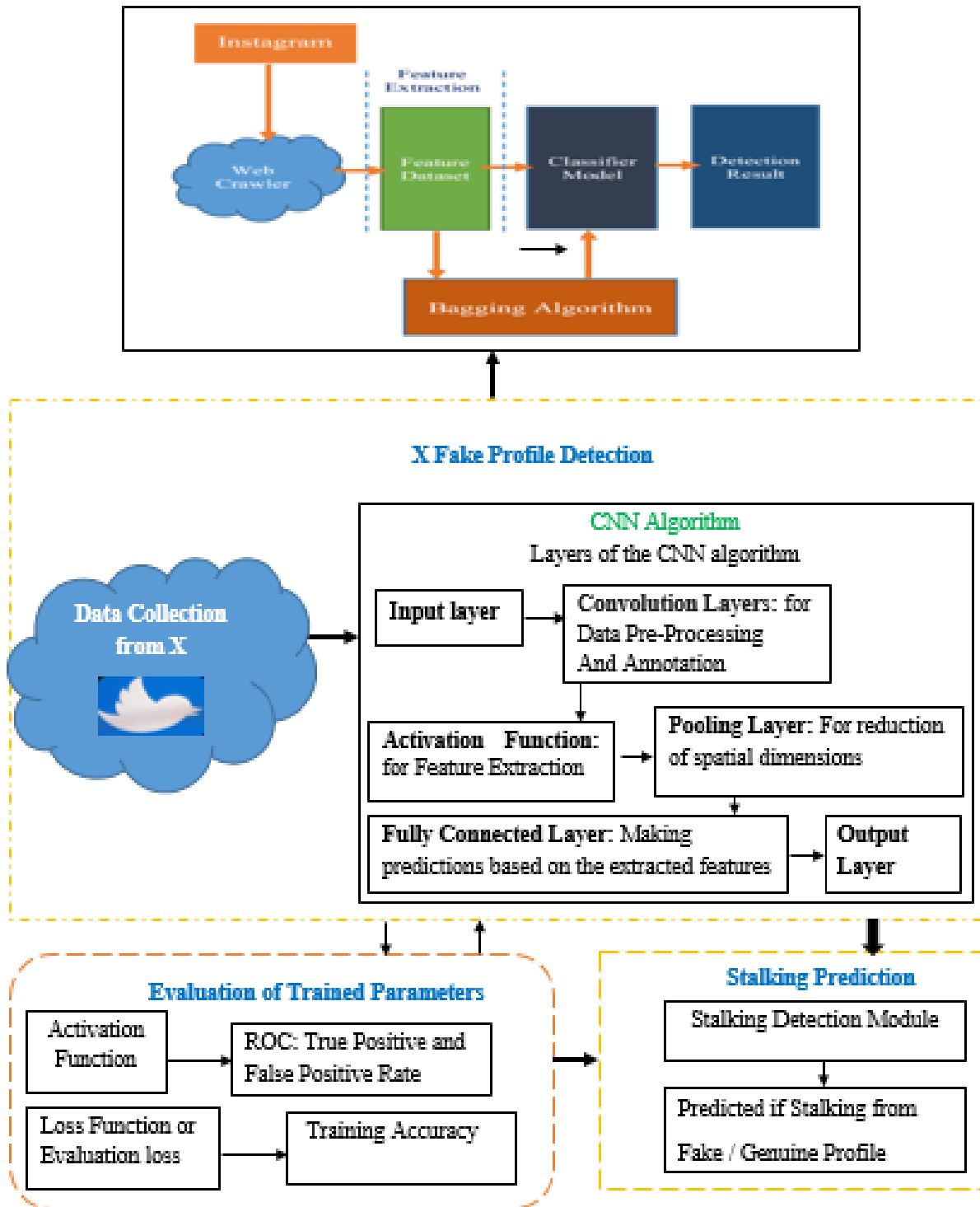


Figure 2: Architecture of the New System

2.4 Mathematical Equations of the New System

Let S be a system having Input (I), Functions (F) and Output (O). $S = \{I, F, O\}$

Where, I is a set of Twitter profiles from different users.

$I = \{\text{Profile 1, Profile 2, \dots, Profile N}\}$

O is the authenticity of account.

