JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FAKE SOCIAL MEDIA ACCOUNT DETECTION USING MACHINE LEARNING

[1]BURIDI AMRUTHA, [2]KOYYANA RAMYA, [3]VANAPALLI SOWJANYA,[4] YAMALA KARTHEEK,

[5] YEGIREDDI PRAVEEN

[1]B. TECH(STUDENT), [2]B.TECH(STUDENT), [3]B.TECH(STUDENT),[4] B.TECH(STUDENT),[5]ASSISTANT PROFESSOR
[1]Department of Computer Science And Engineering,
[1]Satya Institute Of Technology and Management, Vizianagaram, Andhra Pradesh, India

Abstract : Online social media is taking over the globe these days in a number of ways. The amount of people utilizing social media is rapidly rising every day. The primary benefit of social media on the internet is the ease with which we can connect with others and improve our communication with them. This opened up additional avenues for possible attacks, such impersonation and fraudulent information. According to a recent poll, there are far more social media accounts than there are users. This may indicate a rise in phony accounts in recent years. It is challenging for online social media platforms to recognize these fraudulent profiles. Due to the abundance of misleading material and marketing on social media, it is necessary to recognize these bogus accounts. Traditional techniques are unable to reliably differentiate between authentic and fraudulent accounts. The earlier efforts are out of date due to advancements in the creation of false accounts. In order to detect phony accounts, the new models employed a variety of strategies, including automated posting and comments, disseminating misleading material, and bombarding users with spam. As the number of phony accounts on the rise, many algorithms with various features are being used. Algorithms such as Naïve Bayes, Support Vector Machine, and Random Forest, which were once used, are no longer effective in identifying fraudulent accounts. We developed a novel technique in this study to distinguish phony accounts. We used the gradient boosting technique to a three-attribute decision tree. These characteristics include fake activity, engagement rate, and spam comments. We used data science and machine learning together to precisely.

IndexTerms **- machine learning, fake account detection, ANN, Python.**

I. INTRODUCTION

Social media is an essential part of everyone's life in the modern world. Social media is mostly used for sharing news, staying in touch with friends, and other things. Social media users are growing at an exponential rate. Recently, Instagram has become incredibly popular among social media users. Instagram has over a billion active users and has grown to become one of the most popular social media platforms. People with a sizable following on Instagram have been known as Social Media Influencers since the platform's introduction to the social media landscape. Businesses increasingly turn to these social media influencers as a preferred source for advertising their goods and services.

Social media's extensive use has brought society both benefits and drawbacks. The rate at which false information is being shared online through social media fraud is rapidly rising. On social media, fake accounts are the main source of inaccurate information. Businesses that spend large sums of money on social media influencers need to be aware of whether the following those accounts have acquired is natural or artificial. Therefore, a fake account detection program that can reliably determine if an account is false or not is much needed. In this study, we employ machine learning classification techniques to identify phony accounts. The identification of fraudulent accounts is mostly influenced by variables like artificial activity and engagement rate.

EXSITING SYSTEM:

Very few factors are used by the current methods to determine if an account is phony or not. The elements have a significant impact on how decisions are made. A minimal number of elements leads to a significant reduction in decision-making accuracy. The program or application used to identify the false account cannot match the remarkable advancement in the production of phony accounts. The development of phony accounts has rendered the previous approaches outdated. Applications for detecting bogus accounts most frequently employ the Random Forest algorithm. A few drawbacks of the technique include its inability to handle categorical variables with varying numbers of levels efficiently. Furthermore, as the number of trees increases, the algorithm's.

PROPOSED SYSTEM:

The random forest technique is used by the current system to detect bogus accounts. When it has all of the inputs and the right inputs, it is efficient. It gets more challenging for the algorithm to generate the result when part of the inputs are absent. We employed an ANN algorithm in the suggested systems to get over these problems. The ANN algorithm is similar to the random forest algorithm in that it primarily relies on decision trees. Additionally, we modified our approach to identify fraudulent accounts by implementing fresh techniques for account discovery. Spam comments, engagement rate, and fake activity are the techniques employed. The ANN

algorithm makes use of decision trees formed from these inputs.We get an output from this algorithm even in cases where certain inputs are absent. This is the main rationale behind the algorithm's selection. This algorithm's application allowed us to obtain incredibly precise findings**.**