$O = \{\text{Fake, Real}\}$

F is the set of functions used to predict authenticity of the account.

$F = \{F1, F2, F3, F4\}$

Where,

$F1$ is a function for CNN.

F2 is a function for linear activation of DCNN.

F3 is a function for sigmoid activation of DCNN.

F4 is a function for Tan h activation of DCNN.

F1 is a function for CNN and it is given by,

$$w = \sum_{i=0}^n (inp[i]) = (hid[i])$$

$w > t: 1;$

$w < t: 0$

Where,

w = Weight assigned based on equality of input and hidden identities.

t = Threshold kept on calculated weight to detect fake identities.

Here, input (inp) corresponds to the profiles whose class label is to be detected and hidden (hid) profiles corresponds to the training data whose class label is already known.

F2 is a function for linear activation of DCNN and it is given by,

$y = a + k$

Where, $K = \sum wi xi$

X_i = set of features

W_i = weights associated with features

A = bias

F3 is a function for sigmoid activation of DCNN and it is given by,

$$y = 1/(1 + e^{-k})$$

Where, $K = \sum wi xi$

X_i = set of features

W_i = weights associated with features

A = bias

F4 is a function for Tan h activation of DCNN and it is given by,

$$y = \tanh(x) = 2/(1 + e^{-2k}) - 1$$

Where, $K = \sum wi xi$

X_i = set of features

W_i = weights associated with features

A = bias

Algorithm of the New System

The following is the new system algorithm:

Algorithm 2: Convolutional Neural Network Algorithm

Step1: Input Layer: The input layer of the CNN receives the raw input data, which in the case of images is typically represented as a grid of pixel values.

Step 2: Convolution Layers: The convolutional layer is the core component of a CNN. It consists of multiple filters or kernels that convolve over the input data. Each filter extracts local features by performing element-wise multiplications and summations between the filter weights and a small region of the input. This process helps detect different patterns and features in the dataset.

Step3: Activation Function: After the convolution operation, an activation function is applied element-wise to the resulting feature maps. The activation function introduces non-linearities into the network, allowing it to learn complex relationships in the data. Common activation functions include ReLU (Rectified Linear Unit), sigmoid and tanh.

Step4: Pooling Layer: The pooling layer reduces the spatial dimensions of the feature maps while retaining the most important information. It helps reduce computational complexity and makes the network more robust to small spatial variations. Max pooling is a commonly used pooling technique where the maximum value within a local region is retained.

Step 5: Repeat Steps 2-4 (convolution, activation, and pooling) are typically repeated multiple times to create a deep network. Deeper layers can learn higher-level representations of the input data by combining features learned in previous layers.

Step 6: Fully Connected Layer: After several convolutional and pooling layers, the output is usually flattened into a vector and passed to one or more fully connected layers. These layers are similar to those in a traditional neural network and are responsible for making predictions based on the extracted features.

Step 7: Output Layer. The output layer of the CNN produces the final predictions based on the task at hand.

2.5 Use Case Diagram

The Use_case diagram is a graphic that is used to define the core elements and processes that make up a system. The key elements are termed as "actors" and the processes are called "use cases." It shows which actors interact with each use case. Figure 3 shows the use case diagram, which contains the administrator and the user, each has difference function to perform in the model. In the diagram, the user have access to the login and the registration, the user also have access to view output. The administrator control all components of the system without registration. The purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. Use-case diagram is used to show which operations are performed by the user and which operations are performed by the system.

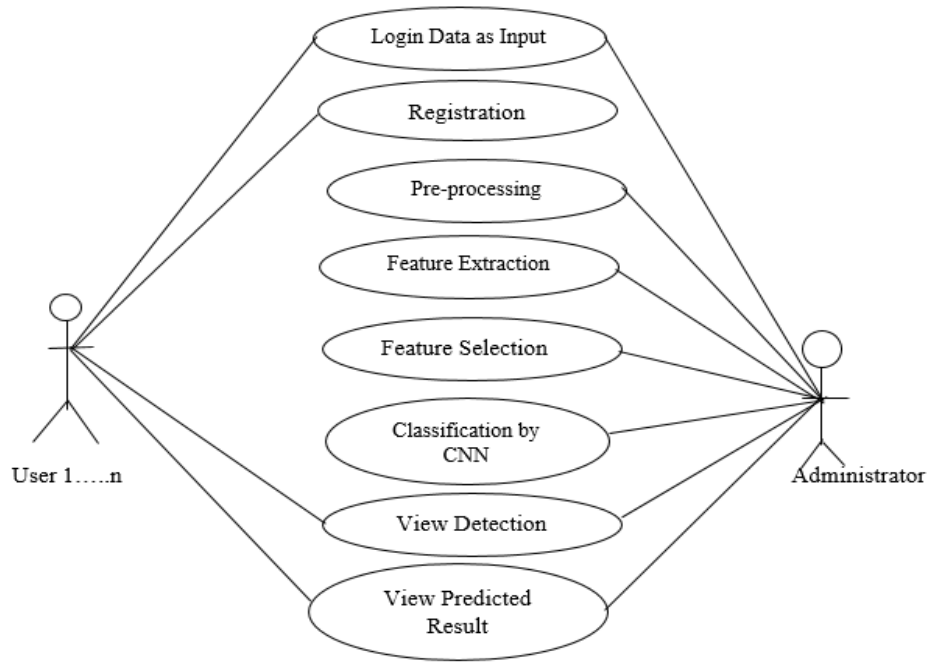


Figure 3: Use-Case Activity Diagram of the New System

2.6 Advantages of the New System

The following advantages of the New System are:

- i. Decrease Fake Account possibilities
- ii. Reduce cybercrime
- iii. It improves accuracy of fake account detection and stalking prediction

3. RESULT

The developed fake profile detection and stalking prediction in twitter using CNN has several output interfaces but only the important interface will be captured and explained. The interfaces are the Login Form, Dashboard, User Registration Page, Training Model interface, Evaluation interface and Prediction interface. User Registration Page: In the user registration page, the user is expected to register after login to have his record in the database so that the system can be recognized. In this page, the user registers all his details before clicking the "Register" button. The Register button is embedded with security restrictions which will detect and deny incorrect user details to the application. After clicking the register button, it will take to the fake profile detection and stalking prediction interface. In the registration page a new user will create an account and fill his details to register successfully while previous user will not create an account but only login his registered details. Figure 4.1 shows the user registration page. Login Form: Figure 4.2 shows the Login Form for the fake profile detection and stalking prediction in twitter. The login form is the first form that is displayed when the user launches the application. The user will be required to enter a valid username, E-mail address and password in the textboxes provided respectively before clicking the "Registration" button. The registration button is embedded with security restrictions which will detect and deny incorrect user details to the application. This form is used for registration of users. Dashboard: After the system does validation on inputs from user and verifies that the user has typed in the correct details, it is then directed to the dashboard (Home Page) where the user

is able to select from the list of options on the sidebar to perform different functions. The dashboard is shown in figure 4.3. The components of the dashboard include model, evaluation, prediction, logout, train model and clear result. Training Model Interface: Before the model will perform any other function such as evaluation, fake profile detection and stalking prediction, the model will be trained with the dataset for proper accuracy in all aspects in test accuracy, true positive rate, false positive rate and the learning curve. The model training is set for ten epochs or ten iterations. In each iteration, it displays the values of the training loss, training accuracy, validation loss and validation accuracy and gives a total training test loss and test accuracy and then saves the CNN model. Figure 4.4 shows the training interface. The result code of the training model is displayed or cited in the appendix. After the training, the model plots a graph showing how the correctness changes during the training process. The training graph is shown in section 4.7. Evaluation Interface: The evaluation interface consists of an evaluation model which displays the values of training test loss and the values of training accuracy in each training epoch or iteration. It also consists of the receiver operating characteristic curve (ROC) which contains the false positive rate and the true positive rate and the value of the ROC for the training is displayed including the graph (see section 4.8). The evaluation also displays the relationship between the training and validation loss graph and the training and validation accuracy graph. The evaluation interface is illustrated in figure 4.5. The values of the dataset features depend on the kind of graph to be plotted. Prediction Interface: In the model, the prediction interface is used to predict a fake X profile and also predict if the user with the information is a stalker or not. This interface consists of several combo boxes where values and fields are entered to give actual predictions. It includes gender, language, statuses count, followers count, friend count, favourites count and listed count. It also contains the prediction result interface where the results are displayed. The predicted result depends on the values entered in the combo boxes. Figure 4.6 and 4.7 show the prediction interface when the profile is fake (User is a Stalker) and when the profile is not fake (User is not a Stalker).