II. IMPLEMENTATION

1. Upload Social Network Profiles Dataset: We will upload the dataset to the application using this module.
2. Preprocess Dataset: Using this module, we will use processing techniques such missing value removal and divide the dataset into train and test sets, with 80% of the dataset used to train the artificial neural network and 20% of the dataset used to assess the prediction accuracy of the ANN.
3. Run the ANN Algorithm: With the help of this module, we can train the ANN algorithm using both train and test data to create a train model, which we can then use to identify phony accounts in fresh datasets.
4.ANN Accuracy & Loss Graph: We use 200 epochs or iterations to train the ANN model, and we plot the accuracy and loss performance of the ANN at each epoch or iteration on the graph.
5. Use ANN to Predict false/Genuine Profile: This module allows us to submit new test data and use an ANN train model to determine if the data is real or false.

2.1 Libraries Used

Pandas: Pandas is a Python computer language library for data analysis and manipulation. It offers a specific operation and data format for handling time series and numerical tables. It differs significantly from the release3-clause of the BSD license. It is a well-liked open-source of opinion that is utilized in machine learning and data analysis.

NumPy: The NumPy Python library for multi-dimensional, big-scale matrices adds a huge number of high-level mathematical functions. It is possible to modify NumPy by utilizing a Python library. Along with line, algebra, and the Fourier transform operations, it also contains several matrices-related functions.

Matplotlib: It is a multi-platform, array-based data visualization framework built to interact with the whole SciPy stack. MATLAB is proposed as an open-source alternative. Matplotlib is a Python extension and a cross-platform toolkit for graphical plotting and visualization.

Scikit-learn: The most stable and practical machine learning library for Python is scikit-learn. Regression, dimensionality reduction, classification, and clustering are just a few of the helpful tools it provides through the Python interface for statistical modeling and machine learning. It is an essential part of the Python machine learning toolbox used by JP Morgan. It is frequently used in various machine learning applications, including classification and predictive analysis.

Keras: Google's Keras is a cutting-edge deep learning API for creating neural networks. It is created in Python and is designed to simplify the development of neural networks. Additionally, it enables the use of various neural networks for computation. Deep learning models are developed and tested using the free and open-source Python software known as Keras.

III. RESEARCH METHODOLOGY

Compile a big dataset of social media accounts, encompassing both fictitious and authentic ones. Social networking sites can be scraped for information, or publically accessible datasets can be used. Eliminate duplicate profiles, unnecessary information, and missing variables from the data to make it cleaner. Transform textual data into numerical format by applying word embeddings or bag-of-words algorithms. From the preprocessed data, extract pertinent variables including the quantity of followers, engagement rate, posting frequency, emoji usage, and linguistic patterns. Depending on the nature of the issue, select an appropriate machine learning method for detecting phony social media accounts. Support vector machines (SVM), Random Forest, K-Nearest Neighbors Algorithm (KNN), Logistic Regression, and Artificial Neural Networks (ANN) are examples of popular algorithms.

Using an appropriate optimization approach such as gradient descent or stochastic gradient descent, train the chosen model on the preprocessed dataset. Divide the dataset into testing and training sets so that the performance of the model can be assessed. Assess the trained model's performance in detecting phony social media accounts by looking at its recall, accuracy, precision, and F1 score on the testing set. For greater performance, tweak the model's hyperparameters using methods like grid search and cross-validation.In order to detect phony social media accounts in real time, deploy the trained model in a production environment. Regularly check its performance and make necessary adjustments to increase its accuracy over time.

3.1 Uploading the data:

A dataset is a grouping of examples, and while utilizing machine learning techniques, we usually require many datasets for various objectives.

• **Training Dataset:** A dataset used to train our model in our machine learning technique.
• **Testing Dataset**: This is a dataset that is not utilized for model training, but rather is used to verify the accuracy of our model. You may refer to it as the validation dataset.