127.0.0.1:5000/register

Fake Profile Detection and Stalking Prediction in Tweeter (X) Using Convolutional Neural Network

Register

Please fill in all boxes

Name

Email

Username

Password

Register

[Login](#)

Figure 4: Registration Page

Fake Profile Detection and Stalking Prediction in Tweeter (X) Using Convolutional Neural Network

Login

Please login to continue

username

Password

Login

[Register](#)

Figure 5: Login Page

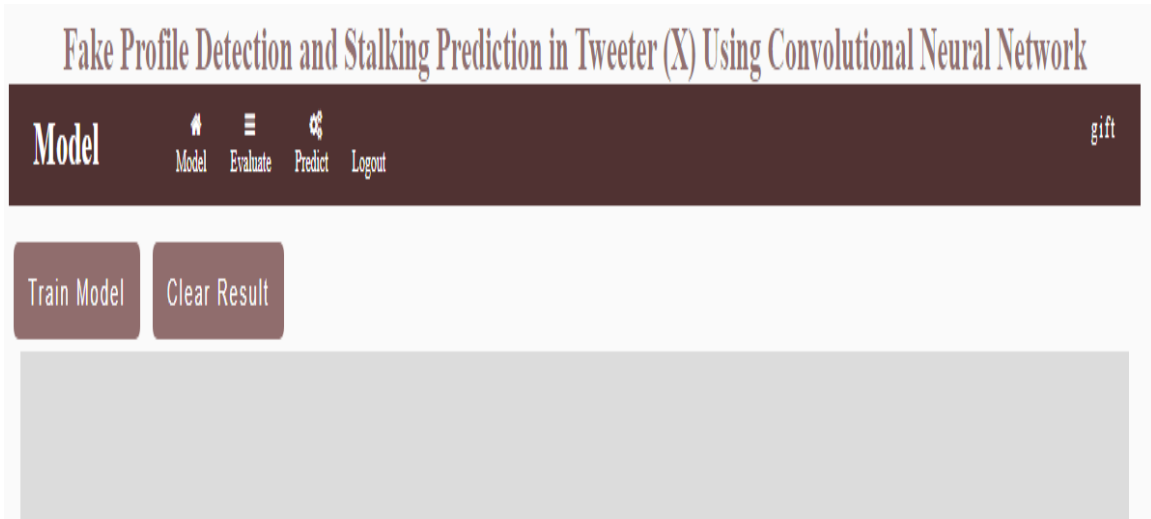


Figure 6: Dashboard of the Fake Profile Detection and Stalking Prediction