3.2 Data and Sources of Data

To train ANN algorithm we are using below details from social networks

Account_Age, Gender, User_Age, Link_Desc, Status_Count, Friend_Count, Location, Location_IP, Status

All fake users main intention is to send friend request to normal users to hack their machine or to steal their data and never they will have many number of posts or have many following friends and their account age also will have less number of years. By

analysing this features Facebook will mark whether user profile is fake or genuine. This Facebook profile data we downloaded from Facebook website and using this data to train ANN model. Below are some values from profile dataset.

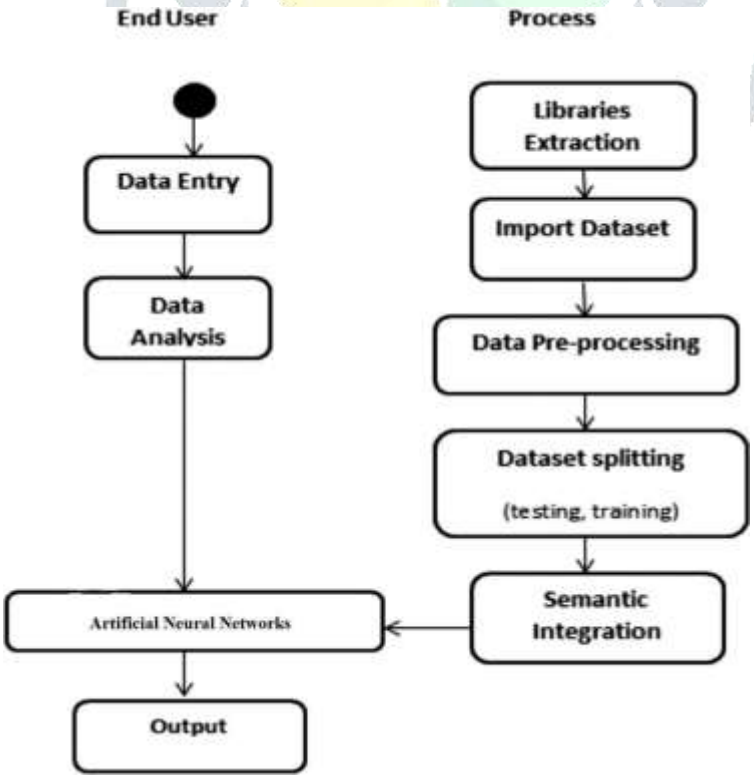
Account_Age, Gender, User_Age, Link_Desc, Status_Count, Friend_Count, Location, Location_IP, Status
10,1,22,0,1073,237,0,0,0
10,0,33,0,127,152,0,0,0
10,1,46,0,1601,405,0,0,0
10,0,25,0,704,380,0,0,0
7,1,34,1,64,721,1,1,1
7,1,30,1,69,587,1,1,1
7,1,36,1,61,782,1,1,1
7,1,52,1,96,827,1,1,1

In above dataset all bold names are the dataset column names and all integer values are the dataset values. As ANN will not take string value so we convert gender values to 0 or 1, if male value is 1 and if female value is 0. In above dataset last column give us information of fake or genuine account if last column contains value 0 then account is genuine otherwise fake. All fake account will have less number of posts as their main intention is to send friend requests not posts, so by analysing this features Facebook mark that record with value 1 which means it's a fake account. We are using above dataset to train ANN model and this dataset saved inside code 'dataset' folder. After building train model we input test data with account details and ANN will give result as fake or genuine. Below are some values from test data.

3.2 Dataset pre-processing:

Finding phony accounts is a crucial first step. This stage involves processing data so that it may be entered into the detection procedure in the proper format. the valuable data that can be obtained

It is crucial that we preprocess our data before putting it into our model since it has a direct impact on its capacity to learn.

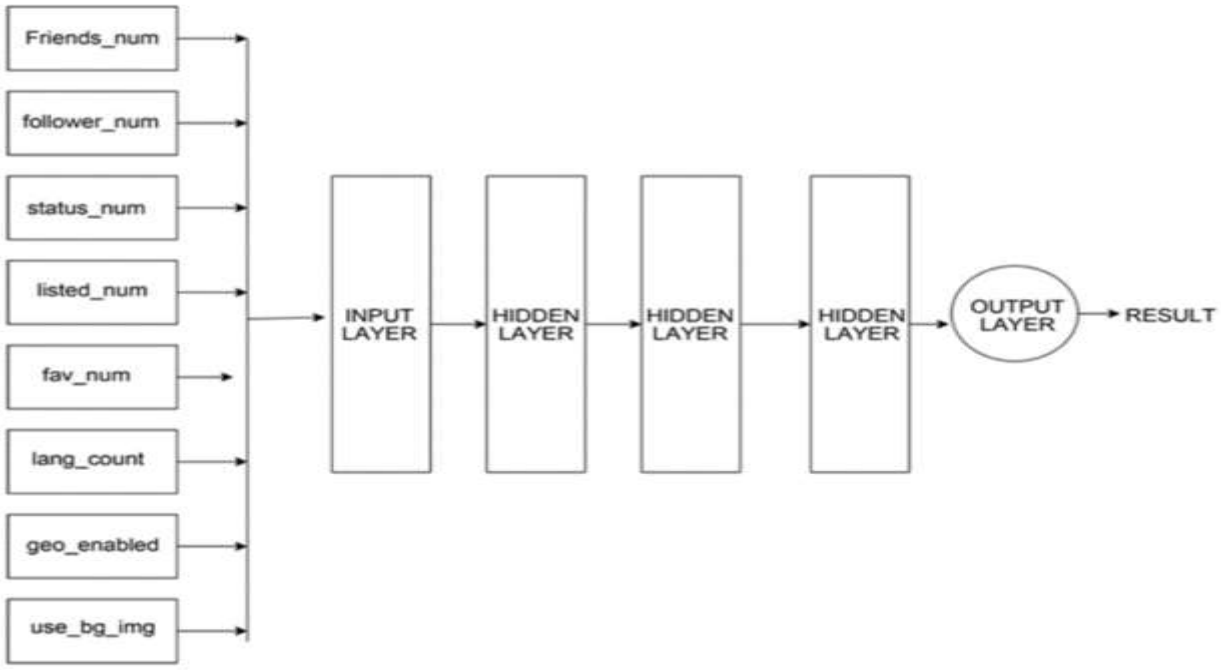


3.3ANN algorithms Details

To demonstrate how to build a ANN neural network based image classifier, we shall build a 6 layer neural network that will identify and separate one image from other. This network that we shall build is a very small network that we can run on a CPU as well. Traditional neural networks that are very good at doing image classification have many more parameters and take a lot of time if trained on normal CPU. However, our objective is to show how to build a real-world convolutional neural network using TENSORFLOW.
Neural Networks are essentially mathematical models to solve an optimization problem. They are made of neurons, the basic computation unit of neural networks. A neuron takes an input (say x), do some computation on it (say: multiply it with a variable w and adds another variable b) to produce a value (say; z= wx+b). This value is passed to a non-linear function called activation function (f) to produce the final output(activation) of a neuron. There are many kinds of activation functions. One of the popular

activation function is Sigmoid. The neuron which uses sigmoid function as an activation function will be called sigmoid neuron. Depending on the activation functions, neurons are named and there are many kinds of them like RELU, TanH.
If you stack neurons in a single line, it's called a layer; which is the next building block of neural networks. See below image with layers.
To predict class label multiple layers operate on each other to get best match layer and this process continues till no more improvement left.
Module Details:
Upload Social Network Profiles Dataset: Using this module we will upload dataset to application
Preprocess Dataset: Using this module we will apply processing technique such as removing missing values and then split dataset into train and test where application use 80% dataset to train ANN and 20% dataset to test ANN prediction accuracy
Run ANN Algorithm: Using this module we will train ANN algorithm with train and test data and then train model will be generated and we can use this train model to predict fake accounts from new dataset.
ANN Accuracy & Loss Graph: To train ANN model we are taking 200 epoch/iterations and then in graph we will plot accuracy/loss performance of ANN at each epoch/iteration.
Predict Fake/Genuine Profile using ANN: using this module we will upload new test data and then apply ANN train model to predict whether test data is genuine or fake**.**