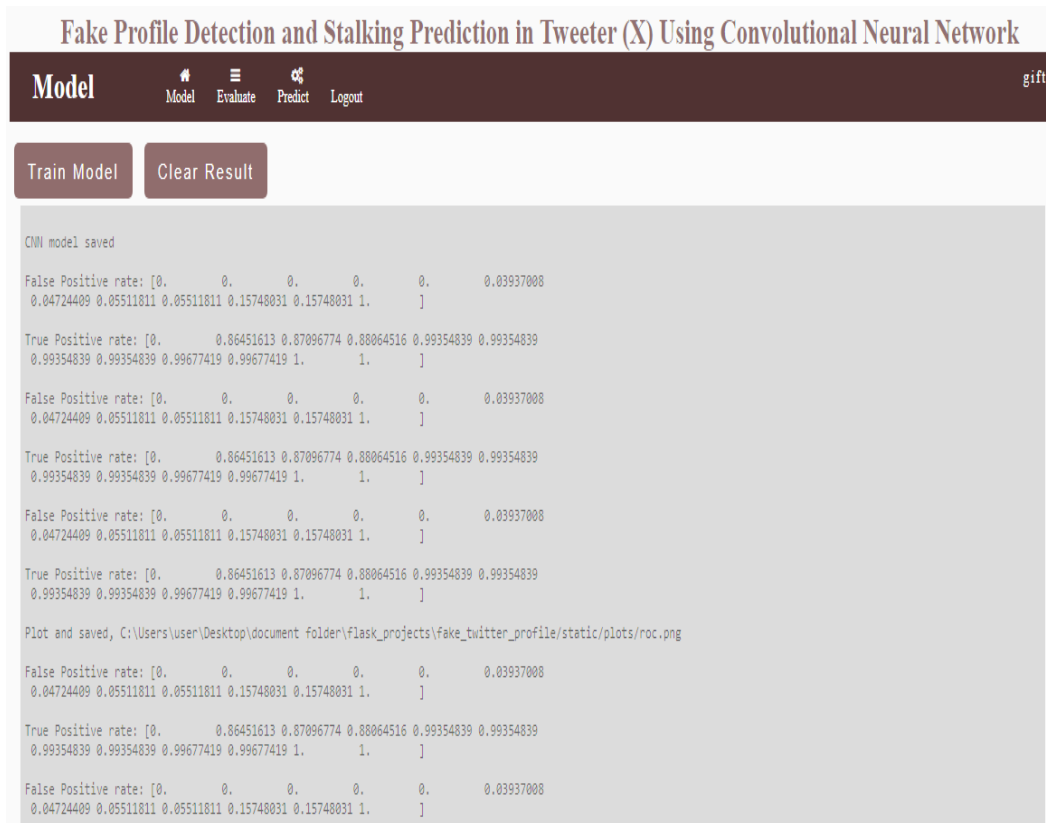


Figure 7: Training Model Interface

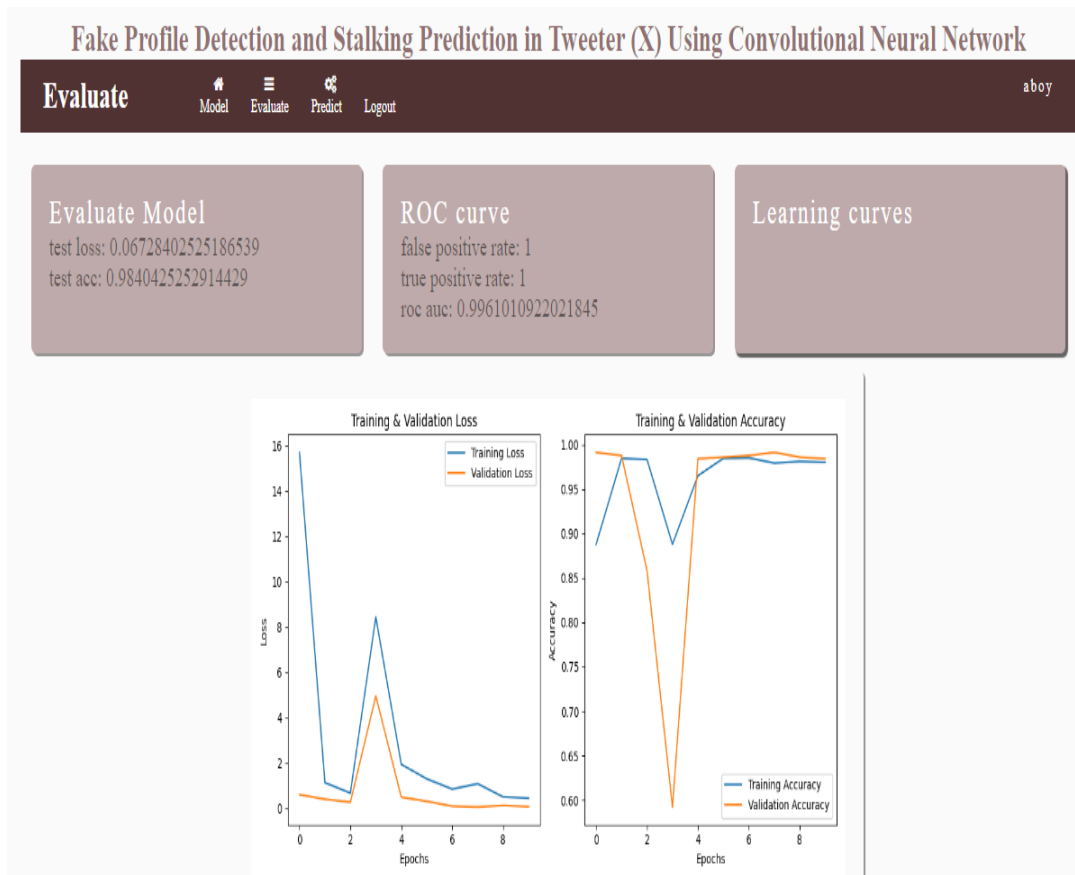


Figure 8: Evaluation Interface of the Model

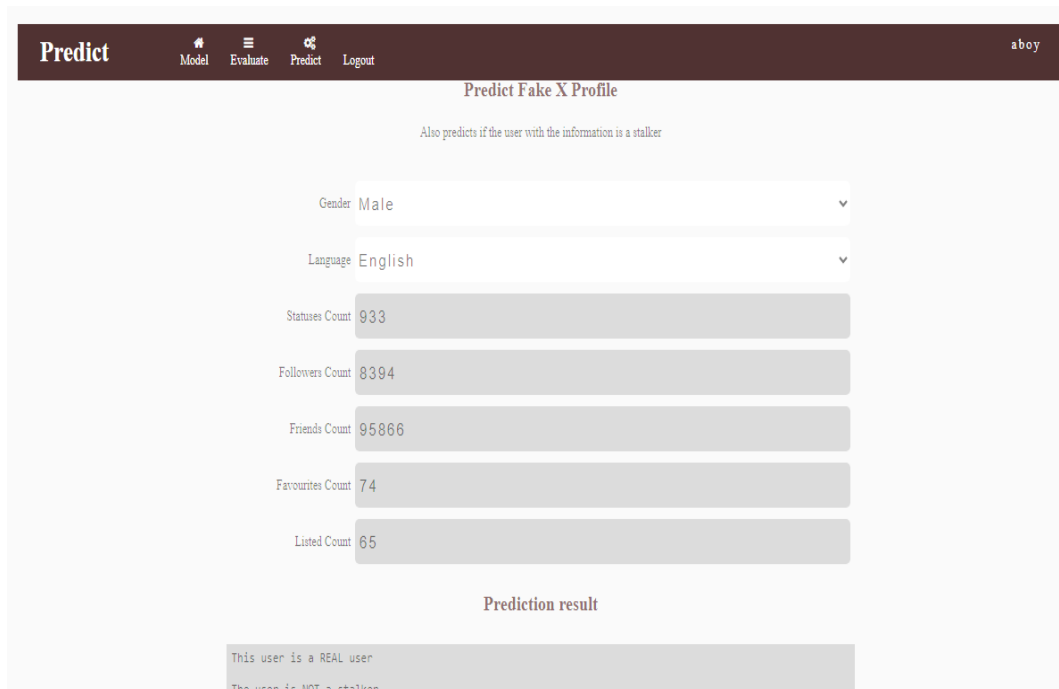


Figure 9: Prediction Interface when the Profile is not Fake (User is not a Stalker)

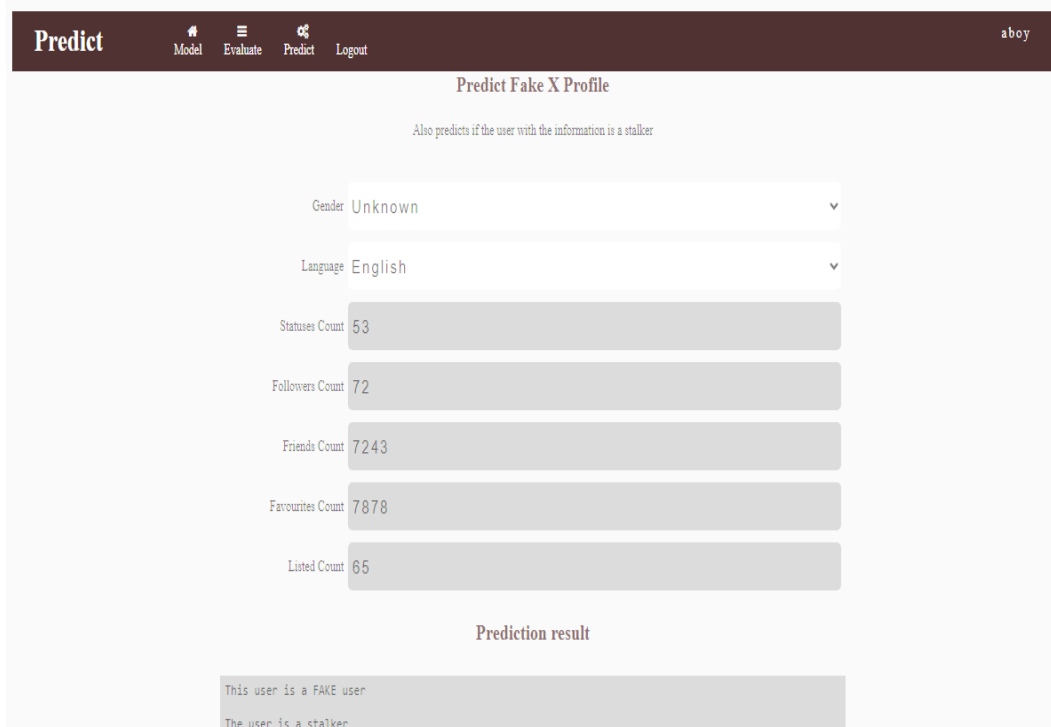


Figure 10: Prediction Interface when the Profile is Fake (User is a Stalker)

4. CONCLUSION

In this study, we have presented an improved approach for fake profile detection and stalking prediction in twitter (X) through the application of convolutional neural network. Convolution neural network (CNN) is a deep learning algorithm used in this work. The improved approach depicts evolutionary learning techniques which is the best model for handling object prediction and recognition issues and may be used to solve difficulties in data classification with a high degree of accuracy. It is frequently used to distinguish objects or perspectives, as well as to detect, segment, and classify images profile including document profile in internet. It is used to detect human activities and stalking activities in any social media function.

5. REFERENCES

- [1] Gupta T., (2019), Fake Profile Identification Using Machine Learning. International Research Journal of Engineering and Technology (IRJET), 6, 1145-1150.
- [2] Rahman (2019), Personal Information Privacy Settings of Online Social Networks and their Suitability for Mobile Internet Devices, International Journal of Security, Privacy and Trust Management (IJSPTM), 2(2), 1 – 17.
- [3] Prathyusha (2021), Forecasting shear stress parameters in rectangular channels using new soft computing methods. Plos One, 15(4): e0229731.
- [4] Nicholson H., (2020), Fake profile detection techniques in large-scale online social networks: A comprehensive review. Computers & Electrical Engineering, 65: 165-177.
- [5] Saied (2020), Fake profile identification on Instagram using Machine Learning techniques. International Research Journal of Engineering and Technology (IRJET), 6(23): 211-323.
- [6] Balakrishnan (2020), utilizing Machine Learning algorithms for automated detection of cyberbullying in Twitter tweets. Applications and Technology Conference (LISAT), Farmingdale, 23(13): 1-7.
- [7] Supraja (2019), pattern detection approach to detect the fake accounts. In this paper, the crawler is used to collect the twitter dataset. International Journal of Scientific Research and Publications, 6(41), 13 – 26.
- [8] Mohammed and Asaad (2020), new approach with dual functions, namely to identify and classify the twitter bots based on ontological engineering and Semantic Web Rule Language (SWRL) rules. International Journal of Advanced Computer Science and Applications, 7(13): 621-732.
- [9] Arun (2017), demonstration of the use of data hiding technique to hide information in profile picture or photo to detect fake profile and is associated with digital forms as cryptography, steganography and watermarking. Journal of Computer and Communications, 4(11): 79-99.
- [10] Alhariri (2020), detection of similar fake accounts on twitter using the random forest algorithm analyze the early detection of similar fake accounts on Twitter using the following features based the confusion matrix: International Conference on Research and Innovation in Information Systems (ICRIIS), 2013 223 – 234.
- [11] Priya G., (2020), Online Social Network: A Threat to Privacy and Security of Human Society, International Journal of Scientific Research and Publications, 5(4), 1 – 6.
- [12] Chibroma A. (2017), Structured Generative Models using the IoT, *Proceedings of the 31st International Conference on Machine Learning, Beijing, China, JMLR: W & CP, V.32, arXiv:1401.0514v2[cs.PL], 1 – 14*