Fig: ANN Algorithm Process

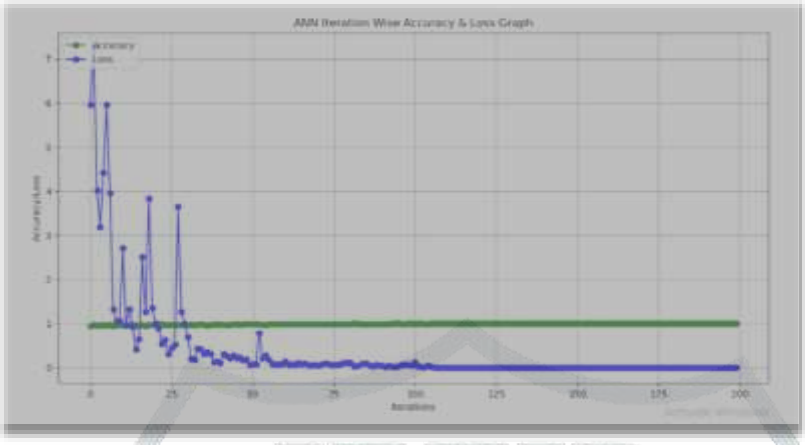


IV. RESULTS AND DISCUSSION

The metrices considered in this project is :
Accuracy

METRICES	RANDOM FOREST	ANN
ACCURACY	0.925	0.99

In above screen click on 'Upload Social Network Profiles Dataset' button and upload dataset.
In above screen selecting and uploading 'dataset.txt' file and then click on 'Open' button to load dataset and to get below screen.
In above screen dataset loaded and displaying few records from dataset and now click on 'Preprocess Dataset' button to remove missing values and to split dataset into train and test part.
In above screen we can see dataset contains total 600 records and application using 480 records for training and 120 records to test ANN and now dataset is ready and now click on 'Run ANN Algorithm' button to ANN algorithm.
In above screen we can see ANN start iterating model generation and at each increasing epoch we can see accuracy is getting increase and loss getting decrease.



In above screen we can see after 200 epoch ANN got 100% accuracy and in below screen we can see final ANN accuracy.
In above graph x-axis represents epoch and y-axis represents accuracy/loss value and in above graph green line represents accuracy and blue line represents loss value and we can see accuracy was increase from 0.90 to 1 and loss value decrease from 7 to 0.1. Now model is ready and now click on 'Predict Fake/Genuine Profile using ANN' button to upload test data and then ANN will predict below result.
In above screen we are selecting and uploading 'test.txt' file and then click on 'Open' button to load test data and to get below prediction result.
In above screen in square bracket we can see uploaded test data and after square bracket we can see ANN prediction result as genuine or fake.

V. CONCLUSION

In this study, we have developed a clever method for identifying phony OSN accounts. We have completely removed the necessity for labor-intensive, time-consuming manual prediction of a phony account by utilizing machine learning algorithms to their fullest potential. The development of phony accounts has rendered the systems that were in place obsolete. The current system is dependent on unstable elements. To improve prediction accuracy in this study, we employed stable variables like fake activity and engagement rate.

REFERENCES

1. "Detection of Fake Twitter accounts with Machine Learning Algorithms" Ilhan aydin,Mehmet sevi, Mehmet umut salur.
2. "Detection of fake profile in online social networks using Machine Learning" Naman singh, Tushar sharma, Abha Thakral, Tanupriya Choudhury.
3. "Detecting Fake accounts on Social Media" Sarah Khaled, Neamat el tazi, Hoda M.O. Mokhtar.
4. "Twitter fake account detection", Buket Ersahin, Ozlem Aktas, Deniz kilinc, Ceyhun Akyol.
5. " a new heuristic of the decision tree induction" ning li, li zhao, ai-xia chen, qing-wu meng, guo-fang zhang.
6. " statistical machine learning used in integrated anti-spam system" peng-fei zhang, yu-jie su, cong wang.
7. " a study and application on machine learning of artificial intellligence" ming xue, changjun zhu.
8. " learning-based road crack detection using gradient boost decision tree" peng sheng, li chen, jing tian.
9. " verifying the value and veracity of extreme gradient boosted decision trees on a variety of datasets" aditya gupta, kunal gusain, bhavya popli.
10. " fake account identification in social networks" loredana caruccio, domenico desiato, giuseppe polese.